E-AGRO MARKETS

Web Analytics

Pedro Campos, João Mendes Moreira e Sara Santos

E-AGRO INNOVATION

agrocluster

Cofinanciado por:

COMPETE 2020

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

Índice

1.
INTRODUÇÃO 7
2.
WEB ANALYTICS – INTRODUÇÃO 8
2.1. Definição e principais métricas 8
2.2. Fundamentos de *web Analytics* para as empresas 11
2.3. Ferramentas de *Web Analytics*: o Google Analytics 12
2.4. *Web Analytics* e a *web* social 16
3.
METODOLOGIA CRISP-DM 18
3.1. Taxonomia básica sobre análise de dados 18
3.2. Apresentação de alto nível da metodologia CRISP-DM 18
3.3. Compreensão do negócio 19
3.4. Compreensão dos dados 21
3.5. Preparação dos dados 22
3.6. Modelação 23
3.7. Avaliação 24
3.8. Implantação 25
4.
RECOLHA DE DADOS 26
4.1. *Web* Crawling 26
4.2. *Cookies* 27
4.3. *Logs* 29
5.
TÉCNICAS DE MODELAÇÃO DE DADOS 31
5.1. Tipos de aprendizagem 31
5.2. Sistemas de recomendação 34
5.3. Trabalhando com textos 41
5.4. Análise de Redes Sociais 47
6.
SEARCH ENGINES 54
6.1 Como funcionam os *search engines*? 54
6.2 *Search Engine Marketing* (SEM) 55
6.3 *Search Engine Optimization* (SEO) 56
7.
Síntese 69
Bibliografia 71

Índice de Figuras

Figura 1. A pirâmide da terminologia de *Web Analytics* 10
Figura 2. Forma como se pode tirar partido dos dados 11
Figura 3. *Dashboard* do *Google Analytics* (A) 14
Figura 4. *Dashboard* do *Google Analytics (*B) 15
Figura 5. *Dashboard* do *Google Analytics* (C) 15
Figura 6. A metodologia CRISP-DM 19
Figura 7. Resultado do método das k-médias, com k=3, usando 14 instâncias com 2 variáveis (idade e nível de educação). Dados artificiais. 32
Figura 8. Técnica de reamostragem *hold-out* 33
Figura 9. Técnica de reamostragem validação cruzada com k partições 33
Figura 10. matriz de confusão para um problema binário (duas classes) e algumas das medidas de avaliação mais utilizadas 33
Figura 11. Não direcionado (a), ponderado (b) onde a espessura de uma relação é proporcional ao seu peso, dirigido (c), e, dirigido e ponderado (d) são tipos de relações em redes 47
Figura 12. Uma rede social de exemplo 48
Figura 13. Parte da rede social da Figura 12 50
Figura 14. Exemplo com três redes 51
Figura 15. A utilização de motores de busca em todo o mundo 55
Figura 16. Pesquisa no Google pela palavra-chave "cabazes de frutas e legumes" 56
Figura 17. Resultados de "envolvimento" no *site* da marca Continente 58
Figura 18. Evolução da pesquisa nos últimos 12 meses por legumes, frutas e azeite 58
Figura 19. Plataforma Ubersuggest - pesquisas por "azeite" nos últimos 12 meses 60
Figura 20. Contagem de palavars em texto sobre a marca "Gallo" 60
Figura 21. Informação no Google para "Club Agrocluster" 61
Figura 22.H1 – "Marketing Digital – ainda faz sentido tratá-lo separado da Estratégia de Marketing Global" 63
Figura 23. H2 "Marketing Digital ou Omnichannel" 63
Figura 24. Ferramenta MOZ bar que permite analisar a qualidade dos *links* 65
Figura 25. Importância dos *links* por tipo 66
Figura 26. Ferramenta para analisar se o site cumpre as boas práticas 68

Índice de Tabelas

Tabela 1. Métricas que podem ser usadas em dados de *social media* 17
Tabela 2. Exemplo de um conjunto de 9 instâncias caracterizadas por 5 variáveis, 4 delas preditivas e uma objetivo (a amarelo) 18
Tabela 3. Exemplo de *feedback* implícito - porta enxertos comprados pelos produtores 35
Tabela 4. Exemplo de feedback explícito - porta enxertos comprados e avaliados pelos produtores 36
Tabela 5. Semelhanças vetor cosseno entre os utilizadores da Tabela 4. Como os valores de similaridade são simétricos, os valores abaixo da diagonal espelhariam os valores acima da diagonal. 38
Tabela 6. Semelhança entre os utilizadores (Tabela 4) pela correlação de Pearson. Como os valores de similaridade são simétricos, os valores abaixo da diagonal espelhariam os valores acima da diagonal. 39
Tabela 7. Conjunto de treino de textos rotulados 43
Tabela 8. Resultados de aplicação de um algoritmo de radicalização 44
Tabela 9. Radicais após remoção das palavras *stop* 45
Tabela 10. Instâncias com os seus radicais em formato estruturado 45
Tabela 11. A matriz de adjacência para a rede da Figura 12 48
Tabela 12. A matriz de adjacência da Tabela 9 ao quadrado mostrando as contagens de caminhos de comprimento dois entre pares de nós 48
Tabela 13. Propriedades básicas dos nós da rede da Figura 12 49
Tabela 14. A matriz de distâncias entre nós para o gráfico da Figura 12 49

Biografia

Pedro Campos

Pedro Campos é Doutorado em Ciências Empresariais pela Universidade do Porto (2008). Atualmente é Professor Auxiliar na Faculdade de Economia do Porto, onde tem lecionado disciplinas nas áreas da Estatística, Análise de Dados, Data Mining e Sistemas de Informação dos vários ciclos de estudos, sendo no presente Diretor do curso de Pós-Graduação em Business Intelligence & Analytics da Porto Business School. É membro do Laboratório de Inteligência Artificial e Análise de Decisão do INESC TEC, sendo responsável pela linha Modelling and Simulation, e ainda Deputy-Director do International Statistical Literacy Project desde 2009, onde desempenha funções internacionais no âmbito da promoção da literacia estatística. Pedro Campos ocupa atualmente o cargo de Vice-Presidente da IASE – International Assotiation for Statistical Education, desde 2017. As suas áreas de interesse são a Ciência dos Dados, Pesquisa de Mercados, Análise de Redes Sociais e Economia Computational (particularmente Modelos Multi-Agentes) com aplicações à Demografia, Redes de Empresas, Marketing e Dinâmica de populações.

João Mendes Moreira

João Mendes Moreira é Doutorado em Ciências da Engenharia pela Faculdade de Engenharia da Universidade do Porto e tem conduzido investigação em temas relacionados com a aprendizagem automática como Machine Learning e Ensemble Learning. Desde 2008, ocupa o cargo de Professor Auxiliar na Faculdade de Engenharia da Universidade do Porto e é igualmente Investigador no Laboratório de Inteligência Artificial e Apoio à Decisão do Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência. Para além de ser avaliador em várias revistas científicas de renome internacional, João Mendes Moreira foi galardoado em 2014 com um “Best Paper Award”, pelo seu artigo "Merging Decision Trees: a case study in predicting student performance", sendo ainda co-autor do livro “A general introduction to data analytics”, editado em 2018.

Sara Santos

Sara Santos é Doutorada em Gestão, na especialidade de Marketing e Estratégia, pela Universidade do Porto, possuindo ainda uma Pós-Graduação em Marketing Digital pelo Instituto Português de Administração e Marketing (IPAM). Atualmente é docente no Instituto Superior de Contabilidade e Administração (ISCA) da Universidade de Aveiro e também no Instituto Politécnico de Viseu, instituição onde lhe foi atribuído o título de especialista em Marketing e Publicidade em 2019. Com mais de 10 anos de experiência profissional, Sara Santos colaborou como Coordenadora e gestora de comunicação e marketing em várias entidades nacionais e internacionais, como o PRODUTECH, a AC Marca e a Bresimar Automação. Atualmente é consultora de Marketing Digital para várias empresas e formadora certificada em gestão de redes sociais, SEO e publicidade online. É ainda autora de vários artigos, capítulos de livros e estudos publicados em revistas científicas e atas de eventos nacionais e internacionais.

1.
Introdução

A *Web* é uma enorme fonte de informação. Devido à autoria díspar das páginas da *web*, as informações encontram-se bastante desorganizadas. Por outro lado, a quantidade de dados gerados e disponibilizados tem crescido exponencialmente e de forma contínua, dando origem ao paradigma do *Big Data*. A quantidade de dados que produzimos todos os dias é realmente impressionante: existem 2,5 quintiliões de *bytes* de dados criados diariamente no nosso ritmo atual, mas esse ritmo está a acelerar com o crescimento da Internet das Coisas (IoT). E admite-se que 90% dos dados existentes em todo o mundo foram gerados apenas nos últimos dois anos.

Quando um utilizador entra num *web site*, ou emite um *tweet*, está a gerar dados. Este livro é sobre o tratamento desses dados.

O *Web Analytics* é a análise de dados qualitativos e quantitativos disponíveis na *Web* para promover uma melhoria contínua da experiência *online* dos utilizadores.

Conforme definido por Delen, Sharda e Turban (2015), o *Web Analytics* é a aplicação de atividades de análise de negócios a processos baseados na *Web*, incluindo comércio eletrónico". Usando algoritmos de *machine learning*, o *software* de *Web Analytics* encontra informações, sequências e padrões ocultos nos dados e utiliza-os para formar novas regras, agrupar utilizadores e prever o comportamento futuro dos clientes, transformando o *Big Data* em conhecimento valioso e oportunidades de negócio (Alghalit, 2015).

Fala-se muito do paradoxo dos dados: muitos dados, mas poucas conclusões. Ou seja, por mais dados a que tenhamos acesso, continuamos a fazer muito pouco uso deles. Este paradoxo, conjugado com o facto de os dados serem, hoje em dia, quase todos gerados na *Web* levou ao aparecimento do *Web Analytics 2.0* (Kaushik, 2018).

Este livro encontra-se organizado da seguinte forma: no Capítulo 2 faz-se uma primeira introdução aos conceitos principais do *Web Analytics*, começando por se definir as principais métricas, os fundamentos de *Web Analytics* para as empresas e o *Google Analytics*. Introduz-se também a análise de dados do *social media*. No Capítulo 3, descreve-se o CRISP-DM: sendo *Web Analytics* uma subárea de análise de dados, CRISP-DM adequa-se igualmente a projetos na área de *Web Analytics*. O Capítulo 4 é dedicado à Recolha de Dados e o Capítulo 5 foca na Análise descritiva, Análise de diagnóstico, Análise preditiva e Análise prescritiva. São também introduzidos algoritmos de Sistemas de Recomendação, Análise de Texto e Análise de Redes Sociais. O Capítulo 6 é dedicado a duas componentes fortes do *Web Marketing* que mexem com *Web Analytics*: *Search engine Marketing* (SEM) e *Search engine Optimization* (SEO). O livro termina com um resumo geral dos temas apresentados.

2.
Web Analytics – Introdução

Este capítulo introduz os conceitos principais em torno do *Web Analytics*. Começa por definir as principais métricas (acessos, visualizações de páginas, visitas, Visitantes únicos, Referenciadores, palavras-chave e frases-chave). De seguida, apresenta os fundamentos de *Web Analytics* para as empresas, focando nos aspetos "Para que serve", "como podem ser recolhidos os dados" e "como deve usado" o *Web Analytics*?

Depois, apresenta uma ferramenta que já se tornou num clássico no *Web Analytics* – o *Google Analytics*. Finalmente, aborda-se uma forma diferente de fazer *Web Analytics* que não se baseia nas métricas definidas anteriormente sobre *clickstream*, mas sim através de dados da *web* provenientes de *social media* (tais como *Clickstream*, *Facebook*, *Instagram*, *LinkedIn*, etc.).

2.1. Definição e principais métricas

Grande parte dos negócios realiza-se hoje através da *Web*. De acordo com informação recente, a indústria do comércio eletrónico gerou cerca de 3,45 triliões de dólares de vendas em 2019. Apesar do investimento crescente das empresas e da sociedade em geral nas ferramentas digitais, existe ainda alguma resiliência que limita a competitividade da nossa economia. O crescimento dessa competitividade precisa de ser reforçado, não só através da produção de novos conhecimentos, mas também da tradução desses conhecimentos em benefícios sociais e económicos para a sociedade, em particular nas áreas que envolvem competências digitais avançadas – tal como o manuseamento e análise de grandes quantidades de dados, designados por *Big Data*. (INCODE, 2019)[1].

O crescimento da Internet fez com que muitos profissionais de marketing adotassem uma abordagem participativa na utilização de dados da *Web*. As ferramentas de gestão e análise de dados (*Analytics*) têm-se adaptado ao crescimento da *Web*, razão pela qual hoje esta informação pode ser recolhida e usada por qualquer pessoa como forma de suporte à decisão para alcançar os objetivos de negócios das empresas. *Web Analytics* é a área que tem por objetivo a recolha, geração de relatórios e análise de dados sobre o comportamento dos visitantes de um *site*. O foco do *Web Analytics* está na identificação de objetivos, no uso dos dados para determinar o progresso para alcançá-los e na condução de estratégias para melhorar a eficácia do marketing.

Como fazer tudo isto?

Há algum tempo atrás, os programas de *software* mais rudimentares apenas contavam o número de páginas visitadas na *World Wide Web*, assim como o número de visitantes que entravam nos *web sites*. As ferramentas gratuitas de análise de dados da *Web* - *software* que analisam o comportamento de visitantes do

[1] Disponível em: - https://www.incode2030.gov.pt/eixos

Internet e WWW

A Internet é o conjunto de ligações entre computadores operadas pelo governo indústria, academia, e empresas privadas. É uma rede de redes mantidas com base em protocolos informáticos. Apesar dos termos Internet e *World Wide Web* (WWW) serem usados de forma semelhante, a *World Wide Web* é apenas um dos muitos serviços que funcionam dentro da internet. Na verdade, a *Web* é um conjunto de documentos interconectados (as páginas *web*, ou que constituem os *web site*s) ligadas por hiper*link*s e URLs.

site - oferecem hoje em dia detalhes minuciosos sobre o quê, quando, onde, e por que razão os visitantes acedem aos *web site*s.

Em particular, podem dar-nos informações sobre as métricas principais, tais como:

- Acessos (*hits*)
- Visualizações de página (*pageviews*)
- Visitas (*visits*)
- Visitantes únicos (*unique visitors*)
- Referenciadores (*referrers*)
- Palavras-chave e frases-chave (*keywords and keyphrases*).

Uma grande parte destes dados provém do denominado *clickstream* ou sequência de cliques que segue a trajetória que um utilizador percorre ao clicar em qualquer objeto numa página *web*. Os dados de *clickstream* são registados em aplicações através de *cookies*.

Iremos agora detalhar um pouco cada uma destas métricas.

Acessos (*hits*)

Esta é uma das métricas mais utilizadas em *Web Analytics*. Mas é, ao mesmo tempo, uma das mais falaciosas, se for mal interpretada. Os acessos aos *web site*s são medidos através do número de ocorrências ou visualizações. As ocorrências ou visualizações são criadas quando o servidor *Web* entrega um ficheiro ao navegador (*browser*) de um visitante. Os ficheiros PDF, de som, *Word*, outros documentos e imagens são alguns exemplos de ficheiros que geram *hits*. Por exemplo, o pedido de uma página com cinco imagens contaria como seis *hits*: um *hit* para a página em si mais um *hit* para cada uma das cinco imagens. Esta tem sido uma métrica popular em *web site*s que esperam pontuar publicidade, mas que, tal como referido, pode enganar a interpretação.

Visualizações de página

Uma exibição de página é registada sempre que um utilizador visualiza uma página no *web site*. A métrica revela quão bem o *web site* capturou o interesse do visitante. Os programas de *Web Analytics* dividem o número de visitantes pelo número de visualizações de cada página para determinar o número médio de páginas que cada visitante visualizou. Se esse número for baixo, isso é sinal de que talvez seja necessário repensar o conteúdo, o design ou a estrutura do *web site*.

Visitas

Às vezes também chamada de sessão ou sessão de utilizador, uma visita descreve a atividade de um utilizador individual no *web site*. Também se pode dizer que uma visita consiste numa série de visualizações efetuadas pelo mesmo visitante. É importante notar que a maioria das ferramentas de *Web Analytics* termina a sessão se o visitante permanecer inativo por 30 ou mais minutos, embora esse limite de tempo geralmente possa ser ajustado nas opções do *software*.

Web Analytics

Web Analytics consiste na recolha, tratamento e análise de dados sobre o comportamento dos utilizadores/visitantes de um *web site*. O foco do *Web Analytics* está na identificação de objetivos do negócio associado ao *web site*, na utilização dos dados para determinar a forma de atingir esses objetivos e na formulação de estratégias para melhorar a eficácia do marketing.

Métricas de *Web Analytics*

Uma métrica é uma medida quantitativa que descreve eventos ou tendências num *web site*. Para ajudar a definir métricas, usam-se medidas de desempenho chave denominadas (KPI – *Key Performance Indicators*) que ajudam a avaliar o desempenho relativamente aos objetivos propostos pela empresa. Os KPIs tendem a ser exclusivos para cada empresa. Exemplos de KPI são o número de visitantes únicos, o valor das vendas num *site* de comércio eletrónico, ou o tempo médio que os utilizadores passam no *site*.

Utilizadores/Visitantes únicos

A métrica de visitantes únicos representa o número de pessoas individuais que visita o *web site*. Cada indivíduo é contado apenas uma vez, pelo que se uma pessoa visitar um *site* diversas vezes no período de análise, esse comportamento pode contar como diversas sessões, embora apenas como um visitante único. A maioria dos programas de análise controla visitantes únicos através do endereço IP, que é uma sequência única de números que identifica um computador ou servidor na Internet. Alguns utilizadores recebem endereços IP dinâmicos do ISP. Isso significa que o endereço IP muda diariamente, ou várias vezes durante o dia. Esses utilizadores podem distorcer o número de visitantes únicos.

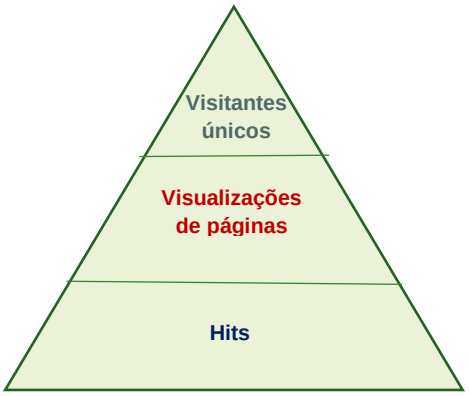
Referenciadores de site

O referenciador do *site*, ou página de referência, é o URL da página de onde o utilizador foi enviado para atingir um *web site*. Um referenciador pode ser um mecanismo de pesquisa, um *blog*, um *banner*, *email*, *site* de e-commerce, anel da *Web*, parceiro de *link* ou algum outro *site*. A ferramenta de *Web Analytics* vai registar o URL exato do *site* que referenciou o tráfego para que se possa medir o sucesso das iniciativas de construção de tráfego.

Palavras-chave e frases-chave

As pessoas usam palavras-chave e frases-chave para pesquisar produtos, serviços e informações na *Web*. Pode-se pagar aos motores de busca (*Google*, *Yahoo*, *Bing*, etc.) para que exibam anúncios nos resultados da pesquisa de um utilizador com base nas palavras-chave ou frases-chave.

Figura 1. A pirâmide da terminologia de *Web Analytics*



Fonte: Sostre e LeClaire, (2007)

A Figura 1 ilustra a forma como estas métricas estão estruturadas em termos de hierarquias.

Mas para além de todas estas métricas, existem outras, um pouco mais complexas, que fornecem resultados importantes para a tomada de decisão, tais como a taxa de conversão (*conversion rate*), a taxa de saída (*exit rate* ou *churn rate*) e a taxa de lealdade (*engagement rate*) que exploraremos de seguida.

Em comércio eletrónico, a conversão significa venda ou encomenda de um determinado bem ou serviço. A Taxa de conversão, expressa em termos de resultados a dividir pelo número de utilizadores únicos, mede a capacidade de venda de um *web site*. Esta métrica é especialmente interessante no comércio eletrónico.

A taxa de saída ou taxa de abandono (*bounce rate*) é o número de pessoas que sai do *web site* antes de executar uma tarefa importante, tal como uma conversão.

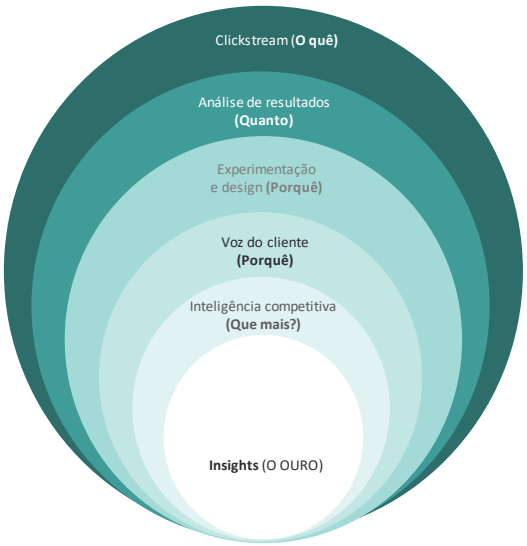
Quanto à lealdade, não existe um consenso claro quanto à forma de a medir, sendo que há quem use a duração da visita ao *web site*, a quantidade de importação de documentos, etc.

2.2. Fundamentos de *web Analytics* para as empresas

Um aspeto a salientar neste livro é a importância do *Web Analytics* para as empresas. Para que serve, como podem ser recolhidos os dados e como deve usado o *Web Analytics*?

Comecemos pelos benefícios do *Web Analytics*. Através da análise do tráfego da *web* num *web site*, é possível melhorar a experiência do utilizador no sentido de incrementar o número de visitas. Em particular, como refere Alghalith (2015), pode-se melhorar a correspondência entre os recursos disponíveis e os interesses os utilizadores; aumentar o valor de cada utilizador; promover a recolha de dados de forma inovadora; testar a relevância do conteúdo do *web site* e da respetiva arquitetura; otimizar o *web site*.

Figura 2. Forma como se pode tirar partido dos dados



Fonte: Alghalith (2015)

Estas são algumas das conclusões ou ações que se podem tomar com base na análise das métricas referidas anteriormente. De seguida, iremos explorar a forma como a recolha, tratamento e obtenção de conclusões podem geridos pelas empresas.

Que dados podem ser recolhidos?

Os dados para *Web Analytics* podem ser gerados por diversas fontes:

- Tráfego direto: corresponde aos utilizadores do *web site* que entram porque digitam diretamente o endereço, ou usam uma URL que dá acesso direto ao *web site*.
- Referenciador do *site*: outros w*eb site*s que enviam tráfego para o nosso *site* como resultados de campanhas, *blogs*, etc.
- Motores de busca: *Google*, *Yahoo*, *Bing* e outros, incluindo tráfego orgânico e pago.
- Outros: campanhas, *email*, marketing direto, etc.

Muitas empresas dispõem de CRM que lhes permite organizar a informação de interação com os clientes proveniente de diversos canais: presencial, *online* (*web*), telefone, etc. Os CRM permitem ajudar a recolher dados úteis para *Web Analytics*.

CRM (*Customer Relationship Management*)

O CRM é um tipo de sistemas de informação que se concentra na recolha, processamento e coordenação de informações sobre interações comerciais com clientes, incluindo contactos, transações comerciais, marketing, vendas e serviços. O uso do *site* de uma empresa pode fornecer diversas informações de marketing que podem ser geridas pelo CRM.

Mas os dados para *Web Analytics* também podem ser gerados por outras formas. Os dados provenientes de *social media*, por exemplo (*Clickstream*, *Facebook*, *LinkedIn*, Instragram, etc,), são importantes fontes de informação que podem ser usadas para tomada de decisão numa empresa. Alguma atenção será dada, mais à frente neste capítulo sobre este tipo de dados.

Que tipos de análises se podem fazer?

Além da análise tradicional do clisktream, existem outros tipos de análise importantes que se podem fazer com os dados da *web*. Olhando para os vários tipos de interações com os utilizadores, a medição de resultados é o mais importante, promovendo medidas que permitam aumentar as receitas, reduzir os custos e melhorar a satisfação e a lealdade do cliente. E tudo o que pudermos fazer no *site* terá a ver com estas medidas, independentemente de se tratar de um *site* de comércio eletrónico ou de *social media*.

De acordo com Kaushik (2015), existem diversos focos de análise consoante a dimensão da empresa.

Como poderemos ver no Capítulo 5 deste livro, existem diversas formas de analisar estes dados. Existem algoritmos que permitem fazer recomendações, previsões, agrupamentos de clientes.

Que decisões se podem tomar?

Um exemplo (retirado de Kaushik, 2018):

Num *web site* de comércio electrónico, não é impossível identificar micro conversões, ou seja, saber exactamente quem é que já converteu. Mas deverá ser mais importante saber qual o valor económico das conversões. A maneira como se pode calcular o valor económico consiste é descobrir o seu valor. Imaginemos uma empresa altamente qualificada. Nessa empresa existe uma *mailing list* de discussão que custa 4€. por endereço de *email*. Poderemos usar esse valor como *proxy* para cada um dos 23.000 assinantes da *mailing list*.

2.3. Ferramentas de *Web Analytics*: o Google Analytics

Hoje em dia, muitas análises se podem fazer com estes dados. Para tal, a paisagem das ferramentas de *Web Analytics* é dominada por *software* que usa principalmente dados recolhidos por ficheiros de *logs* da *web* (ver mais à frente neste capítulo) ou tags JavaScript (Kaushik, 2018). A maioria das empresas usa ferramentas do *Google Analytics*. Mas existem outras ferramentas, tais como *Omniture Site Catalyst*, Webtrends, *Clicktracks* ou *Xiti* para entender o que está acontecendo nos *web sites*.

Usar o Google Analytics

O *Google Analytics* é uma das ferramentas mais usadas de *web Analytics*. Pertence ao universo *Google* e fornece estatísticas detalhadas do tráfego da *web*, sendo usada por mais de 60% dos proprietários de *sites*. O *Google Analytics* ajuda a medir o número de visitas e fornece origens de tráfego e metas de conversão através de relatórios sobre análise de audiência, análise comportamental, e análise de conversão. De forma geral, o

Como começar a usar o *Google Analytics*?

Para começar a usar o *Google Analytics*, deve ter uma conta *Google* e aceder através de *Analytics.Google*.com.
Ou então, através do *browser* "*Chrome*", proceder da seguinte forma:

1. Instalar o *Google Publisher Toolbar* (caso não tenha este ícone, deverá instalar a partir da *Chrome Web Store*
2. No canto superior direito do *Chrome*, clique no ícone *Google Publisher Toolbar*.
3. Clique no ícone e selecione Contas.
4. Nas contas permitidas, selecione Ativar.

Google Analytics permite gerir campanhas de anúncios no *Facebook*, *Google* e outras plataformas. Em termos técnicos, o *Google Analytics* contém linguagem *javascript* em cada página do *site* que o visitante aceder. Este procedimento permite enviar os dados ao *Analytics*, que, por sua vez, os transmite ao proprietário do *web site*.

Mais concretamente, entre outras coisas, o *Google Analytics* permite:

- Analisar quais os dispositivos que as pessoas utilizam para navegar pelo *site* (telemóvel, computador *desktop*, *tablet*, etc.);
- Qual o tempo médio de cada sessão;
- Calcular quais as páginas que geram mais leads para o negócio;
- Descobrir quais as páginas do *web site* com mais ou menos visitas;
- Determinar o que os utilizadores pesquisam no *web site* e o que pesquisaram no *Google* para encontrar o *web site*.

Existem outras ferramentas que permitem fazer *web Analytics*, tais como Adobe *Analytics*, *MixPanel* e *Open Web Analytics*. Cada uma destas aplicações tem pontos positivos e funcionalidades específicas, mas o *Google Analytics*, é provavelmente a ferramenta mais completa de todas.

O *Google Analytics* funciona como um painel de controlo (ou *dashboard*) para gerir as visitas a um *web site*. Através das informações que proporciona, é possível analisar, por exemplo, o desempenho de vendas de cada produto numa loja de *e-commerce*.

Mais do que isso, é possível saber que tipo de ação de marketing digital está a gerar os melhores resultados, assim como o valor desse retorno. O *Google Analytics* tem *templates* pré-definidos de *Web Analytics* contendo as seguintes métricas principais para medir o desempenho de um *web site*, tais como taxa de rejeição (*bounce rate*), duração da sessão, Visualizações de página e Páginas por visita.

O *template* mostrado nas figuras seguintes, sob a forma de um *dashboard*, permite ter uma visão rápida do desempenho de um *web site* relacionado com um produto de *e-learning* destinado a um mercado muito específico.

Na Figura 3 podemos verificar que o *web site* foi consultado por cerca de 240 utilizadores, com picos de consultas entre 2017 e 2018. Cada utilizador visitou, em média, 2 páginas por sessão.

Na Figura 4 podemos também observar a forma como os utilizadores de telemóvel (mobile) se comportaram face aos utilizadores de computadores desktop. A taxa de rejeição em telemóvel foi nula e a duração média da sessão foi superior em telemóvel face a *desktop*.

Os valores de referência para estas métricas são os seguintes:

- Taxa de rejeição (*bounce rate* - idealmente abaixo de 60%);
- Tempo no *site* (acima de 60 segundos indica um visitante comprometido);

- Visualizações de página (quanto mais, melhor);
- Páginas por visita (quanto mais, melhor).

Além das informações anteriores, o *Google Analytics* consegue ainda fornecer informações sobre a proveniência dos utilizadores, tais como o país de origem, fornecedores de serviço, sistema operativo, etc.

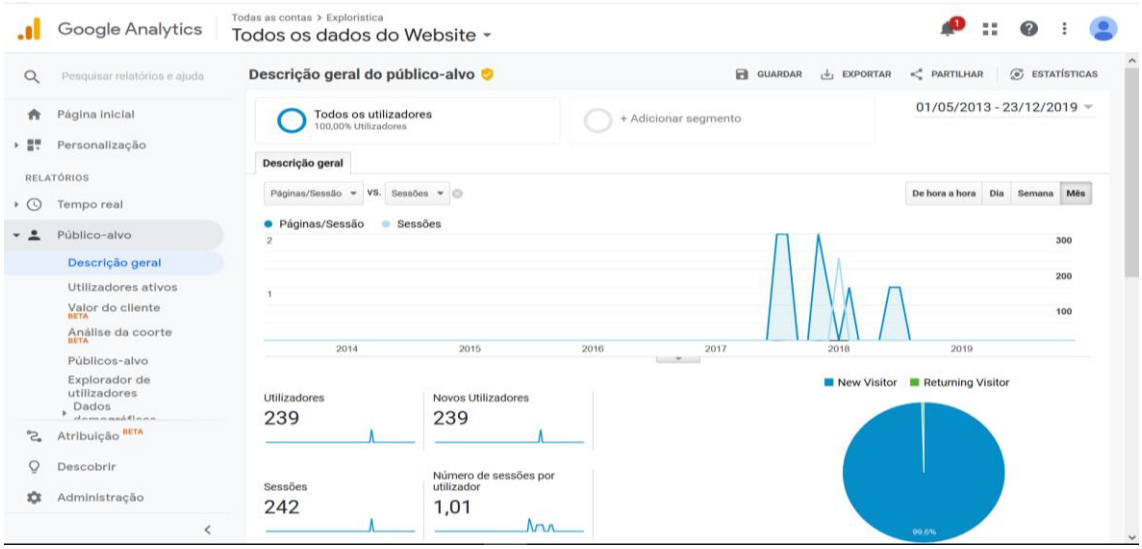
Anúncios online e o Google Analytics

As ferramentas de *Web Analytics* permitem acompanhar diversas formas de anúncios *online* tais como *banners*, *pop-ups*, campanhas de *email*, etc. Seja qual for a forma de fazer publicidade *online*, é necessário identificar antecipadamente uma ferramenta para *Web Analytics*.

Mas além destas formas de publicidade, o *Google Analytics* permite também acompanhar as palavras-chave de pesquisa num motor de busca: as *keywords*. A *Google* oferece uma ferramenta interessante - criador de URL - especificamente para os seus utilizadores do *AdWords*. Para se ter acesso, apenas é necessário preencher um formulário simples. Se a conta do *Google Analytics* estiver associada a uma conta ativa do *Google AdWords*, será possível fazer análise das palavras pesquisadas, por quem, como e quando.

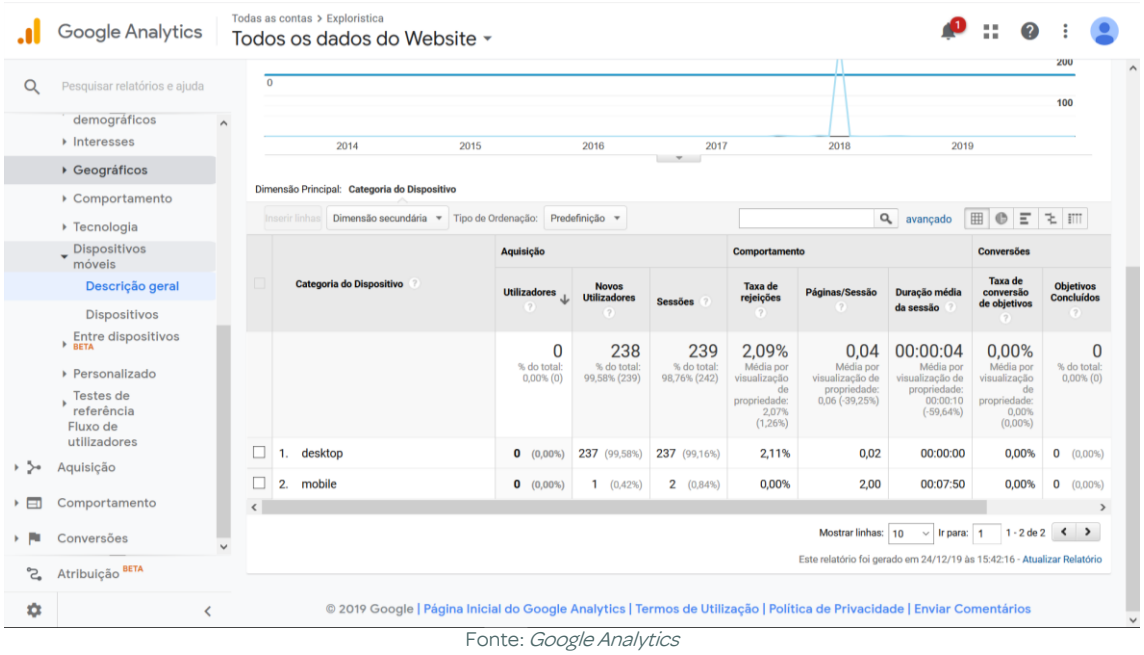
Estas abordagens fazem parte do *Search engine Marketing* (SEM), uma prática de marketing que as empresas usam para promover os seus produtos/serviços na *webpage* de resultados de um motor de pesquisa. Mais à frente neste *ebook* iremos dedicar um capítulo ao *Search engine Marketing*

Figura 3. *Dashboard* do *Google Analytics* (A)



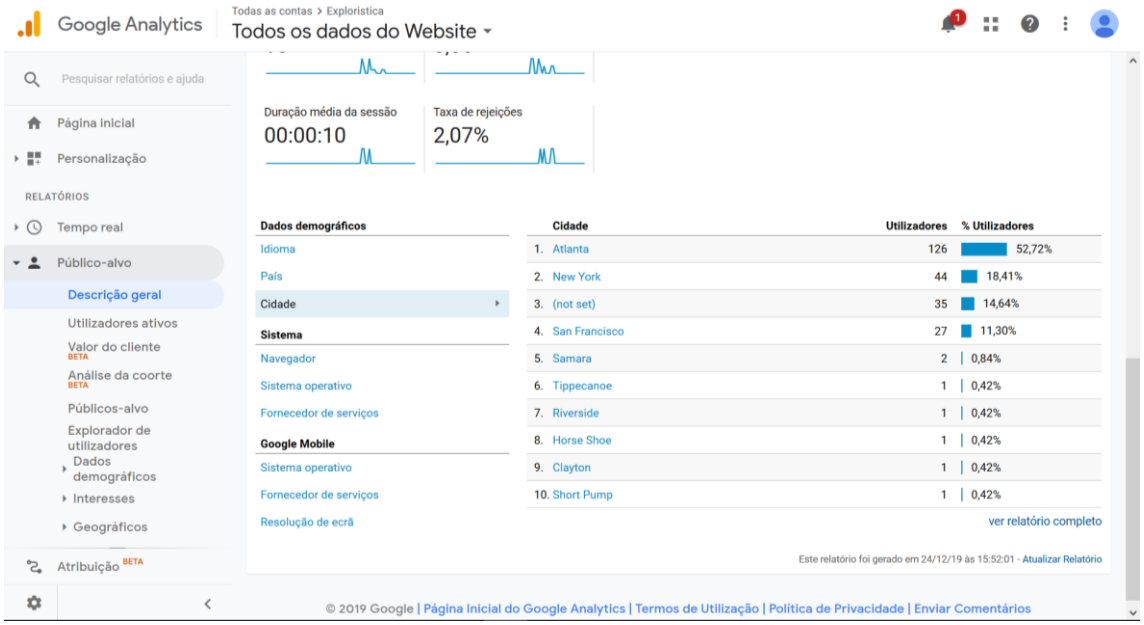
Fonte: *Google Analytics*

Figura 4. *Dashboard* do *Google Analytics* (B)



Fonte: *Google Analytics*

Figura 5. *Dashboard* do *Google Analytics* (C)



Fonte: *Google Analytics*

Anúncios *online* e o *Google Analytics*

As ferramentas de *Web Analytics* permitem acompanhar diversas formas de anúncios *online* tais como *banners*, *pop-ups*, campanhas de *email*, etc. Seja qual for a forma de fazer publicidade *online*, é necessário identificar antecipadamente uma ferramenta para *Web Analytics*.

Mas além destas formas de publicidade, o *Google Analytics* permite também acompanhar as palavras-chave de pesquisa num motor de busca: as *keywords*. A *Google* oferece uma ferramenta interessante - criador de URL - especificamente para os seus utilizadores do *AdWords*. Para se ter acesso, apenas é necessário preencher um formulário simples.

Google Ads e o Search Egine Marketing

A publicidade paga nos motores de busca *online* constitui a área do mercado da publicidade *online* com maior dinâmica. Esse espaço disponível para os anúncios pagos com base em palavras-chave (*keywords*), ou posições, é vendido através de leilões e o seu pagamento é calculado tendo em conta o número de cliques que cada posição recebe. A *Google* promove espaços para a inserção de campanhas publicitárias. É através do *Google* Ads (antigamente denominado *Google AdWords*) que essas campanhas são geridas. O *Search engine Marketing* (SEM), também chamado de *keyword advertising*, foca-se em alocar os anúncios que aparecem nos resultados de pesquisa dos motores de busca às necessidades dos utilizadores.

Se a conta do *Google Analytics* estiver associada a uma conta ativa do *Google AdWords*, será possível fazer análise das palavras pesquisadas, por quem, como e quando.

Estas abordagens fazem parte do *Search Engine Marketing* (SEM), uma prática de marketing que as empresas usam para promover os seus produtos/serviços na *webpage* de resultados de um motor de pesquisa. Mais à frente neste *ebook* iremos dedicar um capítulo ao *Search Engine Marketing*.

2.4. *Web Analytics* e a *web* social

Neste subcapítulo iremos introduzir uma forma diferente de fazer *Web Analytics* que não se baseia nas métricas definidas anteriormente (visitas, páginas visitadas, etc.). Trata-se de dados da *web* provenientes de *social media* (tais como *Clickstream*, *Facebook*, *Instagram*, *LinkedIn*, etc.). Nestes casos, utilizam-se frequentemente aplicações do tipo API (*Application Progamming Interface*) para obter os dados a partir dos quais e podem fazer vários tipos de análise.

Comecemos com o *Clickstream*, por exemplo, que pode ser visto como um serviço de *microblog* que permite que as pessoas se comuniquem com mensagens curtas que correspondem aproximadamente a pensamentos ou ideias. Historicamente, esses *tweet*s limitavam-se a 140 caracteres, embora esse limite tenha sido expandido e possa ser mudar novamente no futuro. Nesse sentido, podemos pensar no *Clickstream* como algo semelhante a um serviço global de mensagens de texto de alta velocidade. A *API to Clickstream* permite-nos obter informação sobre a rede de cada utilizador, usar medidas próprias para avaliar a sua posição na rede, etc.

O mesmo se passa com o *Facebook* ou com o *Instagram*. Ou seja, existem diversas métricas que podem ser analisadas convenientemente neste tipo de dados. Russel (2019) apresenta diversas API para lidar com estas fontes de *social media* e usa *Python* e algoritmos de *data mining* para as analisar, tais como *clustering* ou análise preditiva.

Tabela 1. Métricas que podem ser usadas em dados de *social media*

Description	Segmentation Available	
`engagements`	Total number of engagements	✓
`impressions`	Total number of impressions	✓
`retweets`	Total number of re*tweet*s	✓
`replies`	Total number of replies	✓
`likes`	Total number of likes	✓
`follows`	Total number of follows	✓
`card_engagements`	Total number of card engagements	

Fonte: Russel, 2019. Adaptado de: https://developer.*clickstream*.com/en/docs/ads/*Analytics*/overview/metrics-and-segmentation

3.
Metodologia CRISP-DM

Qualquer projeto na área de análise de dados, que inclui também a análise de dados *web*, requer a utilização de uma metodologia de desenvolvimento que permita estruturar os passos a dar. A metodologia mais comum, apesar de já ter cerca de vinte anos, continua a ser a metodologia mais utilizada na área de análise de dados. É a metodologia CRISP-DM. Este capítulo é sobre esta metodologia. Mas antes de a descrever começaremos por uma muita breve descrição de alguns conceitos relevantes na á área de análise de dados que serão úteis para melhor compreender tanto este capítulo como o Capítulo 5.

3.1. Taxonomia básica sobre análise de dados

Tabela 2. Exemplo de um conjunto de 9 instâncias caracterizadas por 5 variáveis, 4 delas preditivas e uma objetivo (a amarelo)

compri pétala	largura pétala	compri. sépala	largura sépala	variedade
5	3,3	1,4	0,2	setosa
7	3,2	4,7	1,4	versicolor
6,3	3,3	6	2,5	virginica
4,9	3	1,4	0,2	setosa
4,7	3,2	1,3	0,2	setosa
4,9	2,4	3,3	1	versicolor
6,6	2,9	4,6	1,3	versicolor
6,3	3,3	6	2,5	virginica
5,8	2,7	5,1	1,9	virginica

Os dados usados em análise de dados são, na grande maioria dos casos, dados em formato tabular. A Tabela 2 apresenta um exemplo de dados em formato tabular. Cada linha representa um lírio. Designa-se cada linha por instância ou objeto. De cada um desses lírios foram retiradas cinco características. Designam-se essas características por variáveis ou atributos. Em problemas de previsão há, tipicamente uma, mas podem ser mais, variáveis objetivo que são aquelas cujos valores se querem prever. As restantes variáveis designam-se por variáveis ou atributos preditivos. Na Tabela 2, a variável objetivo é a mais à direita, a que se encontra com uma cor diferente.

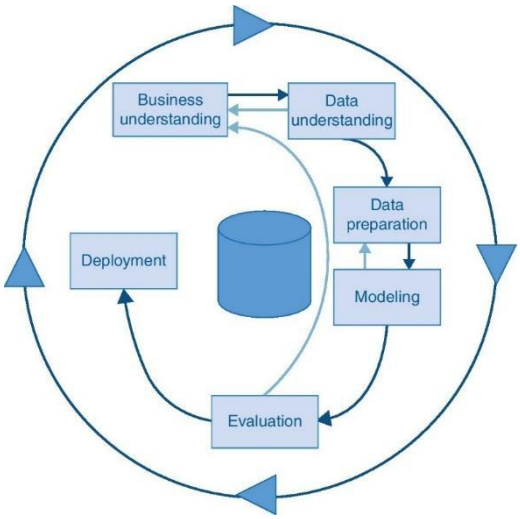
3.2. Apresentação de alto nível da metodologia CRISP-DM

CRoss-Industry Standard Process for Data Mining (CRISP-DM) é uma metodologia para o desenvolvimento de projetos na área de análise de dados. Não é específica da área de *Web Analytics*, mas de acordo com o nosso conhecimento, não há nenhuma metodologia específica de *Web Analytics*. Sendo *Web Analytics* uma subárea de análise de dados, CRISP-DM adequa-se igualmente a projetos na área de *web Analytics*. É composta por 6

fases estruturadas de modo sequencial. Apesar das seis fases, CRISP-DM é visto como um processo contínuo que acompanha em tempo real uma empresa através de sucessivas iterações. Para além disso, apesar das seis fases serem tendencialmente sequenciais, existem habitualmente passos dados atrás para corrigir algum passo mal dado anteriormente, e recomeçar desse ponto mais atrás, como se mostra na Figura 6. As seis fases são:

3. Compreensão do negócio: entender o domínio do negócio, ser capaz de definir o problema do ponto de vista do domínio empresarial e, finalmente, ser capaz de traduzir esse problema de negócios para um problema de análise de dados.
4. Compreensão dos dados: implica a coleta dos dados necessários e sua visualização inicial/sumarização para obter os primeiros insights sobre eles, ou seja, mas não exclusivamente, sobre problemas de qualidade de dados, tais como dados ausentes ou inconsistentes, *outliers*, etc.

Figura 6. A metodologia CRISP-DM



Fonte: adaptada de http://www.crisp-dm.org/

5. Preparação de dados: inclui todas as tarefas necessárias para preparar o conjunto de dados a ser alimentado pela ferramenta de modelação. A transformação de dados, a construção de variáveis, a remoção de *outliers*, o preenchimento de dados em falta ou a remoção de dados incompletos são algumas das tarefas mais comuns feitas na fase de preparação de dados.
6. Modelação: normalmente existem vários métodos para resolver o mesmo problema em análise, implicando muitas vezes algumas tarefas adicionais de preparação de dados que são específicas de cada método. Nesse caso, é necessário voltar à etapa anterior de preparação de dados. A fase de modelação inclui também ajustar os híper-parâmetros para cada um dos métodos escolhidos.
7. Avaliação: resolver o problema do ponto de vista da análise de dados não é o fim do processo. Agora é necessário avaliar a sua utilização do ponto de vista do negócio. Nessa fase, deve-se ter a certeza de que a solução obtida responde positivamente aos requisitos de negócio.
8. Implantação: a integração da solução de análise de dados no processo de negócios é o principal objetivo desta fase. Normalmente, isso implica a integração da solução obtida numa ferramenta de suporte à decisão própria, processo de manutenção do *site*, processo de relatório ou qualquer outra forma considerada adequada.

Apresentam-se, nos restantes subcapítulos deste capítulo, descrições mais alargadas de cada uma das fases da metodologia CRISP-DM.

3.3. Compreensão do negócio

A fase de compreensão de negócio tem as tarefas e respetivas entregas que se seguem:

Determinar os objetivos de negócio

O cliente, ou seja, a pessoa/instituição que lhe pagará pelo projeto ou seu chefe se for um

projeto interno, tem certamente uma boa compreensão do negócio e uma ideia clara dos seus objetivos. O objetivo desta tarefa é tomar nota dos fatores importantes que podem influenciar o resultado final. As entregas devem ser:

- *Background*: a situação sobre o negócio no início do projeto deve ser registada.
- Objetivos de negócios: quando um projeto sobre análise de dados começa, há uma motivação/objetivo por trás dele. A descrição desses objetivos deve ser registada em conjunto com todos os detalhes de negócios relacionados que possam parecer relevantes para o projeto.
- Critérios de sucesso de negócios: o sucesso de um projeto deve ser quantificado tanto quanto possível. Às vezes não é possível fazê-lo devido à natureza subjetiva do objetivo. De qualquer modo, os critérios/processo para determinar o sucesso comercial do projeto devem ser identificados.

Avaliar a situação

Uma vez que uma visão geral sobre o negócio foi feita na tarefa anterior, agora é o momento de detalhar as informações sobre os recursos existentes, restrições, pressupostos, requisitos, riscos, contingências, custos e benefícios. Esta tarefa é toda sobre ela. As entregas são:

- Inventário de recursos: o tipo de recursos relevantes para um projeto de análise de dados são principalmente recursos humanos e computacionais. Repositórios de dados, como bases de dados ou armazéns de dados, sistemas de informação, computadores e outros tipos de *software* são tudo recursos computacionais relevantes.
- Requisitos, pressupostos e restrições: normalmente há requisitos no calendário do projeto e nos resultados, também requisitos legais e de segurança, entre outros. Durante este tipo de projetos, é comum ser necessário fazer suposições sobre, por exemplo, a disponibilidade de dados numa determinada data, alguma mudança esperada no negócio dependente de medidas políticas, entre outros. Todos eles devem ser identificados e registados. Podem existir restrições na disponibilidade e usabilidade dos dados, no tipo de *software* que pode ser usado ou em restrições computacionais para computação de alto desempenho, por exemplo.
- Riscos e contingências: sempre que um risco é identificado, deve-se definir um plano de contingência. Um risco típico é algum tipo de dependência de terceiros que pode atrasar o projeto.
- Terminologia: um glossário de termos tanto para a área de negócios como para a área de análise de dados.
- Custo e benefícios: listar os custos esperados e os benefícios esperados do projeto, preferencialmente de forma quantificada.

Determine os objetivos da análise de dados

O objetivo é passar o problema de uma linguagem de negócios para uma linguagem técnica. Por exemplo, se o objetivo do negócio é "aumentar a fidelidade dos clientes", o objeto de análise de dados pode ser "prever clientes *churn*". As entregas são:

- Objetivos de análise de dados: descrever como os resultados de análise de dados são capazes de responder aos objetivos de negócios. No exemplo anterior, como é que a previsão de clientes *churn* permite aumentar a fidelidade dos clientes.
- Critérios de sucesso de análise de dados: para identificar os critérios para os quais o resultado de análise de dados é considerado bem-sucedido. Usando novamente o mesmo exemplo, um possível critério de sucesso seria prever clientes *churn* com uma precisão de, pelo menos, 60 %.

Produzir o plano do projeto

Fazer o plano necessário para o projeto. As entregas são:

- Plano do projeto: apesar da famosa citação de Dwight D. Eisenhower, "Planos não são nada; planeamento é tudo", é importante fazer um bom plano inicial. Deve conter todas as tarefas a serem feitas, sua duração, recursos, entradas, saídas e dependências. Um exemplo de dependência é, por exemplo: a preparação de dados deve ser feita antes da fase de modelação. Este tipo de dependências são muitas vezes uma causa de riscos devido a atrasos de tempo. Quando há evidências de riscos, as recomendações de ação devem ser escritas no plano. O plano deve descrever as tarefas de cada fase até à fase de avaliação. No final de cada fase, deve ser agendada uma revisão do plano. Na verdade, Eisenhower estava certo!
- Avaliação inicial de ferramentas e técnicas: uma seleção inicial de métodos e ferramentas deve ser feita porque as próximas fases, especialmente a preparação de dados, podem depender de sua escolha.

3.4. Compreensão dos dados

A fase de compreensão de dados tem as seguintes tarefas e respetivas entregas:

Coletar dados iniciais

Os dados para uma inspeção inicial devem ser coletados dos recursos do projeto previamente identificados. Muitas vezes, consultas SQL são usados para fazê-lo. Quando os dados vêm de várias fontes, é necessário integrá-los de alguma forma. Isso pode ser bastante caro. A entrega desta tarefa é:

- Relatório inicial de coleta de dados: a identificação das fontes de dados e todo o trabalho necessário para coletá-los, incluindo todos os aspetos técnicos. Por exemplo, a consulta SQL usada se foi o caso, ou todas as etapas feitas a fim juntar os dados das fontes diferentes, se foi esse o caso.

Descreva os dados

Coletar informações básicas sobre os dados. A entrega é:

- Relatório inicial de coleta de dados: os dados geralmente vêm numa ou mais tabelas de dados. Para cada tabela, o número de instâncias selecionadas, o número de atributos e o tipo de dados de cada atributo deve ser registado.

Exploração dos dados

Veja os dados! Mas deve fazê-lo com os métodos certos. Métodos descritivos de análise de dados são adequados para esta tarefa. A entrega é:

- Relatório de exploração de dados: onde as coisas relevantes descobertas ao explorar os dados são relatadas. Poderia, e em alguns casos deveria, incluir gráficos a fim apresentar visualmente algumas evidências sobre os dados que valem a pena relatar.

Verifique a qualidade dos dados

O objetivo é identificar e quantificar a existência de (1) dados incompletos quando apenas parte do domínio existe, por exemplo, dados de apenas uma propriedade quando o objetivo é estudar dados de toda a região, (2) valores ausentes, ou (3) erros nos dados, como uma amostra de vinho com 110 graus de álcool! A entrega é:

- Relatório de qualidade de dados: onde descobertas relevantes que se fizeram ao verificar a qualidade dos dados são relatadas. Não só os problemas encontrados com os dados devem ser relatados, mas também possíveis soluções para resolvê-los. Normalmente, as soluções para esse tipo de problema implicam um forte conhecimento tanto sobre o assunto de negócios quanto sobre a análise de dados.

3.5. Preparação dos dados

A fase de preparação de dados tem as seguintes tarefas e respetivas entregas:

Selecione os dados

Com base na sua relevância para os objetivos do projeto, a qualidade dos dados e a existência de restrições técnicas, como o volume de dados ou os tipos de dados, os dados são selecionados tanto em termos de variáveis como de instâncias. A entrega é

- Razão para inclusão/exclusão: onde o racional usado para selecionar os dados é descrito.

Limpe os dados

Os métodos que devem ser usados na fase de modelação podem implicar regras mais diretas para a qualidade dos dados. Normalmente, os subconjuntos de dados sem valores ausentes podem ser selecionados, ou técnicas para preencher valores em falta ou remover valores atípicos podem ser aplicadas. A entrega é:

- Relatório de limpeza de dados: descreve como os problemas identificados no relatório de qualidade dos dados da fase de compreensão de dados foram abordados. O impacto das transformações realizadas durante a tarefa de limpeza de dados sobre os resultados na fase de modelação deve ser considerado.

Construir dados

Construção de: (1) novos atributos, (2) novas instâncias, ou (3) valores transformados transformando, por exemplo, uma variável Booleana numa variável numérica 0/1. A entrega é:

- Atributos derivados: são atributos que são obtidos fazendo algum tipo de cálculo a partir de outros atributos existentes. Um exemplo é obter um novo atributo chamado "tipo de dia" com três valores possíveis, "Sábado", "Domingo" ou "dia de trabalho" de outro atributo do tipo *timestamp* (com a data e a hora).
- Gerar registos: novos registos/instâncias são criados. Isso pode ser usado, por exemplo, para gerar instâncias artificialmente de um determinado tipo como uma abordagem para lidar com conjuntos de dados desequilibrados, um problema comum em problemas de classificação.

Integrar dados

Para ter os dados em formato tabular, muitas vezes é necessário integrar dados de diferentes tabelas. A entrega é:

- Juntar dados: um exemplo é a integração de dados referentes a uma parcela, por exemplo, consumo de água por parcela, com dados referentes à propriedade toda, por exemplo receita obtida com a venda de vinho. Ou se junta os consumos de água de todas as parcelas da propriedade ou se distribui de algum modo a receita com o vinho pelas diferentes parcelas.

Formato dos dados

Refere-se a transformações feitas para os dados sem transformar o seu significado, mas que são necessários devido aos requisitos da ferramenta de modelação. A entrega é:

- Dados reformatados: algumas ferramentas têm pressupostos específicos. Por exemplo, a necessidade do atributo a prever ser o último. Existem outros pressupostos.

As entregas da fase de preparação de dados

- Conjunto de dados: um ou mais conjuntos de dados a serem usados na fase de modelação.
- Descrição do conjunto de dados: descreve os conjuntos de dados que serão usados na fase de modelação ou o trabalho de análise principal do projeto.

3.6. Modelação

A fase de modelagem tem as seguintes tarefas e respetivas entregas:

Seleção da técnica de modelação

Na fase de compreensão de negócios, os métodos ou, para ser mais preciso, as famílias dos métodos a serem utilizados já foram identificados. Agora, é necessário escolher quais os métodos específicos que serão utilizados. Como exemplo, enquanto na fase de compreensão de negócios escolhemos, por exemplo, árvores de decisão, agora precisamos decidir se queremos usar CART, C5.0 ou outra técnica.

- Técnica de modelação: descrição da técnica a ser usada.
- Pressupostos dos modelos: vários métodos têm pressupostos sobre os dados, tais como, inexistência de valores em falta, não existência de *outliers*, inexistência de atributos irrelevantes, ou seja, atributos que não são úteis para a tarefa de previsão. Todas os pressupostos existentes devem ser descritos.

Geração do setup experimental

A definição da configuração experimental deve ser feita. Especialmente importante para a análise preditiva de dados, a fim de evitar o sobre ajustamento do modelo.

- *Setup* experimental: descreva a configuração experimental planeada. As abordagens de reamostragem devem levar em consideração a quantidade existente de dados, o quão desequilibrado o conjunto de dados é, ou se os dados chegam como um fluxo contínuo.

Construção do modelo

Use o método para obter um ou mais modelos quando o problema é preditivo ou para obter a descrição desejada quando o problema é descritivo.

- Definição dos valores dos parâmetros de entrada: cada método tem tipicamente vários parâmetros de entrada. Os valores utilizados para cada parâmetro e o processo para defini-los devem ser descritos.
- Modelos: os modelos ou resultados obtidos.
- Descrição do modelo: descrição dos modelos ou resultados, tendo em atenção o quão interpretáveis eles são.

Avaliação do modelo

Normalmente vários modelos/resultados são gerados. É necessário, então, classificá-los de

acordo com uma medida de avaliação escolhida. Essa avaliação é feita principalmente a partir do ponto de vista da análise de dados, mesmo que algumas considerações comerciais também sejam consideradas.

- Avaliação do modelo: resumir os resultados desta tarefa, listar as qualidades dos modelos gerados (por exemplo, em termos de precisão) e classificar a qualidade comparativa dos modelos entre si.
- Configurações de parâmetros revisitado: rever as configurações dos parâmetros, se necessário, de acordo com a avaliação do modelo feita. Isso será usado em futuras iterações na construção do modelo. As iterações entre a construção e a avaliação do modelo para quando o analista de dados entender não haver necessidade de novas iterações.

3.7. Avaliação

A fase de avaliação tem as tarefas e respetivas entregas que se seguem.

Avalie os resultados

Para determinar se a solução obtida atende aos objetivos de negócio e para avaliar possíveis não conformidades do ponto de vista do negócio. Se possível, o teste do modelo num cenário de negócios real seria importante. No entanto, este tipo de solução nem sempre é possível e, sendo possível, os custos para o teste devem ser ponderados.

- Avaliação dos resultados de análise de dados: descrever os resultados de avaliação do ponto de vista empresarial, incluindo uma declaração final sobre se os resultados obtidos atende aos objetivos de negócio inicialmente definidos.
- Modelos aprovados: os modelos que foram aprovados.

Processo de revisão

É o momento de rever todo o trabalho de análise de dados para verificar se ele se enquadra nos requisitos de negócio.

Revisão do processo: resumir o processo de revisão destacando o que está em falta e o que deve ser repetido.

Determine os próximos passos

Em face do processo de revisão, devem ser decididos os próximos passos, passar para a fase de implantação, repetir algum passo voltando para uma fase anterior ou iniciando novos projetos. Tal decisão também depende da disponibilidade de orçamento e de recursos.

- Lista de possíveis ações: liste possíveis ações e para cada ação seus prós e contras.
- Decisão: descrever sobre a decisão de como prosseguir e a lógica por trás disso.

3.8. Implantação

A fase de implantação tem as tarefas e respetivas entregas que se seguem:

Plano de implantação

A estratégia de implantação deve considerar a avaliação dos resultados feitos na fase de avaliação.

- Plano de implantação: resume a estratégia de implantação e descreve o procedimento para criar os modelos/resultados necessários.

Plano de monitoramento e manutenção

Ao longo do tempo, o desempenho dos métodos de análise de dados pode mudar. Por essa razão, é necessário definir uma estratégia de monitoramento, de acordo com o tipo de implantação, e uma estratégia de manutenção.

- Monitorando e plano de manutenção: escreva o plano de monitoramento e manutenção passo a passo, se possível.

Produção do relatório final

Um relatório final está escrito. Pode ser uma síntese do projeto e suas experiências ou uma apresentação abrangente dos resultados de análise de dados.

- Relatório Final: é uma espécie de dossiê com todas as entregas anteriores resumidas e organizadas.
- Apresentação final: é a apresentação da reunião final do projeto.

Revisão do projeto

Uma análise dos pontos fortes e fracos do projeto.

- Documentação da experiência: a revisão está escrita. Deve incluir todas as experiências particulares em cada fase do projeto. Tudo o que possa ajudar futuros projetos de análise de dados.

4.
Recolha de dados

A recolha de dados é uma etapa fundamental em *Web Analytics*. A forma como os dados são recolhidos tem um grande impacto na qualidade dos dados finais. Existem dois tipos de técnicas de recolha de dados: as documentais e não documentais. Nas técnicas documentais o objetivo é a recolha de informação a partir de fontes ou suportes já existentes, é o caso dos dados provenientes de sensores de temperatura em propriedades agrícolas, ou da *World Wide Web*, tais como informação das páginas *web*, textos, etc. Nas técnicas não documentais o investigador realiza observação direta (como por exemplo, o comportamento do crescimento da vinha ou o número de dias em que chove por mês) ou indireta - podendo ser feita, neste caso, através da administração de um questionário.

Com o advento da Internet, começou a ser importante recolher dados através desta via. Existem várias formas de recolher dados da Internet, desde a extração de textos nas páginas *web* (*web* crawling), à possibilidade de instalação de mecanismos para obter informações dos utilizadores (*cookies*), até à recolha de dados para efeitos de caracterização dos utilizadores de um *web site* (*logs* e estatísticas dos motores de busca).

4.1. *Web* Crawling

O *web crawler* ou *bot* é um algoritmo utilizado para analisar o código de um *website* e recolher informações. Estas informações permitem classificar os dados e gerar novo conhecimento. Estes *bots* são usados pelos motores de busca como o *Google*, *Bing* ou outros de modo a apresentar a informação mais relevante aos utilizadores.

Estes algoritmos percorrem toda a *web* recolhendo informações de todas as páginas dos *site*s de forma a apresentarem os resultados mais relevantes cada vez que alguém pesquisa num motor de busca.

Existem também ferramentas que ajudam os gestores a analisar o *site* da empresa de forma a obterem melhorias, alguns exemplos:

- *Oncrawl* - *bot* que realiza auditorias de SEO no *site*.
- *Dyno Mapper* - permite criar automaticamente mapas do *site*.
- *Screaming frog* - que lhe permite melhorar o SEO no *site*.
- *Apify* - permite-lhe analisar a concorrência e tomar decisões importantes no *site*.

Porque é importante o rastreamento *web* nos negócios?

Através deste rastreamento pelos *bots* é possível saber o que é dito sobre a empresa nas notícias, em fóruns, *blogs*, pelos clientes, nas redes sociais, ou outros *site*s.

Vantagens do rastreamento web

1. Informação competitiva

Para saber quais os novos produtos ou serviços oferecidos pela concorrência assim como para comparar preços, um rastreador *web* pode ser muito útil uma vez que lhe

permite recolher essas informações em tempo real e assim poderá analisar e tomar as melhores decisões.

Ao estudar e analisar a concorrência tem acesso a informação competitiva que lhe permite melhorar também os resultados no seu *site,* assim como perceber oportunidades que podem ser exploradas, além de poder encontrar erros que pode transformar em vantagens competitivas.

O objetivo não é copiar a estratégia dos concorrentes, mas sim ter a informação necessária para melhorar e adaptar as suas estratégias.

2. SEO e criação de conteúdo

O *Google* é um dos principais motores de busca e fontes de aquisição de visitas para a maioria dos *site*s. Por isso, é importante conseguir tirar o melhor proveito desse tráfego e gerar vendas.
Através dos *web crawlers* é possível perceber quais as otimizações que podem ser feitas a nível de SEO para melhorar a experiência do utilizador e o *site* ser encontrado mais facilmente nos motores de busca.

São mais de 200 os fatores que influenciam o ranqueamento do *Google*, desde *tags*, estrutura do código, navegação, velocidade de carregamento, dispositivos móveis, e muito mais, que poderá analisar com um *software* de rastreamento para que aumente o tráfego e conversões no seu *site*.

A criação de conteúdos é também um elemento fundamental numa estratégia de marketing digital. Quer seja em *site*s, *blogs* ou redes sociais, a qualidade e relevância do conteúdo deve ser a principal preocupação.

Com um rastreador *web* saberá se o conteúdo que produz tem bom desempenho ou se é necessário otimizar a nível de SEO para atrair e converter melhor o seu público.

Atrair tráfego qualificado é fundamental para qualquer estratégia de marketing uma vez que sem visitas ao *site* e sem retorno sobre o investimento a sua empresa dificilmente sobrevive.

Conforme vai aplicando as otimizações em SEO, também a experiência do utilizador melhora, e o tráfego orgânico irá começar a aumentar.

Ao aplicar esta estratégia os bons resultados irão gerar melhores resultados, e cada otimização gera novas oportunidades de negócio.

3. Monitorizar a reputação da empresa

Mais que analisar os que os concorrentes fazem, é importante descobrir o que clientes, parceiros, colaboradores ou outros dizem sobre a empresa. O rastreamento *web* é possível através de bots que analisam informações deixadas em *site*s, *blogs*, fóruns, redes sociais, entre outros, sobre a sua empresa/marca.

É possível também analisar o sentimento dos utilizadores quando comentam por exemplo nas redes sociais sobre a marca. Assim a empresa saberá o que melhorar e como ir ao encontro das exigências do consumidor.

4.2. Cookies

Quando visita um *site* através do computador ou dispositivos móveis é instalado um pequeno arquivo de texto (*cookie*) que permite durante algum tempo memorizar as ações ou preferências que tem no *site*, para que quando o volte a visitar não tenha que as definir novamente. Os *cookies* armazenam dados como idioma escolhido, nome do utilizador, tamanho dos caracteres, opções de guardar a palavra-passe, entre outros.

Os *cookies* são específicos de cada navegador, ou seja, se armazenar *cookies* no *Firefox* estes não serão mantidos no *Google* Chrome, por exemplo.

Quais os tipos de cookies que existem?

Existem diversos tipos de *cookies* de acordo com a finalidade a que se destinam, por exemplo:

- ***Cookies* sociais**: são utilizados para permitir que o utilizador partilhe o conteúdo dos *sites* nos seus perfis nas redes sociais.
- ***Cookies* publicitários**: permitem direcionar os anúncios do *site* em função dos interesses de cada utilizador, limitando a quantidade de vezes que este vê um anúncio no *site*.
- ***Cookies* analíticos**: são utilizados anonimamente para analisar estatísticas e melhorar o conteúdo do *site*.

Por norma, os *cookies* são utilizados para melhorar a experiência do utilizador no *site* e evitar a repetição de informação.

Vantagens dos cookies

Os *cookies* não ficam armazenados no servidor, mas sim no computador do utilizador, uma vez que tem um reduzido tamanho não ocupam muito espaço na memória.

É possível configurar os *cookies* de modo a expirarem quando a sessão terminar ou definir um período específico. A principal vantagem dos *cookies* é o tempo de armazenamento uma vez que pode persistir no navegador por dias, meses ou até anos e isso facilita a navegação do utilizador que escusa de repetir informação.

Os *cookies* não só guardam os *site*s visitados, mas também informações de formulários o que permite o rápido preenchimento destes.

Para fins de marketing são muitas as empresas a recolherem *cookies*, uma vez que estes permitem segmentar para grupos de produtos, localização, termos de pesquisa e dados demográficos.

Desvantagens dos cookies

Os *cookies* por vezes não são considerados seguros pois são guardados como texto e não são criptografados, pelo que informação sensível não deve ser armazenada como *cookie*. Isto pode representar um possível risco à segurança, porque qualquer pessoa pode abrir e adulterar os *cookies*.

Os *cookies* não podem armazenar informações complexas, pois estão limitados a informações simples em cadeia. Existem muitas limitações no tamanho do texto do *cookie*. O *cookie* individual pode conter uma quantidade muito limitada de informações (não mais que 4 kb).

O utilizador tem a opção de desativar os *cookies* na configuração do navegador em resposta às preocupações de segurança ou privacidade. No entanto, poderá existir problemas para *site*s que os exigem e os *cookies* não funcionam.

Com a ativação de *cookies* os navegadores mantêm o controlo de todos os *site*s que visitou. Também *site*s terceiros podem aceder às informações armazenadas por esses *cookies*. Esses *site*s podem ser anunciantes, ou até outros utilizadores.

Como controlar os cookies?

É possível controlar e/ou apagar os *cookies* se assim pretender. O utilizador pode apagar todos os *cookies* já instalados computador ou dispositivo móvel, ou ativar uma opção disponível nos navegadores de internet que impede a sua instalação. Se preferir apagar ou bloquear a instalação dos *cookies*, poderá ter que configurar manualmente algumas preferências sempre que visitar um *site*. Para mais informações, consulte *aboutcookies.org*.

Cuidados a ter no uso de Cookies

Os *cookies* devem ser usados com cuidado em computadores de utilização partilhada principalmente quando acede a contas de *emails* uma vez que a *password* pode ficar guardada e outro utilizador pode aceder à sua conta. Recomenda-se que apague os *cookies* após usar computadores partilhados.

4.3. *Logs*

Na computação, um ficheiro de log regista eventos que ocorrem num *software* que está a ser executado. Para pesquisa na *Web*, um *log* de transações é um registo eletrónico de interações que ocorreram durante um episódio de pesquisa entre um mecanismo de pesquisa na *Web* (*browser*) e os utilizadores que procuram informações nesse mecanismo de pesquisa na *Web*. Os analistas conseguem usar essa informação e produzir relatórios.

Um log do servidor da *web* é um diário registado num servidor *web* que contém '*hits*' ou registos de todas as solicitações que o servidor recebeu. Os dados recebidos são armazenados anonimamente e incluem detalhes como a hora e a data em que a solicitação foi feita, o endereço IP da solicitação, o URL/conteúdo solicitado e o *user agent* do navegador. Esses ficheiros existem normalmente para auditoria e solução de problemas técnicos do *site*, mas também podem ser extremamente valiosos para a auditoria de SEO (*Search Engine Optimization*).

A aparência de um ficheiro de *logs* depende do tipo de servidor, mas há elementos comus entre todos os ficheiros, tais como os seguintes:

Server IP
User-agent
Timestamp (date & time)
HTTP status code
Method (GET / POST)
Requested URL (aka: URL stem + URL query)
Referrer (the external site, such as Google, from which a user arrives)

Os ficheiros de *logs* são muito úteis em *Analytics*. A *Amazon.com* disponibiliza parte dos seus *logs* para análise aberta dos seus utilizadores. Pode-se, por exemplo, construir uma rede de produtos com base em *reviews* com base nos dados da *Amazon.com* (He & McAuley, 2016). A base de dados disponível contém mais de 150 milhões de *reviews* em produtos de várias categorias, que vão desde "livros e tecnologia" a "artigos de beleza", abrangendo avaliações entre maio de 1996 a julho de 2014.

Ao comparar os resultados da análise da *Web* com os arquivos de *log*, é importante saber por que certas discrepâncias (às vezes realmente enormes) aparecem nos relatórios. O motivo de tais discrepâncias depende das métricas utilizadas pelas análises, que são diferentes das métricas usadas para analisar os arquivos de *log*. A principal diferença entre os dois é que o *Analytics* comum usa dados do lado do cliente para recolher as informações, enquanto para arquivos de *log*, eles contêm informações do servidor.

Essa distinção importante leva a resultados que muitas vezes são diferentes ou fornecem diferentes pistas. De facto, é importante analisar os dados com relação à forma como os dados foram recolhidos. O *Google Analytics*, por exemplo, não regista a atividade dos *web crawlers*. Os arquivos de log registam todas as ocorrências e todos os ficheiros solicitados, independentemente de quem os solicitar. Mas é preciso observar que, se o agente do usuário estiver *logado* no arquivo de *log*, é possível filtrar essas solicitações.

É por isso que é importante usar uma boa ferramenta para analisar os arquivos de *log*.
A *Analytics* usa *cookies* para armazenar as informações da fonte de tráfego. O endereço da Internet e o referenciador também podem extrair todas as informações importantes. Os

arquivos de log, por outro lado, se protegidos, não representam nenhum desses riscos. É por isso que alguns usuários excluem o histórico de navegação e o *cache* de *cookies* na saída do navegador.

5.
Técnicas de modelação de dados

A modelação de dados remete-nos para o tema de ciência de dados que inclui técnicas oriundas de diferentes áreas do conhecimento tais como estatística, inteligência artificial, bases de dados, entre outras. Não se pretende, neste livro, cobrir todo este vasto tema. Há outros livros que o fazem. No entanto, para se falar de análise de dados na *web*, não se pode deixar de abordar os diferentes tipos de aprendizagem existentes. Assim, este capítulo é constituído por quatro secções, o primeiro dos quais faz uma breve taxonomia dos tipos de aprendizagem existentes, fazendo-se nas três secções seguintes a descrição de 3 temas especialmente significativos para lidar com dados *web*.

5.1. Tipos de aprendizagem

A área de ciência de dados é, hoje em dia, classificada em quatro categorias:

- Análise descritiva
- Análise de diagnóstico
- Análise preditiva
- Análise prescritiva

Cada uma destas categorias cobre um conjunto vasto de técnicas que tentam responder a diferentes questões que são descritas de seguida.

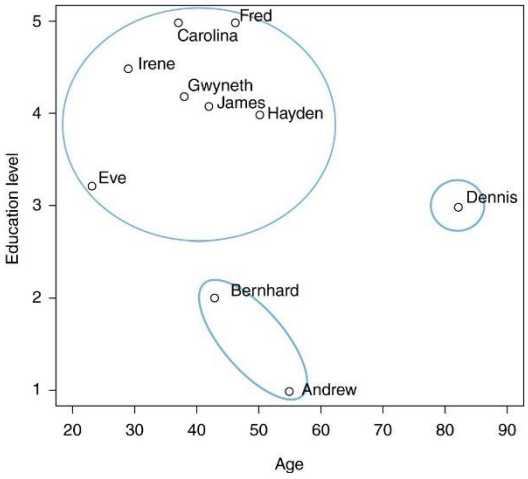
Análise descritiva

A análise descritiva tem como objetivo descrever os dados de modo que sejam entendíveis por nós, humanos. Descreve-se sucintamente algumas das técnicas mais utilizadas de descrição de dados.

A descrição de dados tipicamente remete para alguma forma de sumarização desses dados. A estatística descritiva é uma área da estatística que disponibiliza um conjunto de técnicas que permite descrever amostras de dados. Essas técnicas podem-se sintetizar em: contagens de frequências, estatísticas e gráficos. Existe uma grande variedade de estatísticas e, sobretudo, de gráficos que permitem formas muito apelativas de descrever os dados.

Técnicas de segmentação de dados, mais conhecidas como técnicas de *clustering*, tentam encontrar grupos naturais que existem nos dados. Cada cluster é um grupo de dados em que as instâncias nele contidas são semelhantes entre si e mais dissemelhantes em relação às instâncias contidas nos outros clusters (Figura 7). A técnica de *clustering* mais conhecida é a k-médias mas há muitas outras técnicas para além dessa.

Figura 7. Resultado do método das k-médias, com k=3, usando 14 instâncias com 2 variáveis (idade e nível de educação). Dados artificiais



A análise de padrões frequentes caracteriza-se por usar dados em que cada linha representa uma transação, que pode ser, por exemplo, uma sessão numa loja digital na internet, e cada coluna representa um dos possíveis produtos a adquirir. Cada célula poderá ter, por exemplo a quantidade de cada produto comprado em cada sessão. Existem duas famílias principais de métodos a considerar:

- Análise de padrões frequentes: onde se procuram os conjuntos de itens (produtos, no exemplo) que são frequentemente comprados juntos numa mesma transação (sessão de acesso na loja digital, no exemplo);
- Regras de associação: são regras que indicam que quem compra um determinado conjunto de itens também tende e comprar outro conjunto determinado de itens, em que o primeiro conjunto se designa por antecedente e o segundo por consequente.

Análise de diagnóstico

A análise de diagnóstico tem como objetivo responder à questão de porque é que uma determinada coisa aconteceu. Por exemplo, porque é que em 2019 houve um aumento grande de produção de vinhos brancos DOP do Tejo face aos anos anteriores?

A resposta a esta e outras questões envolvem, tipicamente, a utilização de novas variáveis que permitam explicar o acontecimento em causa que a análise descritiva apenas consegue identificar.

Técnicas de descoberta de subgrupos, que permitam entender quais os subgrupos de instâncias que se diferenciam das restantes instâncias e o que é que as caracteriza. Técnicas preditivas que sejam interpretáveis, i.e., que permitam perceber a relação entre os valores das variáveis preditivas e os valores da variável objetivo. Técnicas que permitam perceber relações entre a variação da variável de interesse com outras variáveis que permitam justificar a referida variável de interesse. Importa, nestas análises, estar bem ciente que a correlação entre duas variáveis não garante necessariamente a existência de causalidade – consequência entre as duas variáveis.

Análise preditiva

Análise preditiva é uma área com muitas aplicações possíveis, nomeadamente na área agrícola. Previsão da produtividade de uma dada cultura, estimativa do índice de área foliar, previsão da data de início de floração, previsão da variedade de uma planta tendo em conta a dimensão e formato das folhas, são apenas alguns exemplos.

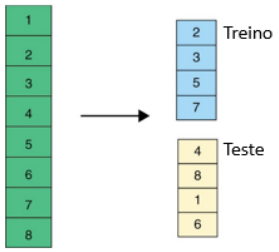
Os conjuntos de dados usados para construir modelos de previsão caracterizam-se por ter uma variável, designada por variável objetivo, que é a variável cujos valores se pretendem prever. Os problemas de previsão designam-se por problemas de classificação quando a variável objetivo é qualitativa nominal (como é exemplo a previsão da variedade de uma planta) e por regressão quando a variável objetivo é quantitativa (como são exemplos a previsão da produtividade de uma dada cultura, a estimativa do índice de área foliar, ou a previsão da data de início de floração). Em problemas de classificação, os diferentes valores que a variável objetivo pode ter designam-se por classes. Há também regressão ordinal (também designada por classificação ordinal) quando o tipo da escala

em que os valores da variável objetivo são expressos é qualitativa ordinal.

Estimação do erro de previsão e medidas de erro

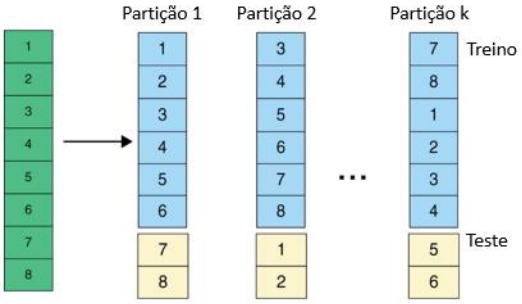
Uma das questões específicas dos problemas de previsão é o da estimação do erro de previsão. Como é que se pode estimar qual o erro de previsão de um método preditivo? A resposta a esta questão é complexa e requer a compreensão do conceito de modelo e de treino. Um modelo é um objeto matemático que resulta da aplicação de um método a um conjunto de treino. O resultado desse treino é um modelo que, dadas novas instâncias, é capaz de prever os valores da variável objetivo dessas instâncias. Para se poder estimar o erro de previsão do modelo obtido em dados que não foram usados para treino, usam-se para testar esse modelo, instâncias com valores conhecidos da variável objetivo, mas que não tenham sido usados para treinar esse mesmo modelo. Só assim é possível estimar o erro preditivo de um modelo por comparação entre o valor previsto e o valor da variável objetivo. Há várias técnicas para separar as componentes de treino e de teste, conhecidas como técnicas de reamostragem. De entre elas destacam-se a técnica *hold-out* (Figura 8) e a validação cruzada (Figura 9).

Figura 8. Técnica de reamostragem *hold-out*



Uma percentagem definida pelo utilizador das instâncias (50% no exemplo) é selecionada aleatoriamente para treino e as restantes instâncias são usadas para teste.

Figura 9. Técnica de reamostragem validação cruzada com k partições



O conjunto das instâncias é dividido aleatoriamente em k grupos de tamanhos semelhantes. Haverá k iterações, sendo um modelo treinado e testado por iteração. Em cada uma das iterações um grupo diferente é usado para teste e os restantes são usados para treino.

Para medir o erro de previsão, há diferentes medidas de avaliação dependendo de o problema ser de regressão ou de classificação. As medidas de avaliação para classificação são, na sua grande maioria, retiradas da matriz de confusão.

Figura 10. Matriz de confusão para um problema binário (duas classes) e algumas das medidas de avaliação mais utilizadas

		Verdadeiro: Positivo	Verdadeiro: Negativo
Previsão	Positivo	Positivo Verdadeiro PV	Positivo Falso PF
	Negativo	Negativo Falso NF	Negativo Verdadeiro NV

$$acerto = \frac{PV + NV}{PV + NF + PF + NV}$$

$$precisão = \frac{PV}{PV + PF}$$

$$sensibilidade = \frac{PV}{PV + NF}$$

$$especificidade = \frac{NV}{NV + PF}$$

As medidas mais comuns que se retiram da matriz de confusão são a percentagem de acerto, a precisão, a sensibilidade e a especificidade tal como se mostra na Figura 10.

As medidas de avaliação para regressão são maioritariamente baseadas no erro quadrático médio, i.e., a média do quadrado da diferença entre o valor previsto e o valor real.

Métodos preditivos

É grande o número de técnicas preditivas existentes tanto para regressão como para classificação. Apresenta-se aqui uma lista de algumas das técnicas mais conhecidas. Sugere-se que o leitor consulte outros livros onde essas técnicas são descritas com detalhe (Gama et al., 2015).

Fica aqui apenas a referência dos nomes de algumas das técnicas preditivas mais conhecidas:

- Regressão Linear Multivariada
- K–vizinhos mais próximos
- Métodos Bayesianos
- Árvores de decisão
- Máquinas com vetores de suporte
- Redes Neuronais Artificiais & *Deep learning*
- Florestas aleatórias
- *AdaBoost*

Análise prescritiva

A análise prescritiva surge como uma fase subsequente de análise de negócios, após as análises descritiva, de diagnóstico e preditiva. A análise prescritiva usa os resultados das análises descritiva e preditiva para sugerir ações, com base nas previsões, mostrando as implicações de cada opção de decisão no negócio.

A análise prescritiva procura antecipar os resultados das previsões (o quê, quando e como irá acontecer) e também por que razão tal acontecerá. A ideia base é a criação de cenários alternativos futuros. Além disso, a análise prescritiva sugere opções de decisão sobre como tirar proveito de uma oportunidade futura ou mitigar um risco futuro. A análise prescritiva usa dados estruturados (números, categorias) e não estruturados (vídeos, imagens, sons, textos) e regras de negócios para fazer previsões e prescrever como se pode tirar proveito desse futuro sem comprometer outras prioridades.

Na agricultura, a chave para aumentar a produtividade agrícola reside na fusão da melhoria acelerada das culturas com a otimização das práticas agrícolas e envolve a análise de vários fluxos de dados. Isso irá acelerar o progresso em direção à agricultura prescritiva, com base em plantas com genótipos específicos.

As empresas que fornecem tecnologia na área da agricultura calculam a maneira mais eficiente de garantir rendimentos máximos sustentáveis para o agricultor. O agricultor pode comprar esses dados num esquema de preços por superfície agrícola utilizada. Os agricultores recebem recomendações prescritivas com base nos seus dados, algo que não teria sido possível no passado (Bennett 2015).

5.2. Sistemas de recomendação

Uma das aplicações de análise de dados mais utilizadas são os sistemas de recomendação (SR) que podemos reconhecer na nossa vida quotidiana: *Netflix* recomenda filmes, *Facebook* recomenda novos amigos para se conectar, *Youtube* recomenda vídeos para ver, *e-shops* recomendam produtos para comprar, a *Amazon* recomendando livros, jornais recomendam artigos para ler, etc.

As empresas querem vender mais itens, aumentar a satisfação e a fidelidade dos utilizadores e, também, entender melhor as necessidades e interesses desses utilizadores. Por outro lado, os utilizadores querem encontrar alguns itens bons com um esforço relativamente pequeno, o que é extremamente importante por causa da sobrecarga de informações que enfrentamos durante a pesquisa e navegação na *web*. Empregar técnicas de recomendação (TR) para catálogos ou lojas *web*, redes sociais e outros aplicativos é um passo razoável para

alcançar os objetivos acima mencionados de fornecedores e utilizadores.

Feedback

O conceito básico na área de TR, além do utilizador e do item, é o *feedback*, mas começamos por falar brevemente sobre os dois primeiros conceitos. Os utilizadores podem ser definidos pelos seus atributos ou características, tais como idade, rendimento, estado civil, educação, profissão ou nacionalidade, mas também pelo desporto preferido ou área de trabalho. Essas informações são muitas vezes obtidas por meio de questionários, no entanto, dada a sensibilidade dessas informações, os atributos do utilizador são geralmente difíceis de obter. Do outro lado da moeda estão os itens, também caracterizados pelos seus atributos como o nome, princípio ativo, quantidade, no caso do domínio de produtos fitossanitários. No caso dos atributos dos utilizadores, essas informações, mesmo não sendo sensíveis, podem, por vezes, terem um custo de obtenção elevado (por exemplo, precisamos obtê-los de descrições textuais, alguém tem que as inserir, etc.).

Quando o utilizador interage com um item, essa interação corresponde a algum tipo de *feedback* sobre o interesse do utilizador nesse item. Dependendo da forma como esse *feedback* foi obtido, distinguimos:

- *Feedback* explícito, i.e. quando o Sistema pergunta diretamente ao utilizador a sua preferência sobre itens, através de, por exemplo, uma escala de *Likert* (e.g. de 1 a 5 estrelas) ou através da ordenação desses itens; e
- *Feedback* implícito, quando a informação sobre as preferências dos utilizadores sobre os itens é obtida através da interação natural com o sistema, e.g. que itens estava o utilizador a ver, ouvir, visualizar, comprar, etc.

O *feedback* explícito é mais preciso. No entanto, coloca alguma carga subjetiva na avaliação. Por outro lado, o feedback implícito não sobrecarrega os utilizadores, mas não é tão preciso. Por exemplo, caso se tenha inscrito para uma palestra de um determinado palestrante, está a dar *feedback*, mesmo que não gostado dessa palestra.

Tarefas de recomendação

Antes de começarmos com a definição de tarefas de recomendação, vamos introduzir algumas notações: seja um utilizador específico do conjunto de n utilizadores. Da mesma forma, corresponderá a um item específico do conjunto de m itens. O *feedback* registado do utilizador u sobre o item i denota-se por r_{ui} e corresponde a um determinado classificação ou *ranking*.

A tarefa de recomendação é, dado o conjunto de utilizadores e itens, bem como os *feedbacks* gravados, aprender um modelo que prevê para cada par de utilizadores e itens (u,i) um valor representando a preferência do utilizador u sobre o item i. Isto soa familiar? Provavelmente sim, uma vez que parece uma tarefa de previsão introduzida no subcapítulo anterior.
Dependendo do tipo de *feedback*, distinguimos duas tarefas de recomendação:

- Previsão da preferência onde o *feedback* é explícito, ou seja, são números que representam preferências de utilizadores sobre itens, e expressam a preferência prevista do utilizador u pelo item i;
- Recomendação onde o feedback está implícito, ou seja, são zeros e uns que representam a ausência ou presença, respetivamente, de interação entre utilizadores e itens, e expressa a probabilidade prevista de um *feedback* implícito 'positivo' do utilizador u para com um item i.

Tabela 3. Exemplo de *feedback* implícito - porta enxertos comprados pelos produtores

	3.309	41-B	101-14	R 110	SO4
Ivo	1	1	1		1
Carlos	1	1		1	1
Irene	1	1	1	1	
Jaime		1	1		1

Vamos coletar as preferências dos nossos agricultores, a fim de sermos capazes de lhes recomendar os melhores produtos. O cenário de recomendação do item é ilustrado na Tabela 3, na qual coletamos informações sobre qual dos cinco porta-enxertos foram comprados por quatro dos nossos agricultores. Os valores de 1 nas células correspondem a *feedback* implícito positivo, por exemplo, Ivo comprou o porta enxerto 3.309, mas não comprou o R 110. A tarefa de recomendação, neste caso, seria aprender a probabilidade de *feedback* positivo dos utilizadores sobre itens que ainda não viram. Por exemplo, saiba a probabilidade de *feedback* positivo de Jaime sobre os porta enxertos 3.309 e R 110. O porta enxertos com maior probabilidade prevista seria aquele que satisfaria mais Jaime.

É de notar que, em caso de *feedback* implícito, apenas o *feedback* positivo é registado. Isso faz sentido assumindo que o facto de um utilizador ter comprado um porta enxertos pode indicar que ele/ela tem preferência por esse porta enxertos em relação a outros. Mas, um *feedback* positivo não significa necessariamente que o utilizador tenha ficado satisfeito com a compra desse item. Esses pressupostos são, no entanto, difusos, motivo pelo qual o *feedback* implícito não pode ser considerado absolutamente preciso.

Tabela 4. Exemplo de feedback explícito - porta enxertos comprados e avaliados pelos produtores

	3.309	41-B	101-14	R 110	SO4
Ivo	1	4	5		3
Carlos	5	1		5	2
Irene	4	1	2	5	
Jaime		3	4		4

Um *feedback* mais preciso é quando pedimos aos utilizadores para avaliarem os itens, como ilustrado na Tabela 4 onde os números nas células correspondem às classificações (número de estrelas) atribuídas aos produtos comprados pelos nossos agricultores. A tarefa, aqui, é o de prever as preferências dos utilizadores por produtos que ainda não compraram. Por exemplo, qual seria a classificação que Jaime daria aos porta enxertos 3.309 e R 110? A classificação mais alta prevista indicaria então que porta enxertos era esperado Jaime gostar mais.

Técnicas de recomendação

Distinguimos três tipos principais de sistemas de recomendação, e suas combinações híbridas. A utilização de cada um deles depende do domínio e dos dados disponíveis, ou seja, da informação sobre utilizadores e itens disponíveis ou se o *feedback* é considerado. Descrevem-se esses três tipos nos pontos seguintes.

Técnicas baseadas no conhecimento

Em Técnicas de Recomendação (TR) baseadas no conhecimento de especialistas, as recomendações são baseadas no conhecimento sobre as necessidades e

preferências dos utilizadores. Nestes tipos de TR, os atributos dos itens (por exemplo, o preço e o tipo do carro, o número de airbags, o tamanho do porta-malas, etc.), os requisitos do utilizador (por exemplo, "o preço máximo aceito do carro é 12.000€" e "o carro deve ser seguro e adequado para uma família',') e o conhecimento de domínio que descreve algumas dependências entre os requisitos do utilizador e as propriedades dos itens (por exemplo, "um carro de família deve ter tamanho de tronco grande')) ou entre os requisitos do utilizador (por exemplo, "se um carro de família seguro é necessário o preço máximo aceito deve ser superior a 20.000€").

Nesses tipos de SR, o processo de recomendação é interativo, o utilizador especifica iterativamente os seus requisitos de acordo com os itens recomendados de acordo com o atual estado da sua 'conversa' com o sistema. As recomendações são obtidas identificando os produtos no catálogo que correspondem aos requisitos do utilizador. Os itens são então ordenados de acordo com sua semelhança com a exigência do utilizador.

A desvantagem da TR baseada no conhecimento é o alto custo para preparar a base de conhecimento subjacente, que é dependente do domínio. Para cada domínio, uma base de conhecimento específica e, portanto, um especialista do domínio, são necessários o que torna esses tipos de SR pouco flexíveis e, consequentemente, menos populares.

Técnicas baseadas em conteúdo

Numa recomendação baseada em conteúdo, são necessários os atributos dos itens e alguns comentários gravados do utilizador. Os interesses do utilizador são aprendidos por técnicas preditivas. Por outras palavras, um modelo do feedback (a variável objetivo) de um determinado utilizador é aprendido com as variáveis preditivas (também conhecidas como as variáveis explicativas) de itens classificados ou ordenados no passado pelo utilizador. Usando o modelo preditivo assim obtido, são realizadas previsões ou ordenações de itens ainda não vistos pelo utilizador. A lista de itens recomendados é então composta com base nessas previsões ou listas ordenadas realizadas. Para cada utilizador, um modelo preditivo separado, de classificação ou de regressão, é aprendido.

A vantagem destes tipos de TR é que os atributos do utilizador não são necessários. Por outro lado, o modelo preditivo é geralmente aprendido a partir de uma pequena quantidade de instâncias, especialmente, no início de utilização do sistema. Além disso, pode acontecer o utilizador classificar alguns produtos específicos, primeiro (por exemplo, apenas um gênero específico, e.g., porta enxertos). Assim, por causa dessas razões, mas também outras, o modelo preditivo aprendido é sensível à superação podendo limitar a recomendação a um espaço específico de itens.

Semelhança vetor cosseno

A medida de semelhança do vetor cosseno é uma medida de semelhança vetorial utilizada nos casos em que os vetores a comparar são esparsos. Para calcular a semelhança de vetores esparsos, os valores em falta em ambos os vetores são substituídos por 0.
Tendo dois vetores $\boldsymbol{x} = (x_1, \dots, x_m)$ e $\boldsymbol{y} = (y_1, \dots, y_m)$, a semelhança do vetor cosseno é calculada como

$$sim^{CV}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{m} x_i y_i}{(\sum_{i=1}^{m} {x_i}^2 \sum_{i=1}^{m} {y_i}^2)^{\frac{1}{2}}}$$

Equação 1.

A semelhança do vetor cosseno é uma medida de semelhança popular na análise de texto e em sistemas de recomendação.

Técnicas de filtragem colaborativa

Técnicas de filtragem colaborativa são as TR mais populares. Reconhecem semelhanças entre utilizadores de acordo com os seus *feedbacks* e recomendam itens preferidos por utilizadores com preferências

semelhantes. As técnicas de filtragem colaborativas podem fornecer bons resultados mesmo nos casos em que não há disponíveis atributos nem de utilizadores nem de atributos. Distinguimos dois tipos principais de técnicas de filtragem colaborativa: técnicas baseadas em vizinhança e baseadas em modelos.
Técnicas baseadas em vizinhança utilizam medidas de semelhança vetorial para calcular semelhanças entre utilizadores ou itens.

Filtragem colaborativa baseada no utilizador

No caso de técnicas de filtragem colaborativa baseadas no utilizador, cada utilizador é representado por um vetor de *feedbacks*. No nosso exemplo da Tabela 3 ou da Tabela 4, o utilizador Jaime seria representado pelos vetores *(?,1,1,?,1)* ou *(?,3,4,?,4)*, respetivamente, correspondente às suas classificações gravadas para os 5 porta enxertos no catálogo. Note, que os vetores do utilizador são geralmente muito esparsos, dado que os utilizadores costumam fornecer algum tipo de *feedback* apenas sobre um pequeno número de itens em comparação com o número de todos os itens no catálogo.

Para prever o *feedback* do utilizador *u* sobre o item *i*, o primeiro passo é obter os *k* utilizadores mais parecidos com o utilizador *u* em termos de *feedbacks*, e que tenham também dado *feedback* sobre o item *i*. O conjunto desses vizinhos *k*-vizinhos mais próximos são denotados por N_i^{uk}.
Em caso de recomendação de item, a probabilidade prevista de um feedback positivo do utilizador *u* sobre o item *i* é calculado como uma semelhança média de cada utilizador $v \in N_i^{uk}$ para com o utilizado *u*

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i^{u,k}} sim(u,v)}{k}$$

Equação 1.

onde *sim(u,v)* é alguma medida de semelhança vetorial, por exemplo, a semelhança vetorial cosseno (ver caixa). Mas qualquer outra medida de distância pode ser usada.

As semelhanças calculadas de vetor cosseno entre os utilizadores da Tabela 3 são mostradas na Tabela 5.

Tabela 5. Semelhanças vetor cosseno entre os utilizadores da Tabela 4. Como os valores de similaridade são simétricos, os valores abaixo da diagonal espelhariam os valores acima da diagonal.

	Ivo	Carlos	Irene	Jaime
Ivo	1,0	0,75	0,75	0,87
Carlos		1,0	0,75	0,58
Irene			1,0	0,58
Jaime				1,0

Com base nesses dados, de acordo com a Equação 1, as probabilidades previstas de *feedback* positivo de Jaime sobre os itens 3.309 e R 110 são calculados do seguinte modo:

$N_{3.309}^{Jaime,2} = \{Ivo, Carlos\}$ e $N_{R\ 100}^{Jaime,2} = \{Carlos, Irene\}$
Continuando os cálculos, temos

$$\hat{r}_{Jaime,3.309} = \frac{sim^{CV}(Jaime, Ivo) + sim^{CV}(Jaime, Carlos)}{2}$$
$$= \frac{0,87+0,58}{2} = 0,725$$
$$\hat{r}_{Jaime,R\ 110} = \frac{sim^{CV}(Jaime, Carlos) + sim^{CV}(Jaime, Irene)}{2}$$
$$= \frac{0,58 + 0,58}{2} = 0,58$$

Assim, Jaime provavelmente vai preferir o porta enxertos 3.309.

Em caso de previsão de classificação, devemos estar cientes do fenómeno chamado viés. As classificações dos utilizadores geralmente são tendenciosas, o que significa que alguns utilizadores, quando fornecem classificações, são mais pessimistas, enquanto outros são mais

otimistas do que a média. Assim, o impacto do viés deve ser levado em consideração ao calcular a semelhança entre utilizadores.

Uma boa escolha seria utilizar algumas medidas de correlação para calcular a semelhança entre os feedbacks de dois utilizadores, como, por exemplo, a correlação Pearson, referida anteriormente, denotada aqui como sim^{pc}. Uma vez que os *feedbacks* não são apenas *0* (em caso de sem *feedback*) ou *1*, como no caso da recomendação de itens, mas números (ver Tabela 2), o modelo que prevê a preferência do utilizador *u* pelo item *i* é

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in N_i^{u,k}} sim(u,v) \times (r_{vi} - \bar{r}_v)}{\sum_{v \in N_i^{u,k}} |sim(u,v)|}$$

Equação 2.

onde $\bar{r}_u$ e $\bar{r}_v$ são as preferências médias dos utilizadores *u* e *v* (calculadas a partir dos dados de treino) respetivamente, $sim\ (u,v)$ é uma medida de semelhança que avalia a semelhança dos utilizadores *u* e *v*, por exemplo a correlação de Pearson acima mencionada $sim^{pc}(u,v)$.

Tabela 6. Semelhança entre os utilizadores (Tabela 4) pela correlação de Pearson. Como os valores de similaridade são simétricos, os valores abaixo da diagonal espelhariam os valores acima da diagonal

	Ivo	Carlos	Irene	Jaime
Ivo	1,0	-0,716	-0,762	-0,005
Carlos		1,0	0,972	0,565
Irene			1,0	0,6
Jaime				1,0

As semelhanças de correlação de Pearson entre os utilizadores da Tabela 4 são mostradas na Tabela 6. Com base nesses dados, de acordo com a Equação 3, as probabilidades previstas de feedback positivo de Jaime sobre o item 3.309 são calculados da seguinte forma (J,I,C e 3 são respetivamente abreviaturas de Jaime, Irene, Carlos e 3.309):

$N_3^{J,2} = \{I, C\}$ e

$\bar{r}_J = \frac{3+4+4}{3} = 3{,}67$

$\bar{r}_I = \frac{4+1+2+5}{4} = 3$

$\bar{r}_C = \frac{5+1+5+2}{4} = 3{,}25$

$$\hat{r}_{J,309} = \bar{r}_J + \frac{sim^{pc}(J,I) \times (r_{I,3} - \bar{r}_I) + sim^{pc}(J,C) \times (r_{C,3} - \bar{r}_C)}{|sim^{pc}(J,I)| + |sim^{pc}(J,C)|}$$

$$= 3{,}67 + \frac{0{,}6 \times (4-3) + 0{,}565 \times (5-3{,}25)}{0{,}6 + 0{,}565}$$

$$= 1{,}36$$

que é a preferência prevista de Jaime pelo item 3.309.

Filtragem colaborativa baseada em itens

Tal como a filtragem colaborativa baseada no utilizador, existem técnicas de filtragem colaborativa baseada em itens. A diferença é que, em vez de considerar a semelhança entre utilizadores, consideram a semelhança entre itens. Assim, as medidas de semelhança vetorial são calculadas a partir das colunas da matriz do utilizador-item (Tabela 3 e Tabela 4).

Observações finais

Há várias questões importantes que se tem de ter em mente ao desenvolver e implementar um sistema de recomendação, alguns dos quais vamos mencionar neste subcapítulo.

Antes de integrar a técnica de recomendação desenvolvida num sistema real (por exemplo, uma *e-shop*), é aconselhável fazer sua avaliação *offline* em dados em que o comportamento do utilizador é simulado. É uma experiência de baixo custo, leva pouco tempo e pode ser feito em larga escala, no entanto, responde apenas a algumas perguntas, tal como o poder preditivo da técnica desenvolvida, o tempo de execução, etc.

O outro, é um passo um pouco mais caro e demorado na avaliação de um sistema de recomendação já que envolve estudos sobre os utilizadores. Em contraste com avaliações *offline*, aqui, as interações com utilizadores reais do sistema são observadas e analisadas por meio de questionários, por exemplo, como eles gostaram das recomendações entregues, se as acharam úteis, etc. Os estudos de utilizadores são geralmente de pequena escala e caros (é necessário recompensar os sujeitos do teste de alguma forma).

A fase final de avaliação de um sistema de recomendação é a avaliação *online* feita da seguinte forma: uma pequena parte do tráfego do sistema é redirecionada para a nova técnica de recomendação desenvolvida e o comportamento dos utilizadores é observado: se o *feedback* de preferências melhora, se permanecem no sistema durante mais tempo, etc. Isso pode ser, no entanto, um pouco arriscado, uma vez que se os clientes não estão satisfeitos com o resultado da nova técnica de recomendação, podemos perder alguns deles. Assim, é bom executar testes *online* após um teste *offline* e apenas se os estudos de utilizadores mostrarem resultados promissores.

Existem várias propriedades que um sistema de recomendação deve ter, que pode ser testado e avaliado em experiências *online* e *offline*, bem como os estudos do utilizador. Além da escalabilidade, robustez e precisão preditiva, o sistema de recomendação deve ter uma boa cobertura, ou seja, ser capaz de recomendar uma grande proporção de itens para uma grande proporção de utilizadores. Em outras palavras, ele não deve estar funcionando apenas para uma pequena parte dos itens e utilizadores. Outra questão é a novidade das recomendações que se preocupa com a questão se o sistema recomenda itens a um utilizador que não os encontraria sozinho. Está de alguma forma ligado à propriedade *serendipity* que se refere ao quão surpreendente são as recomendações para o utilizador. Além disso, vale a pena pensar sobre a diversidade de recomendações que medem o quão colorida a paleta de itens recomendados são.

O problema de começo frio (do inglês *cold-start*) surge quando um novo utilizador ou um novo item aparece no sistema para o qual não há *feedback* registado suficiente. Neste caso, o modo mais fácil é o de recomendar os itens mais populares ao utilizador. Se tivermos algumas informações adicionais disponíveis sobre o utilizador, podemos utilizar alguma base de conhecimento genérico, ou seja, recomendação baseada no conhecimento. Além disso, informações adicionais sobre itens podem ser utilizadas.

Uma nova direção na pesquisa de sistemas de recomendação investiga recomendações baseadas no contexto e em grupo. No primeiro caso, refere-se a qualquer informação adicional para além dos atributos do utilizador e do item que possam ser relevantes para recomendação. Por exemplo, as recomendações de produtos fitossanitários variarão com a época do ano e condições meteorológicas. Em relação a recomendações de grupo, é um tema que terá menos aplicações na área agrícola e, por isso, não será abordado.

Finalmente, temos que notar que um "parceiro" muito importante de qualquer técnica de recomendação é uma boa interface com o utilizador. Mesmo os resultados das melhores técnicas de recomendação podem ser estragados por uma má experiência do utilizador devido a uma má interface. Além disso, temos que ter em mente que cada domínio tem suas especificidades que, se incorporadas a um sistema de recomendação, podem contribuir para a singularidade do sistema desenvolvido. Mais informações e discussão mais profunda sobre sistemas de recomendação e suas aplicações podem ser encontradas em (Ricci2010).

5.3. Trabalhando com textos

Os textos são a forma mais comum de trocar informações na nossa sociedade. Muitas informações preciosas podem ser ocultadas nesses textos. Enquanto os seres humanos podem facilmente extrair informações significativas de um texto, o mesmo não é verdade para o *software* de um computador.

Considere, por exemplo, uma nota simples que escreveu sobre as preferências alimentares de um de seus amigos, Carlos, que é vegetariano. A caixa separada mostra a nota que escreveu na sua rede social há um tempo atrás. Como é que podemos extrair automaticamente conhecimentos úteis deste texto? As técnicas que vimos até agora focam, sobretudo técnicas que só podem ser aplicadas em dados no formato tabular, o que não é o caso do texto referido. Textos, tal como imagens, filmes, sons não seguem o formato tabular. Para distinguir entre esses dois formatos, os dados tabulares são nomeados dados estruturados e os outros formatos de dados são nomeados dados não estruturados.

Sendo uma área muito próxima à análise de dados, a análise de texto fornece várias técnicas desenvolvidas especificamente para extrair conhecimento de textos em bruto escritos em linguagem natural.
Podemos dizer que, embora a análise de dados esteja associada a dados, a análise de texto está associada ao texto Weiss et al., 2015.

A origem da análise de texto remonta a tarefas de índice de documentos na área da recuperação de informações. A recuperação de informações preocupa-se, geralmente, com a recuperação de informações de documentos *online*. É uma área-chave nos motores de busca da *web*, onde é usada a semelhança entre os documentos para recuperar *site*s relevantes da Internet.

A análise de texto é uma área muito ativa da análise de dados. Investiga e fornece ferramentas para extrair conhecimento de textos.

A análise de texto é uma parte importante de várias outras tarefas, como recuperação de informações, deteção de *spam*, análise de sentimentos, sistemas de recomendação e análise na *web*. Para essas aplicações, um aspeto fundamental é como medir a semelhança entre textos.

Como nos tipos de aprendizagem estudados no subcapítulo 5.1., as tarefas de análise de texto podem ser classificadas, não só mas também, como descritivas ou preditivas. As tarefas descritivas de análise de texto incluem procurar grupos de documentos semelhantes, procurar textos sobre questões e palavras semelhantes que frequentemente aparecem juntos em textos. As tarefas preditivas incluem a classificação de documentos em um ou mais tópicos e a identificação de *spam* em *email*s e análise de sentimentos em mensagens curtas.

Este subcapítulo concentra-se em tarefas preditivas de análise de texto, também conhecidas como categorização de texto e classificação de documentos. Os termos texto e documento serão usados com o mesmo significado neste subcapítulo.

Texto sobre preferências de comidas

Carlos gosta muito de jantar em restaurantes chineses. Como Carlos é vegetariano, ele não come carne. A fim de ter proteína suficiente, Carlos está sempre à procura de outros alimentos que têm níveis de proteína semelhantes aos encontrados na carne.

Como já mencionado, a maioria das técnicas de análise de dados espera que os dados estejam num formato tabular atributo-valor, pelo que não podem ser diretamente aplicadas a dados textuais. No entanto, várias técnicas foram desenvolvidas para extrair dados estruturados de textos em bruto.

Assim, um dos primeiros passos na análise de texto é a transformação de textos em dados tabulares, no formato atributo-valor. Neste subcapítulo, descreveremos algumas dessas técnicas e mostraremos como podemos transformar um texto numa tabela atributo-valor.
Para ilustrar como funciona a análise de texto, voltemos à nossa tarefa de classificação automática de mensagens. Suponha que queríamos dividir as mensagens que recebemos em dois grupos: trabalho e família. Para isso, podemos usar ferramentas analíticas de dados, para induzir um modelo capaz de classificar automaticamente as nossas mensagens num desses dois grupos.

O processo preditivo de análise de texto é muito semelhante a um processo de análise de dados. A principal diferença é a transformação de dados não estruturados em dados estruturados através de técnicas de pré-processamento específicas para texto. Em resumo, uma tarefa de análise de texto é composta por cinco fases:

- Aquisição de dados;
- Extração de atributos;
- Pré-processamento dos dados;
- Indução de modelos; e
- Avaliação e interpretação dos resultados.

As últimas três fases são realizadas por técnicas de análise de dados, uma vez que os dados já estarão no formato estruturado. As duas primeiras fases correspondem a tarefas da fase de preparação de dados da metodologia CRISP-DM (ver Capítulo 3). Portanto, vamo-nos concentrar aqui na aquisição de dados e extração de variáveis preditivas.

Aquisição de dados

Em primeiro lugar precisamos de obter um conjunto de dados com instâncias representativas, que são instâncias semelhantes às que acreditamos que receberemos no futuro. É claro que não podemos ter a certeza de como serão as mensagens futuras, mas se pudermos coletar uma quantidade boa de objetos diversos, temos uma boa chance de ter uma amostra representativa. Se tivermos um número muito grande de mensagens e uma boa capacidade de armazenamento e processamento, vamos coletar todas as mensagens de um determinado período, por exemplo, nos últimos 12 meses. Uma coleção de textos ou documentos é conhecida como corpus. Cada texto no corpus será convertido num objeto estruturado.

Se os textos vêm de diferentes fontes, eles podem ter formatos diferentes, tais como formato de texto ASCII ou Unicode, ou o formato da Linguagem de Marcação Extensível (XML), que é o formato padrão de troca de documentos. Um arquivo XML tem palavras-chave, *tags*, usados para marcar algumas partes do documento XML. Essas *tags* podem fornecer informações significativas sobre o conteúdo nessas partes, como o título do documento, os autores, a data, os tópicos e o resumo. As etiquetas podem ser usadas para localizar a parte do documento a ser analisado. Textos que não são linguagem natural, como *email*s e *site*s, são facilmente detetados e podem ser removidos, se necessário.

Extração de variáveis preditivas

Uma vez que todos os textos tenham sido submetidos a esse processo, cada texto ou documento será um fluxo de caracteres, que pode incluir palavras, números, espaços brancos, caracteres de pontuação e caracteres especiais. Como numa tarefa preditiva de análise de dados, separamos um subconjunto de textos para usarmos como conjunto de treino, que são os textos que usaremos para induzir um modelo preditivo, que poderá, por sua vez, ser aplicado a novos textos, após serem pré-processados.

Tokenização

O próximo passo é extrair, para cada texto, uma sequência de palavras do fluxo de caracteres. Neste procedimento, chamado tokenização, cada palavra na sequência é nomeada por *token*. As palavras são detetadas olhando para espaços brancos e

caracteres de pontuação. Se uma palavra aparecer mais de uma vez no texto, o seu *token* aparecerá mais de uma vez na sequência de *tokens*. Este procedimento de representar um texto por um conjunto de *tokens*, onde cada *token* pode aparecer mais de uma vez, é conhecido nas áreas de recuperação de informações e linguagem de processamento natural como saco de palavras (*bag of words* em inglês), Weiss et al., 2015.

Dependendo do contexto do documento, alguns caracteres especiais também podem ser *tokens*.

Para algumas aplicações, uma frase inteira pode ser um *token*.

Para ilustrar como funciona a análise de texto, vamos supor que temos um pequeno conjunto de treino com quatro mensagens curtas recebidas da nossa família e colegas de trabalho. Nas tarefas de classificação de texto, cada texto no conjunto de treino é rotulado por um ou mais tópicos. Para tornar o exemplo mais simples, suponhamos que atribuímos um tópico a cada mensagem de conjunto de treino, que se tornará o rótulo de classe da mensagem. Neste caso, os temas são Família e Trabalho. A Tabela 7 mostra os textos e seu rótulo

Tabela 7. Conjunto de treino de textos rotulados

Mensagem recebida	Classe
Eu gosto das festas de aniversário da minha irmã	Família
Eu gostei da festa da empresa	Trabalho
Eu não os estou a trazer da escola	Família
Vou falar e trazer o contrato	Trabalho
Conversei com outras empresas	Trabalho
A minha mulher está a ter contrações	Família

Nestes exemplos, temos um pequeno número de textos curtos. Na prática, geralmente temos uma grande coleção de textos. Para aplicações de troca de mensagens e análise de sentimentos, os textos são geralmente curtos. Para outras aplicações, textos longos são mais comuns.

Radicalização

Os *tokens* podem ser várias variações de uma palavra, como plural, inflexões verbo e formas de gênero. Assim, se representarmos cada texto por todos os *tokens* que aparecem em todos os textos do conjunto de treino, provavelmente acabaremos com um número muito grande de *tokens*. No entanto, a maioria dos textos terá apenas uma pequena fração de todos esses *tokens*.

Dissemos antes que quando transformarmos os textos numa tabela de valores quantitativos, cada palavra, agora simbólica, se tornará um atributo preditivo. Como apenas uma pequena parte dos *tokens* estará presente em cada texto, os atributos preditivos cujo *token* não aparece no texto de um objeto terão o valor de *0*. Isso tornará a tabela muito dispersa, já que a maioria de seus valores preditivos serão iguais a *0*.

Para evitar um número muito grande de *tokens*, o que pode resultar num conjunto de dados muito disperso, procuramos uma forma base comum capaz de representar muitas das variações de um *token*.

Num dos métodos mais simples, cada *token* é convertido no seu radical, num processo chamado de radicalização (*stemming*, em inglês) usando um algoritmo de contenção, também chamado *stemmer*. Existem diferentes caules para diferentes línguas naturais. Mesmo para a mesma língua, como o inglês, podemos ter caules diferentes. Entre os algoritmos de radicalização, um dos mais comuns é o algoritmo de Porter, 1980.

A contenção é um método muito simples, que geralmente remove apenas afixos de palavras. Um afixo no início de uma palavra é um prefixo e no final de uma palavra é um sufixo.

Aplicando o algoritmo de contenção aos textos da Tabela 7 obtemos os resultados apresentados na Tabela 8 que mostra os radicais de cada objeto, juntamente com a classe desse objeto.

Tabela 8. Resultados de aplicação de um algoritmo de radicalização

Mensagem recebida depois de radicalização	Classe
eu, gostar, de, festa, de, aniversário, de, meu, irmão	família
eu, gostar, de, festa, de, empresa	trabalho
eu, não, o, estar, o, trazer, de, escola	família
ir, falar, e, trazer, o, contrato	trabalho
falar, com, outro, empresa	trabalho
o, meu, mulher, estar, a, ter, contrato	família

Essas operações simples resultaram numa boa redução no número de *tokens* diferentes, representando cada *token* pelo seu radical.

Há também uma variação de contenção, chamada lematização. A lematização é um método mais sofisticado para extrair a forma base comum de palavras, uma vez que utiliza um vocabulário e leva em conta aspetos gramaticais, realizando uma análise morfológica. A lematização de uma palavra devolve a forma de dicionário dessa palavra, chamada lema.

No exemplo anterior houve uma redução do número de *tokens* diferentes. Mesmo assim, ainda temos 23 *tokens* diferentes. Como cada *token* diferente tornar-se-á numa variável preditiva, cada instância terá 23 variáveis preditivas. Uma vez que os textos têm, em média, seis *tokens* diferentes, cerca de 78 % das variáveis preditivas não estarão presentes na maioria dos objetos, o que resultará em 102 valores de 0 nos 138 valores possíveis. Assim, ainda teremos uma tabela esparsa após a conversão para o formato tabular.

A boa notícia é que ainda existem outros procedimentos para reduzir o número de radicais e, como resultado, o número de variáveis preditivas.

O número de radicais pode ser ainda mais reduzido removendo palavras *stop*. As palavras *stop* são muito comuns em textos. Têm baixo poder discriminativo e, provavelmente, não serão úteis para as próximas fases de análise de texto. Alguns exemplos de palavras *stop* são:

- Adjetivos (bom, mau, grande, lindo, etc.);
- Advérbios (aqui, hoje, perfeitamente, muito, etc.);
- Artigos (o, a, um, uma, etc.);
- Negações (nenhum, não, nunca, etc.);
- Pronomes (eu, ele, meu, seu, teu, nosso, etc.);
- Preposições (em, por, de, com, etc.);
- Conjunções (e, mas, ou, com, pois, etc.);
- Verbos frequentes (ser, estar, etc.);
- Qualificadores (pouco, menos, mais, outro, provavelmente, algum, etc.).

É importante observar que a mesma palavra pode ter significados diferentes. Por exemplo, a palavra "para" pode ser, de acordo com o contexto, uma forma do verbo parar, ou uma preposição. Várias palavras podem ser advérbios em alguns textos e adjetivos em outros. Em técnicas simples de deteção de palavras, um dos significados é assumido.

A decisão de quais as palavras *stop* que se devem remover depende de cada aplicação. Por exemplo, em aplicações de análise de sentimento, a presença de adjetivos e negações pode ser uma informação importante. O que geralmente ocorre em aplicações de análise de texto é uma seleção do subconjunto mais adequado da lista de palavras *stop* para a aplicação.

Além disso, alguns dos radicais também podem ocorrer muito raramente no texto e, portanto, poderão não ser úteis para a indução de um modelo preditivo. De acordo com Weiss et al., 2015, estima-se que metade das palavras em um corpus aparecem apenas uma vez. Assim os radicais com frequência muito baixa no corpus podem ser removidos.

Algumas aplicações de análise de texto também usam frases *stop*, onde a frase inteira, que aparece com muita frequência nos textos e tem um poder discriminativo baixo, pode ser removida.

As palavras *stop* podem ser identificadas e removidas antes da extração das variáveis preditivas. No entanto, a sua remoção deve ser feita após todas as transformações anteriores.

Na Tabela 9 pode-se ver o resultado obtido com o nosso exemplo após a remoção das palavras *stop*.

Tabela 9. Radicais após remoção das palavras *stop*

Radicais após remoção	Classe
festa, aniversário, irmão	família
festa, empresa	trabalho
trazer, escola	família
falar, trazer, contrato	trabalho
falar, empresa	trabalho
mulher, contrato	família

Com a remoção das palavras *stop*, reduzimos o número de radicais diferentes de 23 para 9. A maioria das instâncias têm apenas 2 radicais, reduzindo assim o número de valores 0 de 102 para 40.

Conversão para dados estruturados

O próximo passo na análise de texto é transformar a tarefa de análise de texto numa tarefa de análise de dados. Para tal, as informações presentes no tipo não estruturado, o texto, devem ser convertidas no tipo estruturado, tabela com valores quantitativos.

Inicialmente realizamos essa conversão no conjunto de treino de texto, que são os textos que usaremos para induzir um modelo preditivo, que pode então ser aplicado a novos textos, após a sua conversão em valores quantitativos.

O valor da variável preditiva associada a um radical pode ser tão simples quanto um valor binário indicando a presença do radical no texto. Por exemplo, *1* para representar a presença do radical no texto e *0* para a ausência. Este procedimento simplifica a implementação e a análise de dados. No entanto, uma informação importante para a classificação de texto pode ser o número de vezes que cada radical aparece no texto.

Assim, é procedimento comum representar o número de vezes que cada radical aparece na mensagem, método conhecido como saco de palavras, que já foi atrás referido.

No exemplo, uma vez que os textos usados são muito curtos, nenhum radical ocorre mais de uma vez em cada texto após a remoção das palavas *stop*. Assim, todas as variáveis preditivas terão um valor binário, *1* para a presença do radical na instância e *0* no caso de ausência. Só para lembrar, cada texto é transformado em um objeto de atributo valor. A Tabela 10 ilustra o formato tabular obtido.

Tabela 10. Instâncias com os seus radicais em formato estruturado

festa	o	aniversári o	irmã o	empres a	traze r	escol a	fala r	contrat o	mulhe r	class e
1	1		1							F
1				1						T
					1	1				F
					1		1	1		T
				1			1			T
								1	1	F

Saco de palavras é suficiente?

Às vezes, apenas a ocorrência de cada palavra pode ser enganosa. Olhando para o texto sobre preferências de comida, a palavra carne aparece duas vezes mais frequentemente do que a palavra vegetariano. Ao contar as palavras, o nosso método de análise de dados pode acreditar que o Carlos gosta de carne. Além disso, se tivermos um negativo antes de uma palavra, a negação não está a ser levada em conta.

Para uma análise de texto mais sofisticada, técnicas do processamento de linguagem natural podem ser usadas. Apesar de uma interpretação de texto mais precisa, o uso dessas técnicas para grandes textos retarda o processo de análise de texto.

Próximas fases de análise de texto

Uma vez que temos os dados no formato tabular, temos as próximas fases, pré-processamento de dados, indução do

modelo, avaliação e interpretação dos resultados.

As técnicas convencionais de análise de dados são usadas nessas fases.

A técnica de pré-processamento de análise de dados que são frequentemente usadas são técnicas de redução da dimensionalidade. Em aplicações de análise de texto, depois de todas as técnicas de pré-processamento que apresentamos, *tokenização*, contenção e remoção de palavras *stop* e de palavras com frequência muito baixa, ainda podemos ter um grande número de variáveis preditivas e dados muito esparsos. As técnicas de redução dimensional são frequentemente aplicadas a esses dados para reduzir ainda mais o número de atributos preditivos e a esparsidade.

As medidas de avaliação utilizadas para avaliar tarefas de análise de texto são geralmente as mesmas usadas em tarefas de análise de dados. Em algumas aplicações, como por exemplo, ordenação de documentos, medidas de recuperação de informações são também utilizadas.

Tendências

A análise de texto é um problema muito popular na análise de dados. Várias ferramentas e aplicativos de análise de texto foram desenvolvidos e comercializados por muitas empresas.

As tendências atuais na análise de texto incluem a combinação de técnicas de processamento de imagem para extrair conhecimento de documentos e livros impressos antigos, combinação com técnicas de processamento de linguagem natural para compreensão de texto, identificação de expressões idiomáticas de texto e traduções de textos para outros idiomas, descoberta de autoria de textos acadêmicos, identificação de plágio em documentos, livros, notícias e artigos académicos, extração de informações de notícias publicadas pelos *media* para resumir notícias recebidas de diferentes fontes e para fornecer a seleção pessoal de notícias, monitoramento da literatura relacionada à saúde para descobrir novos conhecimentos capazes de melhorar o diagnóstico médico e análise de sentimento de análise de opinião em textos de mensagens curtas.

Outra aplicação frequente é a extração de meta-dados de textos, que são informações importantes presentes no texto. Como exemplo, suponha que queremos filtrar ofertas de emprego. Para tal, os meta-dados extraídos podem ser o nome da empresa, país, endereço, *site*, endereço de *email* e número de telefone, dados de fechamento, desejável, requisitos, salário, habilitações e dados iniciais.

Dois aplicativos muito frequentes são a análise de sentimento e a análise na *web*. Os principais aspetos dessas aplicações são descritos de seguida.

Análise de sentimentos

Um caso especial de análise de texto é a análise de textos curtos trocados através de ferramentas de redes sociais.

Esta análise, conhecida como análise de sentimento ou análise de opinião é geralmente realizada em textos com um número limitado de caracteres. Nestes casos, o uso de saco de palavras leva, geralmente, a bons resultados.

Técnicas de análise de sentimentos têm sido utilizadas para analisar atitudes, avaliações, opiniões e sentimentos dos utilizadores em relação a entidades, eventos, questões, pessoas públicas, produtos, serviços e tópicos. Têm sido usados com sucesso na avaliação de marketing de novos produtos, previsão de lutas entre adeptos de diferentes equipes de futebol e descoberta de tendências de votação em campanhas eleitorais.

Análise de dados na Web

Outro aplicativo de análise de texto comum é a análise de textos de páginas da *web*, chamado *Web Mining*.

Uma coleção muito grande de textos pode ser encontrada em páginas da *Web* na *World Wide Web*, a partir de fontes tão distintas como *logs* e *site*s para instituições académicas e publicações, comércio eletrónico, agências de notícias, jornais, governo.

No entanto, em contraste com os textos comuns, as páginas *Web* são geralmente escritas em formatos especiais, definida por uma linguagem de marcação, por exemplo, a Linguagem de Marcação *HyperText* (HTML). Os idiomas de marcação fornecem informações extras além de texto, como áudios, imagens, vídeos, comentários, meta-dados e híper-*link*s para outras páginas da *Web*, que podem ser usados para extração de conhecimento. Assim, uma vez que o texto é extraído de uma página *web*, técnicas de análise de texto podem ser aplicadas ao texto, usando ou não as informações extras.

Esta informação extra também pode ser um fardo. Primeiro, porque podem conter informações irrelevantes e até mesmo informações que podem prejudicar mais do que ajudar a extração de conhecimento. Em segundo lugar, e uma vez que as páginas *web* podem ter estruturas muito diferentes, porque não é fácil extrair automaticamente dados deles.

5.4. Análise de Redes Sociais

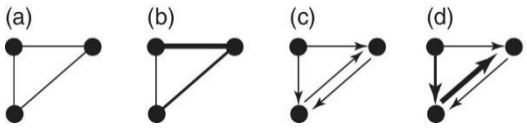
A análise de redes sociais (ARS) ganhou importância nos últimos anos devido à popularidade que diversas redes sociais ganharam. ARS tem as suas raízes na sociologia, no entanto, o estudo das redes vai hoje muito além das ciências sociais. Redes de várias características são estudadas em física, neurociência, economia, ciência da computação e também engenharia, apenas para citar algumas áreas da ciência em que as relações entre várias entidades desempenham papéis importantes.

Uma vez que uma descrição completa de todos os métodos existentes sai fora do âmbito deste livro, apenas alguns conceitos fundamentais serão introduzidos neste subcapítulo considerando tipos simples de redes.

Representando redes sociais

Cada rede é um tipo de gráfico consistindo de nós e relações, também chamadas de bordas, entre os nós. As relações podem ser direcionadas ou não direcionadas, bem como ponderadas ou não-ponderadas como é ilustrado na Figura 11. No caso de existirem vários tipos de relações entre os nós, por exemplo, duas pessoas podem ser conectadas por um "amigo", "família" ou "colega", chamamos a essas relações, multiplex.

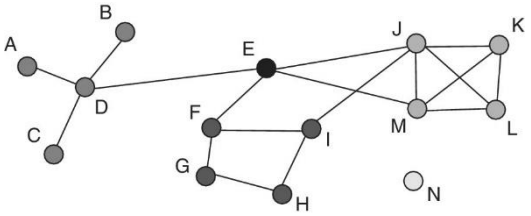
Figura 11. Não direcionado (a), ponderado (b) onde a espessura de uma relação é proporcional ao seu peso, dirigido (c), e, dirigido e ponderado (d) são tipos de relações em redes



Para representar um gráfico direcionado ou não direcionado, uma estrutura adequada é uma matriz chamada de adjacência, denotada aqui como A, cujas linhas e colunas representam nós. Cada célula A_ij na linha i e na coluna j indica se há uma relação do nó i para o nó j. No caso de relações não direcionadas, A é simétrico em relação à sua diagonal.

A Figura 12 apresenta uma rede social exemplo que será utilizada nesta secção para ilustrar os conceitos que vão sendo apresentados. As relações, por uma questão de simplicidade, são não direcionadas e sem pesos e podem representar, digamos, as relações numa qualquer rede social.

Figura 12. Uma rede social de exemplo



Esta rede, ou gráfico, pode também ser representada numa matriz de adjacência, como mostra a Tabela 11. Note-se que, em caso de redes esparsas, pode haver outras formas de representar a matriz de adjacência.

Representar um gráfico numa matriz de adjacência A tem as suas vantagens, uma vez que podemos empregar operações de matriz eficiente e rápida para obter informações úteis sobre o gráfico.

Por exemplo, se multiplicarmos A com ele mesmo, ou seja, elevar A à segunda potência, obtemos o número de caminhos, ou seja, sequências de relações, entre pares de nós que são de comprimento dois.

A matriz de adjacência quadrada da Tabela 11 é mostrada na Tabela 12. A partir desta matriz, podemos ver que há um caminho de comprimento dois entre os nós A e B, através do nó D, ou seja, a sequência de relações entre os nós A e D e os nós D e B. Esta sequência é de comprimento dois. Além disso, há um caminho de comprimento dois de A a D e de D de volta a A, o que pode ser confirmado na primeira célula da primeira fila, primeira coluna da Tabela 12.

Tabela 11. A matriz de adjacência para a rede da Figura 12

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	0	0	0	1	0	0	0	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	1	1	1	0	1	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	1	0	0	0	1	0	0	1	0
F	0	0	0	0	1	0	1	0	1	0	0	0	0	0
G	0	0	0	0	0	1	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	1	0	0	0	0	0
I	0	0	0	0	0	1	0	1	0	1	0	0	0	0
J	0	0	0	0	1	0	0	0	1	0	1	1	1	0
K	0	0	0	0	0	0	0	0	0	1	0	1	1	0
L	0	0	0	0	0	0	0	0	0	1	1	0	1	0
M	0	0	0	0	1	0	0	0	0	1	1	1	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabela 12. A matriz de adjacência da Tabela 9 ao quadrado mostrando as contagens de caminhos de comprimento dois entre pares de nós

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	1	1	1	0	1	0	0	0	0	0	0	0	0	0
B	1	1	1	0	1	0	0	0	0	0	0	0	0	0
C	1	1	1	0	1	0	0	0	0	0	0	0	0	0
D	0	0	0	4	0	1	0	0	0	1	0	0	1	0
E	1	1	1	0	4	0	1	0	2	1	2	2	1	0
F	0	0	0	1	0	3	0	2	0	2	0	0	1	0
G	0	0	0	0	1	0	2	0	2	0	0	0	0	0
H	0	0	0	0	0	2	0	2	0	1	0	0	0	0
I	0	0	0	0	2	0	2	0	3	0	1	1	1	0
J	0	0	0	1	1	2	0	1	0	5	2	2	3	0
K	0	0	0	0	2	0	0	0	1	2	3	2	2	0
L	0	0	0	0	2	0	0	0	1	2	2	3	2	0
M	0	0	0	1	1	1	0	0	1	3	2	2	4	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Da mesma forma, se tomarmos a terceira potência de A, temos o número de caminhos entre pares de nós que são de comprimento três. A quarta potência de A mostrará o número de trajetos entre pares de nós que são de comprimento quatro, e assim por diante.

Propriedades básicas de nós

As propriedades básicas de uma rede são derivadas das propriedades relevantes de seus nós que são o tema principal desta secção. Uma vez que uma rede é determinada pelos seus nós e as relações existentes entre eles, as propriedades básicas dos nós estão relacionadas com essas relações. O número e a estrutura das relações do nó determinam a sua influência ou, por outras palavras, o seu poder na rede.

Tabela 13. Propriedades básicas dos nós da rede da Figura 12

Nó	A	B	C	D	E	F	G
Grau	1	1	1	4	4	3	2
Proximid /	0,196	0,196	0,196	0,25	0,0286	0,25	0,204
Centralid / interméd.	0	0	0	30	37,17	14,5	2,17
Coef. *Clustering*	-	-	-	0	0,17	0	0

Nó	H	I	J	K	L	M	N
Grau	2	3	5	3	3	4	0
Proximid /	0,0196	0,0238	0,0270	0,0217	0,0217	0,0256	0,0054
Centralid / interméd.	1,33	10,67	18	0	0	6,17	0
Coef. *Clustering*	0	0	0,4	1	1	0,67	-

- significa que não está definido

Grau

Esta medida, a mais básica de todas, conta o número de relações do nó. No caso de uma rede com relações não direcionadas, o grau de nó é a soma da linha ou coluna correspondentes na matriz de adjacência.

Os graus dos nós da rede da Figura 12 são mostrados na Tabela 13.

No caso de relações direcionadas, distinguimos os graus-*in* e os grau-*out*. Os graus-*in* de um nó captam o número de nós a partir dos quais há relações apontando para o nó dado. Em sentido inverso, os graus-*out* de um nó é o número de nós para os quais há uma relação com origem no nó em avaliação. Os graus-*in* e os graus-*out* de uma rede podem ser a partir das linhas ou colunas correspondentes, na matriz de adjacência.

Distância

As distâncias entre os nós são características importantes, uma vez que determinam a forma de informação difusa na rede. A distância entre dois nós é calculada como o número mínimo de relações que a informação tem que percorrer para ir de um nó para o outro. Como é ilustrado na matriz de distâncias na Tabela 14, se não houver conexão entre dois nós, o valor da distância calculada é infinito. Note-se que, no caso de um gráfico com bordas não direcionadas, a matriz de distância é simétrica w.r.t. sua diagonal, no entanto, isso não se mantém no caso de gráficos com bordas direcionadas.

Tabela 14. A matriz de distâncias entre nós para o gráfico da Figura 12

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	2	2	1	2	3	4	5	4	3	4	4	3	∞
B	2	0	2	1	2	3	4	5	4	3	4	4	3	∞
C	2	2	0	1	2	3	4	5	4	3	4	4	3	∞
D	1	1	1	0	1	2	3	4	3	2	3	3	2	∞
E	2	2	2	1	0	1	2	3	2	1	2	2	1	∞
F	3	3	3	2	1	0	1	2	1	2	3	3	2	∞
G	4	4	4	3	2	1	0	1	2	3	4	4	3	∞
H	5	5	5	4	3	2	1	0	1	2	3	3	3	∞
I	4	4	4	3	2	1	2	1	0	1	2	2	2	∞
J	3	3	3	2	1	2	3	2	1	0	1	1	1	∞
K	4	4	4	3	2	3	4	3	2	1	0	1	1	∞
L	4	4	4	3	2	3	4	3	2	1	1	0	1	∞
M	3	3	3	2	1	2	3	3	2	1	1	1	0	∞
N	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	0

Proximidade

Esta medida reflete o quão acessível é um nó na rede de tal forma que os valores maiores indicam que esse nó está bem conectado aos outros nós da rede.

Para um determinado nó v, a sua medida de proximidade é

$$proximidade(v) = \frac{1}{\sum_{u \neq v} distância(u, v)}$$

Equação 4.

é calculada a partir da matriz de distâncias como 1 dividido pela soma de distâncias entre o nó dado v e todos os outros nós u≠v na rede.

Caso não haja conexão entre dois nós, em vez do valor infinito, o número de nós na rede é substituído na computação.

A proximidade é sensível ao tamanho da rede, diminuindo com o seu aumento.

O valor de proximidade do nó A da Figura 12, com base nas distâncias introduzidas na primeira fila da Tabela 14, é calculado como 1/(2+2+1+2+3+4+5+4+3+4+4+3+14)=0.0196.

Neste caso, em vez de distância(A,N)=∞, é usado distância(A,N)=14 (a rede tem 14 nós).

Centralidade intermédia

Esta medida é utilizada para avaliar o quão importante a posição de um nó v na rede é, sendo calculada como

$$centralIntermedia(v) = \sum_{u \neq v \neq t} \frac{ncmc_v(u,t)}{ncmc(u,t)}$$

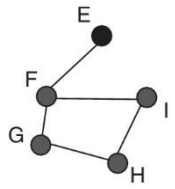
Equação 5.

onde *u* e *t* são pares de nós diferentes de *v*, $ncmc(u,t)$ é o número de caminhos mais curtos do nó *u* para nó *t* e $ncmc_v(u,t)$ é o número desses caminhos mais curtos de *u* para *t* que passam pelo nó *v*.

A centralidade intermédia de um nó mede o grau em que a informação tem de fluir através desse nó.

Uma vez que os cálculos da centralidade intermédia aumentam com o aumento do número de nós na rede, vamos fazer o cálculo de apenas parte da nossa rede social tal como se vê na Figura 13.

Figura 13. Parte da rede social da Figura 12



A centralidade intermédia do nó E é zero, uma vez que nenhum dos caminhos mais curtos entre quaisquer outros pares de nós atravessam E. Agora, vamos calcular a centralidade intermédia para o nó G: $ncmc\ (F,H) = 2$ porque existem 2 caminhos, ambos com distância 2 (ou seja, ambos são mais curtos) entre os nós F e H: o caminho $F \rightarrow G \rightarrow H$ e o caminho $F \rightarrow I \rightarrow H$. No entanto, apenas um deles passa pelo nó G, assim $ncmc_G\ (F,H) = 1$. Da mesma forma, $ncmc(E,H) = 2$ e $ncmc_G\ (E,H) = 1$. Uma vez que não há outros caminhos mais curtos que passem pelo nó G, podemos contar o seu valor de centralidade intermédia de acordo com a Equação 5 como

$$centralIntermedia(G) = \frac{ncmc_G(E,F)}{ncmc(E,F)} + \frac{ncmc_G(E,H)}{ncmc(E,H)} + \frac{ncmc_G(E,I)}{ncmc(E,I)} + \frac{ncmc_G(F,H)}{ncmc(F,H)} + \frac{ncmc_G(F,I)}{ncmc(F,I)} + \frac{ncmc_G(H,I)}{ncmc(H,I)} = \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{0}{1} = 1.$$

Da mesma forma, as centralidades intermédias dos outros nós também podem ser calculadas:

$$centralIntermedia(F) = 3{,}5;$$
$$centralIntermedia(H) = 0{,}5;$$
$$centralIntermedia(I) = 1$$

Coeficiente de agrupamento

Algumas pesquisas em grupo indicam que *triads*, ou seja, três nós conectados formando um triângulo, são formações importantes das quais uma ampla gama de relações sociais interessantes pode ser derivada. O coeficiente de agrupamento mede a tendência de um nó *v* ser incluído numa tríade e pode ser definido como

$$clust_{coef(v)} = \frac{\sum_{u \neq v \neq t} triangulo(u,v,t)}{\sum_{u \neq v \neq t} triplo(u,v,t)}$$

Equação 3.

onde $triangulo\ (u,v,t) = 1$ se os nós u, v e t estão conectados formando um triângulo, caso contrário, $triangulo(u,v,t) = 0$, e, $triplo(u,v,t) = 1$ se os nós u e t estão ambos conectados ao nó v, caso contrário, $triplo\ (u,v,t) = 0$.

Caso $grau(v) < 2$, o coeficiente de agrupamento é igual a zero ou não definido.

O coeficiente de agrupamento do nó E da rede da Figura 12 é calculado da seguinte forma: 4 nós estão conectados a E, formação de 6 triplos no total, de modo que os valores de $triplo(D,E,F)$, $triplo(D,E,M)$, $triplo(D,E,J)$, $triplo(F,E,M)$, $triplo(F,E,J)$ e $triplo(M,E,J)$ são iguais a 1 somando até 6. No entanto, há apenas um triângulo formado, de modo que o valor de $triangulo(M,E,J)$ é igual a 1. Assim, $clust\_coef\ (E) = 1/6 = 0{,}17$.

Propriedades básicas e estruturais das redes

As propriedades acima mencionadas (ver Tabela 13), também chamados como medidas de centralidade dos nós, estão relacionados com nós individuais de uma rede e caracterizam o seu 'poder' ou 'posição' na rede. No entanto, há também algumas propriedades básicas e estruturais relativas à rede toda ou a algumas partes dela. Será esse o tema das secções seguintes.

Diâmetro

O diâmetro de uma rede é definido como o mais longo de todas as distâncias entre os seus nós. Esta medida indica a facilidade com que os nós de uma rede são acedíveis. O diâmetro da rede da Figura 12 é a maior distância presente na matriz de distância, que é igual a 5, ou seja, a distância entre os nós A e H.

Centralidade

Como mostrado na Tabela 13, as pontuações de centralidade são irregulares para os nós do gráfico. Para medir essa desigualdade, as pontuações de centralidade ao nível da rede para a rede N com n nós podem ser calculados como:

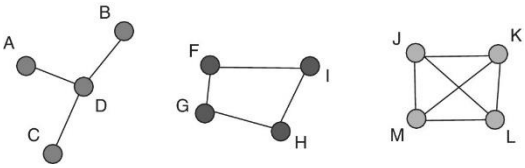
$$C(N) = \sum_{v} \left(\max_{u} c(u) - c(v) \right)$$

Equação 7.

onde $\max_{u} c(u)$ é a pontuação máxima de centralidade de todos os nós u da rede (incluindo o nó v), $c(v)$ é a pontuação centralidade do nó v e c pode ser substituído por medidas de grau, proximidade ou centralidade intermédia.

Vamos discutir as pontuações de centralidade de proximidade das três redes apresentadas na Figura 14.

Figura 14. Exemplo com três redes



As pontuações de proximidade dos nós A, B, C e D são 0,2; 0,2; 0,2; e 0,33, respetivamente, o máximo dos quais é 0,33. Assim, a pontuação de centralidade de proximidade da rede do lado esquerdo é

$$\begin{aligned} C^{proximidade}(esq) &= (0{,}33 - 0{,}2) \\ &+ (0{,}33 - 0{,}2) \\ &+ (0{,}33 - 0{,}2) \\ &+ (0{,}33 - 0{,}33) = 0{,}4. \end{aligned}$$

Para das outras duas redes, as pontuações de proximidade são iguais para todos os seus nós, ou seja, *0,25* para os nós F, G, H e I, e *0,33* para os nós J, K, L e M. Assim, $C^{proximidade}(meio) = 4 \times (0{,}25 - 0{,}25) = 0 = 4 \times (0{,}33 - 0{,}33) = C^{proximidade}(dir)$ onde 'dir' e 'meio' correspondem às redes da direita e do meio, respetivamente. A partir destes resultados, podemos ver que a rede

'esq' é mais centralizada do que as outras duas redes.
Em relação à rede da Figura 6, o seu grau de centralidade é $C^{grau}(N) = 0{,}187$, de proximidade é $\$C^{proximidade}(N) = 0{,}202$ e de centralidade intermédia é $C^{centralIntermedia}(N) = 0{,}395$.

O que significa uma grande pontuação de centralização de uma rede? No exemplo acima vimos alguns extremos, nomeadamente:

- a mais centralizada, a estrela;
- o menos centralizado, o anel; e a
- totalmente conectada.

Numa rede do tipo estrela, um nó é maximamente central, enquanto todos os outros são nós de centralidade mínima. Por outro lado, na rede do tipo anel ou rede totalmente conectada, todos os nós têm igual centralidade. Normalmente, as medidas de centralidade da rede são normalizadas no intervalo [0,1] de tal forma que as pontuações para duas redes são mais fáceis de comparar.

Clicks

Um *click* é um subconjunto de nós de tal forma que quaisquer dois nós no subconjunto estão conectados. Exemplo de *clicks* de tamanho três, ou seja, contendo três nós, na rede da Figura 6 são os seguintes subconjuntos: {E,J,M}, {J,K,L}, {J,K,M}, {J,L,M} e {K,L,M} enquanto há um *click* de tamanho quatro, ou seja, o subconjunto {J,K,L,M}.

Coeficiente de Clustering

Esta medida expressa a probabilidade de que os triplos na rede estejam conectados para formar um triângulo. Ele é calculado de forma semelhante ao coeficiente de agrupamento de nós como a proporção do número de triângulos para o número de triplos conectados na rede. Os coeficientes de agrupamento das redes esquerda e do meio (Figura 14) são iguais a zero, enquanto o coeficiente de agrupamento da rede direita no mesmo exemplo é igual a um. O coeficiente de agrupamento da rede de Figura 12 é 0,357.

Modularidade

A pontuação modular da rede expressa o grau em que uma rede exibe estruturas de cluster (muitas vezes chamadas de comunidades). A alta pontuação de modularidade de uma rede significa que seus nós podem ser divididos em grupos de tal forma que os nós dentro desses grupos estão densamente conectados, enquanto as conexões entre esses grupos não são densas. A modularidade da rede da Figura 12 é 0,44.

Tendências e observações finais

O cenário dos *media* sociais tem aumentado em grande parte nos últimos anos. Várias ferramentas de análise de redes sociais foram desenvolvidas dentro de várias disciplinas de ciência e engenharia. O interesse especial é dedicado à dinâmica das redes sociais, ou seja, como a rede está a evoluir com o tempo.
Neste capítulo, concentrámo-nos na análise de redes sociais mais do que em análise de dados. A razão é que, como discutido no Subcapítulo 5.3 sobre análise de texto, somente após a compreensão dos princípios básicos, como tokenização, radicalização ou saco de palavras, somos capazes de extrair conhecimento do texto e, consequentemente, transformar o texto num formato estruturado para posterior uso das técnicas referidas no Subcapítulo 5.1. Da mesma forma, ao entender as propriedades básicas discutidas sobre nós e redes, podemos ser capazes de extrair recursos úteis da rede, a fim de utilizá-los na análise adicional usando métodos descritivos, de diagnóstico, preditivos ou prescritivos de análise de dados. As direções mais comuns na análise de redes sociais são apresentadas de seguida.

Com a utilização crescente de *site*s de redes sociais, tais como o *Facebook* ou o *LinkedIn*, há uma necessidade de previsão de *link*, uma tarefa intimamente relacionada à classificação e técnicas de regressão discutido na secção referente à Análise Preditiva. O objetivo da previsão do *link* é prever quais as relações que surgirão com mais probabilidade entre os nós da rede. Também está relacionado com o problema de inferir conexões em falta na rede.

Um campo de estudo ativo é o uso de técnicas de análise de texto no contexto da análise de opinião e sentimento com vários casos de uso de aplicativos, como analisar o sentimento e a opinião da sociedade sobre acontecimentos reais ou entender as origens das notícias falsas.

A visualização das redes sociais é outro campo de estudo ao qual muita atenção tem sido dada recentemente. Como foi demonstrado, não é uma tarefa fácil, uma vez que uma boa visualização deve fazer os clusters e também os *outliers* identificáveis, bem como a capacidade de seguir os *link*s.

Outros campos de pesquisa em análise de redes sociais incluem a deteção de comunidades ou a deteção de padrões de interação, fortemente relacionados com técnicas de *clustering* e de análise de padrões frequentes.

6.
Search engines

Este capítulo identifica os conceitos subjacentes aos *Search engine*s: *Search engine Marketing* (SEM) e *Search engine Optimization* (SEO). Começa por explicar as diferenças entre os mesmos e quais as vantagens de uma estratégia de SEO. Posteriormente é apresentada como fazer a análise da concorrência e os fatores a considerar na otimização *on* e *off page*.

Por fim, são apresentados alguns tópicos a ter em conta na análise de resultados em SEO assim como algumas plataformas que o ajudam nesse sentido. Algumas dicas do *Google* assim como regras a seguir são também fundamentais numa estratégia de marketing digital nos *search engines*.

6.1. Como funcionam os *search engines*?

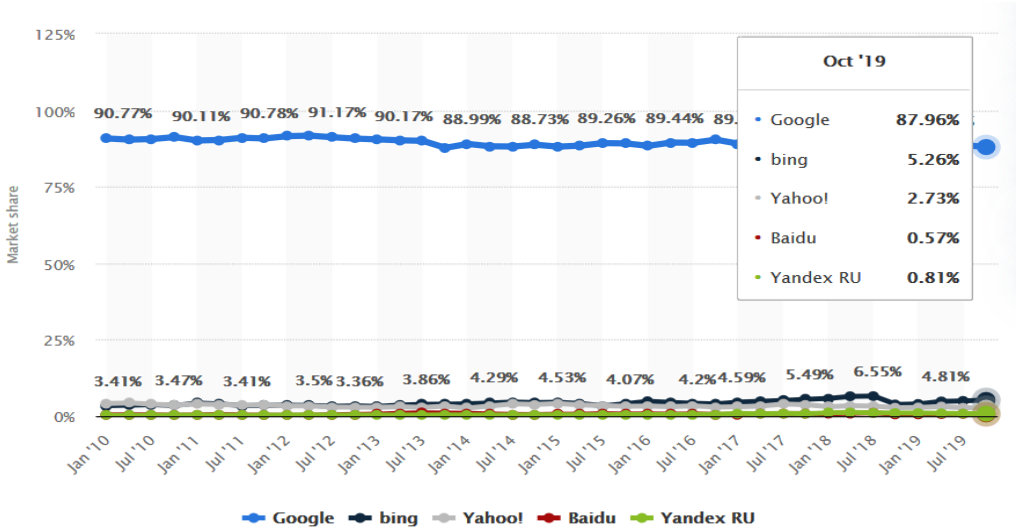
Web search engine ou Internet *search engine* é um *software* criado para realizar pesquisas na *World Wide Web* de forma sistemática através de informações especificadas numa consulta de pesquisa na *web* textual. Os resultados da pesquisa geralmente são apresentados numa linha de resultados, chamados de SERPs (*Search Engine Results Pages*). As informações podem ser uma mistura de *link*s para páginas da *web*, imagens, vídeos, infográficos, artigos, trabalhos de investigação e outros tipos de arquivos. Alguns mecanismos de pesquisa também extraem dados disponíveis em bancos de dados ou diretórios abertos. Ao contrário dos diretórios da *web*, mantidos apenas por editores humanos, os mecanismos de pesquisa também mantêm informações em tempo real executando um algoritmo num rastreador da *web*.

O conteúdo da Internet que não pode ser pesquisado por um mecanismo de pesquisa na *web* é geralmente descrito como deep *web*.

De acordo com o *site Statista* (2019), os 5 principais motores de busca do mundo em termos de participação de mercado são: *Google*, *Bing*, *Yahoo*, *Baidu* e *Yandex*. O *Google* é considerado o melhor motor de busca com uma participação de mercado mundial de 87,96%. De seguida, o *Bing* tem cerca de 5,29%. O *Baidu* possui uma participação de mercado global entre 0,57% mas é o motor de busca mais utilizado na China enquanto o *Yandex* (0,81%) é o mais popular na Rússia (conforme Figura 15).

Os motores de busca não pesquisam apenas páginas, mas também exibem os resultados em função da sua importância, através do funcionamento de algoritmos.

Figura 15. A utilização de motores de busca em todo o mundo



Fonte: *Statista*

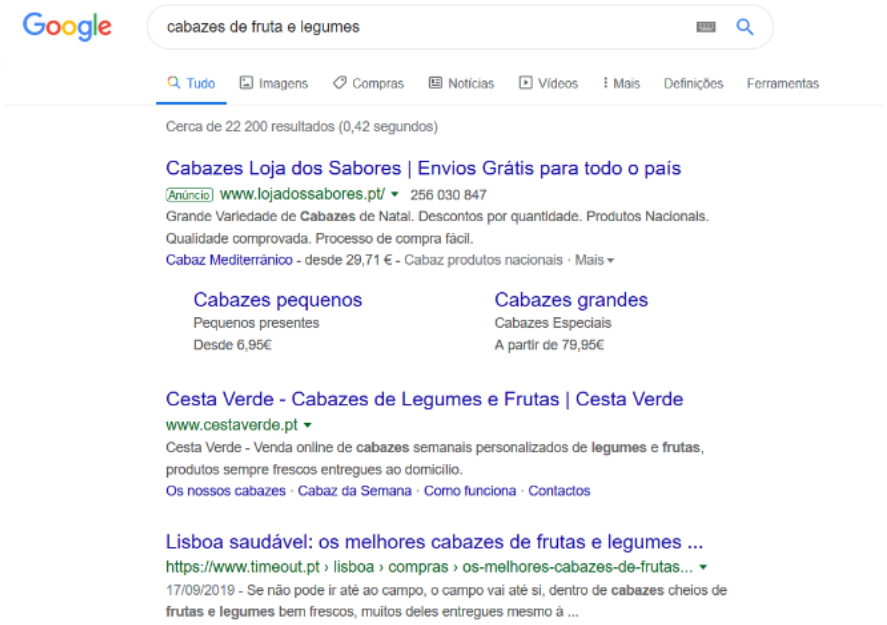
Os motores de busca recolhem os dados através de um *spider* ou *crawler*, que visita automaticamente as páginas e indexa o seu conteúdo. Um *spider* (ou robô) percorre a *web* de *link* em *link*, identificando novas páginas *web* ou páginas *web* alteradas e recolhem o URL de cada uma das páginas. Depois elabora um índice das palavras que encontra e a sua localização. Quando um utilizador faz uma pesquisa, o robô vai procurar nesse índice e o motor de busca apresenta os resultados mais relevantes desse índice.

6.2. *Search Engine Marketing* (SEM)

Cada vez mais, com o objetivo de um *site* ser encontrado nos motores de busca, ganham importância as técnicas de Marketing *Online*. Estas técnicas têm como o objetivo a promoção de um *website* nas páginas de resultados de um motor de busca (como o *Google*, por exemplo) e são conhecidas por SEM - *Search engine Marketing*. Esta é uma das maneiras mais eficazes de expandir negócios num mercado cada vez mais competitivo.

Search Engine Marketing é a prática de marketing de uma empresa que utiliza anúncios pagos que aparecem nas páginas de resultados dos motores de busca (ou *Search Engine* Results - SERPs). Para isso, os anunciantes utilizam plataformas (como o *Google Ads*) onde fazem ofertas por palavras-chave que os utilizadores inserem ao procurar determinados produtos ou serviços. Esses anúncios, geralmente conhecidos pelo termo "anúncios *pay-per-click*" têm vários formatos. Alguns são pequenos anúncios baseados em texto, enquanto outros, como anúncios de lista de produtos (PLAs, também conhecidos como anúncios do *shopping*), são anúncios mais visuais e baseados em produtos que permitem aos consumidores ver informações imediata, como preço e comentários. Quando os utilizadores pesquisam nos motores de busca podem ter acesso a anúncios (identificados como tal) e a resultados não pagos/orgânicos (conforme Figura 16).

Figura 16. Pesquisa no Google pela palavra-chave "cabazes de frutas e legumes"



O ponto forte do *Search Engine Marketing* é que oferece aos anunciantes a oportunidade de colocar os anúncios no momento em que os utilizadores pesquisam e estão mais dispostos a comprar. Nenhum outro meio de publicidade consegue fazer isso, e é por isso que o marketing em motores de busca é tão eficaz e uma forma poderosa de expandir os negócios.

6.3. *Search Engine Optimization* (SEO)

O Search Engine Optimisation (Otimização para os motores de busca) corresponde ao processo de trabalhar um *website*, para que seja indexado e classificado pelos motores de busca, ao mesmo tempo que aumenta a relevância dos *link*s para o *site* a partir de outros *site*s. O SEO é uma prática sem garantia de primeiros resultados, ou seja, nenhuma empresa ou pessoa pode garantir que seu *site* fique em primeiro lugar numa pesquisa. Por sua vez, o *pay-per-click* corresponde à publicidade paga nos motores de busca. O anunciante só paga pelos cliques recebidos dos utilizadores. Este é o modelo utilizado no *Google Ads* e através desta prática a empresa poderá garantir mais facilmente os primeiros lugares (pagos) nos resultados do *Google*.

SEO (Otimização para os motores de busca)

Processo de trabalhar um website, para que seja indexado e classificado pelos motores de busca (como por exemplo o *Google*, *Bing*, entre outros).

Objetivos e vantagens do SEO

O SEO é um processo dinâmico, com objetivos, ações, revisões e interações. Com objetivos a médio/longo prazo é importante perceber quais as vantagens, para o negócio, de uma posição elevada no ranking dos motores de busca:

1) Maior CTR (*Click-through-Rate*)

O CTR é o número de cliques que o anúncio recebeu a dividir pelo número de vezes que o anúncio foi mostrado, expresso em percentagem (cliques ÷ impressões = CTR).

Por Exemplo: se tiver 5 cliques e 1000 impressões, o CTR será de 0,5%.

Ao ter uma estratégia de SEO eficaz vai ajudá-lo a conseguir também melhores resultados em anúncios pagos (menor investimento e melhor posição).

2) Maior envolvimento

Se investir em SEO, os seus atuais ou potenciais clientes irão encontrar mais facilmente o seu *site* na *web*, isto irá traduzir-se numa maior proximidade e confiança na empresa.

3) Mais conversões

O SEO pode levar ao comportamento do utilizador através de conversões: uma compra, um registo, preencher um formulário, subscrição da *newsletter*, fazer um *download*; entre outros.
Isto traz vantagens para a reputação, uma vez que uma posição elevada no ranking dos motores de busca traz:

- Maior visibilidade da empresa/marca/organização;
- Reputação melhorada uma vez que é conferida maior confiança às empresas que se apresentam melhor posicionadas;
- Credibilidade - se aparece nos primeiros lugares os utilizadores consideram que é porque é a "melhor" da área;
- Liderança de mercado- se aparece bem posicionada é porque (à luz do utilizador) tem os melhores produtos/serviços, é líder onde atua;
- Vantagens competitivas, se aparece melhor posicionada que a concorrência é porque tem mais valor a oferecer que os seus semelhantes.

O SEO confere vários benefícios à empresa, mas deve sempre que possível ser complementado com uma estratégia de anúncios pagos. Para uma boa estratégia de SEO é necessário fazer uma análise detalhada da concorrência e definição das palavras-chave do negócio.

Análise da Concorrência

Quer nos resultados naturais/orgânicos quer nos resultados pagos é importante fazer uma análise aprofundada da concorrência. Nos resultados orgânicos devemos perceber quais são os tipos de *site* que o público pesquisa, que resultados aparecem quando fazem pesquisas e o que é oferecido aos utilizadores. Devemos analisar as ofertas que a concorrência esta a fazer assim como o *site* e a otimização que pratica.

Nos resultados pagos devemos identificar qual é a concorrência (que outras empresas estão também a pagar anúncios para as mesmas palavras-chave), quantas empresas são, que ofertas competitivas tem, como fazem as campanhas e perceber se devo rever a minha estratégia.

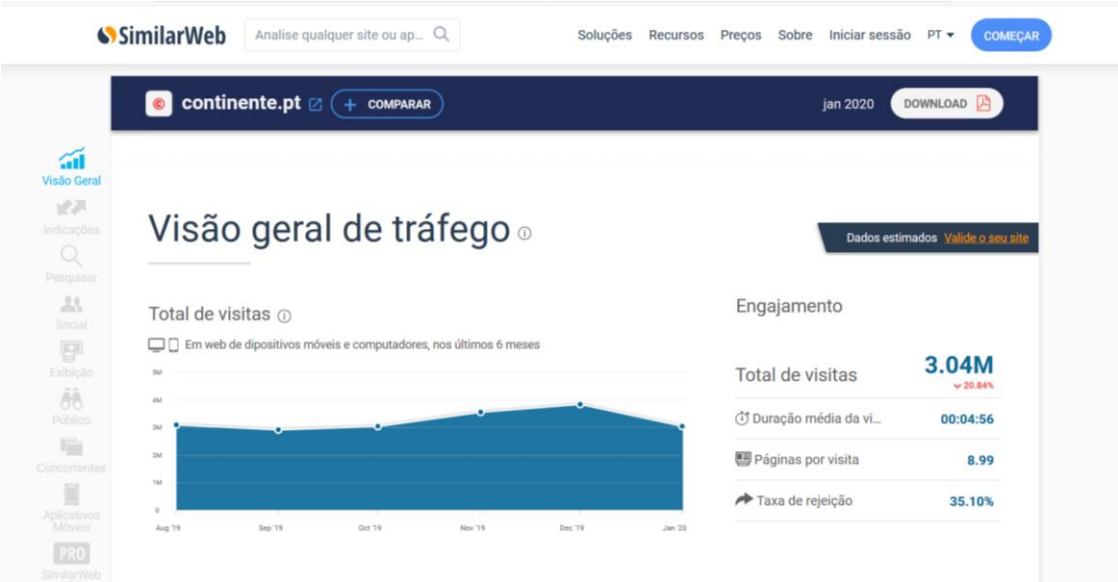
Através do *site* *SimilarWeb* (https://www.similar*web*.com) podemos analisar a posição da empresa e dos concorrentes em Portugal e no mundo, a evolução do nº de visitantes, taxa de rejeição, páginas vistas por utilizador, maior percentagem de palavras-chave pesquisadas para acesso ao *website*, entre outros (ver figura 3). Assim, podemos comparar qual a posição do *site* face aos meus concorrentes.

No *site* What's my SERP (www.whatsmyserp.com/) pode comparar o *site* com a concorrência em termos de posicionamento das palavras-chave nos motores de busca e perceber quais as melhorias que necessita fazer para melhorar o seu posicionamento.

Se utilizar o *Google* Trends (https://www.*Google*.pt/trends) poderá perceber quais são as tendências de pesquisa ao longo do tempo, comparar várias palavras-chave, saber pesquisas por localizações, entre outros. Por exemplo, conforme a Figura 17, pode-se perceber que as pesquisas pela palavra "legumes" nos últimos 12 meses são superiores a "azeite" e "frutas" assim como

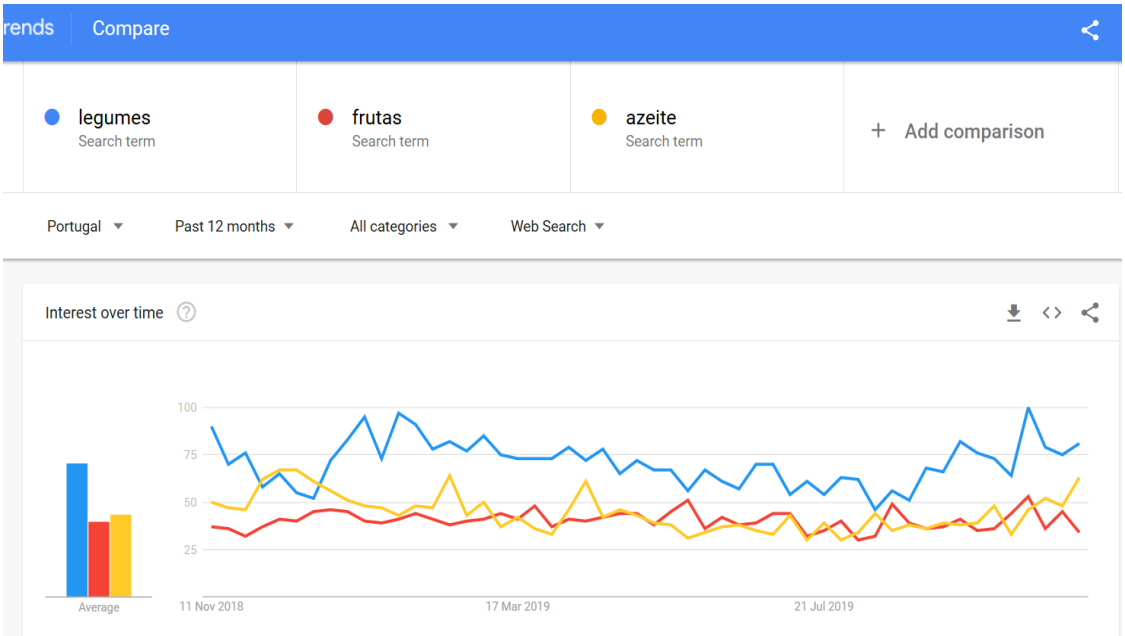
vem aumentando ao longo dos meses (como períodos de elevada procura).

Figura 17. Resultados de "envolvimento" no *site* da marca Continente



Fonte: SimilarWeb

Figura 18. Evolução da pesquisa nos últimos 12 meses por legumes, frutas e azeite



Fonte: *Google* Trends (novembro, 2019)

A estratégia de SEO assenta essencialmente em dois aspetos: otimização *on-page* e *off-page*.

Otimização On-Page

A otimização *on-page* corresponde a técnicas de otimização direcionadas para o *site* ou

blog, ou seja, são melhorias realizadas internamente.

Exemplos:

1. Pesquisa de Palavras-Chave
2. Estrutura e Navegação
3. *Meta Tags (Title, description, keywords...)*
4. Indexação (*Robots.txt, sitemap.xml*)
5. ALT das imagens
6. Subtítulos de conteúdos - *Heading Tags*

1. Pesquisa de palavras-chave

As palavras-chave são as palavras (ou um grupo delas) que descrevem o tema de um *site* ou o assunto de um texto, e são utilizadas pelas ferramentas de busca com o objetivo de apresentar resultados relevantes e precisos. Assim, as palavras-chave (*keywords*, em inglês) são aqueles termos que insere no *site* com o intuito de o descrever para as ferramentas de busca como o *Google*, *Bing* ou outros. Assim quando alguém estiver a procurar por algo que a sua empresa oferece, as ferramentas de busca sabem que o seu *site* deve aparecer nos resultados das buscas.

Exemplo: quando pesquisa "comprar legumes" as empresas que aparecem nos primeiros lugares são as que para o *Google* têm maior relevância para este termo de pesquisa.

O processo de seleção de palavras-chave envolve 4 passos: pesquisa *offline*, pesquisa *online*, validação e definição de prioridades. O SEO não é apenas destinado ao *Google*, sendo um desafio mais alargado. A pesquisa de palavras-chave significa encontrar as palavras/frases mais utilizadas pelos clientes nas suas pesquisas. Classifique as palavras em relação aos concorrentes, tendo em conta o tempo que demora a alcançar o posicionamento desejado e ao custo envolvido.

Pesquise palavras-chave de cauda longa, ou seja, pesquisas de pouco volume, para as quais se pretende uma boa posição. Podem ser palavras aparentemente pouco relevantes, mas que apresentam pouca concorrência e atraem tráfego muito qualificado aos *site*s. *Offline* pode perguntar aos seus clientes, promover inquéritos, fazer brainstormings, entre outros, enquanto *online* existem várias ferramentas que o ajudam nesta pesquisa.

As ferramentas de pesquisa de palavras-chave permitem filtrar as palavras, de acordo com critérios como datas, volumes de tráfego, localização, etc.

No *Google Keyword Planner* (https://AdWords.Google.com/o/KeywordTool) poderá aceder a este tipo de informação incluindo pesquisas por região geográfica, idioma, etc. No entanto, se não tiver uma conta de anúncios paga, esta ferramenta apenas lhe mostrará o intervalo de pesquisas por palavra-chave (por exemplo entre 1.000 a 10.000 pesquisas) sendo necessária uma pesquisa mais exata.

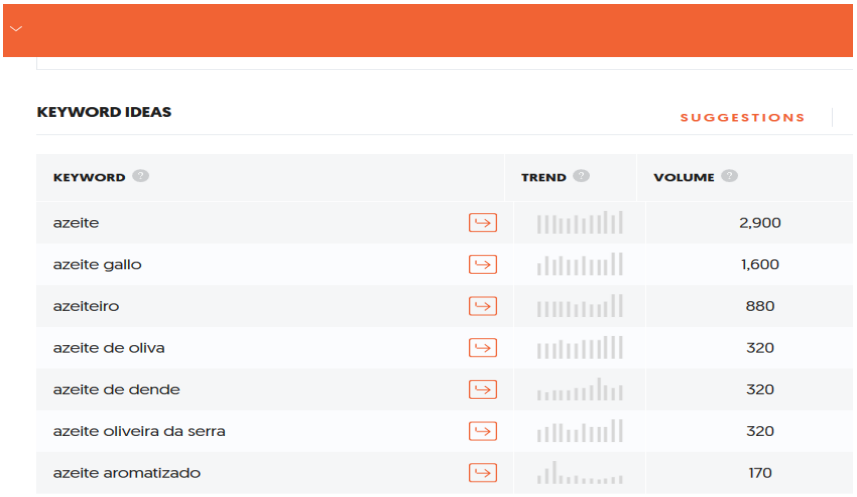
Com a ferramenta *Keyword finder* (https://kwfinder.com/) ou *ubbersugest* (https://neilpatel.com/ubersuggest/) poderá obter gratuitamente sugestões de palavras de acordo com a que introduziu. Na versão paga poderá aceder também ao número de pesquisas por palavra, custo por clique e concorrência no *Google Ads*.

Como pode ver na figura 5, ao utilizar a plataforma "*ubbersugest*" ela mostra-lhe o volume médio de pesquisas por mês para a palavra "azeite".

No entanto, é também muito importante mencionar as palavras-chave no conteúdo da página. Utilize sinónimos e variações das palavras no texto, não abusando da quantidade de palavras no conteúdo. Não existe densidade ideal. Alguns estudos indicam que a densidade de palavra-chave varia entre 1 e 3%, ou seja, a palavra-chave não deve ser repetida mais de duas ou três vezes por cada 100 palavras.

Para analisar a densidade das palavras-chave pode utilizar a ferramentas como o *wordcounter* (http://pt.wordcounter360.com/) que lhe indica para cada texto o número de vezes que aparece cada uma das palavras-chave (ver Figura 19).

Figura 19. Plataforma Ubersuggest - pesquisas por "azeite" nos últimos 12 meses

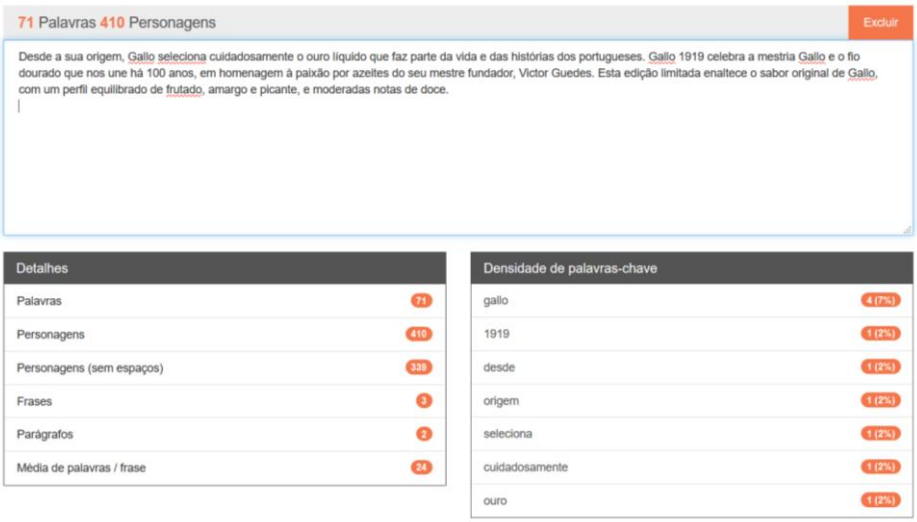


KEYWORD IDEAS — SUGGESTIONS

KEYWORD	TREND	VOLUME
azeite		2,900
azeite gallo		1,600
azeiteiro		880
azeite de oliva		320
azeite de dende		320
azeite oliveira da serra		320
azeite aromatizado		170

Fonte: Ubersuggest

Figura 20. Contagem de palavars em texto sobre a marca "Gallo"



1. Fonte: Ubersuggest

2. Estrutura e Navegação

Para uma correta arquitetura da informação existem 3 ingredientes fundamentais: arquitetura, usabilidade e SEO. Estes 3 fatores devem caminhar em sintonia no desenvolvimento do *site*. Uma boa arquitetura deve ser definida para a estrutura do *site*, sem prejudicar a usabilidade (experiência do utilizador ao visitar o *site*), e sempre acompanhada das melhores estratégias de SEO (para alcançar melhores resultados nos motores de busca).

A arquitetura da informação é tão importante para o *Google* como para o visitante. Por isso deve simplificar a navegação e experiência do utilizador no *site*, para que este não se sinta confuso e que permaneça o mais tempo possível dentro do *site* e já agora se possível volte muitas vezes.

No entanto, para além da questão do utilizador, um *site* com uma arquitetura bem organizada, facilita também igualmente a vida dos motores de busca (*Google*, etc).

Em resumo, a arquitetura da informação é uma das tarefas mais importantes de SEO, baseia-se na estruturação de ambientes de informação e serve para melhorar a experiência e usabilidade do *site* através do design do *layout* e como estruturas informações dentro desse *layout*, com o objetivo de aumentar a conversão do *site*.

Como resultado os robôs do *Google* veem que o *site* tem uma boa arquitetura e podem aumentar a relevância; o *site* fica melhor organizado; a disposição dos conteúdos e/ou produtos/serviços ficam otimizados; os utilizadores do *site* ficam satisfeitos por encontrarem o que precisam e têm uma boa experiência que resulta em mais conversões no *site*.

3. Meta-Tags

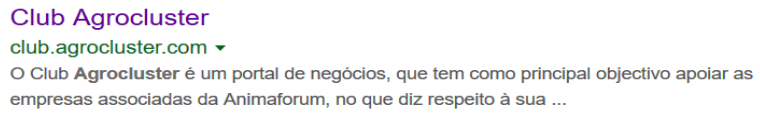
Uma página *web* é feita em HTML (*Hyper Text Mark-up Language*) - uma linguagem de programação para páginas *web* e contém elementos cruciais para SEO. Uma meta-tag relevante é um elemento HTML, que fornece informação sobre a página.

Exemplos de meta-tags são: a title tag, meta-description e meta-*keywords*.

Title Tag – Título da Página

A *Title Tag* é parte do HTML e é uma das *tags* mais importantes em termos de SEO, uma vez que é dos itens que mais influencia o posicionamento no *Google*. O objetivo do algoritmo do *Google* é conseguir mostrar às pessoas os resultados mais relevantes nas suas pesquisas. Se fizermos uma pesquisa no *Google* e nenhum dos resultados estiver relacionado com a palavra-chave que pesquisamos, vamos ter uma péssima experiência com o *Google*. O título dos resultados ajuda-nos a encontrar rapidamente o que procuramos. Cada página deverá ter uma *Title Tag* única, que reflita o conteúdo específico da página. O conteúdo da *Title Tag* aparece no topo do *browser* e é o principal *link* na maioria dos resultados de pesquisa. Na Figura 21 podemos ver que quando pesquisamos no *Google* por "Club Agrocluster" a página tem como título "Agro Cluster" e como descrição "O club Agrocluster é um portal de negócios..."

Figura 21. Informação no Google para "Club Agrocluster"



Meta-description – Descrição da Página

A *Meta description* faz a descrição do que a página contém e como funciona como atrativo de clientes, na página de resultados dos motores de busca. O *Google* procura mostrar sempre aos utilizadores a informação mais relevante para a sua pesquisa, da forma mais concisa possível. Prioritariamente, procura exibir o texto contido na *meta description* da página, mas se o algoritmo *Google* definir que outra informação é mais relevante é essa informação que será mostrada. É cada vez mais comum, por exemplo, *sites* de *e-commerce* terem informações de produtos exibidos com os preços na descrição. Se a palavra-chave procurada não se encontrar na *meta description*, há a possibilidade do *Google* procurar uma parte do conteúdo da página onde se encontra esse texto ao invés de exibir o texto da *meta description*. A *meta description* deve ser informativa, interessante e ao mesmo tempo *call-to-action*. Não deve ter mais de 160 caracteres.

Meta-Keyword

A *Meta Keyword* é uma lista de palavras-chave que estão incluídas no *site*. Uma vez que a

meta keyword não é mostrada ao utilizador em qualquer momento, seja nos resultados de pesquisa, seja no conteúdo do *site*, não há qualquer razão lógica para que o *Google* a considere como um fator de relevância. Apesar de não ter valor para os motores de busca, alguns sistemas de *social bookmarking* utilizam-nas para classificar o conteúdo da página (delicious.com). Cabe a si decidir que pretender colocar ou não as *keywords* no *website*.

4. Indexação (Robots.txt, *site*map.xml)

O *robots.txt* é um arquivo de texto com instruções para rastreadores de mecanismos de busca. Ele indica que áreas que os rastreadores de um *site* podem pesquisar. Através deste arquivo de texto simples, pode excluir facilmente domínios inteiros, diretórios completos, um ou mais subdiretórios ou arquivos individuais do rastreamento do mecanismo de pesquisa. No entanto, este arquivo não protege contra acesso não autorizado. O *Google* recomenda criar *sitemaps* com no máximo 50.000 URLs. Porém, é recomendado arquivos com no máximo 10.000 URLs. O sitemap.xml pode ser criado de 2 formas:

- Aplicativos – através de *site*s como o *GsiteCrawler* é possível rastrear o *site* simulando o robô dos motores de busca, armazenando as URLs encontradas e criando automaticamente o arquivo *site*map.xml;
- Ferramentas *online* para gerar *sitemaps* - *Site*s como www.xml-sitemaps.com ou http://www.auditmypc.com/free-sitemap-generator.asp criam sitemaps.xml sem a necessidade da instalar um programa no computador.

Como submeter no Google um Sitemap.xml?

Uma vez gerado o(s) arquivo(s) sitemap, este deve ser colocado preferencialmente no diretório principal do *site*. Em seguida, deve informar o *Google* da localização e nome do arquivo. Para isso deve aceder à ferramenta *Webmasters* do *Google* na área de submissão de *sitemaps.xml* e acompanhar o *status* de indexação das páginas e arquivos submetidos.

5. ALT das imagens

O ALT é o comando utilizado para identificar o texto alternativo que é mostrado quando a imagem não é carregada. A sintaxe de imagens em HTML é:

O atributo ALT é utilizado para facilitar a leitura de imagens em *site*s pelo *Google*, assim, é importante para o ranqueamento do *site* no *Google Images*. A função principal desta *tag* é oferecer uma descrição alternativa para a imagem e funciona, também, como texto âncora quando a imagem é usada como *link*.

O ALT *text* é mostrado pelos *browser*s quando por algum motivo não mostram as imagens de determinados *site*s, ou enquanto estas imagens são carregadas ou, até mesmo, por não estarem mais disponíveis, sendo mostrado o texto no lugar de tais imagens desativadas.

Deve adicionar a palavra-chave também na imagem através do ALT. Utilize sempre as palavras-chave no nome do arquivo, evite nomes genéricos como "imagem1.jpg" ou "DSC046651.JPG". Utilize hífens em vez de *underscores* quando o nome for mais que uma palavra.

Exemplo: azeite-virgem-extra.jpg

Nas imagens deve utilizar formatos padrão como: jpg, png e gif.

A qualidade da imagem pode interferir na velocidade de carregamento, ou seja, quanto maior a qualidade mais devagar o carregamento da página e isso pode influenciar a permanência ou não do utilizador no *site*. A resolução da imagem quanto maior for, melhor. No entanto, também pesa no carregamento do *site*. É importante encontrar

o equilíbrio entre qualidade e tamanho da imagem.

6. Subtítulos de conteúdos - *Heading Tags*

As *Heading Tags* (H1, H2, H3, etc.) são recursos de programação HTML utilizados para destacar títulos e subtítulos numa página. H1 é a abreviação do inglês para *Header* 1, ou Cabeçalho 1, logo, é o mais importante dos headers. Conceitualmente, o H1 possui um destaque maior, uma fonte maior, e é geralmente o elemento de texto mais visível da página. Assim como nos negritos de uma página e o *Title*, a *tag* H1 é um importante elemento que o *Google* utiliza para determinar o principal assunto abordado numa página, visto que o título de uma página conceitualmente define seu conteúdo.

No fundo, servem como marcação de capítulos e subcapítulos de um livro. Como podemos ver na Figura 22 o título principal (H1) remete para o tema/palavra-chave principal do artigo (Marketing Digital) enquanto o título seguinte (H2) especifica as diferentes abordagens.

Figura 22.H1 – "Marketing Digital – ainda faz sentido tratá-lo separado da Estratégia de Marketing Global"

Figura 23. H2 "Marketing Digital ou Omnichannel"

A tag <h1> deve ser usada apenas uma vez e deve conter um resumo do que se trata a página.
Cada página tem um h1 único.

Exemplo:
<h1> SEO: Otimize o seu *site* </h1>
Aqui é definido o tema principal da página. A página é sobre SEO.
<h2> Estratégias de SEO</h2>

Esta é uma subsecção de H1, com *keywords* relacionadas com a página. Isto ajuda não só a aparecer nos motores de busca como ao utilizador classificar o que é o que o *site* oferece.

Otimização off-page

A Optimização *off-page* preocupa-se essencialmente com melhorar o ranking do *site* nos motores de busca. Aspetos a verificar para melhorar o ranking do *site*:

1. *Link Building*
2. Formato de *links*
3. Marketing de conteúdo
4. *Social Linking*

1. *Link*building

Link building é o processo de aquisição de *hiperlinks* de outros *sites* para o *site* da empresa. Um *hiperlink* (geralmente chamado

de *link*) é uma maneira de os utilizadores navegarem entre as páginas da Internet. Os mecanismos de pesquisa usam *link*s para rastrear a *web*. Eles rastrearão os *link*s entre as páginas individuais do *site* e os *link*s entre *site*s terceiros. Existem muitas técnicas para criar *link*s, no entanto o *link* building é uma das partes mais difíceis do trabalho de SEO.

Existem três tipos de angariação de *link*s: 1) *link*s naturais (alcançados por *site*s e páginas que pretendem *link*ar para o nosso conteúdo ou empresa); 2) *link building* manual (contacto direto com *blog*gers/*site*s para que estes publiquem posts ou corrijam *link*s existentes, mas que estejam quebrados); 3) *Link*s não editoriais, criação própria (inseridos em diretórios de grande qualidade, como o DMOZ ou outras listagens de negócios locais).

Para trabalhar o *link building* do *site* deve: 1) Fazer com que os clientes coloquem *link*s para o *site*; 2) Ter um *blog* com informação e conteúdos interessantes/relevantes; 3) Ter conteúdos que motive a partilha, que se tornem virais; 4) Chamar a atenção da imprensa, *blog*gers e media, para o lançamento de novos produtos, ofertas, promoções.

É importante estar presente *online*, especialmente nas principais redes sociais (*Facebook*, *Instagram*. *YouTube*, *Clickstream*, etc.), pois uma referência pode ser mais rapidamente encontrada aí.

O que são backlinks de qualidade?

Conseguir *backlinks* de qualidade de forma natural é a forma correta para aumentar o posicionamento da página vai evitar ser penalizado pelo *Google*. São considerados *link*s de qualidade os que são em *website*s com uma autoridade de Domínio (DA) bastante elevada. Além disso, é fundamental também a autoridade da página (PA) bastante elevada. Mas o que é a qualidade? É uma medida de 0 a 100, criado pelo algoritmo da MOZ e baseado em fatores de ranqueamento do *Google*. São dados relativamente fiáveis e de confiança e pode fazer *download* da MOZ bar para analisar a qualidade dos *link*s: http://moz.com/tools/seo-toolbar (ver figura 24).

Link Baiting

Esta estratégia que consiste em criar um conteúdo de tal forma interessante que naturalmente atrai *link*s externos editoriais de qualidade. O termo terá origem no mundo do SEO, referindo-se a conteúdos muito eficientes na angariação de *link*s externos atrativos, com efeitos positivos a três níveis:

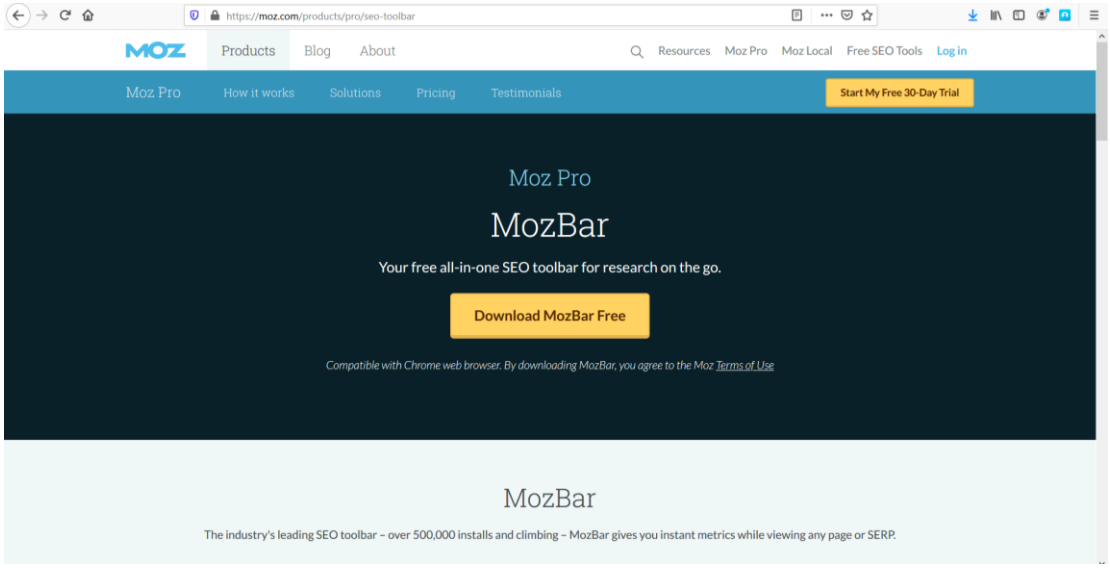
1. Rankings de pesquisa melhorados por os conteúdos serem atraentes, mais utilizadores irão *link*ar para eles, com impacto direto no SEO e por isso será melhor a classificação orgânica.
2. Aumento de tráfego e melhoria nas conversões.
   Poder gerar-se um efeito de viralidade, com os conteúdos a serem partilhados em maior quantidade.
   Daqui resulta um aumento natural de tráfego no *site* e em consequência, maiores são as possibilidades de conversão.
3. Exposição a novos públicos.
   A viralidade indicada antes tem ainda outro fator de interesse: o aumento da exposição do *site* a outros públicos. Em resultado, o topo do funil de conversões alarga-se, aumentando o seu alcance e gerando ainda mais possibilidades de conversão.

Resumidamente, o *Link Baiting* deve considerar vários pontos, na preparação do seu conteúdo:

- Envolver os decisores e influenciadores;
- Ter conteúdos fáceis de compreender;
- Oferecer qualquer coisa, sem custos;
- Abordar temas muito importantes e apelativos;
- Incluir conteúdos muito personalizados;
- Afirmar-se como o conteúdo de referência;

- Ser visualmente atrativo (infografias…);
- Utilizar botões sociais, para partilha rápida.

Figura 24. Ferramenta MOZ bar que permite analisar a qualidade dos *links*



Fonte: Moz

Ao pensar nestes tópicos quando cria o seu conteúdo irá fazer com o mesmo seja indexado e posicionamento mais favoravelmente no *Google*.

Como angariar links?

Existem muitas formas distintas de angariar *link*s para um *website*. Exemplos: 1) Editorial - *Link*s fornecidos de forma natural em virtude da qualidade do conteúdo do *site* (o desafio além de criar o conteúdo é fazer com que ele chegue ao conhecimento das pessoas certas). 2) Manual - *Link*s ganhos na sequência de um esforço e contacto do gestor com o *site* que fornece o *link*. 3) Autocriados - *Link*s criados pelo próprio gestor em *site*s que permitem a sua edição (wikis, fóruns, comentários de *blogs*…). São pouco valorizados e podem em alguns casos ser vistos como *web* SPAM. 4) Pedir *Link*s a Parceiros - Angariar *link*s junto de clientes, fornecedores, associações a que se pertence, etc; 5) Assinatura de *Email*s - Colocar na assinatura dos *email*s da empresa de forma clara o endereço do *site*. 6) Contactar *site*s que usam as nossas imagens - Para pedir que mencionem a fonte com um *link*. 7) Menções à Marca - Procurar por *site*s que já mencionam a marca, mas que por alguma razão não fizeram o *link*. 7) *Link*s Quebrados - Procurar por *link*s para o nosso *site* que por alguma razão não estejam funcionais e pedir que sejam corrigidos.

2. Formatos de *link*s

Um *link* (*hyperlink*) é uma referência a outro *website* ou página, que os utilizadores podem seguir, de maneira a chegarem a esse *website*. Os *link*s têm dois elementos: *link* de texto (texto que aparece na página); e *link* URL (o destino para onde o utilizador será levado ao clicar no *link*).

Um *link* de entrada (inbound *link*) é uma ligação de um *site* para outro enquanto m *link* interno é uma ligação de uma página de um *site* para outra página dentro do mesmo *site*.

- *Link* de entrada – ex.: *Blog* de vida saudável tem um *link* para o *site* de produtos hortícolas.
- *Link* interno – ex.: A *homepage* de um *site* produtos hortícolas tem um *link* para a página de alfaces.

Os *link*s de entrada são importantes como fator de determinação de rankings que reconhece o valor dos *link*s de entrada, ao conferir autoridade e credibilidade.

Os *website*s que recebem muitos *link*s, ou seja, que estão presentes em muitas fontes externas, são mais propensos a ter boas classificações por parte dos motores de busca, uma vez que o conteúdo e informação são de qualidade. Os *link*s internos são importantes para ajudar a navegação e os motores de busca a reconhecer o conteúdo.

Apesar de não ser o único fator de classificação usado pelos motores de busca, os *link*s de entrada (*inbound link*s) são também muito importantes.

Por outro lado, ao estar presente em muitas fontes externas, mais pessoas encontram o *link* direto para o *website*, estando mais propensas a clicar para conhecer o conteúdo do *website*.

Existem diferenças de qualidade entre os *link*s de entrada obtidos em diretórios genéricos, artigos, *blogs* até *site*s especializados. Para os motores de pesquisa, quanto mais difícil for conseguir angariar *link*s em determinados *site*s (*site*s de prestígio) maior a importância a ser dada. Como podemos ver na figura 7 *link*s em diretórios e fóruns tem uma qualidade baixa pois conferem pouca credibilidade ou confiança.

Figura 25. Importância dos *links* por tipo

No entanto, são também vários os fatores que causam problemas no *link building*:

- Conteúdos em *flash;*
- Vendedores de *link*s;
- Excesso de palavras-chave;
- Conteúdos duplicados;
- *Link*s quebrados;
- *Tags* de *No Follow*.

Numa estratégia de SEO *off-page* é importante também considerar a criação de conteúdo e a otimização deste.

3. Marketing de conteúdo

O marketing de conteúdo atrai potenciais clientes e transforma-os em clientes, criando e partilhando conteúdo gratuito de valor. O marketing de conteúdo ajuda as empresas a criar lealdade à marca ao longo do tempo, fornece informações aos consumidores e cria disposição de comprar produtos da empresa no futuro. Esta forma relativamente nova de marketing não envolve vendas diretamente. Em vez disso, cria confiança e relacionamento com o público. Os motores de busca valorizam os *site*s que apresentem conteúdos únicos, relevantes e atualizados. As atualizações de conteúdos podem incluir um conjunto variado de aspetos: texto, imagens, vídeos, *slides*, etc.

A apresentação de conteúdos otimizada para SEO inclui cabeçalhos de página e de parágrafos, textos a negrito, *link*s, imagens e vídeos. Deve procurar garantir a consistência nos conteúdos ao longo das páginas de um *site*. A densidade de palavras-chave é bastante importante e corresponde à percentagem de vezes que uma *keyword* aparece numa página, em relação ao número total de palavras nessa página.

A densidade de palavras ótima corresponde ao encontrar o equilíbrio entre escrever para humanos e escrever para os motores de busca.

4. Social *Link*king

Social linking significa utilizar a atividade nas redes sociais e ligação a *site*s de confiança nas contas sociais para ajudar a classificar o tráfego orgânico do seu *site* nos mecanismos de pesquisa. Esta pode ser uma ferramenta poderosa para o seu negócio e não deve ser ignorada.

O aumento da atividade nas páginas nas redes sociais permite levar mais pessoas ao seu *site* e, também melhorar a classificação em SEO (conduzindo também a mais vendas). Os gostos e envolvimento com a marca nas redes sociais irá ajudar a aumentar uma base de fãs leais. Usar estas plataformas para vincular a conteúdo útil no *site*, como publicações no *blog*, permite que mais pessoas tenham acesso à sua marca e informação. Quanto mais altos os níveis de envolvimento, maior a classificação de SEO.

Também a autoridade social é importante, se os mecanismos de pesquisa perceberem que as pessoas estão a partilhar os seus *link*s nas redes sociais, estes serão considerados autênticos e o seu *site* terá uma classificação mais elevada para o *Google*.

Os robôs de pesquisa do *Google* visitam o seu *site* aproximadamente uma vez por semana. Portanto, depois de criar uma nova página e colocar o conteúdo, é necessário aguardar para ser indexado. Nas redes sociais, o conteúdo é publicado a cada segundo, o que facilita a otimização mais rápida dos seus conteúdos pelos motores de busca.

Análise de resultados em SEO

A análise analítica pode ajudar na elaboração de relatórios e na medição da performance de SEO de um *site*. Na ferramenta de *Analytics* pode avaliar o nível de tráfego, o conteúdo consultado, objetivos alcançados, etc.
As métricas são essenciais para medir o desempenho do SEO.

Quais os critérios de comparação de desempenho?

- *Site* visível para palavras selecionadas?
- Número de páginas indexadas?
- Ranking do *site* para as palavras selecionadas?
- Aspetos técnicos de SEO completos?
- Que concorrentes estão melhor classificados?
- Existe tráfego orgânico para páginas chave do *site*?
- Existem conversões nas páginas chave, feitas por tráfego orgânico?

Mantenha um registo do desempenho destes indicadores, para perceber a evolução das atividades de SEO.

Como aumentar o desempenho de SEO?

- Deve manter um registo semanal e mensal da performance do *site*, em comparação com as referências.
- Sempre que aconteçam eventos relevantes que possam conduzir tráfego orgânico para o *site*, deve registá-los como: atividades de RP, marketing, concursos, etc.
- Avaliar o impacto das atividades, detetar tendências e preparar ações de resposta em caso de necessidade.
- Verificar mensalmente se o *site* continua a ser encontrado para as palavras e frases chave.

Como está o meu site?

Algumas ferramentas que o ajudam a fazer o check-up de como está o seu *site* a nível de SEO:
https://*website*.grader.com
http://seo*site*checkup.com/
https://www.woorank.com/pt

No SEO a análise de resultados deve ser constante e contínua e para isso existem várias ferramentas que o podem ajudar. Através do *Google Analytics* (https://www.*Google*.com/*Analytics*/) por exemplo poderá ter acesso ao número de visitantes, páginas vistas, taxa de rejeição, localização, *browser* entre outros que lhe permitiram melhorar as estratégias de SEO.

Os indicadores de desempenho mais comuns em SEO são:

- Relatórios de tráfego orgânico
  - Número de visitantes
  - Page views
  - CTR (*Click-Through-Rate*)
- Envolvimento com o conteúdo
  - *Downloads*
  - Conversões
  - Vendas
  - Envolvimento com o *website*

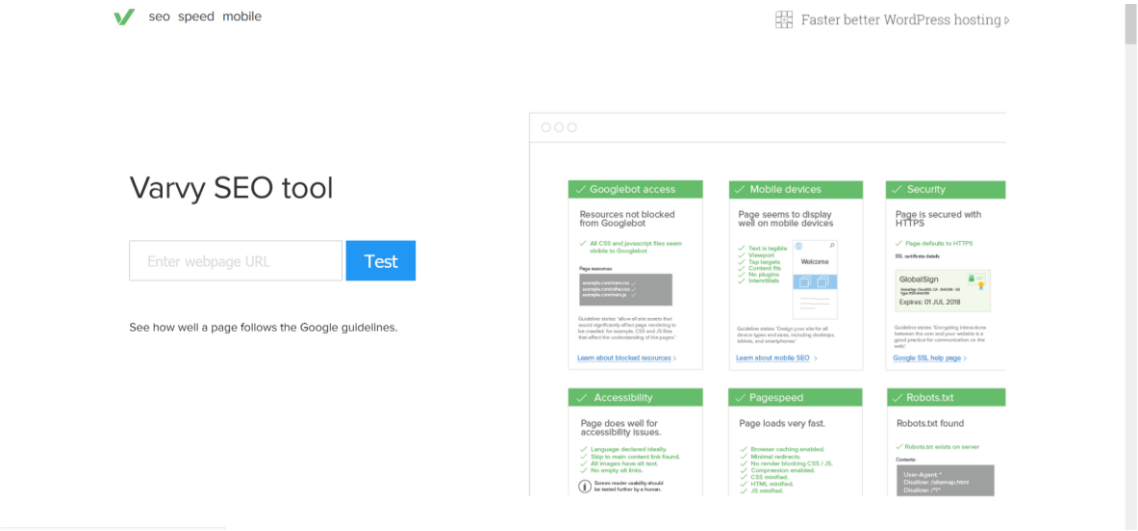
Por outro lado, deve conhecer as diretrizes e boas práticas do *Google* para não ser penalizado.

Diretrizes do *Google*:
bit.ly/directrizes-*Google*

Ferramenta para analisar se o seu *site* cumpre as boas práticas:
https://www.feedthebot.com/
Guia básico de otimização do *Google*:
bit.ly/guia-otimização

A ferramenta *Webmaster Tools* do *Google* fornece relatórios sobre indexação, classificação, penalizações, erros de pesquisa, localização e segmentação geográfica, etc. assim como permite inserir o mapa do *site*. O *Webmaster Tools* permite aos proprietários do *site* ver como este interage com os motores de busca e é essencial numa estratégia de SEO.

Figura 26. Ferramenta para analisar se o site cumpre as boas práticas



Fonte: Google Webmasters Tools

7.
Síntese

Chegados ao fim do livro, importa sintetizar os conteúdos que foram sendo expostos neste *ebook* sobre *Web Analytics*.

No Capítulo 2 introduziram-se os conceitos principais em torno do *Web Analytics*. Começamos por definir as principais métricas (acessos, visualizações de páginas, visitas, visitantes únicos, referenciadores, palavras-chave e frases-chave). De seguida, apresentámos os fundamentos de *Web Analytics* para as empresas, focando nos aspetos "Para que serve", "como podem ser recolhidos os dados" e "como deve usado" o *Web Analytics*?

Depois, apresentámos uma ferramenta que já se tornou num clássico no *Web Analytics* – o *Google Analytics*. Finalmente, abordou-se uma forma diferente de fazer *Web Analytics* que não se baseia nas métricas definidas anteriormente sobre *clickstream*, mas sim através de dados da *web* provenientes de *social media* (tais como *Clickstream*, *Facebook*, *Instagram*, *LinkedIn*, etc.).

Sendo este um livro sobre *Web Analytics*, tema este indissociável da ciência dos dados ou análise de dados, importa saber uma metodologia para desenvolver projetos nesta área. A metodologia CRISP-DM, descrita no Capítulo 3, não sendo uma metodologia específica de *Web Analytics*, é uma metodologia adequada para a área da ciência dos dados e, como tal, também adequada para *Web Analytics*. Esta metodologia é composta por seis fases e apesar de já ter cerca de vinte anos, continua a ser a metodologia mais utilizada. É especialmente adequada para desenvolver projetos nos temas descritos no Capítulo 5.

No Capítulo 4 abordámos as várias formas de recolher dados da *Web*, desde a extração de textos nas páginas *web* (*web crawling*), à possibilidade de instalação de mecanismos para obter informações dos utilizadores (*cookies*), até à recolha de dados para efeitos de caracterização dos utilizadores de um *web site* (*logs* e estatísticas dos motores de busca). No Capítulo 4 introduziu-se também o conceito de *web crawler*, um algoritmo utilizado para analisar o código de um *website* e recolher informações. De seguida abordaram-se os *cookies*, que armazenam dados sobre os utilizadores e, finalmente, os ficheiros registam a atividade num ficheiro de texto.

O Capítulo 5 é capítulo do livro dedicado à análise de dados na *Web*. Mas antes de abordar questões específicas da *Web*, começa-se por definir as grandes famílias de problemas de análise de dados. Depois descrevem-se três tipos de problemas analíticos que são especialmente relevantes no contexto *web*, nomeadamente: sistemas de recomendação, análise de texto e análise de redes sociais.

No Capítulo 6 abordamos as técnicas SEM - *Search Engine Marketing* que têm como o objetivo a promoção de um *website* nas páginas de resultado de um motor de busca (como o *Google*, por exemplo). Neste sentido as empresas podem utilizar anúncios pagos que aparecem nas páginas de resultados dos motores de busca (ou *Search Engine Results* - SERPs) ou técnicas de *Search Engine Optimisation* (Otimização para os motores de busca). Estas correspondem ao processo de trabalhar um *website*, para que seja indexado e classificado pelos motores de busca, ao mesmo tempo que aumenta a relevância dos *link*s para o *site* a partir de outros *site*s.

Quer nos resultados naturais/orgânicos quer nos resultados pagos é importante fazer uma análise aprofundada da concorrência.

A otimização *on-page* corresponde a técnicas de otimização direcionadas para o *site* ou *blog*, ou seja, são melhorias realizadas internamente através da 1) Pesquisa de Palavras-Chave; 2) Estrutura e Navegação; 3) Meta Tags (Title, description, *keywords*...); 4) Indexação (Robots.txt, *site*map.xml); 5) ALT das imagens; 6) Subtítulos de conteúdos - *Heading Tags*, entre outros.

A otimização *off-page* preocupa-se essencialmente com melhorar o ranking do *site* nos motores de busca. Entre os aspetos a verificar para melhorar o ranking do *site* estão: 1) *Link Building*; 2) Formato de *link*s; 3) Marketing de conteúdo e 4) *Social Linking*.

Para ter sucesso *online* é importante aplicar estes conceitos ao seu *website* e medir os respetivos resultados.

Bibliografia

Avinash Kaushik (2009), *Web Analytics* 2.0, The Art Of *Online* Accountability And Science Of Customer CentricityJohn Wiley and Sons

Ramesh Sharda, Dursun Delen, Efraim Turban, Business Intelligence: A Managerial Perspective on *Analytics*, 3rd Edition, 2014 |Pearson

Pedro Sostre, Jennifer LeClaire, 2007, *Web Analytics* For Dummies on Amazon.com. Paperback: 384 pages Publisher: For Dummies; 1st edition

Alghalit, N. (2015), *Web Analytics*: Enhancing Customer Relationship Management, Journal of Strategic Innovation and Sustainability Vol. 10(2) 2015

Mikhail Klassen, Matthew A. Russell, 2019, Mining the Social *Web*, 3rd Edition , Publisher(s): O'Reilly Media, Inc.

He & McAuley, 2016 Amazon Data available at: http://jmcauley.ucsd.edu/data/amazon/

Gama, João et al., Extração de conhecimento de dados: data mining, Edições Sílabo, 2015 (3ª edição).

Porter, M., An algorithm for suffix stripping. Program, 14 (3), 130–137, 1980.

Ricci, F. et al., Recommender Systems, Handbook, Springer-Verlag, 2010 (1ª edição).

Statista (2019). Worldwide desktop market share of leading *search engine*s from January 2010 to July 2019. Acedido a 3 de novembro de 2019, em: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Weiss, S.M. et al., Fundamentals of Predictive Text Mining, Springer, 2015 (2ª edição).

Witten, I.H., Frank, E., and Hall, M.A., Data Mining: Practical *Machine learning* Tools and Techniques, Morgan Kaufmann, 2011 (3ª edição).

E-AGRO
MARKETS

E-AGRO
INNOVATION

agrocluster

Cofinanciado por:

COMPETE 2020

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Europeu
de Desenvolvimento Regional