MBA USP ESALQ

Processamento de dados textuais: aplicação da biblioteca NLTK como ferramenta analítica

Vinícius Andrade Lopes
João Vitor Matos

INTRODUÇÃO

- Interpretação de dados textuais
- Compreensão de dados
  a. Diversos idiomas existentes no mundo;
  b. Categorias de livros;
  c. Linguagem informal;
  d. Caracteres especiais.
- Informações significantes
- Entendimento da regra de negócio
- Diversidade de dados

Desenvolver uma base de dados estruturada, sem erros de formatação e com os devidos parâmetros definidos, facilita a objetividade da visualização e análise das informações (Wickham, 2016).

INTRODUÇÃO

- Inteligência artificial
- Processamento de Linguagem Natural – PLN
    a. Utiliza preceitos linguísticos como classe de palavras para realizar as análises;
    b. Substantivos;
    c. Verbos;
    d. Adjetivos;
    e. Pronomes;
    f. Estruturas gramáticas.
- Desenvolvimento computacional de linguagens naturais
- Compreender conteúdos descritos por humanos
- Informações válidas

O objetivo deste trabalho é utilizar técnicas de processamento de linguagem natural nos “datasets” dos livros e filmes da saga Harry Potter com a finalidade de extrair informações relevantes, como frequência de palavras e padrões de linguagem, que possam auxiliar na análise comparativa entre as versões literárias e cinematográficas.

MATERIAIS E MÉTODOS

A estratégia de pesquisa utilizada no desenvolvimento deste projeto será exploratória, visando apresentar de forma clara e concisa, didática e prática, a implementação das técnicas de processamento de linguagem natural disponibilizadas pela biblioteca “Natural Language Toolkit” [NLTK].

- NLTK – “Natural Language Toolkit”
  a. Biblioteca;
  b. Técnicas e recursos de processamento estatístico de linguagem natural;
  c. Implementada em diversos softwares;
  d. Recursos léxicos;
  e. Análise de raciocínio semântico.
- Desenvolvimento computacional de linguagens naturais
- Compreender conteúdos descritos por humanos

MATERIAIS E MÉTODOS

COLETA DE DADOS

Para realizar a comparação entre os dados textuais da saga Harry Potter, foi necessário coletar as informações dos livros e das obras cinematográficas.

- Para a extração de informações dos livros, foram utilizados arquivos no formato .txt
- Para a extração de informações dos filmes, foram utilizado datasets contendo os diálogos dos personagens que foram adaptados para os filmes da saga

Ambas as bases de dados estão em inglês.

Nomes dos livros/filmes da saga:

1. The Philosopher's Stone
2. The Chamber of Secrets
3. The Prisoner of Azkaban
4. The Goblet of Fire
5. The Order of the Phoenix
6. The Half Blood Prince
7. The Deathly Hallows

MATERIAIS E MÉTODOS

PRÉ-PROCESSAMENTO

Para melhorar a análise dos dados textuais, ambas as bases de dados foram submetidas à mesma etapa de pré-processamento.

- Remoção de dados desnecessários
  - Rodapé, no caso dos livros;
  - “Stopwords”;
  - Caracteres especiais.
- Colocar todos os textos em minúsculo
- Aplicar as etapas do processamento de linguagem natural, disponibilizadas pela biblioteca NLTK

Estágios de análise do processamento de linguagem natural
Fonte: Adaptado de Dale et al. (2000)

RESULTADOS E DISCUSSÕES

COLETA DE DADOS

- Identificação de arquivos
- Google Colab + Google Drive
- “Python”

```
from google.colab import drive
drive.mount('/content/drive')

import os
_path_books_drive = '/content/drive/MyDrive/Harry_Potter/'
list_books_hp = sorted(os.listdir(_path_books_drive))
list_books_hp

>>>["Book 1 - The Philosopher's Stone.txt",
 'Book 2 - The Chamber of Secrets.txt',
 'Book 3 - The Prisoner of Azkaban.txt',
 'Book 4 - The Goblet of Fire.txt',
 'Book 5 - The Order of the Phoenix.txt',
 'Book 6 - The Half Blood Prince.txt',
 'Book 7 - The Deathly Hallows.txt']
```

Leitura de arquivos .txt armazenados no Google Drive
Fonte: Resultados originais da pesquisa

```
books_hp_no_processed[0]

>>> '/ \n\n\n\n\nTHE BOY WHO LIVED \n\nMr. and Mrs. Dursley, of number
 ↪  four, Privet Drive, \nwere proud to say that they were perfectly
 ↪  normal...'
```

“Print” de uma parte do livro Harry Potter e a Pedra Filosofal sem nenhuma aplicação de pré-processamento
Fonte: Resultados originais da pesquisa

MBA USP ESALQ

RESULTADOS E DISCUSSÕES

TRATATIVA DE DADOS

- Implementação de expressões regulares (ReGex)
- Limpeza dos dados
    a. “Stopwords”.
- Padronização dos dados
- Processamento de dados
- “FreqDist”: Mensura a frequência que cada palavra aparece dentro de um contexto

```python
def processing_text(str_text, _stopwords):
  if not str(str_text).isdigit():
    text_format = re.findall(r'\b[A-zÀ-úü]+\b', str_text.lower())

    if _stopwords:
      no_stopwords = [words for words in text_format if words not in _stopwords and not len(words) == 1]
    else:
      no_stopwords = text_format

  return(' '.join(no_stopwords))

#Função responsável por tratar as informações textuais dos livros.
books_hp_no_processed = []
books_hp_processed = []
_footer_removed = []

for book in list_books_hp:
  if book.endswith('.txt'):
    with open(_path_books_drive + book, encoding='utf8') as f:
      _book = ''
      for line in f:
        if not line.startswith("Page |"):
          _book += line
        else:
          _footer_removed.append(line)

      books_hp_no_processed.append(_book)
      books_hp_processed.append(processing_text(_book, english_stopwords))
```

```
1  books_hp_processed[0]
2
3  >>> 'boy lived mr mrs dursley number four privet drive proud say
   ↪  perfectly normal...'
```

“Print” de uma parte do livro Harry Potter e a Pedra Filosofal com a aplicação do pré-processamento

Fonte: Resultados originais da pesquisa

MBA USP ESALQ

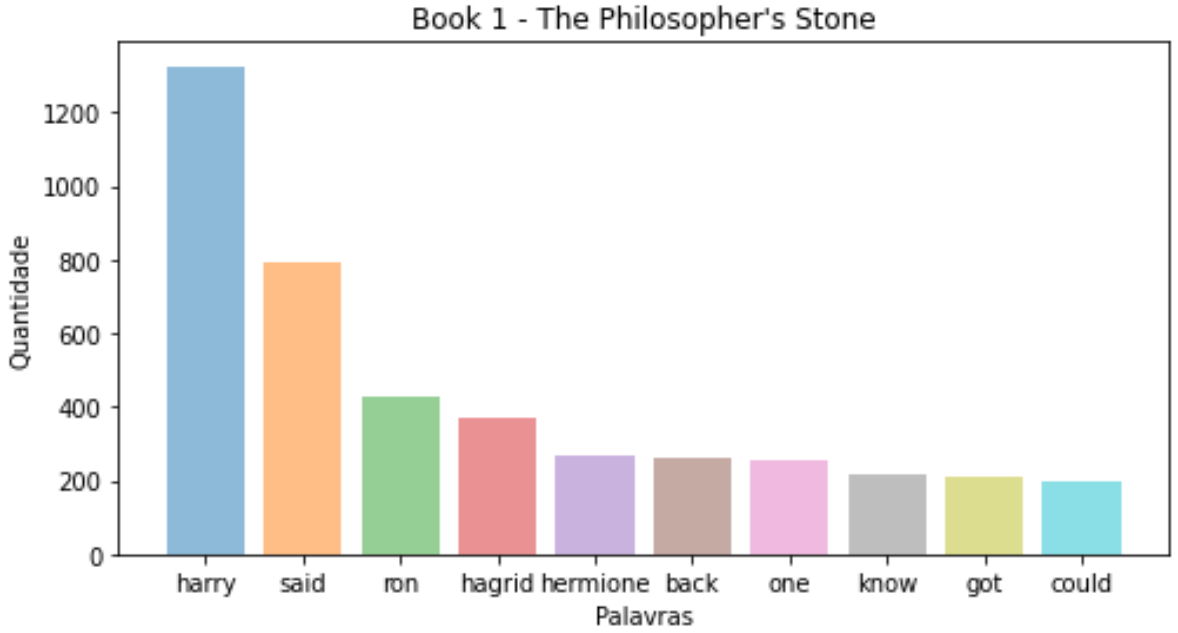
RESULTADOS E DISCUSSÕES

VISUALIZAÇÃO DE DADOS

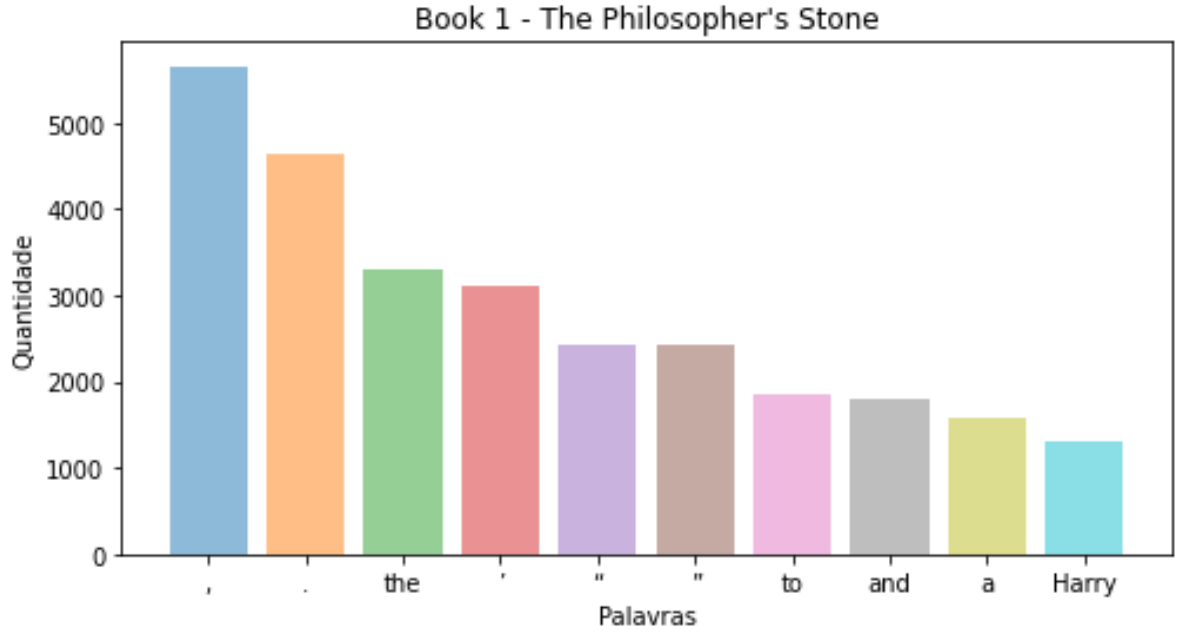
```
print(prob_hp_noprocess_philosopher_stone.most_common(5))
print(prob_hp_philosopher_stone.most_common(5))

>>>[(',', 5658), ('.', 4639), ('the', 3312), ('’', 3111), ('"', 2437)]

>>>[('harry', 1325), ('said', 794), ('ron', 429), ('hagrid', 370),
 ('hermione', 269)]
```

Output gerado após o processamento textual do livro Harry Potter e a Pedra Filosofal, limitado às cinco palavras que mais se repetem, onde é possível representar a diferença das informações não tratadas e a aplicação do pré-processamento.
Fonte: Resultados originais da pesquisa

COM A APLICAÇÃO DE PRÉ-PROCESSAMENTO



SEM A APLICAÇÃO DE PRÉ-PROCESSAMENTO



RESULTADOS E DISCUSSÕES

ANÁLISE DE DADOS

- Similaridade
  - Medida que avalia o grau de semelhança entre dois ou mais elementos linguísticos;
  - Categorização;
  - Classificação de sentimentos.
- Concordância
  - Permite analisar como as palavras são usadas em diferentes contextos;
  - Concordância com outras palavras em uma frase.
- Bigramas
  - Palavras consecutivas;
  - Sequência de duas palavras consecutivas em um texto;
  - Padrões de uso de palavras.

```
analytics_hp_chamber_secrets.similar('basilisk')

>>>said snake serpent
```

Output gerado após a aplicação da técnica de similaridade no livro Harry Potter e a Câmara Secreta, com os dados textuais pré-processados.
Fonte: Resultados originais da pesquisa

```
analytics_hp_chamber_secrets.concordance('basilisk')

>>>deadly basilisk known also king serpents snake

>>>deadly venomous fangs basilisk murderous stare fixed beam eye

>>>suffer instant death spiders flee basilisk mortal enemy basilisk

>>>monster chamber basilisk giant serpent hearing voice
```

Output gerado após a aplicação da técnica de concordância no livro Harry Potter e a Câmara Secreta, com os dados textuais pré-processados.
Fonte: Resultados originais da pesquisa

```
analytics_hp_chamber_secrets.collocations()

>>>professor mcgonagall; uncle vernon; mrs weasley; chamber secrets;
↪  fred george; headless nick; madam pomfrey; nearly headless; harry
↪  potter; gilderoy lockhart; moaning myrtle; aunt petunia; hospital
↪  wing; mrs norris; common room; sorting hat; professor sprout;
↪  crabbe goyle; ron hermione; great hall
```

Output gerado após a aplicação da técnica de bigramas no livro Harry Potter e a Câmara Secreta, com os dados textuais pré-processados.
Fonte: Resultados originais da pesquisa

MBA USP ESALQ

RESULTADOS E DISCUSSÕES

LIVROS VS FILMES

Nome do livro/filme	Filme	Livro	Porcentagem
The Philosopher's Stone	4801	39907	12,03%
The Chamber of Secrets	5429	45096	12,04%
The Prisoner of Azkaban	4752	56366	8,43%
The Goblet of Fire	3943	98844	3,89%
The Order of the Phoenix	4595	132023	3,48%
The Half Blood Prince	5383	85961	6,23%
The Deathly Hallows	7773	100537	7,73%

Tabela 1. Quantidade de palavras em cada fonte de dados, com a aplicação da etapa de pré-processamento, e percentual de redução na frequência de palavras dos filmes se comparado com os livros.
Fonte: Resultados originais da pesquisa

Nome do livro/filme	Filme	Livro	Porcentagem
The Philosopher's Stone	12273	101227	12,12%
The Chamber of Secrets	13243	111538	11,87%
The Prisoner of Azkaban	12027	142306	8,45%
The Goblet of Fire	9360	247523	3,78%
The Order of the Phoenix	11501	332518	3,46%
The Half Blood Prince	13849	220235	6,29%
The Deathly Hallows	20186	255569	7,90%

Tabela 2. Quantidade de palavras em cada fonte de dados, sem a aplicação da etapa de pré-processamento, e percentual de redução na frequência de palavras dos filmes se comparado com os livros.
Fonte: Resultados originais da pesquisa

RESULTADOS E DISCUSSÕES

LIVROS VS FILMES

Nome do filme	Pré-processado	Texto Original	Porcentagem
The Philosopher's Stone	4801	12273	38,12%
The Chamber of Secrets	5429	13243	41,00%
The Prisoner of Azkaban	4752	12027	39,51%
The Goblet of Fire	3943	9360	42,13%
The Order of the Phoenix	4595	11501	39,95%
The Half Blood Prince	5383	13849	38,87%
The Deathly Hallows	7773	20186	38,51%
TOTAL:	36676	92439	39,68%

Tabela 3. Percentual de “stopwords” e caracteres especiais removidos na base de dados relacionadas aos filmes, após a etapa de pré-processamento dos dados.
Fonte: Resultados originais da pesquisa

Nome do livro	Pré-processado	Texto Original	Porcentagem
The Philosopher's Stone	39907	101227	39,42%
The Chamber of Secrets	45096	111538	40,43%
The Prisoner of Azkaban	56366	142306	39,61%
The Goblet of Fire	98844	247523	39,93%
The Order of the Phoenix	132023	332518	39,70%
The Half Blood Prince	85961	220235	39,03%
The Deathly Hallows	100537	255569	39,34%
TOTAL:	558734	1410916	39,60%

Tabela 4. Percentual de “stopwords” e caracteres especiais removidos na base de dados relacionadas aos livros, após a etapa de pré-processamento dos dados.
Fonte: Resultados originais da pesquisa

RESULTADOS E DISCUSSÕES

EXPLICAÇÃO

LIVROS:

- Descrição detalhada de lugares
- Descrição de personagens
- Contextualizar uma ação ou reação de cenas específicas
- Descrição de animais místicos
- Funcionamento de uma magia
- Emoções nos diálogos

FILMES:

- Esses cenários são substituídos por imagens, não sendo necessário descrever, de forma textual, suas características

Outra razão que impacta diretamente na redução da frequência de palavras está relacionada aos diálogos nos filmes de Harry Potter, que são mais sucintos quando comparados aos dos livros (Burkhardt e Gauvain, 2013).

“O compromisso de Max em preservar a integridade de meus livros é importante para mim, e estou ansiosa para fazer parte desta nova adaptação que permitirá um grau de profundidade e detalhes apenas proporcionados por uma longa série de televisão” (Rowling, 2023).

CONCLUSÃO

- As ferramentas disponibilizadas pela biblioteca NLTK são robustas o suficiente para gerar análises avançadas. Porém, é necessário conhecimento para implementar essas funcionalidades;
- Entender toda a estrutura/formatação dos dados que serão analisados é de suma importância para a etapa de pré-processamento das informações;
- Um fluxo robusto de extração, transformação e armazenamento de dados será decisivo para definição das etapas analíticas;
- Foi possível constatar o funcionamento analítico da biblioteca em mais de um idioma;
- Com a aplicação das técnicas de similaridade, concordância e bigramas, foi possível entender o significado de algumas palavras, mesmo para quem não teve contato com o universo Harry Potter.

REFERÊNCIAS

- Burkhardt, J. M., & Gauvain, M. C. (2013). Comparing the Books to the Movies: A Study of Harry Potter and the Philosopher's Stone. Children's Literature in Education, 44(3), 257-273;
- Dale, Robert; MOISL, Hermann; SOMERS, Harold (Ed.). Handbook of natural language processing. CRC press, 2000;
- Desmond, John, and Peter Hawkes. Adaptation: Studying film and literature. McGraw-Hill Humanities Social, 2006;
- Max Orders First Ever “Harry Potter” Television Series. Warner Bros. Discovery Inc, 2023. Disponível em: https://press.wbd.com/us/media-release/max/max-orders-first-ever-harry-potter-television-series. Acesso em: 21/05/2023.
- Wickham, Hadley. "Data analysis." ggplot2. Springer, Cham, 2016. 189-201.

MBA USP ESALQ

REPOSITÓRIO

GitHub: https://github.com/Wiryco

MBA USP ESALQ