

IPBeja

INSTITUTO POLITÉCNICO
DE BEJA

Escola Superior de Tecnologia e Gestão
Licenciatura em Engenharia Informática

Sistema de apoio à promoção do turismo rural

**Sistema de apoio à promoção do património no contexto turístico
baseado em Data Mining**

Gonçalo Amaro

*Pedro Tomás
Vítor Abreu*

Beja, 9 de Fevereiro de 2022

INSTITUTO POLITÉCNICO DE BEJA

Escola Superior de Tecnologia e Gestão

Licenciatura em Engenharia Informática

Sistema de apoio à promoção do turismo rural

**Sistema de apoio à promoção do património no contexto turístico
baseado em Data Mining**

Gonçalo Amaro

Pedro Tomás

Vítor Abreu

Orientado por :

Doutora Isabel Brito, IPBeja

Relatório de Projeto Final, realizado na cadeira de Projeto Integrado, apresentado na
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Beja

Resumo

Sistema de apoio à promoção do turismo rural

Sistema de apoio à promoção do património no contexto turístico baseado em Data Mining

Tendo em vista uma sustentável, duradora e benéfica relação entre o turismo e o património histórico-cultural há que seguir uma estratégia. Existem várias estratégias, porém a tomada na elaboração deste trabalho foi a monitorização de fluxos de visitantes. Esta monitorização pode ser realizada através de um sistema de informação como o TripAdvisor, Zoomato e Booking, que foram os escolhidos na elaboração do trabalho, o armazenamento dos dados, ou seja, o desenvolvimento de uma base de dados em SQL, o processamento de dados usando técnicas para normalizar e analizar textos, assim como a extração das "keywords" e análise de opinião que seriam mais tarde úteis na elaboração de gráficos como medida para uma fácil visualização dos resultados acerca dos resultados obtidos, indicando muitos aspectos interessantes acerca das preferências turísticas dentro dos patrimónios.

Abstract

Sistema de apoio à promoção do turismo rural

Sistema de apoio à promoção do património no contexto turístico baseado em Data Mining

With a view to a sustainable, lasting and beneficial relationship between tourism and the historical-cultural heritage, a strategy must be followed. There are several strategies, however the decision taken in the elaboration of this work was the monitoring of visitor flows. This monitoring can be carried out through an information system such as TripAdvisor, Zoomato and Booking, which were chosen in the elaboration of this work, the storage of data, that is, the development of a database in SQL, the processing of data using techniques to normalize and analyze texts, as well as the extraction of "keywords" and opinion analysis that would later be useful in the elaboration of graphs as a measure for an easy visualization of the results about the results obtained, indicating many interesting aspects about tourist preferences within the assets.

Agradecimentos

O relatório de projeto final da cadeira de Projeto Integrado decorre de uma experiência que exigiu trabalho e esforço para alcançarmos os objetivos pretendidos. Como tal, agradecemos a disponibilidade, acompanhamento atento e colaboração demonstrados pela professora Isabel Brito, agradecemos também o seu apoio, incentivo, confiança e essencialmente por nos ter guiado para uma melhor execução do trabalho.

Índice

Índice de Figuras

Índice de Tabelas

Índice de Listagens

Capítulo 1

Introdução

1.1 Objetivo do trabalho

Este trabalho tem como principal objetivo a monitorização de fluxos de visitantes para identificar boas práticas, tendo sempre foco na criação de uma benéfica, sustentável e duradoura relação entre o turismo e o património. Para essa monitorização ser realizada, será necessário desenvolver um sistema de informação que recolha, armazena, processa e comunica dados e informação sobre as visitas ao património dos turistas nacionais e estrangeiros. Para que a essa monitorização seja realizada devidamente, será necessário que sejam realizadas algumas etapas, das quais seriam:

1. A recolha de dados sobre as visitas dos turistas nacionais e estrangeiros ao património cultural de Beja, destacando-se dessa mesma recolha, as atrações, hóteis e restaurantes, e tendo como origens as fontes, "*TripAdvisor*", "*Booking*", "*Zoomato*". Para que a recolha de dados fosse realizada das devidas fontes foi necessário recorrer ao conceito de *webscraping*;
2. O armazenamento de dados numa base de dados SQL, para mais tarde facilitar o gerenciamento de toda a informação para as seguintes etapas;
3. O processamento dos dados, iniciando-se com a sua normalização e mais tarde passando á sua análise recorrendo ao *sentiment analysis* e à *Keyword extraction*;
4. A elaboração de gráficos usando a biblioteca *Matplotlib* e o *software PowerBI*;

O presente relatório encontra-se organizado na seguinte forma: na secção 2 descreve-se a fase de investigação; na secção 3 é descrito como foi realizado o processo de *webscraping* nos *websites* mencionados, assim como a estratégia pensada e dividida pelos elementos do grupo; na secção 4 é explicado o processo de normalização/formatação desenvolvido; na secção 5 mostramos os métodos escolhidos na extração das *keywords*; na secção 6 falamos de todo o processo por detrás dos *sentiment analysis*; na secção 7 de como foram gerados

1. INTRODUÇÃO

os gráficos ao decorrer da elaboração do projeto; na secção 8 de como e do porquê de termos reorganizado o projeto e também acerca da base de dados gerada; na secção 9 finalmente começamos a analisar os dados obtidos e por fim, na secção 10 são apresentadas as conclusões relativas à elaboração do presente trabalho.

1.2 Métodos utilizados

Para a extração dos dados ser realizada devidamente, tal como referido, foi utilizado o conceito de *webscraping*. O conceito de *webscraping* é simplesmente a recolha de dados de uma forma automatizada, no nosso caso foi utilizada a linguagem *Python* juntamente com algumas bibliotecas como a "*BeautifulSoup4*". É importante também referir que durante a elaboração do trabalho algumas ideias e conceitos foram surgindo, o que colmatou com algumas possíveis dúvidas e percurso que inicialmente o grupo pensaria que iria tomar tal como a utilização dos processos *ETL* (*extract, transform, load*). O grupo tinha como ideia levar o conceito á risca como ideia inicial, porem, com o decorrer do trabalho foi verificado que alguns conceitos não se encaixariam da melhor forma, mais em concreto, a fase de *load*. As fases de *extract* e *transform* encaixam-se na perfeição, já que o trabalho consiste na extração de informação de *websites* (*extract*) e a sua normalização/formatação (*transform*). No entanto o processo de *loading* só acabaria por ser usado como um processo secundário para outro, mais concretamente para a realização dos gráficos dentro do software *PowerBI* uma vez que esta ferramenta é altamente útil para a realização de gráficos usando como base grandes quantidades de informações.

1.3 Descrição e motivação

O tema deste trabalho é bastante interessante do ponto de vista turístico e mais tarde financeiro, já que a partir dele é possível analisar as opiniões dos turistas (nacionais ou estrangeiros) e baseando-se nisso, ter noção de quais pontos turísticos, hoteis ou restaurantes cativam mais a atenção do público e também estudar os pontos fortes e fracos de cada um deles. É possível também verificar se determinados locais têm tendências a manter, aumentar ou diminuir o número de turistas com o decorrer dos anos. Estes valores são bastante importantes para um país como Portugal que usa o turismo como forte fonte de rendimento, e uma vez que no presente trabalho o foco é o alentejo, que é altamente movimentado nas épocas balneares, mais importante a análise das informações recolhidas se tornam.

1.4 Divisão de tarefas

Uma vez que o trabalho era realizado em grupo, foi decidido previamente que existiriam etapas onde ocorreria separação de tarefas por cada elemento do grupo, como por exemplo

ao realizar o *webscraping* dos *websites* pretendidos. Para além da tarefas divididas entre os elementos do grupo, foi utilizado uma ferramenta para gerenciar as tarefas de cada elemento respetivamente, que foi o *Trello*. O *Trello* é tão simples como uma ferramenta de gerenciamento de projetos, é uma plataforma versátil e pode ser usada para acompanhamento de tarefas pessoais ou para organizar projetos que envolvam equipas/grupos com maior número de pessoas. Para além do *Trello* o grupo também utilizou o *GitHub* como ferramenta de gerenciamento do trabalho e repositório do mesmo, sendo a principal ferramenta no controlo de versões dos trabalhos realizados por cada elemento do grupo. Foram também utilizadas outras tecnologias no decorrer do trabalho como algumas bibliotecas específicas para algumas partes, como a *BeautifulSoup4* ou o *Yake!* ou até mesmo a ferramenta *PowerBI* que já foi mencionado, porém estas serão faladas mais adiante no decorrer do trabalho.

1.5 Ambientes virtuais *Python*

Neste projeto usamos ambientes virtuais *Python*. Um ambiente virtual é uma forma de ter várias instâncias paralelas do interpretador de *Python*, cada uma com diferentes conjuntos de pacotes e diferentes configurações. Cada ambiente virtual contém uma cópia do interpretador de *Python*, incluindo cópias dos seus utilitários de suporte como o *pip*. Estes contêm também uma zona para instalação de pacotes/bibliotecas localmente (dentro do ambiente virtual), sendo esta a razão principal pela qual foi decidido usá-los.

Capítulo 2

Investigação

2.1 Objetivo desta investigação

Com esta investigação foi determinado como objetivo definir o curso de ação que iríamos ter para o trabalho, assim como esclarecer alguns conceitos mais teóricos associados ao mesmo. Assim sendo, foram obtidos todos os dados de *posts* e comentários relacionados com *providers* de acesso, entretenimento, refeições e estadia diretamente ligados ao património cultural do alentejo. Estes a serem analisados e classificados, criando assim um modelo de possíveis sentimentos e procura que o comércio local tem o interesse em fornecer aos visitantes. O foco principal foram os *posts* e comentários em português.

2.2 Websites escolhidos

Os *websites* selecionados assim como os métodos de obtenção de dados, e também de analisar os mesmos foi decidida previamente pelo grupo e todos os elementos decidiram usar as mesmas ferramentas para cada função que lhes fora determinado. Após exaustiva procura pelos *websites* preferenciais e discussão entre todos os elementos do grupo, para a utilização de um conjunto deles onde fosse possível se realizar uma boa análise e obtenção dos dados, com especial foco na linguagem portuguesa, reunimos alguns possíveis candidatos que serão abordados no ponto seguinte.

Foi decidido também que dentro das opções que serão referidas em seguida, teremos preferência na utilização de um específico, uma vez que temos intenções de dar alguma prioridade às informações relacionadas com o turismo rural alentejano, sendo assim e devido a uma maior procura e número de resultados que o *website* oferecia, inicialmente iríamos utilizar o *website* *TripAdvisor* como primeira opção.

2.3 Sites pesquisados

Foram usados os sites:

2. INVESTIGAÇÃO

- *TripAdvisor*
- *Booking*
- *Zomato*
- *Google Maps*

O foco principal é o *TripAdvisor* visto que este oferece a maior variedade de conteúdo (hotéis, restaurantes e outros estabelecimentos), no entanto como uma plataforma é pouco, decidimos adicionar *Booking* e *Zomato* à lista para uma maior e mais ampla rede de hotéis e restaurantes.

Foi também considerado o *Google Maps*, mas este apresentou um novo set de problemas que vão ser descritos já de seguida.

2.4 (Im)possibilidade de uso de *APIs*

Em nenhum dos *websites* testados foi observada uma facilidade na obtenção de acesso às suas *APIs*, apenas alguns (3/4) ofereceram acesso à documentação da(s) mesma(s) facilmente. A maioria requer um contacto, que foi tentado e continuou sem resposta durante quase todo o processo de elaboração do trabalho(contactos iniciados por Amaro, entre dia 26 e 29 de Out, e dia 11 de Nov).

Dentro dos *websites* que oferecem documentação foi observado que todos subdividem os seus serviços de *API* em 3 ou 4 *APIs* para casos de uso específicos (reservas, dados, etc) em vez de uma com *endpoints* que oferecem solução para todos os casos.

A anomalia aqui é o *Google Maps* que é o único que facilita o acesso à *API* (mas paga, temos de ver os créditos disponíveis no *Cloud Platform*), e de elevada dificuldade em *scraping* pelo óbvio.

2.5 Alternativas

Sendo que é impossível o uso das *APIs* (que facilitariam o trabalho) temos de recorrer a outras técnicas para obter os dados.

2.5.1 *Web-crawling* VS *web-scraping*

Web crawling, também conhecido como Indexação, é usado para indexar as informações na página usando bots também conhecidos como *trackers*. O *tracker* é essencialmente o que os motores de busca fazem. É uma questão de visualizar uma página como um todo e indexá-la. Quando um *Bot* rastreia um *website*, ele passa por todas as páginas e todos os *links*, até a última linha do *website*, em busca de qualquer informação.

O *web scraping*, também conhecido como extração de dados da *web*, é semelhante ao *web crawling*, pois identifica e localiza os dados de destino das páginas da *web*. A principal diferença é que, com o *web scraping*, sabemos quem identificou o conjunto de dados exatamente, por exemplo, uma estrutura de elemento *HTML* para páginas da *web* que estão a ser corrigidas, da qual os dados precisam ser extraídos.

Com isto visto, *web scraping* é o nosso alvo, visto que minimiza lixo e é direcionado. No entanto devemos olhar para:

Vantagens de *web scraping*

1. Mais rápido: É possível manusear grandes quantidades de dados que poderiam levar dias ou semanas a serem processados através do trabalho manual, com o uso do *scraping* podemos reduzir substancialmente o esforço e aumentar a velocidade de decisão.
2. Confiável e consistente: Ao fazer o trabalho manual é muito fácil de haver erros, por exemplo, erros tipográficos, informações esquecidas ou inserção nas colunas erradas. O uso do *web scraping* garante consistência e a qualidade dos dados.
3. Ajuda a reduzir a carga de trabalho.
4. Menor custo: Uma vez implementado o *scraping*, o custo total da extração de dados é significativamente reduzido, especialmente quando comparado ao trabalho manual.
5. Manutenção básica: Fazer o *scraping* de dados geralmente não requer muita manutenção.

Desvantagens de *web scraping*

1. Baixa proteção: Se os dados na *web* são protegidos, o uso do *scraping* também pode se tornar um desafio e aumentar os custos.
2. Dados estruturados: Não vai ser possível fazer scraping a 1000 *websites* diferentes pois cada *website* tem uma estrutura completamente diferente. Será necessário haver alguma estrutura básica que seja diferente em determinadas situações.

2.5.2 Necessidade de *Web Scraping*

Com o acima dito, é necessário recorrer a soluções de *Web Scraping*.

A comunidade internauta reparou no mesmo, visto que em reação ao observado existem dezenas de projetos e tutoriais de *Web Scraping* das variadas plataformas de turismo e reservas. Infelizmente, as mesmas não ajudam no processo e cada uma tem uma forma de atuação bastante diferente.

2.6 Bibliotecas de *Python* para *Web Scraping*

Para fazer *Web Scraping* vamos usar *Python*, pela sua facilidade de uso e multifaceta “*Development speed is more important than execution speed*”. Com *Python* também temos as opções de criar cadernos *Jupyter* onde o próprio código e os resultados são “encadernados” com parágrafos de texto fazendo o próprio projeto o seu pequeno relatório de progresso e resultados; como também a criação de ambientes virtuais (*containers*), onde os pacotes usados ficam registrados e instalados localmente, garantindo assim a portabilidade.

Para tal linguagem existem 5 grandes bibliotecas para a resolução deste caso:

- *Requests*
- *BeautifulSoup*
- *Scrapy*
- *lxml*

Cada uma tem diferentes vantagens e desvantagens. Caso nenhuma destas tivesse resultado, teríamos usado Selenium, que é uma biblioteca mais completa e poderosa que as listadas, visto que é uma ferramenta de testes de automação. Porém tem um nível de complexidade maior e requer um *setup* inicial maior e mais trabalhoso, requer *WebDrivers* para a execução das tarefas, pode ser complicada com Firefox, sendo preferencial usar *Chromium-based Browsers* como o Chrome e o novo Edge (necessário ainda o *ChromeDriver* e o *EdgeDriver*).. Tentaremos evitar essa, a todo o custo, pela sua complexidade e extras desnecessários às nossas necessidades e custo temporal do setup inicial.

Para cada website pode ser necessário usar bibliotecas diferentes por necessidade ou por obtenção de informações/blog posts/etc... que facilitem ou melhorem o output desejado.

2.6.1 Tratamento de output

Os outputs do conteúdo *scraped* podem vir em *.xml* ou *.csv* (ou outras mas essencialmente essas duas). Tentámos ao máximo usar *.csv*, e transformar qualquer *.xml* em *.csv*, visto que um maior número de ferramentas gráficas (Excel, PowerBI, etc...) e/ou bibliotecas de *Python* (*Pandas*, *matplotlib*, *etc*) para a análise de dados tratam melhor ficheiros separados por vírgulas.

Foi também feita uma análise e extração de *keywords* nos textos das descrições e *reviews*. Para tal existem variados algoritmos que podemos usar, alguns "clássicos" outros até de *machine learning*.

Inicialmente todas as informações (dados e meta-dados) são dados como relevantes, após consideração e ponderação durante análises iniciais do decorrer do estudo poderemos descartar dados que não consideremos relevantes. No entanto nada nos impede de tentar

prever ou imaginar quais esses serão e posteriormente avaliar o nosso julgamento para ver o que foi aprendido.

2.7 Análise

2.7.1 Algoritmos de mineração de texto

Os algoritmos de análise de texto podem ser considerados ferramentas de mineração de texto, isto é, o processo de descoberta de conhecimento potencialmente útil e inicialmente desconhecido, ou seja, a extração de conhecimento útil utilizando bases textuais.

O processo de mineração de texto é dividido em quatro etapas bem definidas:

- Seleção
- Pré-processamento
- Mineração
- Assimilação

Na seleção, os documentos relevantes devem ser escolhidos e mais tarde processados. No pré-processamento ocorrerá a conversão dos documentos em uma estrutura compatível com minerador, bem como ocorrerá um tratamento especial do texto. Na mineração, o minerador irá detetar os padrões com base no algoritmo escolhido. E por fim, na assimilação, os utilizadores irão utilizar o conhecimento gerado para apoiar as suas decisões.

A etapa pré-processamento pode ser dividida em quatro tarefas:

- Remover *Stop Words*
- Compilação
- Normalização de sinônimos
- Indexação

Na etapa Remover *stopwords* os termos com pouca ou nenhuma relevância para o documento serão removidos. São palavras auxiliares ou conjetivas, ou seja, não são discriminantes para o conteúdo do documento.

Na etapa seguinte, compilação, realiza-se uma normalização morfológica, ou seja, as palavras são reduzidas ao seu radical, serão combinadas em uma única representação. A radicalização pode ser efetuada com o auxílio de algoritmos de radicalização, sendo os mais utilizados o algoritmo de *Porter (Porter Stemming Algorithm)* e algoritmo de *Orengo (Stemmer Portuguese ou RLSP)*.

2. INVESTIGAÇÃO

Após a compilação, na etapa de normalização de sinônimos, os termos que possuem significados similares serão agrupados em um único termo, por exemplo, as palavras ruído, tumulto e barulho serão substituídas ou representadas pelo termo barulho.

E, por fim, na etapa indexação atribui-se uma pontuação para cada termo, garantindo uma única instância do termo no documento. No processo de atribuição de pesos devem ser considerados dois pontos:

- Quanto mais vezes um termo aparece no documento, mais relevante ele é para o documento
- Quanto mais vezes um termo aparece na coleção de documentos, menos importante ele é para diferenciar os documentos

2.7.2 Algoritmos Clássicos

Os algoritmos clássicos são instruções passo a passo de modo a que, dada uma entrada específica, é possível rastrear e determinar exatamente a saída.

- Algoritmos clássicos especificam as regras exactas para encontrar a resposta geral
- Um algoritmo clássico usa código e dados para prever a resposta correta para uma pergunta
- Algoritmo clássico produz uma saída com base nas etapas descritas no algoritmo
- Em algoritmos clássicos, normalmente é necessário um grande número de exemplos para determinar a que distância a equação está da equação desejada. É por isso que “big data” é uma grande negócio hoje em dia
- Um algoritmo clássico não dá uma solução depois de chegar a uma solução optima para um problema

2.7.3 Algoritmos de *machine learning*

Os algoritmos de *machine learning* são partes de código que ajudam as pessoas a explorar, analisar e localizar o significado em conjuntos de dados complexos.

Os algoritmos de *machine learning* utilizam parâmetros baseados em dados de preparação, um subconjunto de dados que representa o conjunto maior. À medida que os dados de preparação se expandem para representar o mundo de forma mais realista, o algoritmo calcula resultados mais precisos.

Algoritmos diferentes analisam os dados de diversas formas. Geralmente, são agrupados consoante as técnicas de *machine learning* para as quais são utilizados:

- Aprendizagem supervisionada

- Aprendizagem não supervisionada
- Aprendizagem por reforço

Aprendizagem supervisionada

Na aprendizagem supervisionada, os algoritmos fazem previsões com base num conjunto de exemplos etiquetados fornecidos por si. Esta técnica é útil quando sabe como deverá ser o resultado. Por exemplo, fornece um conjunto de dados que inclui populações de cidades por ano nos últimos 100 anos e deseja saber qual será a população de uma cidade específica dentro de quatro anos. O resultado utiliza etiquetas que já existem no conjunto de dados: população, cidade e ano.

Aprendizagem não supervisionada

Na aprendizagem não supervisionada, os pontos de dados não são etiquetados. O algoritmo etiqueta-os ao organizar os dados ou ao descrever a sua estrutura. Esta técnica é útil quando não sabe como deverá ser o resultado. Por exemplo, fornece dados de cliente e deseja criar segmentos de clientes que gostam de produtos semelhantes. Os dados que está a fornecer não são etiquetados e as etiquetas no resultado são geradas com base nas semelhanças descobertas entre os pontos de dados.

Aprendizagem de reforço

A aprendizagem por reforço utiliza algoritmos que aprendem com resultados e decide a ação a realizar em seguida. Após cada ação, o algoritmo recebe comentários que o ajudam a determinar se a escolha feita foi correta, neutra ou incorreta. Por exemplo, se estivermos a criar um carro autónomo, queremos que este cumpra a lei e mantenha as pessoas seguras. À medida que o carro ganha experiência e um histórico de reforço, aprende a permanecer dentro da faixa, a não ultrapassar o limite de velocidade e a travar quando encontrar peões.

Existem muitos tipos diferentes de algoritmos de *machine learning*. Contudo, por norma, os casos de utilização destes algoritmos enquadram-se numa destas categorias.

- Algoritmos de classificação de duas classes (binários) dividem os dados em duas categorias. São úteis para perguntas com apenas duas respostas possíveis mutuamente exclusivas, incluindo perguntas de sim/não
- Algoritmos de classificação multiclasse (multinomial) dividem os dados em três ou mais categorias. São úteis para perguntas com três ou mais respostas possíveis mutuamente exclusivas
- Algoritmos de deteção de anomalias identificam os pontos de dados que estão fora dos parâmetros definidos para o que é considerado “normal”

2. INVESTIGAÇÃO

- Algoritmos de regressão preveem o valor de um novo ponto de dados com base em dados históricos
- Algoritmos de séries temporais mostram as alterações a um determinado valor ao longo do tempo. Com a análise e a previsão de série temporal, os dados são recolhidos a intervalos regulares ao longo do tempo e utilizados para fazer previsões e identificar tendências, sazonalidade, periodicidade e irregularidade
- Algoritmos de *clustering* dividem os dados por vários grupos ao determinar o nível de semelhança entre os pontos de dados
- Algoritmos de classificação utilizam cálculos de previsão para atribuir dados a categorias predefinidas

2.7.4 Decisão sobre o tipo de algoritmo

Após a grande análise dos tipos e subtipos de algoritmos de mineração de texto (análise de texto), é óbvia a escolha em algoritmos de *machine learning* com aprendizagem não supervisionada para a execução da nossa análise; a questão está em qual serão usados visto que muitos deles para os nossos casos poderão ter de ser sujeitos a pré-processamento; o que removeria as nossas requeridas dimensões de análise textual. No entanto vai continuar a haver pré-processamento como algoritmos de redução de dimensionalidade, a diferença comparado com a frase anterior é quão a pré-modelação não afetará negativamente os resultados e possíveis associações.

Algoritmos considerados

Dos variados algoritmos vistos e disponíveis na internet ou em bibliotecas de *Python* (como *SciKitLearn*, *Tensorflow*, *Keras*), tomamos a decisão de considerar os seguintes algoritmos como candidatos a uso e/ou pertencentes aos grupos de algoritmos usados para comparação de resultados:

- *LDA (Latent Dirichlet Allocation)*: um modelo de distribuição gaussiana, muito usado por empresas de software sobre *feedback* e *bug reports* para associação de resultados do *QA*,
- *K-Means Clustering*: muito usado para fazer *clusters* de *keywords* em redes sociais,
- *KNN (K-Nearest Neighbor)*: usado para agrupar dados relacionais com os contactos, relatórios, correspondência e *emails* em empresas,
- *SVM (Support Vector Machines)*: este é usado nos mesmos lugares que regressões lineares, porém mais rápido ou poderoso, é usado para agrupar pontos como texto com imagens ou tópicos de texto em sites de vendas de 2^a mão.

No entanto existe um algoritmo de *machine learning* semi-supervisto (aprendizagem não supervista mas ele faz a sua auto-supervisão; logo é questão de semântica). Este é o:

- LSTM (*Long-short term memory*): que é um tipo de RNN (rede neural recorrente) com ou sem um *layer* CNN (um *layer* de convulsão).

Poderá ser usado este método visto que não só “está na moda” como existe muito material de apoio por esse motivo, isto se houver tempo extra para experimentar e se essa experiência produzir melhores resultados comparativamente a um SVM por exemplo...

Pode-se notar que aqui foram escolhidos algoritmos de *machine learning* já comuns a grupos que realizaram tarefas semelhantes e que são algoritmos simples e rápido não sendo mais um algoritmo de uma família de algoritmos (mais complexos ou não, mas que têm muita variedade), tais como redes neurais (à exceção do RNN LSTM), algoritmos genéticos ou algoritmos lineares (como regressões lineares).

Porém no decorrer do trabalho o algoritmo LSTM não funcionou da melhor maneira, e optámos por outra técnica que será explicada e mostrada mais adiante no relatório.

Capítulo 3

Webscraping

3.1 Planeamento

3.1.1 Divisão de Tarefas

Para a realização desta fase de trabalho o grupo decidiu dividir as tarefas e organizá-las a partir da plataforma "Trello" (<https://trello.com/b/PIApdmTA/pi-2021-22>), uma plataforma de gestão de projeto estilo *board* do qual podemos aplicar Kanban, Scrum ou outra AGILE *management framework*.

A divisão de tarefas decidida foi a distribuição de cada um dos três "websites" por cada um dos elementos do grupo, pela ordem de dificuldade em congruência linear com o tempo extra-curricular disponível de cada elemento.

Assim sendo, o site "TripAdvisor" foi realizado pelo aluno Gonçalo Amaro, o "Booking" pelo Pedro Tomás e o "Zomato" pelo Vítor Abreu. No final verificou-se que todos conseguiram aceder aos seus devidos "websites" e adquirir as informações possíveis via *scraping*.

3.1.2 Tecnologias Usadas

Na realização do *web-scraping* foi desenvolvido um ambiente virtual de Python 3 para realizar os *scripts* que iriam recolher as informações.

Como forma de organizar todos os pacotes e possíveis actualizações de bibliotecas dentro do código também foi gerado um ficheiro .txt denominado "requirements" que actualizávamos e usávamos sempre que um dos elementos do grupo iria realizar o seu trabalho.

A linguagem optada para a construção dos *scripts* foi o Python já que é uma das mais acessíveis linguagens de programação disponíveis devido à sua simples *syntax* e também pela vasta quantidade de bibliotecas disponibilizadas, das quais temos uma dúzia que são bastante úteis para a realização deste projecto.

Para finalizar, todos os ficheiros foram guardados em formato .csv uma vez que, como explicado no relatório de progresso anterior, é um formato que é aceite e nos facilita a

3. WEBSRAPING

manipulação de dados (ETL), a alimentação dos dados ao algoritmo de *machine learning*, e pode ser usado com ferramentas de análise de dados e geradores de tabelas (como o PowerBI).

Ambientes Virtuais de Python

Neste projecto usamos Ambientes Virtuais de Python. Um ambiente virtual é uma forma de ter várias instâncias paralelas do interpretador de Python, cada uma com diferentes conjuntos de pacotes e diferentes configurações.

Cada ambiente virtual contém uma cópia discreta do interpretador de Python, incluindo cópias dos seus utilitários de suporte como o pip. Estes contêm também uma zona para instalação de pacotes/bibliotecas localmente (dentro do ambiente virtual), sendo esta a razão principal pela qual foi decidido usá-los.

Tendo introduzido a razão, consegue-se perceber o óbvio: sendo este um trabalho de grupo e que posteriormente poderá ser testado pelos docentes ou futuros alunos, ao usar ambientes virtuais podemos fazer “pip freeze” para um ficheiro de texto do qual facilita a portabilidade e transmissão de requerimentos do projecto.

Para a criação destes ambientes virtuais foi instalado o virtualenvwrapper o qual traz o virtualenv como dependência e um *set* de extensões para o mesmo, tal como sempre recomendado pelo Professor José Jasnau Caeiro, todos os anos na disciplina de Linguagens de Programação.

Bibliotecas de Python

Como dito previamente, na explicação pelo qual o uso de Python, foi referida a grande quantidade de bibliotecas que nos são facilmente fornecidas pelo pip.

Dentro deste repositório existe (perto de) uma dúzia de bibliotecas que nos permitem facilmente completar as nossas tarefas deste projecto. Dessa dúzia, para esta etapa, foram usadas:

- BeautifulSoup4, uma biblioteca que facilita a extração de informações de páginas da web, fornecendo expressões para iterar, pesquisar e modificar a árvore de análise;
- lxml, uma biblioteca Python que permite fácil manuseio de arquivos XML e HTML;
- requests, uma biblioteca HTTP elegante e simples para Python, construída de raiz para ser fácil de usar;
- pandas, uma ferramenta de manipulação e análise de dados de código aberto rápida, poderosa, flexível e fácil de usar;
- jupyter, um meta-pacote o qual traz (como dependências) o sistema Jupyter (em especial os cadernos), o *kernel* IPython e outros.

Com estes pacotes temos um mapa de actuação para esta etapa de projecto (de webscraping): abrir um ambiente virtual (e instalar bibliotecas), abrir um caderno de Jupyter, importar as bibliotecas das quais usamos “requests” para ir buscar a nossa página, fazer *parsing* da pagina via “lxml”, criar um objecto “Soup” com o conteúdo *parsed*, fazer *scraping* e iterar pelos *scrapes* dos quis criamos *dataframes* de “pandas” e exportarmos os mesmos em *.csv* para uso futuro.

3.2 Booking

3.2.1 Estratégia

Para a realização do “web-scraping” no “website” da Booking.com, inicialmente foi necessário a filtragem pelos hotéis apenas na localidade de Beja, uma vez ser o local que o grupo em conjunto decidiu optar para realizar todas as pesquisas num sítio em comum. Após ter o Booking a apresentar todos os resultados para os hotéis de Beja, foi recolhido o link que redireciona especificamente para esses resultados. Para aceder ás informações específicas de cada elemento da página e mais tarde aceder aos mesmos para retirar a informação pretendida, foi usado a ferramenta de “inspeccionar a página” e assim descobrir os nomes das classes e todos os outros elementos que continham conteúdo importante para o projecto, como o nome dos hotéis, preço, classificação, número de comentários e alguns outros detalhes que pudessem ser úteis.

Em seguida foi necessário realizar o “web-scraping” das *reviews* de cada hotel, a realização desta parte foi um pouco mais difícil uma vez que para as *reviews* serem bem recolhidas era fulcral que o “web-scraping” fosse realizado usando outro link, ou seja, foi retirado do site o prefixo de um novo link que seria o “<https://www.booking.com/reviews/pt/hotel/>” e baseando nos hotéis já retirados foi colocado o nome de cada um á frente do mesmo, criando assim um novo link que seria usado na realização do “web-scraping” após a criação de um novo link para cada hotel, os processos foram semelhantes aos anteriormente feitos.

Para finalizar, os resultados foram todos guardados em ficheiros *.csv* para uma mais fácil visualização.

3.2.2 Desenvolvimento

Aqui detalha-se o processo de desenvolvimento do ”webscraping” do ”website” Booking.

Hotéis

Inicialmente foi feita a filtragem de apenas os hotéis de Beja.

3. WEBSRAPING

No código foi implementado as bibliotecas BeautifulSoup para facilitar a tarefa de realizar o “web-scraping”.

A partir do “website” ao inspeccionar a página era possível retirar os *headers* que eram valores necessários na realização do “web-scraping”. Também é realizado o pedido HTTP e juntou-se a informação com a biblioteca “BeautifulSoup”.

Foram criados diferentes *arrays* para receber as informações e posteriormente colocada a respectiva informação em cada um deles.

Devido a alguns “arrays” conterem mais informação, possivelmente devido a algum tipo de informação adicional que possa estar em algum hotel especificamente, para prevenir erros, foram reduzidos ao tamanho do *array* mais curto.

Por fim todos os resultados contidos nos “arrays” foram guardados num ficheiro *.csv* denominado “listtable.csv”.

Construção dos links para realizar o “web-scraping” das “reviews” de cada hotel.

Foi realizado o pedido “HTTP” e juntado á biblioteca “BeautifulSoup” para aceder ás “reviews” de cada site e todos os valores foram salvos no formato *.csv*.

No final, temos esta tabela representativa do hotéis *scraped* ordenada e representativa dos *scrapes* “hotelXX.csv”.

3.2.3 Resultado

Aqui exemplifica-se um ficheiro *.csv*, “hotel18.csv”, que contem os *reviews* do hotel “Quinta do Castelo”.

3.3 TripAdvisor

O TripAdvisor é uma empresa americana de viagens *online* que opera *web* e *mobile apps* com conteúdo *user generated* e um “website” de comparação de preços, dos quais se pode fazer com hotéis, locais atractivos (como monumentos, parques, museus, etc..) e restauração.

Este como sendo um produto/serviço que oferece acesso a três categorias distintas (hotéis, restaurantes e atracções), foi dividido em três partes que representam as três categorias.

Este “website” é conhecido pelas suas tentativas de dificultar os processos de *scraping*, o qual foi observado, mas resolvido a custo de tempo. Felizmente encontramos um “website” chamado “*Worth Web Scraping*” o qual nos mostrou como fazer na página de hotéis do TripAdvisor o *scraping* da tabela de referência e dos *reviews*.

3.3.1 Estratégia

Após um *scouting* inicial às páginas das três categorias, foi observado as seguintes peculiaridades:

- as páginas das três categorias são diferentes no seu *layout* e organização;
- os nomes das classes nos “span”, “div” e outros elementos são *random generated* e mudam de acordo com a sessão aberta ou cookie;
- existem representações repetidas, estes são os *posts sponsored* pelo próprio “website”;
- as “subpáginas” que nos retornam os *reviews*, são de comprimentos diferentes de acordo com a categoria de *listing*;
- as “subpáginas” que nos retornam os *reviews*, usam múltiplos de cinco ou dez na *query*;
- as “subpáginas” que nos retornam os *reviews* mostram por defeito os que estão na linguagem referente ao domínio (.pt, .com, etc..) sem *query parameter* para alterar,
- *links* com *query parameters* que representem uma “subpágina” não existente não dão erro 404 (Page not found), mas redirecionam para a primeira;
- quando tentamos extrair o total de *reviews* apenas conseguimos o total dos totais e não o total por linguagem, impedindo assim de fazer uma conta para saber qual o múltiplo de cinco ou dez que seria a última “subpágina”.

Assim sendo, a estratégia que foi usada, embora extremamente má em termos de tempo despendido e extracções redundantes, era a única que assegurava que se conseguia extrair todos os *reviews*. Essa estratégia foi:

- criar uma lista de *links* para 200 ou 400 “subpáginas” (de acordo com o *listing* daquela categoria com mais *reviews* em português);
- extrair incluindo os repetidos para um *array/list/arraylist*;
- usar compreensão de listas através de *sets/dicionários/tuples* que possam ser ordenados para remover repetidos e não perder ordem;
- transpor esses dados para um *dataframe* de “pandas” e exportá-lo para *.csv* para uso futuro.

3.3.2 Desenvolvimento

Aqui iremos detalhar o processo longo do *webscraping* da plataforma TripAdvisor e as suas três principais categorias.

3. WEBSRAPING

Atracções

Para o desenvolvimento do *webscraping* das Atracções de Beja, foi aberto um caderno de Jupyter no qual começamos por fazer o *import* das bibliotecas e desactivar o aviso da falta de certificado SSL (após a introdução do trabalho em Inglês).

Seguidamente, foi feita a configuração do *request* onde qual fazemos download da página *web* pretendida. Estes *headers* foram extraídos do *browser* do computador usado, Microsoft Edge (Chromium).

Após fazer *request* e verificar o status code (vazio ou 200 para OK), foi criado um objecto “Soup” com o parsing (via “lxml”) da página *requested*.

Para a criação da tabela de referência das atracções fazemos um ciclo que nos vão fazer *scrape* aos nomes.

Sendo que agora podemos simplesmente através destes *arrays* criados fazer um *dataframe* der “pandas” via um dicionário de Python com os variados *pandas* referidos. Seguidamente exportamos o *dataframe* para um ficheiro *.csv*.

Agora um ciclo que retira os “HTML tag” onde contem um “href” com uma parte do *link* que nos possibilita (criar o *link* inteiro e) visitar a pagina de *reviews*.

Essas páginas têm determinadas restrições faladas nas secções anteriores e a sua solução. A qual aqui em baixo representada, cria uma enormidade de *links* por local. Dos quais *links* agora serão *scraped* (incluindo os *reviews* repetidos e excepto os que contem “desde” e “euros”) e seguidamente tratados (remoção de repetidos) indo seguidamente para um (dicionário e transformado num) *dataframe* de “pandas”, o qual é imediatamente exportado com o número do atracção referente na tabela de referência.

Hotéis

Para o desenvolvimento do *webscraping* dos Hotéis de Beja, foi aberto um caderno de Jupyter no qual começamos por fazer o *import* das bibliotecas e desactivar o aviso da falta de certificado SSL (após a introdução do trabalho em Inglês).

Seguidamente, foi feita a configuração do *request* onde qual fazemos download da página *web* pretendida. Estes *headers* foram extraídos do *browser* do computador usado, Microsoft Edge (Chromium).

Após fazer *request* e verificar o status code (vazio ou 200 para OK), foi criado um objecto “Soup” com o parsing (via “lxml”) da página *requested*.

Para a criação da tabela de referência dos hotéis fazemos um grupo de ciclos que nos vão fazer *scrape* aos nomes, *ratings*, número total de *reviews* e preços. Sendo que este número de *reviews* não nos vale de muito tal como previamente referido.

Sendo que agora podemos simplesmente através destes *arrays* criados fazer um *dataframe* der “pandas” via um dicionário de Python com os variados *pandas* referidos. Seguidamente exportamos o *dataframe* para um ficheiro *.csv*.

O qual gerou uma tabela de hotéis como referência.

Mesmo que o numero total de *reviews* não nos seja relevante o “HTML tag” onde é retirado contem um “*href*” com uma parte do *link* que nos possibilita (criar o *link* inteiro e) visitar a pagina de *reviews*.

Essas páginas têm determinadas restrições faladas nas secções anteriores e a sua solução. O que cria uma enormidade de *links* por local.

Dos quais *links* agora serão *scraped* (incluindo os *reviews* repetidos) e seguidamente tratados (remoção de repetidos) indo seguidamente para um (dicionário e transformado num) *dataframe* de “pandas”, o qual é imediatamente exportado com o número do hotel referente na tabela de referência.

Restaurantes

Para o desenvolvimento do *webscraping* dos Restaurantes de Beja, foi aberto um caderno de Jupyter no qual começamos por fazer o *import* das bibliotecas e desactivar o aviso da falta de certificado SSL (após a introdução do trabalho em Inglês).

Seguidamente, foi feita a configuração do *request* onde qual fazemos download da página *web* pretendida. Estes *headers* foram extraídos do *browser* do computador usado, Microsoft Edge (Chromium).

Após fazer *request* e verificar o status code (vazio ou 200 para OK), foi criado um objecto “*Soup*” com o parsing (via “*lxml*”) da página *requested*.

Para a criação da tabela de referência das atracções fazemos um ciclo que nos vão fazer *scrape* aos nomes e as partes de *href* contidas nos “*href*” para a criação dos links dos *reviews*.

Sendo que agora podemos simplesmente através destes *arrays* criados fazer um *dataframe* der “pandas” via um dicionário de Python com os variados *pandas* referidos. Seguidamente exportamos o *dataframe* para um ficheiro *.csv*. O qual gerou uma tabela de restaurantes como referência.

Agora um ciclo que vai buscar os as partes de *links* onde do ciclo anterior que nos possibilita (criar o *link* inteiro e) visitar a pagina de *reviews*.

Essas páginas têm determinadas restrições faladas nas secções anteriores e a sua solução. A qual aqui em baixo representada, cria uma enormidade de *links* por local.

Dos quais *links* agora serão *scraped* (incluindo os *reviews* repetidos) e seguidamente tratados (remoção de repetidos) indo seguidamente para um (dicionário e transformado num) *dataframe* de “pandas”, o qual é imediatamente exportado com o número do restaurante referente na tabela de referência.

3.3.3 Resultado

Aqui apresenta-se os resultados do *webscrape* do TripAdvisor. Os quais representam um exemplar dos dez primeiros *reviews* do primeiro hotel, atracção e restaurante, respectivamente.

3.4 Zomato

A Zomato é um serviço de busca de restaurantes para quem quer sair para jantar, buscar comida ou pedir em casa. A Zomato possui duas secções: guia de restaurante e blog. Previamente, havia uma secção de eventos, já descontinuada.

O guia de restaurantes Zomato permite ao usuário buscar informações relacionadas a restaurantes, bares, cafés, pubs e casa nocturnas. As informações fornecidas geralmente incluem o nome do estabelecimento, telefones de contacto, endereço, cardápio, fotografias, avaliações e mapas de localização.

3.4.1 Estratégia

As páginas da *web app* do Zomato usam um “parallax” de *scrolling* infinito (até não haver mais restaurantes) e as classes dos “HTML tags” mudam por sessão e/ou *rendering*, logo aqui a estratégia é literalmente fazer “download” da página web e fazer o *scrape* a partir do *parsing* dessa página.

3.4.2 Desenvolvimento

Aqui é representada uma aproximação do desenvolvimento deste *scraping*.

Restaurantes

Primeiramente foi feito o *import* das bibliotecas.

Depois pegando no código dos colegas como *template*, adaptou-se para usar uma página previamente descarregada.

Fazemos um ciclo de *scraping* dos nomes dos locais de consumo.

E agora dois ciclos, um para as classes com nome gerado no *prerender* e outra pós *render*; vamos buscar os tipos/classes de restaurantes, e outros dois ciclos do mesmo motivo, para ir buscar os preços. Pelo mesmo motivo criamos dois ciclos; que vão buscar os links das páginas dos *reviews*, o qual extraímos todos os “tags” de parágrafos porque que sempre que se corria o código gerava uma classe nova. Logo, estas extracções desapontantes, vão sofrer ETL.

3.4.3 Resultado

Os resultados deste *scrape* foram desapontantes no minimo devido à infeliz *random generated* nome da classe, que é gerado por cada vez que se usa a página. Estes resultados vão sofrer muito ETL posterior.

Capítulo 4

Pré-processamento

4.1 Pré-processamento de dados

Esta etapa consiste em remover todos os caracteres especiais, que não são letras, números ou espaços. As quais podemos considerar diacrítico como caracteres especiais ou não. Dependendo do caso, podemos remover todos os caracteres especiais, ou apenas os que não são diacríticos.

Nos nossos casos de analise textual, houve necessidade de remover todos os caracteres especiais, mas os diacríticos podem ser úteis para a nossa análise, em especial nos passos seguintes, extracção de palavras-chave e analise de sentimentos, foram descartados ou usados respectivamente, pelo motivo da precisão da análise em questão.

4.1.1 Metodologia

Para remover os caracteres especiais, foram criados dois *scripts*, um para limpar os ficheiros *.csv* e outro tratar do texto em si, que consiste em remover os caracteres especiais, fazer a normalização dos caracteres via NFKD e aplicar a remoção de acentos, caso seja necessário (apenas alterando o ultimo passo de *encoding/decoding*).

Estes não requerem bibliotecas externas, podem ser executados em *Python 3*, e as suas bibliotecas standard.

Limpar ficheiros

O *script* “*trimer.py*” é responsável fazer *trimming* (remover espaços em branco), remover linhas em branco e colocar aspas duplas em cada *review*.

Limpar texto dos *reviews*

O *script* “*normalize*” é responsável por remover caracteres especiais, fazer a normalização dos caracteres via NFKD e aplicar a remoção de acentos, caso seja necessário (apenas alterando o ultimo passo de *encoding/decoding*).

4.1.2 Execução

Executamos o *script* “*trimer.py*” para limpar os ficheiros *.csv* e o *script* ‘normalize’ para limpar o texto dos *reviews*. Estes aceitam um caminho para uma pasta com os ficheiros *.csv* e itera sobre os ficheiros, executando as funções de limpeza.

Após a execução da limpeza textual e normalização sem remoção de acentos, os tenhamos ambas as versões. Para a versão com remoção de acentos, foi necessário usar o pacote “*Unidecode*” para aplicar a remoção de acentos.

Com estas duas versões podemos obter os resultados óptimos para a nossa análise textual.

4.1.3 Resultados

Os ficheiros *.csv* foram limpos e normalizados sem remoção de acentos e com remoção de acentos.

Capítulo 5

Extracção de Palavras-Chave

5.1 Extracção de Palavras-Chave

A extracção de palavras-chave é uma técnica de extracção de informações que consiste em extrair palavras-chave de um texto.

Geralmente, as palavras-chave são utilizadas para identificar o conteúdo de um documento, ou seja, para identificar o que o documento contém. No nosso caso, as palavras-chave são utilizadas para identificar pontos fulcrais da recepção dum cliente turístico num estabelecimento turístico de Beja (hotel, atracção, restaurante, etc).

Com estas palavras-chave, é possível identificar o que o cliente quer, ou seja, o que ele quer ver, o que ele quer comer, o que ele quer fazer, ou o nível de satisfação com o serviço prestado.

5.1.1 Metodologia

Para a extracção de palavras-chave, utilizamos uma biblioteca de *Python* chamada *YAKE!*. Esta é um pipeline de processamento de linguagem natural, que utiliza um algoritmo personalizado descendente do *Naïve Bayes* para extraer palavras-chave de um texto.

Para a extracção de palavras chaves de cada *review*, utilizamos o *YAKE!* iterativamente por cada *review*, alimentando-o com o seu conteúdo textual, após o pre-processamento textual de todas as *reviews* (a qual especificamente foi usada a versão sem diacríticos).

O output da alimentação do *YAKE* é um par de palavras-chave e seus respectivos pesos, que são armazenados num dicionário. O dicionário é ordenado pelo valor de seus pesos, e o número de palavras-chave extraídas é limitado ao numero de *reviews* que foram utilizadas para a extracção por cada estabelecimento.

YAKE!

Esta biblioteca é um software livre, e pode ser obtida via *pip* ou via GitHub, que é uma ferramenta de extracção de palavras-chave desenvolvida por autores portugueses (e

5. EXTRACÇÃO DE PALAVRAS-CHAVE

um japonês) da Universidade do Porto, Politécnico de Tomar, e da Universidade da Beira Interior (e da Universidade de Kyoto).

O sua utilização pode ser simples, basta instalar a biblioteca e executar o seguinte comando: `yake.KeywordExtractor(lan="pt").extract_keywords(text)`

No entanto esta pode ser mais complexa caso seja necessário optimizar a extracção de palavras-chave, com os seus variados parâmetros opcionais.

Este funciona da seguinte forma: quando recebe um texto, vai testar todas as palavras do texto, com uma determinada formula, e guardar o peso da palavra, e a palavra em si, no final expõe um dicionário com as palavras-chave e seus respectivos pesos.

$$A\ formula\ falada\ anteriormente\ é:\ S(kw) = \frac{\prod_{w \in kw} S(kw)}{TF(kw) * \sum_{w \in kw} S(w)}$$

Mais especificamente este modulo é uma forma mais delicada e avançada de um classificador *Naïve Bayes* o qual será mais e melhor explicado no capítulo seguinte onde procedemos ao desenvolvimento de um para efeitos de análise de sentimentos.

Execução

Foi criado um *notebook* de *Jupyter* o qual contém o código usado para a extracção de palavras via *YAKE!*. Cada bloco de código está sobreposto por um bloco de *markdown*, que é um comentário. Os comentários são usados para explicar o que cada bloco de código faz.

As rotinas de extracção de palavras-chave são: iterativamente, para cada ficheiro *.csv* dentro da pasta indicada, importar via *DataFrame* de *pandas*. O qual *DataFrame* é um conjunto de dados, como uma tabela de dados, e contém uma coluna com os *reviews*. O *DataFrame* é iterado na sua coluna única, e o seu conteúdo é passado para o *YAKE!*. O qual exporta o resultado para um ficheiro *.csv*, que é um ficheiro *.csv* com uma coluna com as palavras-chave e outra com o seu peso.

O *notebook* e o código estão disponíveis nos anexos.

5.1.2 Resultados

Os resultados obtidos detêm um sentimento misto no grupo. Estas palavras chave por vezes não são palavras únicas, mas são expressões que são frequentes. Muitas palavras únicas aparecem lado a lado das expressões, o que dá uma noção de repetição ou confirmação de resultado.

Capítulo 6

Análise de sentimentos

6.1 Analise de Sentimentos

A analise de sentimentos é uma técnica de extracção de informações que consiste em extrair sentimentos de um texto.

Geralmente, os sentimentos são utilizados para identificar o sentimento de um texto, ou seja, para identificar o que o texto contém. No nosso caso, os sentimentos são utilizados para identificar a satisfação do cliente com o serviço prestado em variados serviços turísticos de Beja (hotéis, atracções, restaurantes, etc).

Com base no texto, o sentimento é extraído através de um algoritmo que identifica a intensidade do sentimento. No nosso caso, a classificação é binária, ou seja, o sentimento é positivo ou negativo. Consideramos uma não-reclamação como positiva.

A percentagem de sentimentos positivos e negativos é necessária para identificar a satisfação do cliente com o serviço prestado. A satisfação do cliente é uma medida de qualidade de serviço.

6.1.1 Metodologia

Para fazer uma analise de sentimento textual, utilizamos *machine learning*, a qual criamos um modelo de classificação de texto binário ou ternário, no nosso caso, o modelo é um classificador de sentimentos binário.

Este modelo é treinado através da alimentação de um conjunto de dados de treinamento e um conjunto de dados de teste, ao modelo matemático criado ou importado. Estes modelos matemáticos podem ser probabilísticos ou não. Os dados de treinamento são utilizados para treinar o modelo matemático. Os dados de teste são utilizados para testar o modelo matemático.

Após o treinamento do modelo matemático, e os resultados do teste (matriz de confusão, precisão, exactidão, etc), o modelo é utilizado para classificar um texto. O resultado da classificação é o sentimento do texto, ou seja, positivo ou negativo.

6. ANÁLISE DE SENTIMENTOS

Para pre-processar o texto a ser classificado (e os textos de treinamento e teste), utilizamos o pacote de linguagem natural do *Python*, ou seja, a biblioteca NLTK, para aplicar algumas transformações ao texto, como remover pontuação, remover *stopwords*, etc.

Este passo é o mais importante para o modelo matemático ser bem treinado e o mais importante para o modelo ser bem classificado.

Pre-processamento

Para fazer o pre-processamento do texto, utilizamos as ferramentas do NLTK, das quais em especial os que fazem (ou determinam) as *stopwords*, fazem *stemming*, *tokenização*, etc.

Mais especificamente *stopwords* portuguesas do *corpus* de *stopwords* do NLTK, o *SnowballStemmer*, é utilizado para remover *stopwords* e os sufixos das palavras deixando apenas a raiz da palavra (*the stem*), e o *Vectorizer* que é utilizado para transformar o texto em um vector (ou matriz) de características (termos numéricos).

Este ultimo é preferencial que se use o *CountVectorizer*, pois ele conta o número de vezes que um termo aparece no texto com inteiros, e não com *floats*, como o *TF-IDF* faz. Sendo o ultimo mais preferencial para outras tarefas (não classificação).

O *Stemming* é para nós a fase mais importante do processo de pre-processamento, pois é o que removemos os sufixos das palavras. Para muitos propósitos, e em especial este, a conjugação dos verbos atrapalha a classificação e a aprendizagem do modelo (ou até a nossa). Um exemplo da sua *desnecessidade* é a enorme diferença entre o Inglês e o Português.

Os verbos em inglês são conjugados com substantivos, e os substantivos são conjugados com verbos. Porém, os verbos em português são conjugados com adjetivos, e os adjetivos são conjugados com verbos. Mais a enorme variação entre pessoas e tempos, em inglês variamos de três formas para seis (nove se separarmos *he/she/it*) pessoas para três tempos (um presente e dois passados, futuros e condicionais são modificações de um presente), já em português variamos de seis formas para seis pessoas (oito se separarmos *ele/ela* e *eles/elas*) para seis tempos (um presente, três passados, um futuro e um condicional).

Estas conjugações tem a mesma acção e uma enorme semelhança frásica: a raiz do verbo (*the stem*). As conjugações são desnecessárias e demasiado complexas.

Sendo assim, ao aplicar *stemming*, reduzimos a dimensão da matriz/vector, e simplificamos a linguagem, de maneira inteligente.

6.1.2 Tentativas

Foram feitas três tentativas de classificação de sentimentos. A primeira foi um fracasso completo, foi tentada a criação de um modelo sequencial com *layers LSTM* com *embedding* e de convulsão. À falta de conhecimento prévio, e à falta de informações simples e palpáveis com acesso fácil na *internet*, não foi possível fazer a classificação de sentimentos com este

modelo. Este modelo foi tentado com o uso da biblioteca de *TensorFlow*, que é uma biblioteca de código aberto.

A segunda tentativa foi muito mais bem sucedida que a primeira. Foi usado para o algoritmo de classificação de sentimentos o *Naïve Bayes (Multinomial)*, e como dados de treino e teste, foram utilizados os dados de treino e teste disponíveis no *Kaggle*, estes usavam *reviews* de filmes e produtos de *e-commerce* em português do *Brasil*, já que em português de Portugal não foi possível encontrar.

Modelo Sequencial LSTM com *Embedding*

Este modelo não chegou a completar qualquer fase de treinamento, pois não foi possível criar um modelo que funcionasse com o *dataset* de treino, a quantidade absurda de variáveis e modelações junto com a falta de conhecimento prévio impossibilitaram que um modelo funcional fosse criado, inclusive com *trimming* ou *truncation* dos *inputs*.

Execução

Após a importação do pacote de *TensorFlow*, foi criado um modelo sequencial com o uso de *embeddings*, ou seja, um modelo que utiliza *embeddings* para representar os inputs. Os *layers* LSTM e Dropout foram utilizados para aumentar a capacidade de aprendizagem do modelo. E os *layers* de saída são os softmax e categorical_crossentropy.

Cada um destes *layers* detinha parâmetros que teriam de ser ajustados para que o modelo fosse capaz de classificar os sentimentos. O qual requeria algum apoio não disponível, ou seja, o modelo não foi capaz de aprender a classificar sentimentos.

Resultados

Não foi possível executar a tentativa de classificação de sentimentos com este modelo.

Modelo *Naïve Bayes* do *Scikit-Learn* com *Datasets PT-BR* do *Kaggle*

Nesta tentativa de classificação de sentimentos, foi utilizado o pacote *Scikit-Learn*, e foi utilizado um par de *datasets* de *reviews* do *Kaggle*, que é um *dataset* de *reviews* de filmes e produtos de *e-commerce* em português do *Brasil*, já que em português de Portugal não foi possível encontrar.

Foi criado um objecto (que deriva da nossa classe *StemmerTokenizer*), que foi utilizado para aplicar o *Stemming* e *Tokenização* aos dados de treino e teste. Reduziu-se a quantidade de dados, e aumentou a capacidade de aprendizagem do modelo. A necessidade de aplicar *stemming* à *Tokenização* veio da necessidade de reduzir a dimensão da matriz/vector, e simplificar a linguagem, de maneira inteligente.

Sendo assim, ao aplicar *stemming*, reduzimos a dimensão da matriz/vector, e simplificamos a linguagem, de maneira inteligente.

6. ANÁLISE DE SENTIMENTOS

Naïve Bayes Um classificador *Naïve Bayes* é um classificador simples e probabilístico baseado em aplicar a teoria de *Bayes* com assumpções fortes (naïve) de independência. *Naïve Bayes classifiers* são simples e fáceis de entender, requerendo nenhuma etapa de *prunning* para evitar o *overfitting*. Contudo, eles não são mais poderosos do que outras técnicas avançadas, como árvores de decisão ou vector de suporte, muito menos que máquinas neurónios.

Como funciona um classificador *Naïve Bayes*? Estes classificadores são baseados em aplicar a teoria de *Bayes* com assumpções fortes (*naive*) de independência. A hipótese *naïve* diz que os *features* são independentes uns dos outros. Isso significa que os *features* são condicionais independentes a partir da classe. Indústria matemática pode ser usada para mostrar que os *features* são independentes a partir da classe, portanto a classe é a mais provável de saída.

A ideia principal do *Naïve Bayes* é calcular a probabilidade de cada classe, dado os *features*. A probabilidade de uma classe é calculada por multiplicar as probabilidades de cada *feature* dado a classe.

Este é um exemplo simples de um classificador *Naïve Bayes*: recebe um vector de *features* e um rótulo de classe e retorna a probabilidade da classe dado os *features*. No problema de classificação de texto, o rótulo de classe é a classe real do texto. Se inserirmos um vector de *features* do texto e o rótulo de classe, o classificador *Naïve Bayes* retorna a probabilidade da classe dado os *features*.

Outros exemplos de classificadores *Naïve Bayes* são: análise de sentimento, detecção de *spam*, e classificação de texto. Como para nosso caso (o problema de análise de sentimento), temos um vector de *features* do texto e o rótulo de classe é o sentimento do texto. O conjunto de classe é um conjunto binário de positivo e negativo. Recebe um vector de *features* do texto e o rótulo de classe e retorna a probabilidade de zero a um de classe, sendo o mais próximo a um o mais positivo o texto é.

Execução

Foi criado um *notebook* para a execução da tarefa. Este *notebook*, contém o código do modelo *Naïve Bayes* do *Scikit-Learn*, e as referencias para *datasets* de treino e teste. Com blocos alternados de código e *markdown* que comentam a execução do modelo, foi possível executar a tentativa de classificação de sentimentos com este modelo.

As rotinas executadas foram: a criação de um objecto *StemmerTokenizer*, a criação de um modelo *Naïve Bayes* do *Scikit-Learn*, o treinamento do modelo, em que de fazia *drop* a inúmeras colunas, os testes e métricas de avaliação do modelo, a importação iterativa dos *.csv* que contem os *reviews* dos estabelecimentos em que iterativamente foi aplicada uma limpeza e normalização, e a execução do modelo nos *reviews*, que quando classificados, geraram um arquivo *.csv* com os resultados.

Resultados

Este modelo foi capaz de classificar sentimentos com sucesso. A classificação foi bem sucedida, e foi possível classificar os sentimentos de um conjunto de *reviews*, e gerar um arquivo *.csv* com os resultados. Porém a classificação foi pouco precisa, e a precisão real foi baixa, apesar de ser um modelo bem sucedido e de ter sido bem treinado e avaliado. A sua *accuracy* foi de entre 91% e 87%, dependendo dos *runs*.

Calculamos que o problema está na diferença entre o português de Portugal e o português de *Brasil*, o qual o último é a língua dos *datasets* de treino e teste, o qual o modelo foi bastante preciso; e o primeiro a língua materna dos estabelecimentos e dos clientes, o qual o modelo foi pouco preciso.

Pipelines de Transformers do HuggingFace e Modelos BERT da Google fine-tuned

Como os resultados finais do modelo *Naïve Bayes* não foram satisfatórios, foi criado um *pipeline* de *transformers* do *HuggingFace*, com um modelo BERT da Google *fine-tuned* para classificar sentimentos de forma binária.

Esta biblioteca pode usar *PyTorch* ou *TensorFlow* como backend de computação, cada um destes tem suas vantagens e desvantagens. Por defeito ele usa *PyTorch* na maioria dos seus *pipelines*, mas pode ser configurado para usar *TensorFlow*.

Estes *pipelines* do *HuggingFace* são muito mais complexos, são muito mais eficientes, e também são muito flexíveis e fáceis de usar. São definidos com um texto que determina qual é o pipeline, e os parâmetros pedidos. No nosso caso foi utilizado o pipeline de classificação de sentimentos em que tinha dois parâmetros, o *tokenizer* e o modelo, mais especificamente um *AutoModelForSequenceClassification* e um *AutoTokenizer*, os quais foram buscar o modelo *pretrained* e o *tokenizer* do [gchhablani/bert-base-cased-finetuned-sst2](https://huggingface.co/gchhablani/bert-base-cased-finetuned-sst2)

O que é um Transformer

Um *Transformer* é um tipo de rede neural que é capaz de aprender funções complexas de dados. Ele funciona através de transformar os dados de entrada em uma nova representação, que pode então ser usada para fazer previsões em novos dados.

Transformers têm muitas aplicações em linguagens de processamento natural, processamento de imagens e visão computacional. Eles foram bem-sucedidos em diversos domínios, incluindo:

- Sumários de texto (*Bert*, *DistilBert*, *RoBERTa*, *XLNet*)
- Captação de imagens (*XLNet*)
- Tradução de imagens (*Bert*, *DistilBert*, *RoBERTa*)
- Respostas a perguntas (*Bert*, *DistilBert*, *RoBERTa*)

6. ANÁLISE DE SENTIMENTOS

- *Chatbots (Bert, DistilBert, RoBERTa)*
- *Classificação (Bert, DistilBert, RoBERTa)*
- *Etc.*

O primeiro *Transformer* introduzido foi o *Bert* model. Foi desenvolvido pela Google. É um modelo de sequência-para-sequência que tem um *encoder* e um *decoder*. O *encoder* mapeia uma sequência de *tokens* de entrada para uma sequência de estados ocultos. O *decoder* toma o output do *encoder* e tenta mapeá-lo de volta para a sequência original de *tokens*.

A *NVIDIA* e a *Facebook* desenvolveram *XLNet* e *DistilBert*. Eles são similares ao *Bert*, mas tem uma arquitectura diferente e um conjunto de pesos diferentes. *XLNet* é um modelo de sequência-para-sequência que tem um *encoder* e um *decoder*. O *encoder* mapeia uma sequência de *tokens* de entrada para uma sequência de estados ocultos. O *decoder* toma o output do *encoder* e tenta mapeá-lo de volta para a sequência original de *tokens*. *DistilBert* é similar ao *XLNet*, mas é treinado em um subconjunto muito menor do que o data.

Como funciona um *Transformer*?

A funcionalidade de um *Transformer* é transformar uma sequência de *tokens* de entrada em uma sequência de *tokens* de saída. Os *tokens* de entrada são geralmente palavras, mas eles também podem ser outros tipos de *tokens*, como imagens de captação. Os *tokens* de saída são geralmente iguais aos *tokens* de entrada, mas eles podem ser diferentes.

Ele funciona através de transformar os *tokens* de entrada em uma nova representação, que pode então ser usada para fazer predições em novos dados. Os *layers* de sequência são chamados de *encoder* e de sequência. Elas tem x *layers* de *encoder* e y *layers* de *decoder*. x e y são geralmente iguais, mas eles podem ser diferentes. Os *layers* podem ser diferentes tamanhos.

Especificamente o *Transformer* é um modelo de sequência-para-sequência. Ele usa *Embeddings* para os *tokens* de entrada e posiciona-os antes de serem enviados para o primeiro conjunto de *layers*. O primeiro conjunto de *layers* tem um *self-attention* mecanismo que toma os *tokens* de entrada e transforma-os em uma nova representação que é adicionada e normalizada antes de ser enviada para o segundo conjunto de *layers* onde os primeiros passos são iguais aos anteriores mas eles usam os outputs anteriores como inputs. Essa nova *layer* combina os inputs com os outputs antigos e os outputs são enviados para o próximo conjunto de *layers*. Isso é repetido até que o output seja o mesmo que o input. Então ele passa por um *layer linear regression* e um *softmax*. O output é o final output.

Layers em específico

Os três mais importantes tipos de *layers* para bem explicar são:

- A *softmax function* é uma função comum em redes *neurais*. Ela toma um vector e retorna um vector com o mesmo tamanho. A *softmax function* é usada para normalizar o output da rede. Ela é usada para garantir que o output é uma distribuição probabilística.
- Uma regressão linear é uma função que toma um vector e retorna um vector. Ela é usada para fazer previsões usando um meio probabilístico simples e comum.
- A função de *self-attention* é feita usando uma combinação de regressão linear e a *softmax*, com ou sem paralelismo (*multi-head attention*), a função de *self-attention* é usada para imprimir uma importância para o conjunto de palavras que está sendo avaliado.

Execução

No nosso *notebook*, foi criado um pipeline de *transformers* do *HuggingFace*, com um modelo BERT da Google *fine-tuned* para classificar sentimentos de forma binária. E foi criado um *notebook* para a execução da tarefa. Este *notebook*, contém o código do pipeline de *transformers* do *HuggingFace*, e as referencias para o modelo *pretrained* e o *tokenizer* desse modelo.

Com blocos alternados de código e *markdown* que comentam a execução do pipeline, foi possível executar a classificação de sentimentos com este pipeline. Iterativamente fomos buscar os *.csv* dos *reviews* dos *estabelecimentos*, e foi aplicado o pipeline de *transformers* do *HuggingFace* sequencialmente a cada *review*, e ao final, foi gerado um arquivo *.csv* com os resultados do local.

Este pipeline foi o mais rápido das três tentativas de classificação de sentimentos, e foi o mais preciso das três tentativas de classificação de sentimentos.

Resultados

Os resultados foram bem sucedidos, e foi possível classificar os sentimentos de um conjunto de *reviews*, e gerar um arquivo *.csv* com os resultados. A classificação foi bem sucedida e bastante precisa, e foi possível classificar os sentimentos de um conjunto de *reviews*, e gerar um arquivo *.csv* com os resultados.

De acordo com a documentação deste modelo (e do BERT original da Google), a classificação foi bem precisa, e a precisão foi de entre 97% e 98%. Dependendo da tarefa o BERT pode até chegar a quase 100% de precisão, neste caso desce devido à natureza binária da classificação.

Embora não sejam 100% precisos, o BERT foi o mais preciso das três tentativas de classificação de sentimentos e os resultados foram bastante satisfatórios.

6. ANÁLISE DE SENTIMENTOS

6.1.3 Ponderação dos Resultados finais

Os resultados obtidos podem não ser 100% precisos, mas ainda assim, podem ser bastante satisfatórios. O qual podemos considerar como um resultado final, e que pode ser usado como um indicador de qualidade do estabelecimento em avaliação.

Os resultados finais estão em pequenas tabelas de amostra nos anexos.

análise sentimental:

naive bayes explica mat por detras (estatistica/regressão linear) pipelines -> BERT
-> Transformer (explicar que tb usam um vectorizer com um stemming/lematization e explicar como funciona a matematica de um transformer e como foi alimentado/treinado o BERT da Google) Um transformer tal como o LSTM é um RNN (rede neural recorrente)

O extrator de palavras do YAKE tb é um naive baye

é qs igual à 2^a tentativa de analise sentimental (naive bayes) so que em vez de ser uma classificação binaria (pos / neg) gera palavras chave (keywords) (classificação multinomial)
keywords + usadas por mes

Capítulo 7

Geração de gráficos

7.1 Geração de gráficos

Para proceder à análise dos dados, é necessário que os dados sejam organizados num formato que permita a leitura e a visualização dos dados facilitada a humanos. Foram gerados os gráficos para a análise de sentimentos e palavras-chave em *sets* de totais e de forma temporal.

Para os gráficos de totais, foi utilizado o pacote *matplotlib* e o *WordCloud* em que foram usados todos os dados disponíveis (das três plataformas). Já os gráficos de análises temporais foram gerados apenas com os dados do *TripAdvisor*, visto que não só eram os mais completos e extensos de todas as categorias de estabelecimento, como também foi possível extrair as datas de criação dos *reviews*.

7.1.1 Gráficos de totais

Com gráficos de totais, queremos dizer que nos dados apresentados e nas análises não é considerado o desenvolvimento temporal, mas sim o desenvolvimento de um total de *reviews*, ou seja, o mês e ano são descartados.

7.1.2 Metodologia

Para a geração dos gráficos de totais, foi utilizado o pacote *matplotlib* e o *wordcloud* num *script Python*, que iterava sobre os dados disponíveis e gerava os gráficos para cada estabelecimento.

Os tipos de gráficos gerados são: gráficos circulares de sentimentos, gráficos de palavras-chave e nuvens de palavras-chave. Os quais demonstram a quantidade de sentimentos positivos e negativos, as dez palavras-chave mais frequentes e as nuvens de palavras-chave limitadas até cem palavras.

7. GERAÇÃO DE GRÁFICOS

Execução

Para cada tipo de estabelecimento, de cada plataforma, foi accionada as rotinas de geração dos gráficos. As quais foram exportadas em formato de imagem. Iterando sobre os tipos de estabelecimento e as plataformas, em que itera sobre os dados disponíveis, segmentados em sentimentos e palavras-chave, que passam por uma e duas funções respetivamente. Gerando e exportando *.jpg* dos gráficos e nuvens.

Resultados

Como se podem verificar nestas três imagens (figura: ??, figura: ??, figura: ??), os gráficos de totais apresentam um desenvolvimento de sentimentos e palavras-chave, que demonstra um sentimento total positivo no nosso turismo, e palavras-chave que apresentam a satisfação com o serviço ou uma característica do estabelecimento.

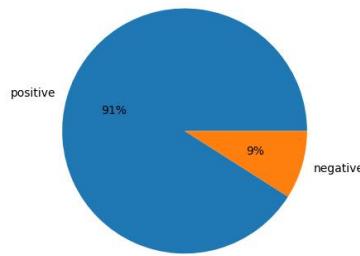


Figura 7.1: Gráfico circular gerado baseando-se nos *sentiments* dados da plataforma *Tripadvisor* referente ao hotel 21

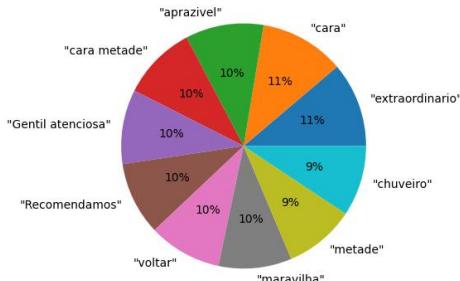


Figura 7.2: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 21



Figura 7.3: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 21

7.1.3 Gráficos temporais

Com gráficos temporais, queremos dizer que nos dados apresentados e nas análises é considerado o desenvolvimento temporal, quer isto dizer que as *reviews* mostradas ao longo do tempo tornando possível verificar as datas em que foram escritas e a quantidade que cada hótel/ restaurante/ atração recebeu ao longo dos anos e meses.

7.1.4 Metodologia

Para a geração dos gráficos temporais, foi utilizado o *software PowerBI*, que utilizava os ficheiros .csv gerados e organizados previamente vindos da base de dados, contendo todas as informações em ficheiros únicos. Os gráficos gerados são: gráficos circulares e de tabelas. Os quais demonstram a quantidade de *sentiments* e *keywords* usadas ao longo do tempo por cada hótel, divididos por anos e meses e também por cada hótel.

Execução

Os ficheiros .csv que contêm as informações relativas a todas as *keywords* e *sentiments*, foram importados para o *software powerBI* e posteriormente organizados da maneira que o grupo achou mais conveniente para que os gráficos ficasse o mais apresentáveis e visivelmente mais fáceis para analisar os dados. Posteriormente os gráficos circulares e de tabelas foram exportados para .jpg e guardados.

Resultados

Como se podem verificar nestas três imagens (figura: ??, figura: ??, figura: ??), os gráficos temporais apresentam um desenvolvimento de *sentiments* e *keywords* ao longo do tempo bastante positivo revelando-se um ótimo ponto para o nosso turismo e *keywords* que mostram bastante agrado.

7. GERAÇÃO DE GRÁFICOS

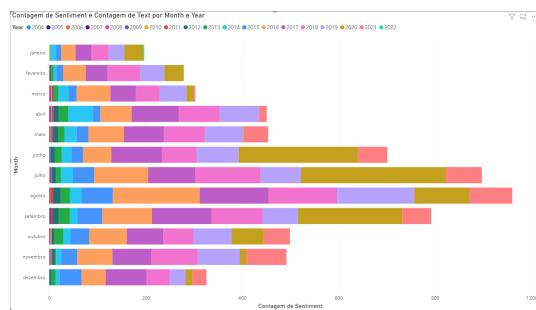


Figura 7.4: Gráfico de tabelas gerado baseando-se em todos os *sentiments* dados da plataforma *Tripadvisor* ao longo dos anos

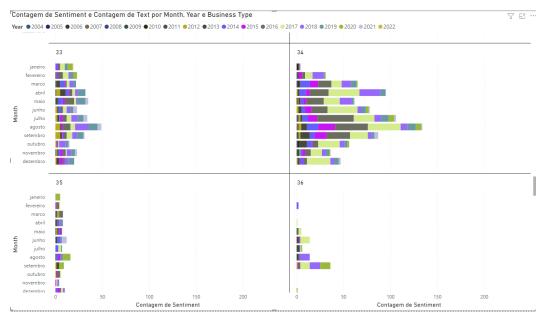


Figura 7.5: Gráfico de tabelas gerado baseando-se em todos os *sentiments* da plataforma *Tripadvisor* referente a cada hotel com o decorrer dos anos

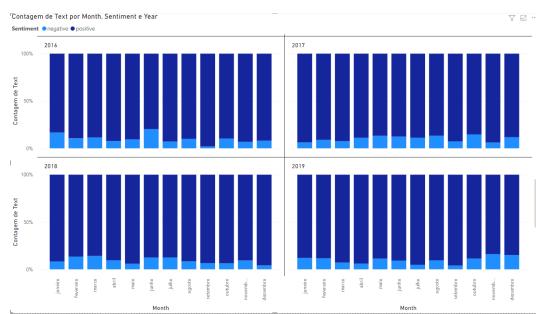


Figura 7.6: Gráfico de tabelas gerado baseando-se na quantidade de *sentiments* da plataforma *Tripadvisor* positivos e negativos ao longo do tempo para cada hotel

Capítulo 8

Reorganização e DB storage

8.1 Reorganização dos dados (*TripAdvisor* apenas)

Para as análises temporais e para o transporte de dados, foi necessário criar uma base de dados organizada e relacional, que garantisse a integridade dos dados e a sua coerência.

Para isso foram usados os pacotes *pandas* para o *import* dos *.csv* em *DataFrames* e *sqlite3* para a criação da base de dados *SQLite3*.

8.1.1 Metodologia

Inicialmente, foi criada uma base de dados *SQLite3*, que será usada para armazenar os dados. A base de dados *SQLite3* foi criada e as tabelas foram criadas, com os campos correspondentes a cada coluna do ficheiro *.csv*, dos quais os dados foram importados via *pandas*.

Os *DataFrames* de *pandas* oferecem uma maneira de aceder e manipular dados, e também fornecem uma maneira de criar e manipular tabelas. Todas as operações de manipulação de dados são feitas através de funções de *DataFrame*.

Seguidamente, é empurrado os dados dos *DataFrames* correspondentes às tabelas para a base de dados. Com esta base de dados, é possível fazer consultas e manipulações de dados, tal como também exportar os dados organizados para ficheiros *.csv* de forma a que possam ser usados em outros programas, tais como o *R*, o *Excel*, o *PowerBI*, etc.

8.1.2 Execução e Resultados

Foi feito um *script* para executar a criação da base de dados *SQLite3*, e para o *import* dos dados dos *DataFrames* para a base de dados. O qual foi executado com sucesso, como podemos ver na base de dados e nos ficheiros *.csv*.

Capítulo 9

Análise dos dados obtidos

9.1 Resultados obtidos

Tendo em vista os gráficos gerados por meio das bibliotecas *Matplotlib*, *Wordcloud* e também os do *software PowerBI* foi possível reunir um grande conjunto de informação para realizar uma análise dos melhores pontos turísticos que o património cultural alentejano pode oferecer e assim melhorar possíveis pontos negativos e positivos, ou então prever quais as atrações/hóteis/restaurantes que mais cativam os turistas.

Assim sendo, foram gerados gráficos para cada hótel/ restaurante/ atração dos *websites Tripadvisor*, *Booking* e *Zoomato* usando as bibliotecas mencionadas com o intuito dos resultados serem mais facilmente visíveis com as *keywords* mais usadas para mencionar cada um dos pontos turísticos, assim como um mapa da cidade de Beja que contem todas as *keywords* e as percentagens entre *sentiments* positivos e negativos.

Por fim foram gerados também gráficos com as informações ao longo do tempo dos mesmos *websites* referidos, acerca dos *sentiments* de cada ponto turístico utilizando o *software PowerBI*.

9. ANÁLISE DOS DADOS OBTIDOS

9.1.1 Resultados de totais

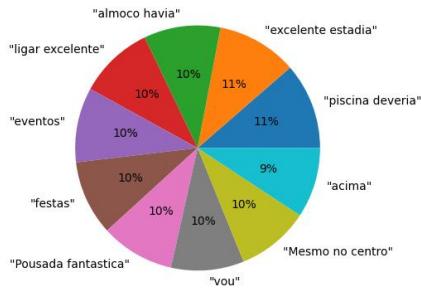


Figura 9.1: Gráfico circular com as *keywords* mais usadas para o hotel 0



Figura 9.2: Mapa de Beja com as *keywords* mais usadas para o hotel 0 no seu interior

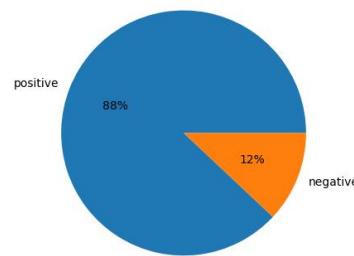


Figura 9.3: Gráfico circular a representar a diferença entre os *sentiments* positivos e negativos

Nas figuras apresentadas acima são mostrados alguns resultados gerados pelas bibliotecas mencionadas, dos quais podemos notar que existe uma maioria para a quantidade de *sentiments* positivos em relação aos negativos e que a maior parte das *keywords* são também positivas. Porém, estes valores são retirados no momento em que a extração dos dados foi realizada e não é possível verificar à medida do tempo como esses valores foram surgindo. Os valores entre restaurantes/hóteis/atrações é bastante semelhante entre si e então foi decidido que só iria ser mostrado alguns exemplos da realização desta etapa e todos os resultados ficariam mostrados nos anexos.

9.1.2 Resultados temporais

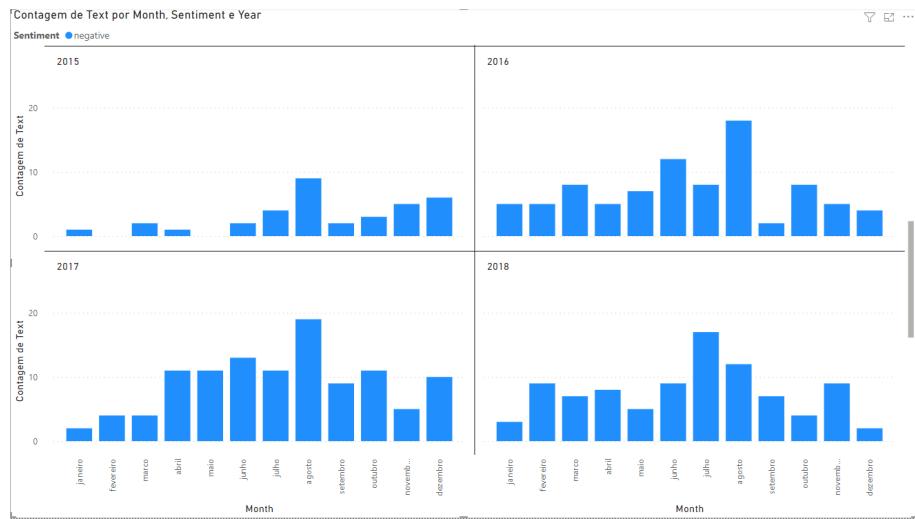


Figura 9.4: Gráfico de tabelas com a quantidade de *sentiments* negativos ao longo do ano de cada hotel

9. ANÁLISE DOS DADOS OBTIDOS

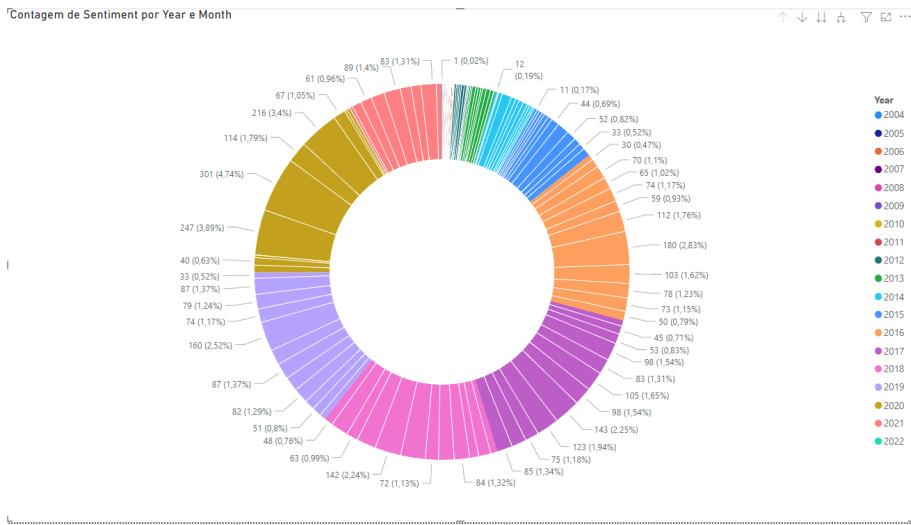


Figura 9.5: Gráfico circular com a quantidade de *reviews* ao longo do ano

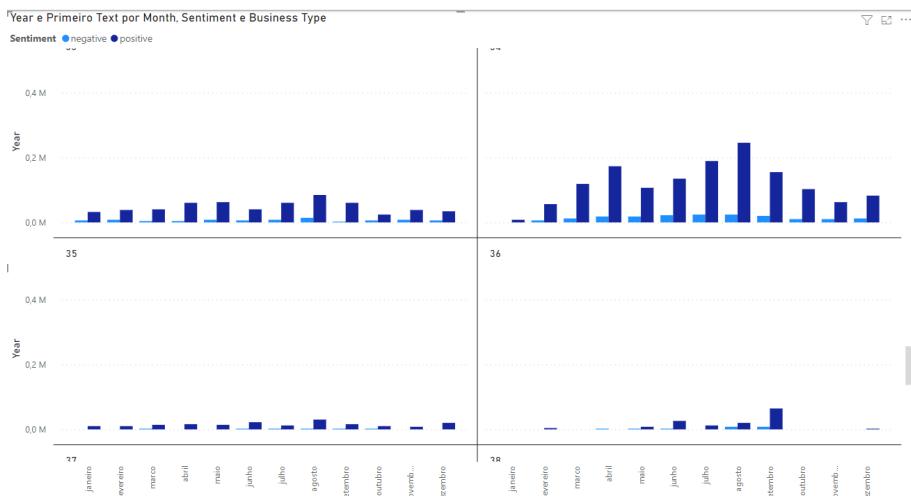


Figura 9.6: Gráfico de tabelas com a quantidade de *sentiments* ao longo do ano por cada estabelecimento

As figuras apresentadas desta vez foram elaboradas com o valor temporal bastante visível, sendo possível verificar desta vez uma evolução com o decorrer do tempo dos valores apresentados, assim como a forte diferença entre a quantidade de *sentiments* escritos em meses onde um grande número de pessoas adere aos serviços, como no verão, Páscoa ou Natal.

9.2 Análise dos resultados

Graças a estes gráficos podemos concluir que as opiniões acerca do turismo cultural na zona alentejana é bastante positiva, mostrando uma enorme maioria de comentários positivos contra uma pequena quantidade de comentários negativos (Figura: ??). Podemos também notar que existem muitas mais pessoas a dar as sua opiniões em meses como junho, julho e agosto, muito possivelmente devido á abertura das épocas balneares que movem grandes grupos de turistas nacionais e estrangeiros a fazerem férias pelas zonas costeiras que o Alentejo consegue fornecer com enorme facilidade graças ás magnificas praias na sua zona costeira. Por fim também é possível notar a evolução no número de opiniões com o decorrer dos anos e com a popularidade que o *website* vai conseguindo, já que no começo o número de opiniões é baixo, porem com o passar dos anos começa a subir em elevado número.

É interessante também realçar um detalhe acerca de um gráfico em específico que o grupo decidiu não passar em branco. O gráfico da figura ??.

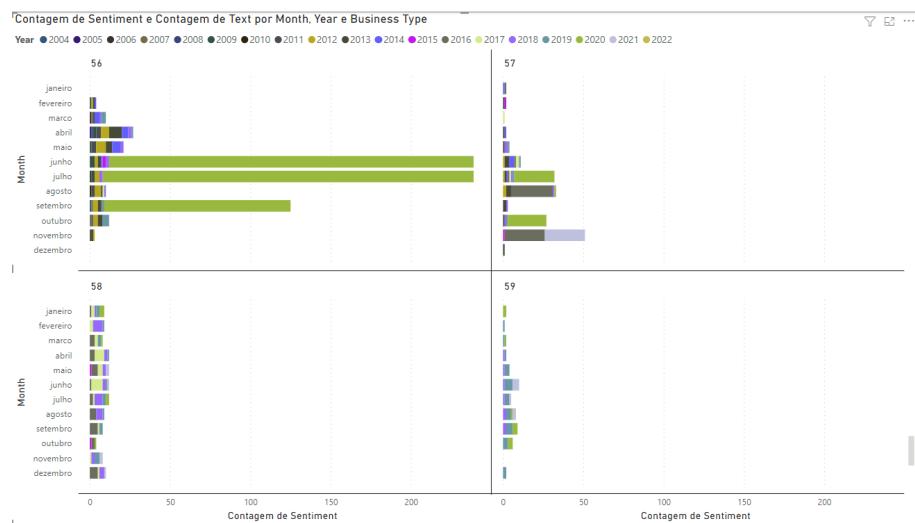


Figura 9.7: Gráfico de tabelas com a quantidade de *sentiments* ao longo do ano por cada estabelecimento com valores diferentes

O estabelecimento com o número 56 contém valores interessantes nos meses de junho e julho. São interessantes uma vez que durante o ano de 2020 o país se encontrava em confinamento devido á COVID-19, não sendo possível que ajuntamentos fossem realizados, porem foi verificado através do gráfico que o mesmo não se parece verificar já que existe uma brutal subida no número de pessoas a dar a sua opinião acerca da estadia que realizou, o que dá a entender que esse estabelecimento continuou a realizar as suas tarefas com normalidade ao contrário de outros que provavelmente seguiram as normas recomendadas.

Capítulo 10

Conclusão

Uma analise de dados apresenta sempre uma oportunidade de crescimento, em especial num meio de turismo rural (ou semi-rural) em que os habituais clientes podem (e vão) usar redes sociais e plataformas de comentário de prestação de serviço, para agregar dados comportamentais e percepção publica das ofertas turistas.

Neste trabalho foi apresentado um agregado exemplar desse tipo de dados e analise o qual foi feito na melhor das formas, dado o conhecimento a nós atribuído pela escola e o tempo disponível. O qual apresentou pouco de informação nova, mas colabora com a percepção publica que se sente nos meios de comunicação verbal subjectiva em locais públicos, ou seja, existe a tendência de haver um maior numero de clientes durante as típicas alturas de ferias, referimos o ano novo, a pascoa e em especial o verão.

O que foi dito acima foram as tendências observadas na analise dos dados obtidos nos gráficos gerados para analise temporal, com os dados da plataforma mais completa deste estudo, o *TripAdvisor*. O único caso mais em especial, foi a explosão de *reviews* em Junho e Julho de 2020, que coincide com a abertura da época balnear após desconfinamento (COVID-19) no dia 6 de Junho de 2020, que embora fazendo *cross-referencing* com os dados das palavras-chave da mesma analise temporal, não nos dê informação relevante, a coincidência é meramente atractiva e oferece uma fácil explicação.

Sendo assim podemos concluir que foi um trabalho bem sucedido, mesmo que não ofereça novos *insights*, mas corrobora as noções e conhecimento publico da área, tal como os dados das analises anuais oferecidas pelo *Booking*. Graças a isto obtemos uma noção da metodologia desta área profissional e os nossos avaliados terão uma noção da qualidade de trabalho produzida por nós.

Sumarizamos assim que nesta secção de turismo rural (e semi-rural), da região de Beja, tem por norma uma boa prestação de serviço, ou pelo menos essa noção é transposta na mente popular de quem visita, e que não existem anomalias na distribuição de visitantes nem de tendências sentimentais ao longo do ano nem ao longo das décadas, mantendo-se previsível mas agradável.

Apêndices

Apêndice I

Jupyter Notebook da Extração de Palavras-Chave

Nas páginas seguintes está incluída uma renderização do *notebook* de *Jupyter* em que se detalha (em inglês) os passos das rotinas de extração de palavras-chave, comentando de forma simples o funcionamento das funções de código usadas.

Keyword Extraction

In this notebook, we will use the [Keyword Extraction](#) technique to extract keywords from text. We will use the [YAKE!](#) library to extract keywords from text.

[YAKE!](#) is a library that can be used to extract keywords from text made by portuguese authors (and a japanese) from Polytechnic Institute of Tomar, University of Beira Interior, University of Porto, INESC TEC (and Kyoto University).

Importing the libraries

Here we import the libraries we will use.

```
In [ ]: import os
import warnings
import yake
import pandas
```

Functions

Now we define the functions we will use. We will use the following functions:

- `extract_keywords`: to extract keywords from text.
- `extract_keywords_chunks`: to extract keywords from text using chunks.
- `import_csv`: to import a csv file.
- `import_csv_chunks`: to import a csv file using chunks.
- `get_keywords`: to get the keywords from a list of keywords.
- `get_keywords_chunks`: to get the keywords from a list of keywords using chunks.
- `get_keywords_dir`: to get the keywords from a directory of files.
- `get_keywords_dir_chunks`: to get the keywords from a directory of files using chunks.
- `get_keywords_zomato_dir`: to get the keywords from zomato folder.

Function: extract_keywords

This function extracts keywords from text using YAKE! library. It takes as input a text and returns a list of keywords.

```
In [ ]: def extract_keywords(df):
    keywords = []
    for i in range(0, len(df)):
        review = df["Avaliacoes"][i]
        keywords.append(yake.KeywordExtractor(lan="pt").extract_keywords(review))
    return keywords
```

Function: import_csv

This function imports a csv file. It takes as input a csv file and returns a dataframe.

```
In [ ]: def import_csv(path):
    df = pandas.read_csv(path, encoding="utf-8")
    return df
```

Function: get_keywords

This function gets all the csv files in a directory and returns a list of keywords from the dataframes.

```
In [ ]: def get_keywords(path):
    keywords = []
    for file in os.listdir(path):
        if file.endswith(".csv") and not file.startswith("list"):
            df = import_csv(path + "/" + file)
            keywords.append(extract_keywords(df))
    return keywords
```

Function: get_keywords_dir

This function gets all the csv files in a directory and get a list of keywords from the dataframes, then it writes the keywords in a csv file.

```
In [ ]: def get_keywords_dir(path1, path2, name):
    current_dir = os.getcwd()
    path = current_dir + path1
    keywords = get_keywords(path)
    # print(keywords)
    i = 0
    for keyword in keywords:
        # print(keyword)
        with open(
            current_dir + path2 + name + str(i) + ".csv",
            "w",
        ) as f:
            f.write("Expressao, Frequencia\n")
            for k in keyword:
                for word in k:
                    f.write(
                        str(word)
                        .replace("(", "")
                        .replace(")", "")
                        .replace("\u2010", "-")
                        + "\n"
                    )
        i += 1
```

Function: get_keywords_zomato_dir

This function gets all the csv files from the zomato directory and get a list of keywords from the dataframes, then it writes the keywords in a csv file.

```
In [ ]: def get_keywords_zomato_dir(path1, path2, name):
    current_dir = os.getcwd()
    path = current_dir + path1
    keywords = get_keywords(path)
    # print(keywords)
    i = 0
    restaurantes = [0, 1, 2, 12, 13, 14, 15]
    for keyword in keywords:
        # print(keyword)
        with open(
            current_dir + path2 + name + str(restaurantes[i]) + ".csv",
            "w",
        ) as f:
            f.write("Expressao, Frequencia\n")
            for k in keyword:
                for word in k:
                    f.write(
                        str(word)
                        .replace("(", "")
                        .replace(")", "")
                        .replace("\u2010", "-")
                        + "\n"
                    )
        i += 1
```

Execution

Now we execute the code. We will use the last set functions to get the keywords from the reviews.

```
In [ ]: warnings.filterwarnings("ignore")

get_keywords_dir("../scrapes/booking/hotels", "/booking/hotels/", "hotel")
get_keywords_zomato_dir(
    "../scrapes/zomato/restaurantes",
    "/zomato/restaurantes/",
    "restaurante",
)
get_keywords_dir(
    "../scrapes/tripadvisor/restaurants",
    "/tripadvisor/restaurants/",
    "restaurant",
)
get_keywords_dir(
    "../scrapes/tripadvisor/activities",
    "/tripadvisor/activities/",
    "place",
)
get_keywords_dir("../scrapes/tripadvisor/hotels", "/tripadvisor/hotels/", "hotel")
```

[Loading IPython/Jupyter output/CommonHTML /fonts/TeXfontdata is]

Apêndice II

Jupyter Notebook da Analise de Sentimentos via Naïve Bayes Classifier

Nas paginas seguintes está incluída uma renderização do *notebook* de *Jupyter* em que se detalha (em inglês) os passos das rotinas de extracção de palavras-chave, comentando de forma simples o funcionamento das funções de código usadas.

Sentiment Analysis

In this notebook, we will use a dataset of movie reviews to train a model to predict whether a review is positive or negative. We will use the [ultc-movies.csv](#) to train the model (this is not on the Github repo due to its size).

To train the model, we will use the `sklearn.feature_extraction.text.CountVectorizer` to create a bag of words representation of the reviews with a `nltk.stem.SnowballStemmer` to stem the words, and `sklearn.naive_bayes.MultinomialNB` as our machine learning model.

We will use the `sklearn.metrics.classification_report` to evaluate the model.

Importing the libraries

Here we import the libraries we will use.

```
In [ ]: import os
import re
import warnings
import unicodedata
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report
from nltk import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer
```

Creating a custom tokenizer

We'll need a custom tokenizer that stems the words in our reviews. We will use the `nltk.stem.SnowballStemmer` to stem the words.

This tokenizer will be used to create a bag of words representation of the reviews.

```
In [ ]: class StemmerTokenizer:
    def __init__(self):
        self.stemmer = SnowballStemmer("portuguese")
    def __call__(self, doc):
        return [self.stemmer.stem(t) for t in word_tokenize(doc)]
```

Functions

Now we will define the functions we will use. We will use the following functions:

- `create_dataframe`: to create a dataframe from the csv file
- `load_directory`: to load the data from the directory
- `get_training_data`: to get the training data
- `drop_useless_columns`: to drop the columns that we don't need
- `get_csv`: to get the csv file
- `filter_string`: to filter the string
- `integer`: to convert the string to an integer
- `get_count_vectorizer_with_stopwords`: to get the count vectorizer with stopwords
- `normalize`: to normalize the data
- `emotion_from_int`: to get the emotion from the integer
- `predict`: to predict the emotion

Function: `create_dataframe`

This function will create a dataframe from the csv file.

```
In [ ]: def create_dataframe(filename):
    nan_value = float("NaN")
    df = pd.read_csv(filename)
    df.replace("", nan_value, inplace=True)
    df.dropna(how="any", inplace=True)
    return df
```

Function: load_directory

This function will load the data from the directory.

```
In [ ]: def load_directory(directory):
    files = []
    for filename in os.listdir(directory):
        if filename.endswith(".csv") and not filename.startswith("list"):
            files.append(filename)
    return files
```

Function: get_training_data

This function will get the training data.

```
In [ ]: def get_training_data():
    df = create_dataframe(
        "models/utlc_movies.csv"
    ) # original_index, review_text, review_text_processed, review_text_tokenized, polarity, rating, kfold_polarity, kfold_rating
    df = drop_useless_columns(df)
    df = df.rename(columns={"polarity": "Class", "review_text_processed": "Data"})
    df = df.reindex(columns=["Class", "Data"])
    return df.sample(frac=1).reset_index(drop=True)
```

Function: drop_useless_columns

This function will drop the columns that we don't need.

```
In [ ]: def drop_useless_columns(df):
    df.drop(
        [
            "original_index",
            "review_text",
            "review_text_tokenized",
            "rating",
            "kfold_polarity",
            "kfold_rating",
        ],
        axis=1,
        inplace=True,
    )
    return df
```

Function: get_csv

This function will get the csv file.

```
In [ ]: def get_csv(path):
    df = create_dataframe(path)
    return df
```

Function: filter_string

This function will filter the string.

```
In [ ]: def filter_string(df, column):
    ret = (
        df[column]
        .apply(lambda x: re.sub("[^a-zA-Z]", " ", x))
        .apply(lambda x: re.sub(" +", " ", x))
        .apply(lambda x: x.strip())
        .apply(lambda x: x.lower())
        .apply(lambda x: normalize(x))
        .values
    )
    return ret
```

Function: integer

This function will convert the string to an integer.

```
In [ ]: def integer(x):
    return int(x)
```

Function: get_count_vectorizer_with_stopwords

This function will get the count vectorizer with stopwords.

```
In [ ]: def get_count_vectorizer_with_stopwords():
    return CountVectorizer(
        tokenizer=StemmerTokenizer(),
        ngram_range=(1, 2),
        stop_words=stopwords.words("portuguese"),
    )
```

Function: normalize

This function will normalize the data.

```
In [ ]: def normalize(text):
    text = (
        unicodedata.normalize("NFKD", text)
        .encode("ascii", "ignore")
        .decode("utf-8", "ignore")
    )
    return text
```

Function: emotion_from_int

This function will get the emotion from the integer.

```
In [ ]: def emotion_from_int(x):
    if x == 0:
        return "Negative"
    elif x == 1:
        return "Positive"
    else:
        return "Unknown"
```

Function: predict

This function will predict the emotion.

```
In [ ]: def predict(model, vec, directory):
    files = load_directory(directory)
    for filename in files:
        df_predict = get_csv(directory + "/" + filename)
        predict_data = filter_string(df_predict, "Avaliacoes")
        # print("Loaded data to predict")
        reviews = []
        sentiments = []
        for review in predict_data:
            sentiment = emotion_from_int(
                model.predict(vec.transform([review]).toarray())[0]
            )
            # print("Model predicts that: " + str(review) + " is " + str(sentiment))
            reviews.append(str(review))
            sentiments.append(str(sentiment))
        with open(
            directory.replace("scrapes", "sentimentanalysis") + "/" + filename,
            "w",
            encoding="utf-8",
        ) as f:
            f.write("Sentiment, Review\n")
            for i in range(len(reviews)):
                f.write("'" + sentiments[i] + "', '" + reviews[i] + "'\n")
```

Execution

Now we will execute the code.

```
In [ ]: warnings.filterwarnings("ignore")
df = get_training_data()
```

```

df_test = df.sample(frac=0.1, random_state=42).reset_index(drop=True)

df = df[:15000]
df_test = df_test[:5000]

train_data, train_class = (
    filter_string(df, "Data"),
    df["Class"].apply(lambda x: integer(x)).values,
)
print("Loaded training data")

test_data, test_class = (
    filter_string(df_test, "Data"),
    df_test["Class"].apply(lambda x: integer(x)).values,
)
print("Loaded test data")

vec = get_count_vectorizer_with_stopwords()
print("Created vectorizer")

train_data = vec.fit_transform(train_data).toarray()
print("Transformed training data")

test_data = vec.transform(test_data).toarray()
print("Transformed test data")

model = MultinomialNB()
print("Created model")

model.fit(train_data, train_class)
print("Trained model")

print(
    "Tested model: " + str(model.score(test_data, test_class) * 100) + "%" + " accuracy"
)

```

Prediction of the emotion of the reviews of various establishments of various types and platforms and export the results to a csv file.

```
In [ ]: predict(model, vec, "../scrapes/booking/hotels")
predict(model, vec, "../scrapes/zomato/restaurante")
predict(model, vec, "../scrapes/tripadvisor/hotels")
predict(model, vec, "../scrapes/tripadvisor/restaurants")
predict(model, vec, "../scrapes/tripadvisor/activities")
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Apêndice III

Jupyter Notebook da Analise de Sentimentos via BERT Transformer Pipelines

Nas paginas seguintes está incluída uma renderização do *notebook* de *Jupyter* em que se detalha (em inglês) os passos das rotinas de extracção de palavras-chave, comentando de forma simples o funcionamento das funções de código usadas.

Sentiment Analysis (using Transformers and PyTorch)

This notebook demonstrates how to use `transformers` to perform sentiment analysis. These transformers use the PyTorch library to perform the actual computation.

Imports

Here we import the required packages.

```
In [ ]: import os
import warnings
import pandas
import torch
from transformers import pipeline
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
```

Functions

Here we define the functions that we will use in this notebook. We will use these functions to perform sentiment analysis. Which are:

- `get_pipeline`: This function creates a pipeline that will be used to perform sentiment analysis.
- `get_dir`: This function returns the directory where the data is stored.
- `get_reviews`: This function returns the reviews.
- `get_sentiments`: This function returns the sentiments of the reviews.
- `export`: This function exports the sentiment analysis.
- `analyse_directory`: This function analyses the sentiment analysis of the reviews in a directory.

Function: `get_pipeline`

This function creates a pipeline that will be used to perform sentiment analysis.

```
In [ ]: def get_pipeline():
    model = AutoModelForSequenceClassification.from_pretrained(
        "gchhablani/bert-base-cased-finetuned-sst2"
    )
    tokenizer = AutoTokenizer.from_pretrained(
        "gchhablani/bert-base-cased-finetuned-sst2", do_lower_case=False
    )
    senti_pipeline = pipeline(
        "sentiment-analysis", model=model, tokenizer=tokenizer, truncation=True
    )
    return senti_pipeline
```

Function: `get_dir`

This function returns the directory where the data is stored.

```
In [ ]: def get_dir(path):
    files = []
    for file in os.listdir(path):
        if file.endswith(".csv") and not file.startswith("list"):
            files.append(file)
    return files
```

Function: `get_reviews`

This function returns the reviews.

```
In [ ]: def get_reviews(df):
    reviews = []
    for i in range(0, len(df)):
        reviews.append(str(df["Avaliações"][i]))
    return reviews
```

Function: `get_sentiments`

This function returns the sentiments of the reviews.

```
In [ ]: def get_sentiments(reviews, senti_pipeline):
    sentiments = []
    for review in reviews:
        sentiments.append(str(senti_pipeline(review)[0].get("label")))
    return sentiments
```

Function: export

This function exports the sentiment analysis.

```
In [ ]: def export(sentiments, reviews, path, name):
    df = pandas.DataFrame({"sentiment": sentiments, "review": reviews})
    path = path.replace("../scrapes/", "/")
    df.to_csv(str(path) + str(name), index=False)
```

Function: analyse_directory

This function analyses the sentiment analysis of the reviews in a directory.

```
In [ ]: def analyse_directory(path, senti_pipeline):
    files = get_dir(path)
    for file in files:
        df = pandas.read_csv(path + file, encoding="utf-8")
        reviews = get_reviews(df)
        sentiments = get_sentiments(reviews, senti_pipeline)
        export(sentiments, reviews, path, file)
```

Execution

Here we execute the notebook. First we create the pipeline.

```
In [ ]: senti_pipeline = get_pipeline()
```

Then we get the directory where the data is stored. Then we get the reviews and the sentiments. Finally we export the sentiment analysis.

```
In [ ]: current_dir = os.getcwd()
path = current_dir + "../scrapes/"
analyse_directory(path + "booking/hotels/", senti_pipeline)
analyse_directory(path + "zomato/restaurantes/", senti_pipeline)
analyse_directory(path + "tripadvisor/hotels/", senti_pipeline)
analyse_directory(path + "tripadvisor/activities/", senti_pipeline)
analyse_directory(path + "tripadvisor/restaurants/", senti_pipeline)
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Apêndice IV

Jupyter Notebooks do Webscraping dos hotéis do Booking

Nas páginas seguintes está incluída uma renderização dos *notebooks* de *Jupyter* em que se detalha (em inglês) os passos das rotinas de extração de palavras-chave, comentando de forma simples o funcionamento das funções de código usadas.

O primeiro detalha a busca dos hotéis e o segundo apenas os *reviews* de cada um dos hotéis encontrados no primeiro.

Importing some libraries that we need to make the webscrapping of the booking.com

```
In [ ]: from bs4 import BeautifulSoup
import requests
import pandas as pd
```

We need to create the request to make the website send the information: To make that we use the library `requests` and the `BeautifulSoup` and the inspect tool to extract que name of the classes and the headers needed

```
In [ ]: headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4649.116 Safari/537.36'
url = "https://www.booking.com/searchresults.pt-pt.html?aid=375654&label=msn-jrwrFdUb9zKNuCHIkGmz2g-8074541083442"
response = requests.get(url, headers=headers)
soup = BeautifulSoup(response.content, 'lxml')
```

And the the arrays to receive the information from the website like the name of the hotels, the ratings... Every hotel are from Beja.

```
In [ ]: hotel = []
badge = []
title = []
reviews = []
price = []
for item in soup.select('.fb3c4512b4'):
    try:
        hotel.append(item.select('.fde44d7ef')[0].get_text().strip())
        badge.append(item.select('.9c5f726ff')[0].get_text().strip())
        title.append(item.select('.192b3a196')[0].get_text().strip())
        reviews.append(item.select('.le6021d2f')[0].get_text().strip())
        price.append(item.select('.e885fdc12')[0].get_text().strip())
    except Exception as e:
        print('')
```

Only in case any of the arrays are different in size

```
In [ ]: # bad code
length = len(price)
if length > len(reviews):
    length = len(reviews)
if length > len(hotel):
    length = len(hotel)
if length > len(badge):
    length = len(badge)
if length > len(title):
    length = len(title)
```

Saving the respective information and extracting to .csv

```
In [ ]: d1 = {'Hotel': hotel[:length], 'Classificação': badge[:length],
          'Suma': title[:length], 'Avaliações': reviews[:length], 'Preço': price[:length]}
df = pd.DataFrame.from_dict(d1)
print(df)
df.to_csv('listtable.csv')
```

The same libraries were used when webscraping was performed for hotels

```
In [ ]: from bs4 import BeautifulSoup
import requests
import pandas as pd
```

In this case it took a larger amount of headers to access the site information.

```
In [ ]: headers = {
    "Access-Control-Allow-Origin": "*",
    "Access-Control-Allow-Methods": "GET",
    "Access-Control-Allow-Headers": "Content-Type",
    "accept": "*/*",
    "accept-encoding": "gzip, deflate",
    "accept-language": "en-GB,en;q=0.9,en-US;q=0.8",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.122 Safari/537.36"
}
url = "https://www.booking.com/searchresults.pt.pt.html?aid=375654&label=msn-jrwrFdUb9zKNuCHIkGmz2g-8074541083442"
response = requests.get(url, headers=headers)
soup = BeautifulSoup(response.content, 'lxml')
```

To access the respective comments of each hotel it was necessary to build a new link to be performed the webscraping, for this was used a common part for all links and added the name of the hotel, to access your comments specifically

```
In [ ]: reviews_links = []
for link in soup.findAll('a', {'class': 'fb01724e5b'}):
    a = link['href']
    hotel = a.split('/')[5].split('?')[0]
    a = 'https://www.booking.com/reviews/pt/hotel/' + hotel
    reviews_links.append(a)
```

In this last part, the comments relating to each website were extracted and the .csv of each of them was created.

```
In [ ]: count = 0
allreviews = []

for link in reviews_links:
    try:
        response2 = requests.get(link, headers=headers)
        soup2 = BeautifulSoup(response2.content, 'lxml')
        for r in soup2.findAll('span', {'itemprop': 'reviewBody'}):
            try:
                rev = r.text
                allreviews.append(rev + '\n')
            except:
                pass
    except:
        pass
    count += 1
    if allreviews != []:
        seen = set()
        allreviews = [item for item in allreviews if not(
            tuple(item) in seen or seen.add(tuple(item)))]
        dfr = pd.DataFrame.from_dict({'Avaliações': allreviews})
        print(dfr)
        dfr.to_csv('hotel' + str(count) + '.csv')
        allreviews = []
```


Apêndice V

Jupyter Notebook do Webscraping dos reviews de Restaurantes do Zomato

Nas páginas seguintes está incluída uma renderização do *notebook* de *Jupyter* em que se detalha (em inglês) os passos das rotinas de extração de palavras-chave, comentando de forma simples o funcionamento das funções de código usadas.

Zomato

teste 123

```
In [ ]: from bs4 import BeautifulSoup
import requests
import pandas as pd

texto

In [ ]: headers = {
    # "Access-Control-Allow-Origin": "*",
    # "Access-Control-Allow-Methods": "GET",
    # "Access-Control-Allow-Headers": "Content-Type",
    # "accept": "/*",
    "User-Agent": "Mozilla/5.0 (Linux; Android 6.0; Nexus 5 Build/MRA58N) AppleWebKit/537.36 (KHTML, like Gecko)"
}

# url = "https://www.zomato.com/beja/beja-restaurants"
# response = requests.get(url, headers=headers)
# response.status_code
#soup = BeautifulSoup(response.content, 'lxml')
soup = BeautifulSoup(open(r"C:\Users\vitor\Desktop\pi2021\projeto\webscrape\scrapes\zomato\restaurantes\zomato-t

In [ ]: restaurant = []
for name in soup.findAll('h4',{'class':'sc-1hp8d8a-0'}):
    restaurant.append(name.text.strip())
print(len(restaurant))

In [ ]: type = []
for name in soup.findAll('p',{'class':'jaKOQh'}):
    print(name)
    type.append(name.text.strip())
for name in soup.findAll('p',{'class':'kegdaG'}):
    print(name)
    type.append(name.text.strip())
print(len(type))

In [ ]: price = []
for p in soup.findAll('p',{'class':'ftdqla'}):
    price.append(p.text.replace('€ para dois','').strip())
for p in soup.findAll('p',{'class':'kOONhy'}):
    price.append(p.text.replace('€ para dois','').strip())
print(len(price))

In [ ]: #bad code

length = len(price)
if length > len(type):
    length = len(type)
if length > len(restaurant):
    length = len(restaurant)

In [ ]: d1 = {'Restaurante': restaurant[:length], 'Tipo':type[:length], 'Preço':price[:length]}
df = pd.DataFrame.from_dict(d1)
print(df)
df.to_csv('listtable.csv')

In [ ]: reviews_links = []
for link in soup.findAll('a', {'class': 'ieKty'}):
    a = link['href']
    reviews_links.append(a.replace('/info', '/reviews'))
for link in soup.findAll('a', {'class': 'jjSACU'}):
    a = link['href']
    reviews_links.append(a.replace('/info', '/reviews'))
# print(reviews_links)

In [ ]: count = 0
allreviews = []
```

```
for link in reviews_links:
    try:
        response2 = requests.get(link, headers=headers)
        soup2 = BeautifulSoup(response2.content, 'xml')
        for r in soup2.findAll('p'): # sempre a mudar a class, vai sofrer ETL
            try:
                rev = r.text
                # print(rev)
                allreviews.append(rev + '\n')
            except:
                pass
    except:
        pass
    count += 1
if allreviews != []:
    seen = set()
    allreviews = [item for item in allreviews if not(
        tuple(item) in seen or seen.add(tuple(item)))]
    dfr = pd.DataFrame.from_dict({'Avaliações': allreviews})
    print(dfr)
    dfr.to_csv('restaurante' + str(count) + '.csv')
    allreviews = []
```


Apêndice VI

Jupyter Notebooks do Webcrawing dos reviews dos hotéis, atracções e restaurantes do TripAdvisor

Nas paginas seguintes está incluída uma renderização dos *notebooks* de *Jupyter* em que se detalha (em inglês) os passos das rotinas de extracção de palavras-chave, comentando de forma simples o funcionamento das funções de código usadas.

O primeiro extraiu os hotéis, o segundo as atracções e o terceiro os restaurantes (e os seus respectivos *reviews*).

TripAdvisor

Webscraping

Imports

First we start with the imports. We need essentially three (or four) main libraries to work this out; these are:

- requests (to fetch the website)
- lxml (a faster html parser to speed bs4)
- bs4 (a.k.a beautiful soup, a web scraping library)
- pandas (a maths oriented data(set) manipulation library)

Since requests uses urllib3 as a dependency, we can import it first to configure it to suppress the annoying warning about the "insecure" connection (lack of SSL).

```
In [ ]: import urllib3
urllib3.disable_warnings(urllib3.exceptions.InsecureRequestWarning)
import requests
from bs4 import BeautifulSoup as soup
import pandas as pd
```

Request configuration

We need to configure our request, specially in this case, since TripAdvisor won't send us a webpage if we at least not try to emulate a real browser. First we configure our headers, ripping the main headers from our browser, as seen in the Developer tools (F12) in Chromium (we used the new Microsoft Edge).

Then we request the webpage (Hotels in Beja, Portugal) with our headers attached, a timeout to stop if it takes too long (something is wrong), and verification is disabled (SSL). If the status code is OK (200), doesn't print.

After that we create a BeautifulSoup4 scrapable object with the html content of the page, using lxml.

```
In [ ]: headers = {
    'Access-Control-Allow-Origin': '*',
    'Access-Control-Allow-Methods': 'GET',
    'Access-Control-Allow-Headers': 'Content-Type',
    'accept': '/*',
    'accept-encoding': 'gzip, deflate, br',
    'accept-language': 'en-GB,en;q=0.9,en-US;q=0.8',
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.110 Safari/537.36'}
url = "https://www.tripadvisor.pt/Hotels-g189102-Beja_Beja_District_Alentejo-Hotels.html"
req = requests.get(url,headers=headers,timeout=5,verify=False)
req.status_code
bsobj = soup(req.content, 'lxml')
```

Scraping

Now we start scraping. First we start getting hotel names.

```
In [ ]: hotel = []
for name in bsobj.findAll('div',{'class':'listing_title'}):
    hotel.append(name.text.strip())
print(len(hotel))
```

Then their ratings.

```
In [ ]: ratings = []
for rating in bsobj.findAll('a',{'class':'ui_bubble_rating'}):
    ratings.append(rating['alt'])
print(len(ratings))
```

The number of reviews (they have a big issue).

```
In [ ]: reviews = []
for review in bsobj.findAll('a',{'class':'review_count'}):
    reviews.append(review.text.strip())
print(len(reviews))
```

This number is referring to the TOTAL number of reviews. However, we can only scrape a single language, that's dependent of the domain/locale (.com .pt); there's no query parameter to change either the number here, or the sorting of reviews.

Now we get the prices.

```
In [ ]: price = []
for p in bsobj.findAll('div',{'class':'price-wrap'}):
    price.append(p.text.replace('€','').strip())
print(len(price))
```

Some of these will be empty, since the price is gotten via phone call.

Now, we get to the weird part: Reviews. These reviews are handled by the hotel page in weird ways:

- only four shown per subpage
- each subpage is counted via a multiple of five
- any multiple of five non-existent will not give 404, but redirect to the first subpage
- language selection via scraping is non-existent, no query parameters, only radio buttons with random labels, defaults chosen via domain/locale (.com .pt)

So the strategy found is:

- create a monstrous amount of subpage links (about 200)
- scrape all reviews, even repeated via the redirect to the first subpage
- later use sets, or dictionaries to remove duplicates

That was done with this awful looking but functional code.

```
In [ ]: links = []
for review in bsobj.findAll('a',{'class':'review_count'}):
    try:
        a = review['href']
        a = 'https://www.tripadvisor.pt'+ a
        c = a[:a.find('Reviews')+7] + ' -or' + a[(a.find('Reviews')+7):]
        links.append(c)
    for i in range(5,1000,5):
        b = a[:a.find('Reviews')+7] + '-or' + str(i) + a[(a.find('Reviews')+7):]
        links.append(b)
    except:
        pass
print(links)
```

Now, we get the smallest length number of the arrays regarding to the hotels, hoping we remove some of the (repeated) sponsorships.

```
In [ ]: # bad code
length = len(price)
if length > len(reviews):
    length = len(reviews)
if length > len(hotel):
    length = len(hotel)
if length > len(ratings):
    length = len(ratings)
```

And then create the ID table of the hotels with the most basic information in a pandas DataFrame, and export that one to a .csv file that we can use in Excel, PowerBI, ML libraries like Keras, Tensorflow, SciKitLearn can use.

```
In [ ]: d1 = {'Hotel':hotel[:length],'Estrelas':ratings[:length],'Avaliações':reviews[:length],'Preço':price[:length]}
df = pd.DataFrame.from_dict(d1)
print(df)
df.to_csv('listtable.csv')
```

Now the most horrible of codes presents you with the creations of various, separated .csv files with the scraped reviews, that we can use.

It iterates all the links and since there's 200 links per hotel, every 200 we use some list comprehension magic with sets to remove duplicates and export the DataFrame to a useful .csv file.

```
In [ ]: count = 0
count2 = 0
allreviews = []
for link in links:
    try:
        html2 = requests.get(link,headers=headers)
        bsobj2 = soup(html2.content,'lxml')
        for r in bsobj2.findAll('q'):
            try:
                rev = r.span.text.strip()
                allreviews.append(rev + '\n')
            except:
                pass
    except:
        pass
    count += 1
    if count == 200:
        df = pd.DataFrame(allreviews)
        df.to_csv('Reviews'+str(count2)+'.csv')
        allreviews = []
        count2 += 1
        count = 0
```

```
        except:  
            pass  
    except:  
        pass  
    count += 1  
    if count == 200:  
        seen = set()  
        allreviews = [item for item in allreviews if not(tuple(item) in seen or seen.add(tuple(item)))]  
        dfr = pd.DataFrame.from_dict({'Avaliações':allreviews})  
        print(dfr)  
        dfr.to_csv('hotel' + str(count2) + '.csv')  
        allreviews = []  
        count = 0  
        count2 += 1
```

TripAdvisor

Webscraping

Imports

First we start with the imports. We need essentially three (or four) main libraries to work this out; these are:

- requests (to fetch the website)
- lxml (a faster html parser to speed bs4)
- bs4 (a.k.a beautiful soup, a web scraping library)
- pandas (a maths oriented data(set) manipulation library)

Since requests uses urllib3 as a dependency, we can import it first to configure it to suppress the annoying warning about the "insecure" connection (lack of SSL).

```
In [ ]: import urllib3
import requests
import re
from bs4 import BeautifulSoup as soup
import pandas as pd
urllib3.disable_warnings(urllib3.exceptions.InsecureRequestWarning)
```

Request configuration

We need to configure our request, specially in this case, since TripAdvisor won't send us a webpage if we at least not try to emulate a real browser. First we configure our headers, ripping the main headers from our browser, as seen in the Developer tools (F12) in Chromium (we used the new Microsoft Edge).

Then we request the webpage (Attractions in Beja, Portugal) with our headers attached, a timeout to stop if it takes too long (something is wrong), and verification is disabled (SSL). If the status code is OK (200), doesn't print.

After that we create a BeautifulSoup4 scrapable object with the html content of the page, using lxml.

```
In [ ]: headers = {
    "Access-Control-Allow-Origin": "*",
    "Access-Control-Allow-Methods": "GET",
    "Access-Control-Allow-Headers": "Content-Type",
    "accept": "*/*",
    "accept-encoding": "gzip, deflate, br",
    "accept-language": "en-GB,en;q=0.9,en-US;q=0.8",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.122 Safari/537.36"
}
url = (
    "https://www.tripadvisor.pt/Attractions-g189102-Activities-Beja_Beja_District_Algentejo.html"
)
req = requests.get(url, headers=headers, timeout=5, verify=False)
req.status_code
bsobj = soup(req.content, "lxml")
```

Scraping

Now we start scraping. First we start getting attraction names.

```
In [ ]: place = []
for name in bsobj.findAll("span", {"name": "title"}):
    place.append(re.sub(r"\b\d+\b", "", name.text.strip())[2:])
print(place)
```

Some of these will be empty, since the price is gotten via phone call.

Now, we get to the weird part: Reviews. These reviews are handled by the attraction page in weird ways:

- only ten shown per subpage
- each subpage is counted via a multiple of ten
- any multiple of five non-existent will not give 404, but redirect to the first subpage
- language selection via scraping is non-existent, no query parameters, only radio buttons with random labels, defaults chosen via domain/locale (.com .pt)

So the strategy found is:

- create a monstrous amount of subpage links (about 400)

- scrape all reviews, even repeated via the redirect to the first subpage
- later use sets, or dictionaries to remove duplicates

That was done with this awful looking but functional code.

```
In [ ]:
links = []
for review in bsoobj.findAll("a", {"href": re.compile(r'#REVIEWS')}):
    try:
        a = review["href"]
        a = "https://www.tripadvisor.pt" + a
        c = a[: (a.find("Reviews") + 7)] + "" + a[(a.find("Reviews") + 7):]
        links.append(c)
        for i in range(10, 4000, 10):
            b = (
                a[: (a.find("Reviews") + 7)]
                + "-or"
                + str(i)
                + a[(a.find("Reviews") + 7):])
            )
        links.append(b)
    except:
        pass
# print(links)
```

And then create the ID table of the attractions with the most basic information in a pandas DataFrame, and export that one to a .csv file that we can use in Excel, PowerBI, ML libraries like Keras, Tensorflow, SciKitLearn can use.

```
In [ ]:
length = len(place)
d1 = {
    "Attraction": place[:length]
}
df = pd.DataFrame.from_dict(d1)
print(df)
df.to_csv("listtable.csv")
```

Now the most terrible of codes presents you with the creations of various, separated .csv files with the scraped reviews, that we can use.

It iterates all the links and since there's 400 links per attraction, every 400 we use some list comprehension magic with sets to remove duplicates and export the DataFrame to a useful .csv file.

```
In [ ]:
count = 0
count2 = 0
allreviews = []
for link in links:
    try:
        html2 = requests.get(link, headers=headers)
        bsoobj2 = soup(html2.content, "lxml")
        for r in bsoobj2.findAll("span", {"class": "NejBf"}): # as of 7Dez, because in 6Dez it was "class": "cSoN"
            for rev in r:
                try:
                    rv = rev.text
                    if "desde" not in rv and "€" not in rv:
                        allreviews.append(rv + "\n")
                except:
                    pass
    except:
        pass
    count += 1
    if count == 400:
        seen = set()
        allreviews = [
            item
            for item in allreviews
            if not (tuple(item) in seen or seen.add(tuple(item)))]
    dfr = pd.DataFrame.from_dict({"Avaliações": allreviews})
    print(dfr)
    dfr.to_csv("place" + str(count2) + ".csv")
    allreviews = []
    count = 0
    count2 += 1
```

TripAdvisor

Webscraping

Imports

First we start with the imports. We need essentially three (or four) main libraries to work this out; these are:

- requests (to fetch the website)
- lxml (a faster html parser to speed bs4)
- bs4 (a.k.a beautiful soup, a web scraping library)
- pandas (a maths oriented data(set) manipulation library)

Since requests uses urllib3 as a dependency, we can import it first to configure it to suppress the annoying warning about the "insecure" connection (lack of SSL).

```
In [ ]: import urllib3
import requests
import re
from bs4 import BeautifulSoup as soup
import pandas as pd
urllib3.disable_warnings(urllib3.exceptions.InsecureRequestWarning)
```

Request configuration

We need to configure our request, specially in this case, since TripAdvisor won't send us a webpage if we at least not try to emulate a real browser. First we configure our headers, ripping the main headers from our browser, as seen in the Developer tools (F12) in Chromium (we used the new Microsoft Edge).

Then we request the webpage (Restaurants in Beja, Portugal) with our headers attached, a timeout to stop if it takes too long (something is wrong), and verification is disabled (SSL). If the status code is OK (200), doesn't print.

After that we create a BeautifulSoup4 scrapable object with the html content of the page, using lxml.

```
In [ ]: headers = {
    "Access-Control-Allow-Origin": "*",
    "Access-Control-Allow-Methods": "GET",
    "Access-Control-Allow-Headers": "Content-Type",
    "accept": "*/*",
    "accept-encoding": "gzip, deflate, br",
    "accept-language": "en-GB,en;q=0.9,en-US;q=0.8",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.122 Safari/537.36"
}
url = "https://www.tripadvisor.pt/Restaurants-g189102-Beja_Beja_District_Algentejo.html"
req = requests.get(url, headers=headers, timeout=5, verify=False)
req.status_code
bsobj = soup(req.content, "lxml")
```

Scraping

Now we start scraping. First we start getting restaurant names.

```
In [ ]: place = []
prelinks = []
for name in bsobj.findAll("div", {"class": "OhCyu"}):
    place.append(re.sub(r"\b\d+\b", "", name.span.text.strip())[2:])
    prelinks.append(name.span.a["href"])
```

Some of these will be empty, since the price is gotten via phone call.

Now, we get to the weird part: Reviews. These reviews are handled by the restaurant page in weird ways:

- only ten shown per subpage
- each subpage is counted via a multiple of ten
- any multiple of five non-existent will not give 404, but redirect to the first subpage
- language selection via scraping is non-existent, no query parameters, only radio buttons with random labels, defaults chosen via domain/locale (.com .pt)

So the strategy found is:

- create a monstrous amount of subpage links (about 400)
- scrape all reviews, even repeated via the redirect to the first subpage

- later use sets, or dictionaries to remove duplicates

That was done with this awful looking but functional code.

```
In [ ]:
links = []
for pre in prelinks:
    try:
        a = "https://www.tripadvisor.pt"
        c = a + "" + pre
        d = c[: (c.find("-Reviews-") + len("-Reviews-") - 1)]
        e = c[(c.find("-Reviews-") + len("-Reviews-") - 1) :]
        links.append(c)
        for i in range(10, 4000, 10):
            b = d + "-or" + str(i) + e
            links.append(b)
    except:
        pass
```

And then create the ID table of the attractions with the most basic information in a pandas DataFrame, and export that one to a .csv file that we can use in Excel, PowerBI, ML libraries like Keras, Tensorflow, SciKitLearn can use.

```
In [ ]:
length = len(place)

d1 = {"Restaurant": place[:length]}
df = pd.DataFrame.from_dict(d1)
print(df)
df.to_csv("listtable.csv")
```

Now the most terrible of codes presents you with the creations of various, separated .csv files with the scraped reviews, that we can use.

It iterates all the links and since there's 400 links per restaurant, every 400 we use some list comprehension magic with sets to remove duplicates and export the DataFrame to a useful .csv file.

```
In [ ]:
count = 0
count2 = 0
allreviews = []
for link in links:
    try:
        html2 = requests.get(link, headers=headers)
        bsobj2 = soup(html2.content, "lxml")
        for r in bsobj2.findAll("p", {"class": "partial_entry"}):
            for rev in r:
                try:
                    rv = rev.text.strip()
                    allreviews.append(rv + "\n")
                except:
                    pass
    except:
        pass
    count += 1
    if count == 400:
        seen = set()
        allreviews = [
            item
            for item in allreviews
            if not (tuple(item) in seen or seen.add(tuple(item)))]
    dfr = pd.DataFrame.from_dict({"Avaliações": allreviews})
    dfr.to_csv("restaurant" + str(count2) + ".csv")
    print(dfr)
    allreviews = []
    count = 0
    count2 += 1
```

Apêndice VII

Código do *script* de *Python* que fez o *trimming* dos ficheiros *.csv*

Aqui abaixo está representado o código de *Python* que fez o que fez o *trimming* dos ficheiros *.csv webscraped* para uso posterior sem potenciais erros.

Apêndice VIII

Código do *script* de *Python* que fez a normalização NFKD dos ficheiros

.csv

Aqui abaixo está representado o código de *Python* que fez a normalização NFKD dos ficheiros *.csv webscraped* para a sua utilização com os módulos de *keyword extraction* e análise sentimental poderem ser feitos.

Apêndice IX

Código do *script* de *Python* que reorganizou os dados numa base de dados relacional

Aqui abaixo está representado o código de *Python* que reorganizou os dados numa base de dados relacional, assim podendo exporta noutras *.csv* de forma mais organizada e/ou relacionar os dados garantindo uma maior e melhor integridade e coerência.

Apêndice X

Código do *script* de *Python* que gerou os gráficos não temporais dos hotéis, atracções e restaurantes do *TripAdvisor*, *Booking* e *Zomato*

Aqui abaixo está representado o código de *Python* que gerou os gráficos circulares e as nuvens de palavras com de análise não temporal de todos os estabelecimentos de todos os tipos de todas as plataformas. O qual usa essencialmente “matplotlib” e “wordcloud” para a geração dos gráficos (mais uma máscara com o formato da região de Beja).

Apêndice XI

Código do *script* de *Python* que gerou os gráficos temporais dos hotéis, atracções e restaurantes do *TripAdvisor*

Aqui abaixo está representado o código de *Python* que gerou os gráficos circulares e as nuvens de palavras com de análise temporal de todos os estabelecimentos da plataforma *TripAdvisor*. O qual usa essencialmente “matplotlib” e “wordcloud” para a geração dos gráficos (mais uma máscara com o formato da região de Beja).

Anexos

Anexo I

Anexo referente às imagens dos gráficos utilizados no relatório (Tripadvisor-Hóteis)

Uma vez que as imagens que retratam os gráficos elaborados são muito extensas, serão apenas mostradas algumas delas e em cada secção apontado o link que redirecciona para o GitHub do projecto onde será possível aceder a cada imagem respectivamente assim como ao ficheiro *.pbix* que contém os gráficos realizados no *PowerBI*. Inicialmente serão expostos os gráficos totais, anteriormente mostrados no capítulo 7 (Geração de gráficos) exclusivamente para o website *Tripadvisor* e acerca dos hóteis.

Acesso a todos os gráficos originados: GitHub.

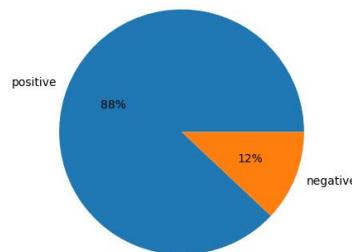


Figura I.1: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente ao hotel 0

I. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (TRIPADVISOR-HÓTEIS)



Figura I.2: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 0

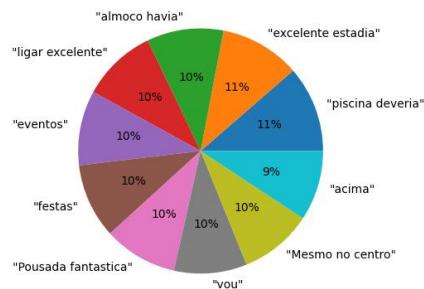


Figura I.3: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 0

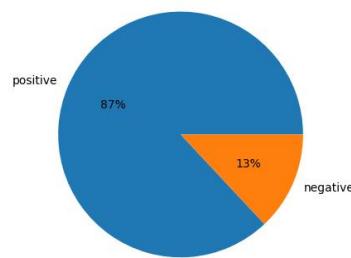


Figura I.4: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente ao hotel 8



Figura I.5: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 8



Figura I.6: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 8

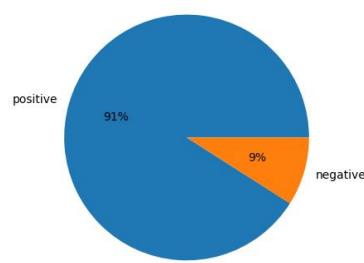


Figura I.7: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente ao hotel 21

I. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO
(TRIPADVISOR-HÓTEIS)



Figura I.8: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 21

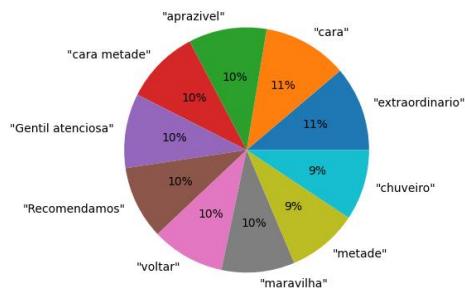


Figura I.9: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao hotel 21

Anexo II

Anexo referente às imagens dos gráficos utilizados no relatório (Tripadvisor-Restaurantes)

Acesso a todos os gráficos originados: GitHub.

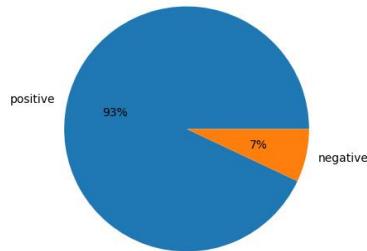


Figura II.1: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente ao restaurante 0

II. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (TRIPADVISOR-RESTAURANTES)

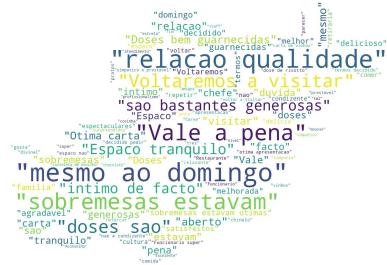


Figura II.2: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao restaurante 0



Figura II.3: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao restaurante 0

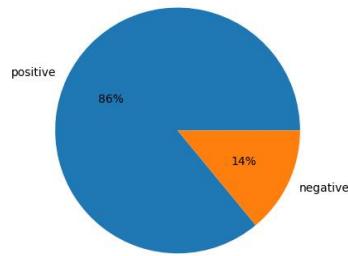


Figura II.4: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente ao restaurante 8



Figura II.5: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao restaurante 8

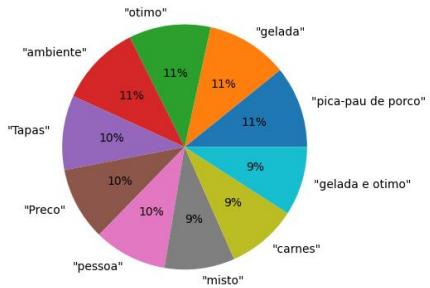


Figura II.6: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao restaurante 8

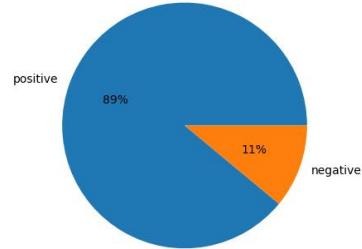


Figura II.7: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente ao restaurante 21

II. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (TRIPADVISOR-RESTAURANTES)



Figura II.8: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente ao restaurante 21

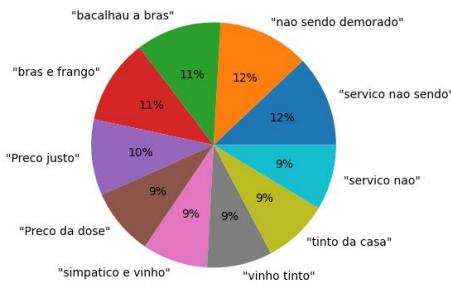


Figura II.9: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente ao restaurante 21

Anexo III

Anexo referente às imagens dos gráficos utilizados no relatório (Tripadvisor-Actividades)

Acesso a todos os gráficos originados: GitHub.

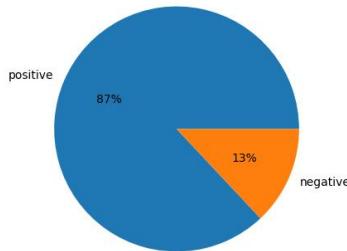


Figura III.1: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente à actividade 0

**III. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO
(TRIPADVISOR-ACTIVIDADES)**



Figura III.2: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente à actividade 0

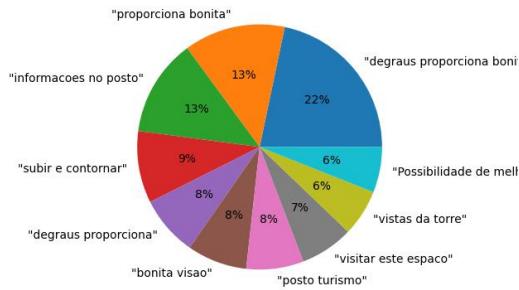


Figura III.3: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente à actividade 0

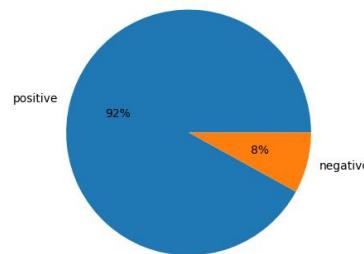


Figura III.4: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente à actividade 8



Figura III.5: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente à actividade 8

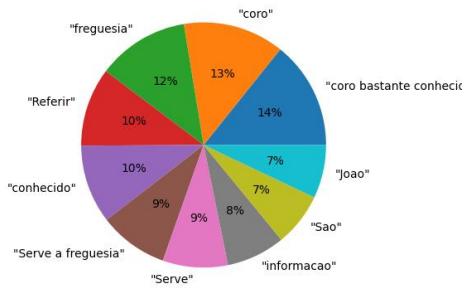


Figura III.6: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente à actividade 8

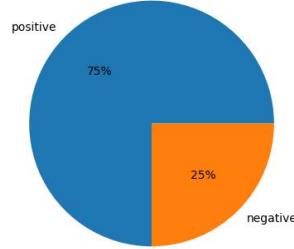


Figura III.7: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* referente à actividade 21

III. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (TRIPADVISOR-ACTIVIDADES)

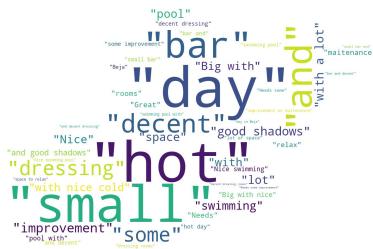


Figura III.8: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Tripadvisor* referente à actividade 21

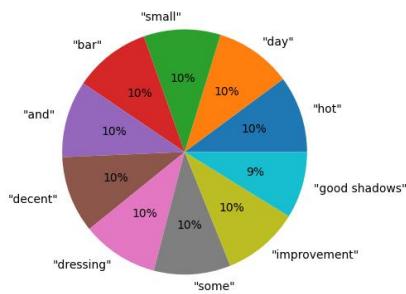


Figura III.9: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Tripadvisor* referente à actividade 21

Anexo IV

Anexo referente às imagens dos gráficos utilizados no relatório (Booking)

Acesso a todos os gráficos originados: GitHub.

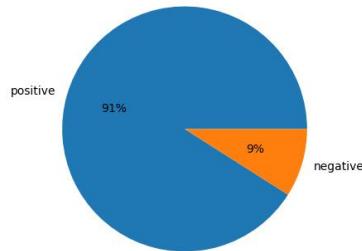


Figura IV.1: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Booking* referente ao hotel 1

IV. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (BOOKING)



Figura IV.2: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Booking* referente ao hotel 1



Figura IV.3: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Booking* referente ao hotel 1

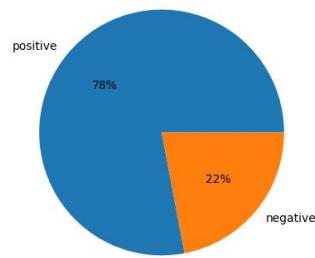


Figura IV.4: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Booking* referente ao hotel 5



Figura IV.5: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Booking* referente ao hotel 5



Figura IV.6: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Booking* referente ao hotel 5

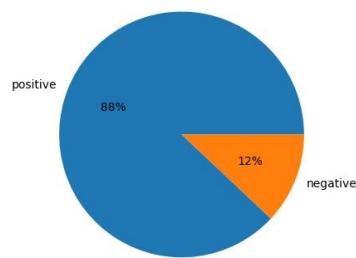


Figura IV.7: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Booking* referente ao hotel 21

IV. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (BOOKING)



Figura IV.8: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Booking* referente ao hotel 21

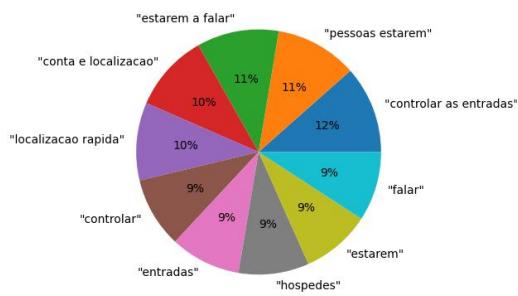


Figura IV.9: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma Booking referente ao hotel 21

Anexo V

Anexo referente às imagens dos gráficos utilizados no relatório (Zoomato)

Acesso a todos os gráficos originados: GitHub.

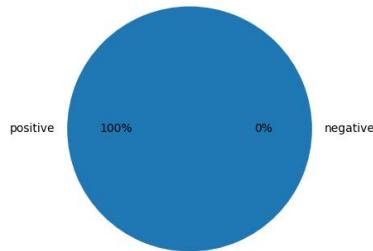


Figura V.1: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Zoomato* referente ao restaurante 0

V. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (ZOOMATO)



Figura V.2: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Zoomato* referente ao restaurante 0

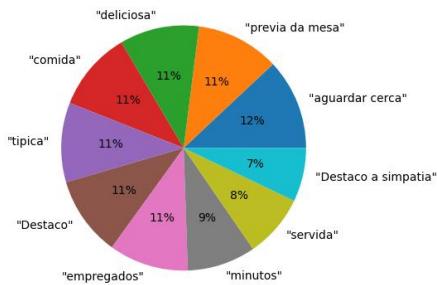


Figura V.3: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma Zoomato referente ao restaurante 0

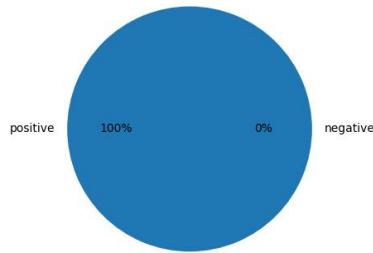


Figura V.4: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Zoomato* referente ao restaurante 2



Figura V.5: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Zoomato* referente ao restaurante 2



Figura V.6: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma Zoomato referente ao restaurante 2

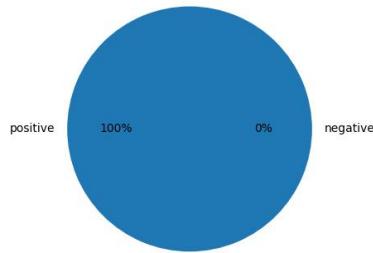


Figura V.7: Gráfico circular gerado baseando-se nos *sentiments* mais usados da plataforma *Zoomato* referente ao restaurante 12

V. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO
(ZOOMATO)



Figura V.8: Gráfico de palavras-chave e nuvens de palavras-chave contendo as *keywords* mais usadas da plataforma *Zoomato* referente ao restaurante 12

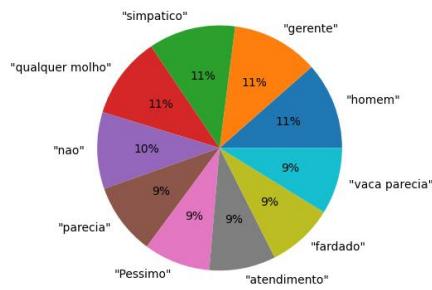


Figura V.9: Gráfico circular gerado baseando-se nas *keywords* mais usadas da plataforma *Zoomato* referente ao restaurante 12

Anexo VI

Anexo referente às imagens dos gráficos utilizados no relatório (Tripadvisor - gráficos temporais)

Acesso aos ficheiros *PowerBI* criados e as imagens retiradas do mesmo:

1. GitHub gráficos PowerBI 1.
2. GitHub gráficos PowerBI 2.
3. GitHub PowerBI screenshots.

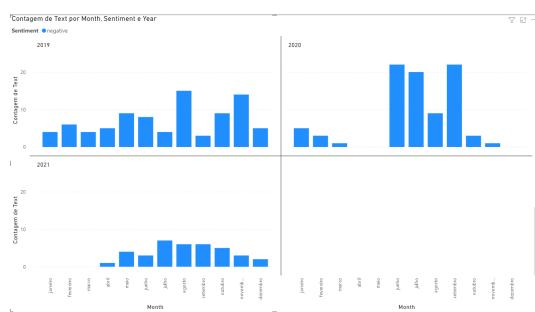


Figura VI.1: Gráfico de barras gerado baseando-se nos *sentiments* mais usados da plataforma *Tripadvisor* com níveis temporais sobre os *sentiments* negativos ao longo do tempo

VI. ANEXO REFERENTE ÀS IMAGENS DOS GRÁFICOS UTILIZADOS NO RELATÓRIO (TRIPADVISOR - GRÁFICOS TEMPORAIS)

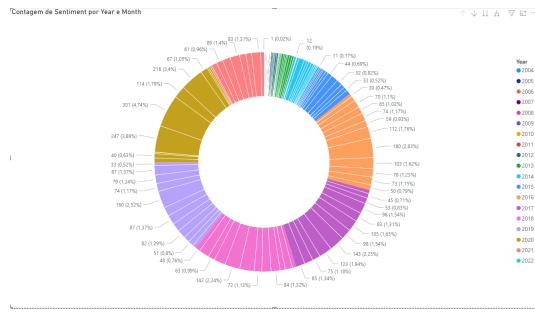


Figura VI.2: Gráfico circular gerado baseando-se no número de *sentiments* da plataforma *Tripadvisor* ao longo dos anos

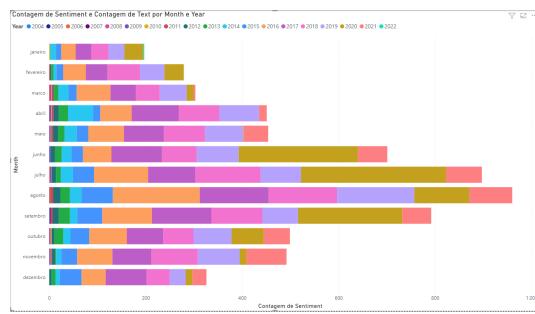


Figura VI.3: Gráfico de barras gerado baseando-se no número de *sentiments* da plataforma *Tripadvisor* ao longo dos anos

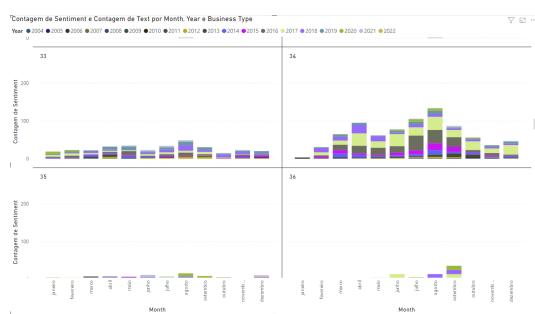


Figura VI.4: Gráfico de barras gerado baseando-se nos *sentiments* mais usados e em cada hotel da plataforma *Tripadvisor* ao longo dos anos

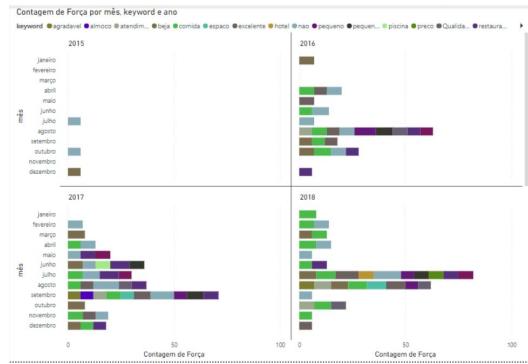


Figura VI.5: Gráfico de barras gerado baseando-se nas *keywords* mais usadas dos hóteis da plataforma *Tripadvisor* ao longo dos anos

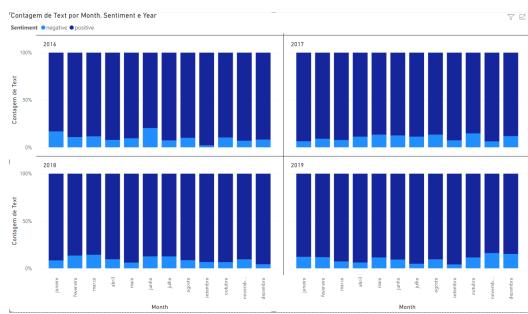


Figura VI.6: Gráfico de barras gerado baseando-se nos *sentiments* positivos e negativos dos hóteis da plataforma *Tripadvisor* ao longo dos anos

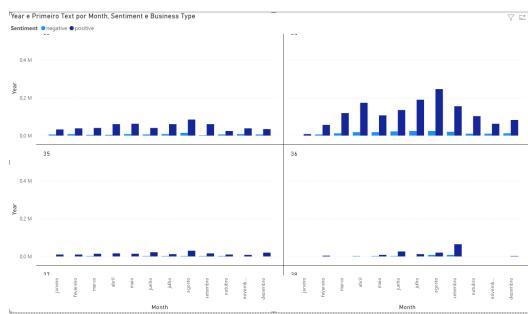


Figura VI.7: Gráfico de barras gerado baseando-se nos *sentiments* positivos e negativos da plataforma *Tripadvisor* ao longo dos anos em cada hotel