Link

https://www.kaggle.com/headsortails/treemap-house-of-horror-spooky-eda-lda-features

Explicar o problema de negócio apresentado no Case

Será preciso analisar um trecho de textos e verificar qual a probabilidade do texto ser de Edgar Allan Poe, Mary Shelley ou HP Lovecraft.

A análise de texto é algo bem subjetivo, por isso será necessário encontrar padrões de linguagem nos textos, utilizando métodos estatísticos facilmente encontrados dentro do R, para então oferecer uma análise mais precisa e tentar mostrar as chances de um texto ser de um autor ou de outro.

Autor: Heads or Tails

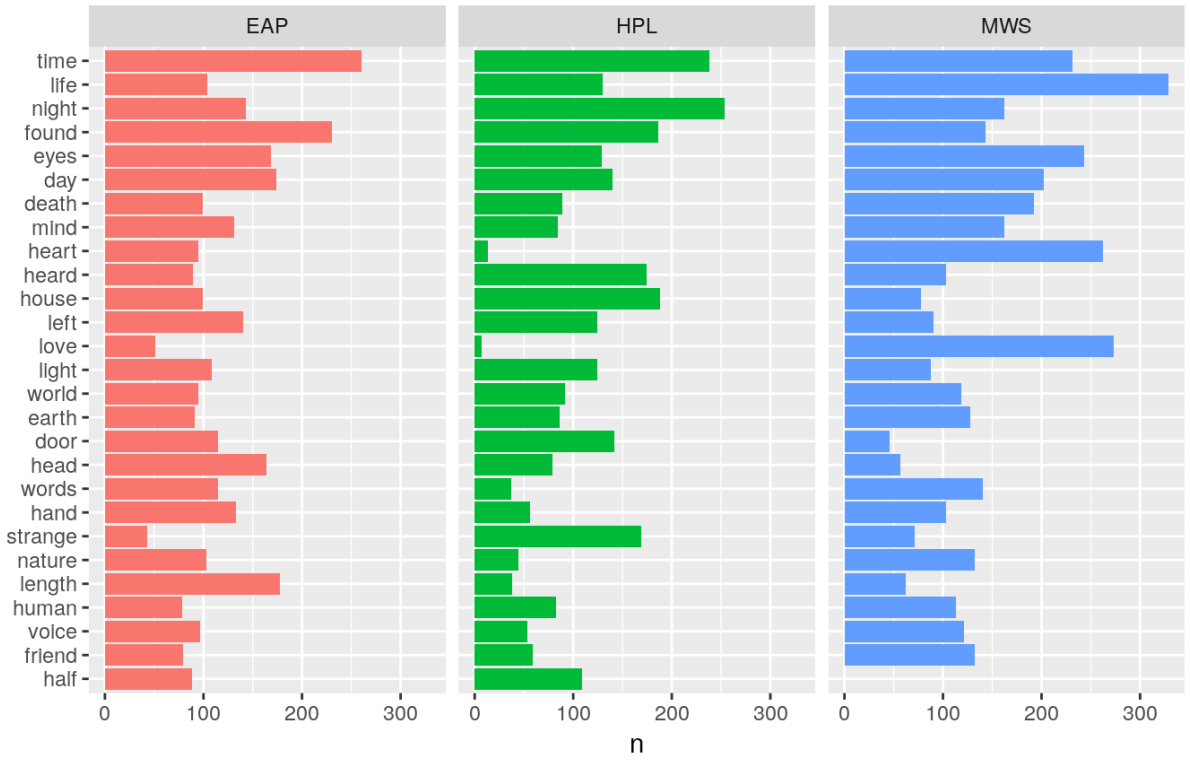
Mencionar como R foi usado e o que foi possível fazer, graças à sua adoção

O R foi usado nesse projeto por possuir várias funcionalidades que facilitam o tratamento e a análise dos textos. Através do algoritmo foi possível identificar:

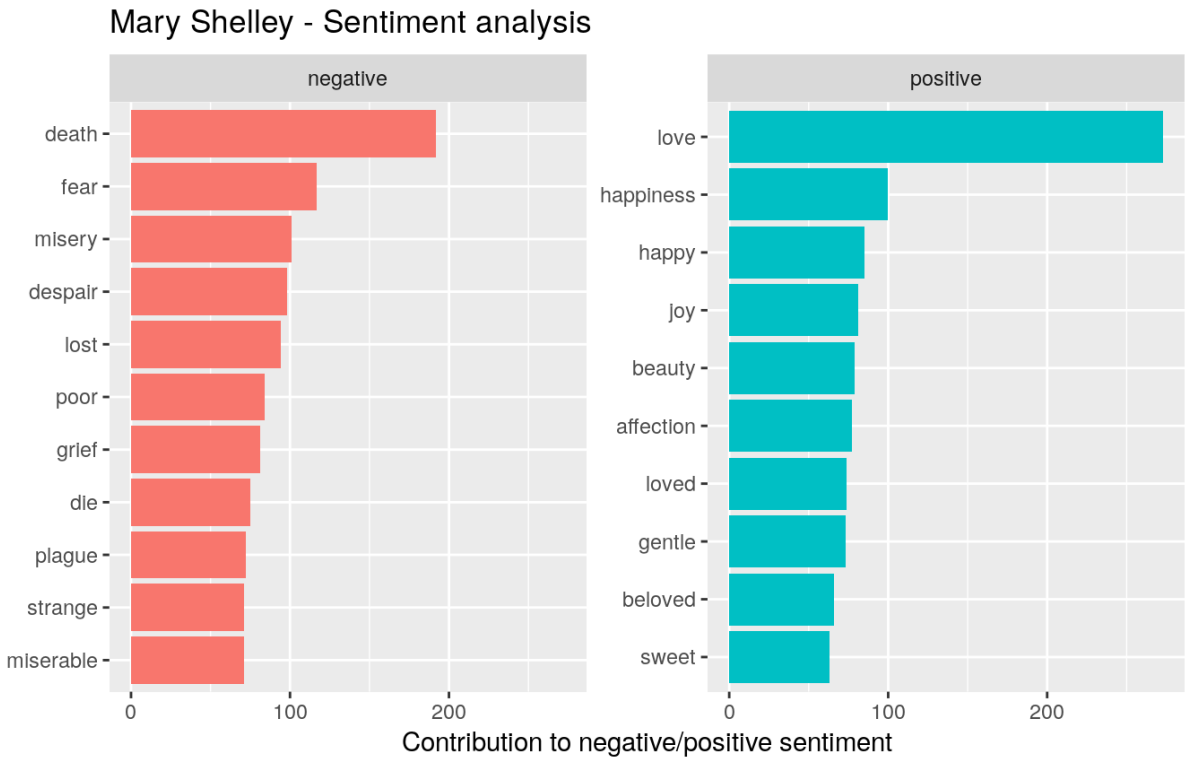
- Quais palavras cada autor mais usa
- Qual o comprimento das frases de cada autor
- Quem utiliza palavras mais longas
- A relação entre palavras
- Análise de sentimentos

Mostrar “insights” obtidos durante a análise

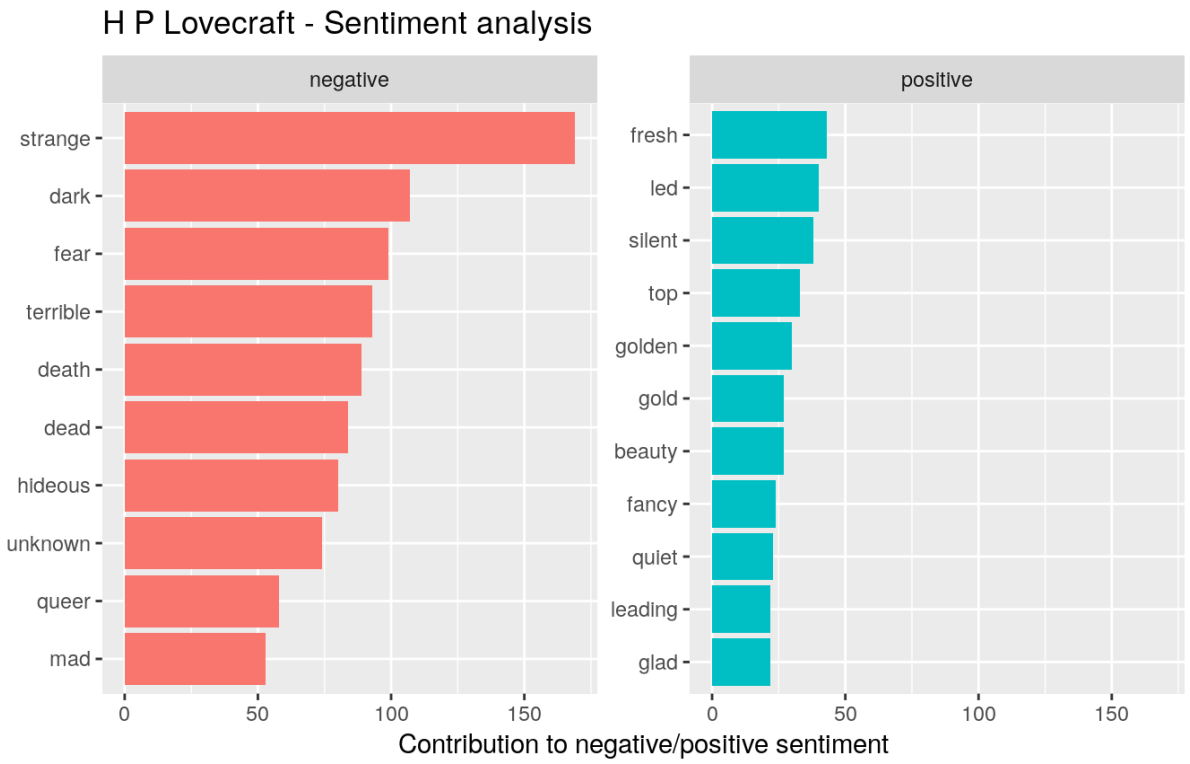
Foi possível observar quais autores são mais positivos ou mais negativos. As análises foram individuais, mas os gráficos mostram que a única mulher entre os autores pesquisados é mais positiva, também usando a palavra "amor" mais frequentemente. Além disso, todos os autores usam frequentemente a palavra "tempo".



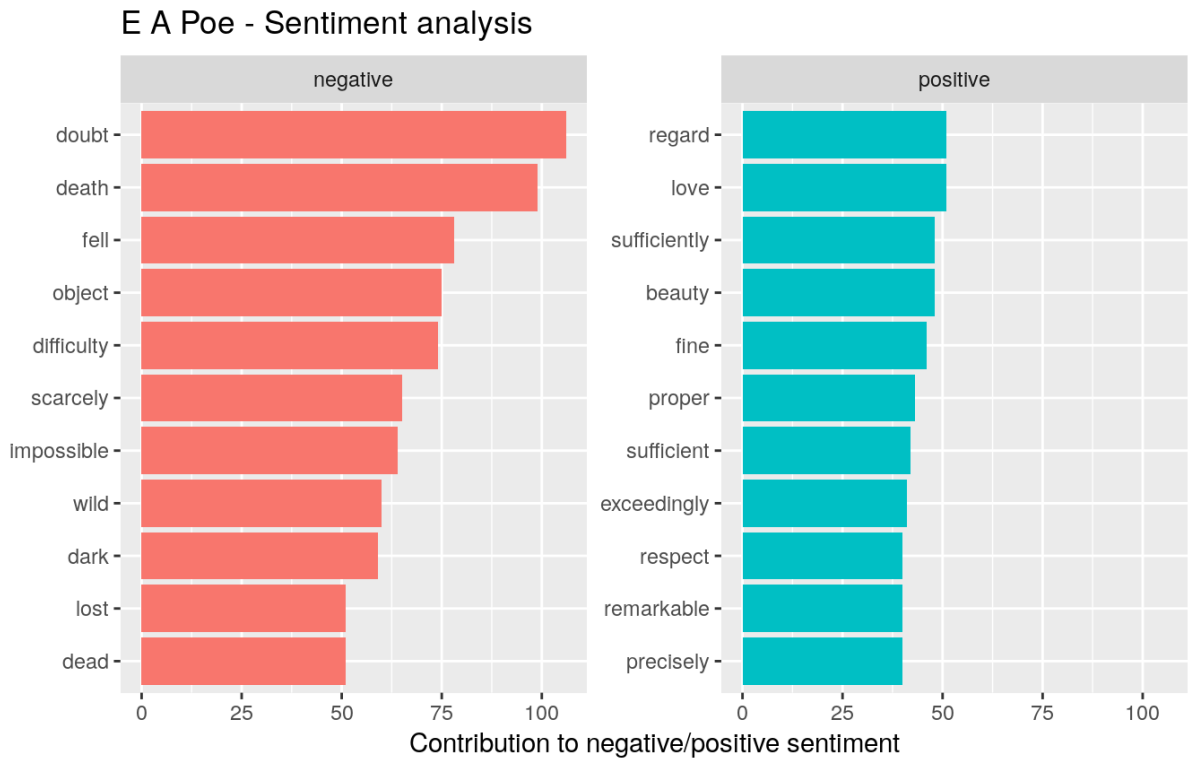
Os textos de Mary Shelley tem um equilíbrio entre palavras positivas e negativas, apesar a palavra "amor" ser usada mais frequentemente. “Morte” é claramente sua palavra negativa mais comum, talvez por conta de fatos que aconteceram em sua vida.



As histórias de H. P. Lovecraft não possuem um clima muito positivo, como podemos notar no gráfico. Não à menção às palavras "amor" ou "felicidade" no top 10 de palavras positivas. Sua palavra negativa mais frequente é "estranho", mas é o que poderíamos esperar do autor que revolucionou o gênero de terror inserindo elementos fantásticos, típicos das histórias de fantasia.

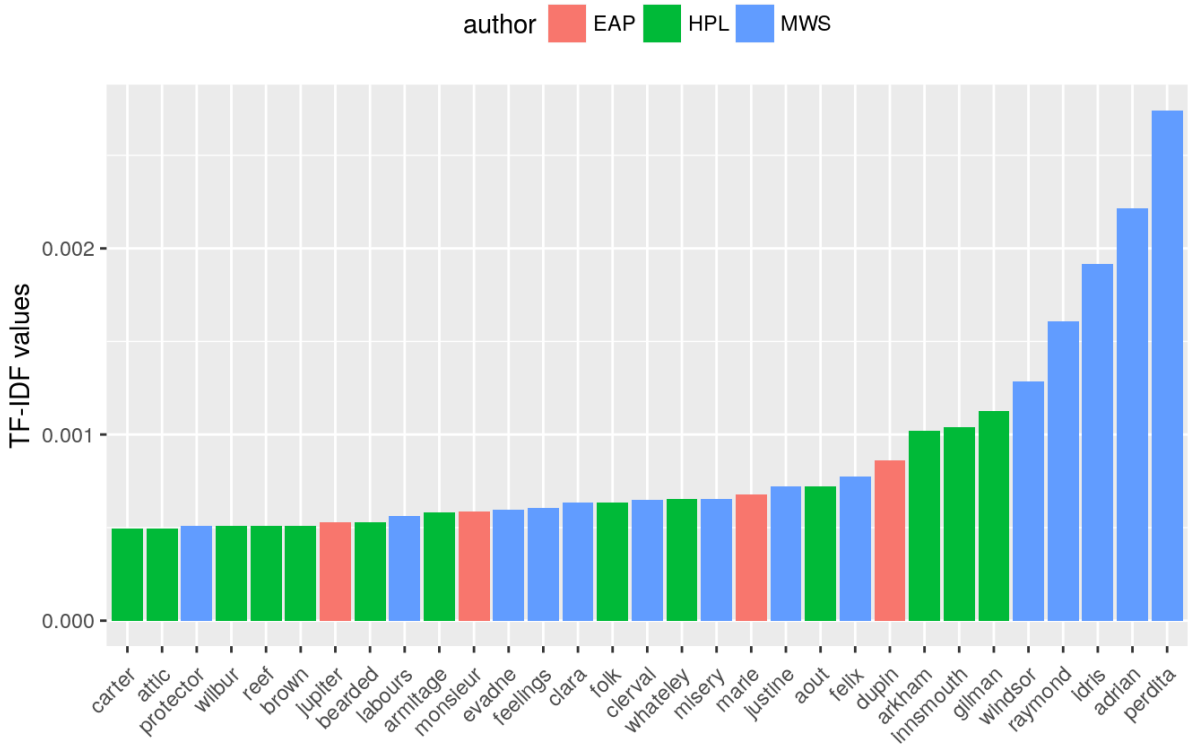


Edgar Allan Poe, apesar de ser um autor mais negativo, possui um equilíbrio maior em seus textos que H. P. Lovecraft. É interessante notar que a palavra negativa de Poe seja dúvida, sugerindo histórias de horror mais imaginativas.



Também foi possível identificar palavras mais características de cada autor, como por exemplo Mary Shelley que usava muito a palavra italiana "perdita" (perdido em italiano), talvez por conta da morte de seus dois filhos que aconteceram em Veneza e Roma, o que a fez mergulhar numa depressão e talvez tenha tentado expor seus sentimentos através de seus livros.

Os outros autores também possuem palavras características, mas não as usam com a frequência que Mary usa. Já Poe possui apenas 4 palavras mais frequentes.



Mostrar, em slide, como é uma instrução em R e seu resultado no relatório

Tratamento: Eliminar pontos / Deixar todas as letras em minúsculo / Identificar palavras de parada (e/próximo)

```
t1 <- train %>% unnest_tokens(word, text)

t1 <- t1 %>%

anti_join(stop_words, by = "word")
```

Palavras: Separar palavras / Organizar em nuvem de palavras / A nuvem mostra as palavras mais usadas de todos os autores

```
t1 %>%

count(word) %>%

with(wordcloud(word, n, max.words = 50, color = c("purple4", "red4", "black")))
```

Autores: Criamos uma nuvem de palavras mais populares de cada autor

```
t1 %>%

filter(author == "MWS") %>%

count(word) %>%

with(wordcloud(word, n, max.words = 30, color = "purple4"))
```

Dizer como um trabalho similar poderia ter sido feito em outra(s) ferramenta(s)

Várias funções usadas para para fazer a análise dos autores também podem ser encontradas em Python, já que ambas utilizam o Pandas para tratar dados e treinar modelos.

R	**Python**
count()	count()
anti_join()	anti_join()
unnest_tokens()	unnest_tokens()
filter()	filter()

Assim como no R, é fácil plotar dados na tela com o Python e todos os gráficos mostrados no estudo que foi realizado, também poderiam ser mostrados se o autor do trabalho tivesse usado essa linguagem de programação.