TRABALHO DE MINERAÇÃO DE TEXTO

MBA – Big Data & Bussiness Analytics

Prof. Gustavo Mirapalheta

Alunos:

Ana Paula Puddu, Fábio Monteiro, Lucas Sena, Marcos Soares e Wagner Fonseca

Introdução:

***Neste trabalho, vamos realizar um estudo aprofundado sobre algumas obras da escritora britânica* Agatha Christe.**

Nascida na cidade de Torquay, no sul da Inglaterra, em 1890, Agatha se tornou uma das maiores escritoras do gênero Romance Policial.

Curiosidades sobre a autora

Rainha do Crime

Sua vasta obra de Romances Policiais a rendeu o apelido de “Rainha do Crime”, pelos enredos criativos e enigmáticos.

Guiness Book

Agatha Christie é a autora com o maior número de obras vendidas, com mais de 4 bilhões de livros. Só fica atrás de William Shakespeare e da Bíblia.

Realeza

Em 1971, foi condecorada como Dama-Comendadora da Ordem do Império Britânico, a maior honraria concedida para as mulheres pela família real de lá.

Aposta

Seu primeiro livro, “O Misterioso caso de Styles”, foi escrito após uma aposta com sua irmã, que duvidava que Agatha conseguiria escrever um livro e publicá-lo.

68 anos sem parar

A peça “A ratoeira” (The Mousetrap) foi ao palco pela primeira vez em 1952 e só parou de ser exibida por conta da pandemia, em 2020. Foram 68 anos ininterruptos e 10 milhões de ingressos vendidos.

Obras

Agatha foi responsável por escrever mais de 100 livros em seus 85 anos de vida.

Obras analisadas neste trabalho

Com uma obra tão extensa, realizamos o trabalho a seguir focado em seis obras (em inglês) da escritora britânica:

- The Mysterious Affair at Styles (1920)
- The Murder on the Links (1923)
- The Man in the Brown Suit (1924)
- The Hunter's Lodge Case (1924, parte do livro de contos Poirot Investigates)
- The Case of the Missing Will (1924, parte do livro de contos Poirot Investigates
- The Plymouth Express (1978, parte do livro de contos Poirot's Early Cases)

Livros Escolhidos

The Mysterious Affair at Styles (1920)

Lançado em outubro de 1920, foi o primeiro livro da autora. Sua edição de lançamento tinha 296 páginas.

The Murder on the Links

Terceira obra da escritora, o livro de 1923 tinha 298 páginas em sua edição.

The man in the Brown Suit

Obra lançada em agosto de 1924, com edição inicial em capa dura e 312 páginas.

Contos Escolhidos

The Mystery of Hunter's Lodge

Parte da coletânea de contos chamada Poirot Investigates, que foi lançada em março de 1924, este conto é o quarto da obra e toma 14 páginas do livro.

The Case of the Missing Wil

Este conto é o 11º conto da coletânea Poirot Investigates, com 11 páginas dedicadas a estória.

The Plymouth Express

Apesar de ter sido publicado em 1923 no jornal The Sketch, este conto se tornou famoso somente em 1974, com o lançamento da coletânea Poirot's Early Cases.

1

Preparação:

Importação, tratamento e modelagem dos dados para análise

Gutenbergr:

- Para o nosso trabalho utilizamos a Biblioteca gutenbergr para importar os livros da Agatha Christie
- Essa biblioteca permite a importação de mais de 60.000 livros direto para o R para que possa ser manuseado em projetos de analise de texto
- Com isso adicionamos suas obras a um data frame com duas colunas, onde cada linha representa uma linha do livro

A partir dos livros importados, buscamos juntar os dataframes em uma única base para se iniciar o processo de Tokenização

Tokenização:

Como funciona:

O processo de tokenização consiste em estruturar o Data frame importado, onde cada palavra representa uma linha, para um data frame onde cada linha representa uma palavra de seu respectivo livro. Com isso, conseguimos quantificar o número de vezes em que cada palavra apareceu no livro e também a frequência de uma palavra em cada um dos livros, quando se trata de uma análise comparativa. Esse processo é essencial para a maioria das analises realizadas em Mineração de Texto.

Tokenização:

Stop_Words

Para que as análises sejam efetivas e façam sentido, é importante filtrar aquelas palavras que não agregam significado ao texto e sim estão lá somente para complementar o texto ou reforçar um significado. Para isso utilizamos a biblioteca stop_words e realizamos um anti_join() com as palavras que pertencem a esse conjunto

47.863 linhas

Ao combinar todos os livros em um DataFrame

13.034 Palavras

Após o processo de Tokenização

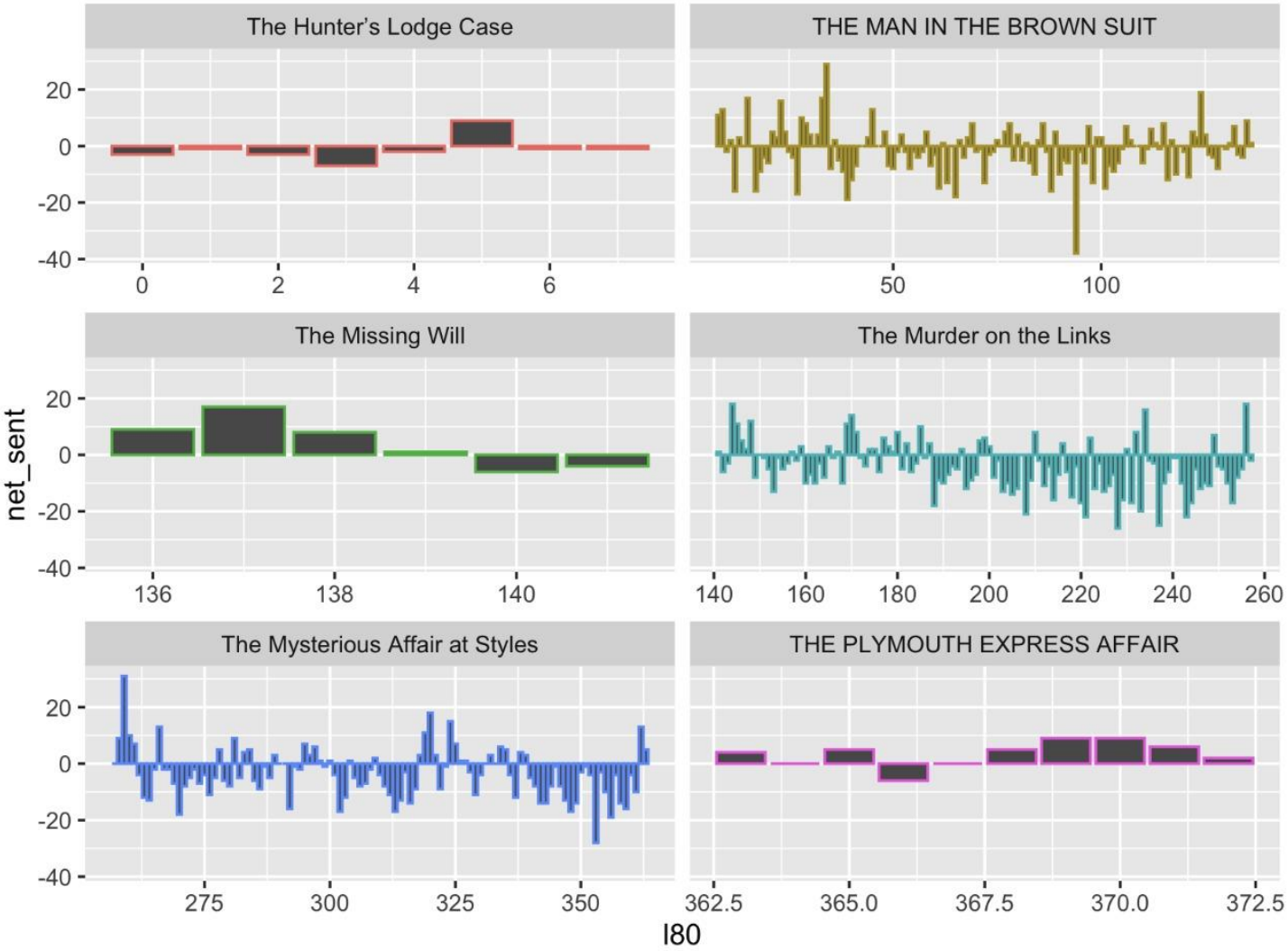
06

Livros em nossa análise

1

Análise de Sentimentos:

Estrutura e sentimentalização das obras selecionadas



Acima de zero:
termos positivos

Abaixo de zero:
termos negativos

Sentimentalização (get_sentiments (“bing”)

Após realizar a categorização de cada um dos seis livros analisados em agrupamentos, contendo 80 linhas cada cluster, empregamos técnicas de análise de sentimentos (get_sentiments de "bing") para categorizar a quantidade de ocorrências de palavras positivas ou negativas.

Conseguimos montar gráficos individualmente pelos títulos ao se analisar os clusters categorizados pelas classificações das palavras. Assim, conseguimos saber em quais clusters existentes das subdivisões de cada livro podem ser considerados como trechos do livro com termos "tristes/negativos" ou "felizes/positivos".

Ao colocarmos esses valores em ordem, traçarmos a sequencia do mapeamento das classificações dos conjuntos de palavras dos clusters de cada livro. Conseguimos ter um ideia em qual parte do livro a história é triste e em quais partes a história é feliz (ou trazem mais termos associados a positividade ou negatividade).

Sentimentalização (get_sentiments (“bing”)

Nos livros “The man in the Brown Suit" e "The Mysterious Affair at Styles“, nota-se a incidência maior de palavras positivas no começo das obras, mas que logo após os primeiros capítulos os termos negativos se sobressaem. Que é exatamente o que acontece em romances policiais, onde a contextualização é feita no começo e, logo depois, aparecem as descrições de enigmas, crimes e suspense de cada história. No fim, é observado um comportamento crescente de palavras qualificadas como "positivas”, o que nos leva a deduzir que, apesar de ser uma história complicada, o final seria "bom“, onde a ocorrência de muitas palavras "boas“ sugerem um final feliz.

Já no conto “The Plymouth Express“, notamos pouca oscilação entre termos positivos e negativos, e isso se deve à dinâmica do conto: logo no início, uma jovem rica é encontrada morta em um vagão de um trem, e suas joias foram roubadas. Para descrever a personagem e os envolvidos na história (pais, marido, etc.) são utilizados muitos termos positivos, e a mensagem principal é de que nem todo criminoso parece ser uma pessoa com atributos negativos. Uma análise fria dos dados pode induzir ao erro.

Sentimentalização (get_sentiments (“bing”)

Já nos demais três títulos "The Hunter’s Lodge Case", “The murder on the links" e "The Missing Will", notamos nos clusters uma ocorrência de variações de palavras positivas e negativos bem alternada no decorrer das obras, histórias com altos e baixos, como são os bons thrillers de Agatha Christie.

2

Distribuição de palavras

Nuvens de palavras mais utilizadas nas obras analisadas

Palavras mais recorrentes

Insights:

Poirot: trata-se de Hercule Poirot, protagonista principal dos livros de Agatha e um dos detetives mais famosos do universo Ficcional

Renauld: Personagem principal de The Murder on the Links e vítima no conto.

Inglethorp: sobrenome da vítima principal do livro “The Mysterious Affair at Styles

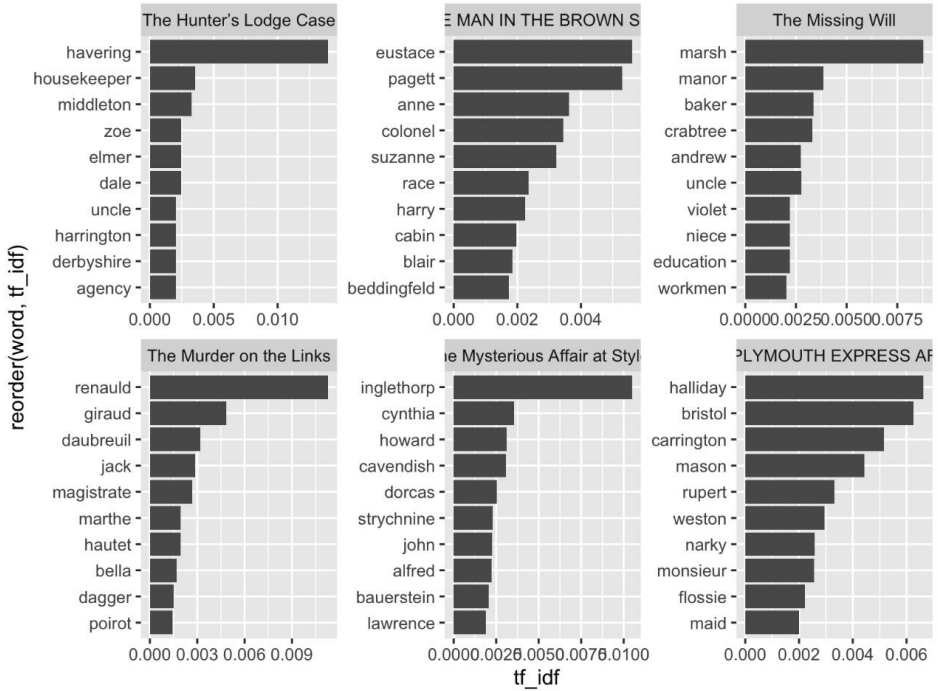
Palavras positivas x palavras negativas

Insights:

Termos em inglês classificados como positivos podem confundir, como “well”, “like” e “right”, uma vez que estão associados a outras expressões do idioma

Entre os termos em vermelho, negativos, destacam-se “crime”, “murder” (assassinato), doubt (dúvida” e death (morte).

Ocorrências de termos por obra



Insights:

Todos os termos mais citados em cada uma das obras são referentes aos nomes (no caso, sobrenomes) dos personagens principais de cada estória.

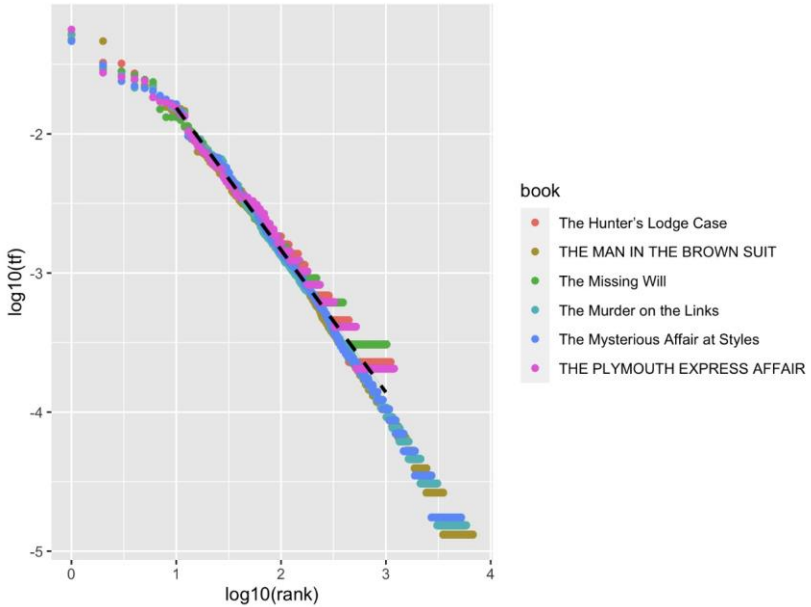
Palavras positivas x palavras negativas

Insights:

Termos em inglês classificados como positivos podem confundir, como “well”, “like” e “right”, uma vez que estão associados a outras expressões do idioma

Entre os termos em vermelho, negativos, destacam-se “crime”, “murder” (assassinato), doubt (dúvida” e death (morte).

Aplicação da Lei de Zipf



Relação de frequência das palavras em um ranking, considerando palavras mais utilizadas.

As mesmas palavras aparecem com maior frequência nos mesmos livros, entre outros motivos, por se tratar sempre de romances policiais e vindas da mesma autora.

O conto The Plymouth Express é o que mais foge dessa regra pela dinâmica da história, onde uma grande virada acontece quase no fim (um plot twist, como se diz popularmente).

“Ganhar uma guerra é tão desastroso quanto perdê-la.”

Agatha Christie

Obrigado!