UNIVERSIDADE
FEDERAL DO CEARÁ

AV1 - NLP

Arielly Gonçalves

Objetivo

Analisar como diferentes escolhas de **pré-processamento** e **extração de atributos** influenciam o desempenho de um modelo de **classificação de texto.**

Dados Utilizados

B2W um conjunto de avaliações de produtos com mais de 130.000 avaliações de usuários.

A B2W Digital, uma das mais proeminentes empresas de comércio eletrônico latino-americanas.

Texto/Comentários:

- **íntegra**
- Processado

Tem duas características-alvo:

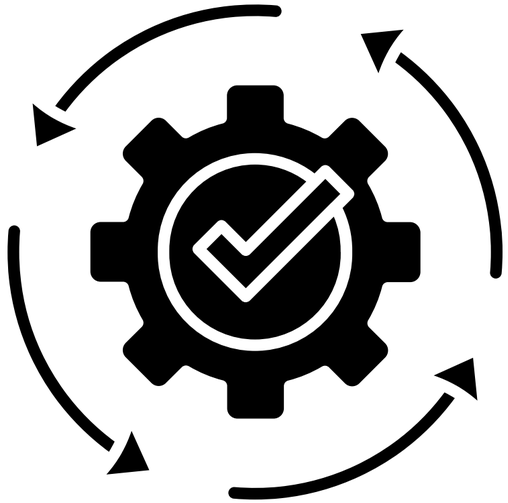
- **polaridade**
- classificação

Etapas de desenvolvimento

UNIVERSIDADE FEDERAL DO CEARÁ

Etapa 1:

preprocessing.py



Etapa 3:

classification.py

Etapa 2:

atribute_extraction.py

Etapa 1: preprocessing.py

Limpeza básica

Normalização com Enelvo

Remoção de dígitos, menções e links. (regex)

Tokenização

Separar o texto em tokens (spaCy)

Stopwords

Remoção de Stopwords(NLTK)

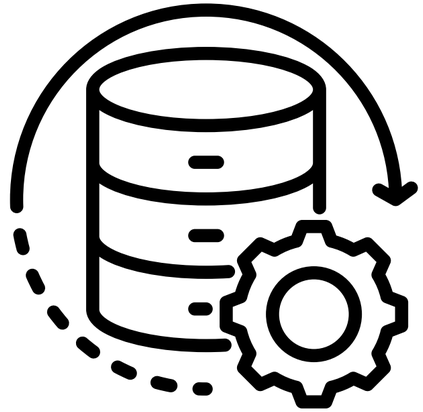
Lematização

Palavras no inifinitivo e singular (spaCy)

Stemming

Radical das palavras(NLTK)

Aplicar as técnicas de extração de atributos



Etapa 2:

atribute_extraction.py

Universidade Federal do Ceará

Após limpeza e lemmatização, passa para as representações vetoriais:

Análise estatística

- nº de tokens
- nº de tokens unicos
- media de tamanho
- nº de caracteres
- nº simbolos

CountVectorizer (Bag-of-Words)

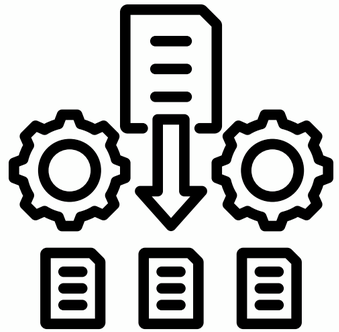
Transforma texto em vetores de contagem.

TF-IDF

Destaca palavras mais relevantes do documento.

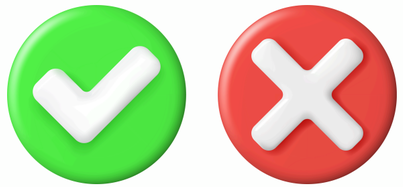
Word2Vec

Gera embeddings densos usando contexto semântico.



Etapa 3:

classification.py



UNIVERSIDADE FEDERAL DO CEARÁ

Classificar e Avaliar

MultinomialNB

Classificador probabilístico usado principalmente textos vetorizados em contagens.

Treino e Teste

test_size=0.2, random_state=42

Comparações

Desempenho com e sem pré-processamento

Diferentes técnicas de extração de atributos

lematização versus stemming.

Métrica

Acurácia: A acurácia mede a proporção de previsões corretas.

Funciona muito bem em problemas binários

Resultados

UNIVERSIDADE FEDERAL DO CEARÁ

Remoção todas as linhas onde a coluna polarity está vazia (NaN).

Seleção de uma amostra do dataset (7.000 linhas aleatórias)

Texto na Íntegra + Polaridade

Resultados

item a

Técnica	Preprocessamento	Acuracia
Analise Estatística	Sem	0.735000
CountVectorizer	Sem	0.902857
TF-IDF	Sem	0.905000
Word2Vec	Sem	0.774286
Analise Estatística	Com	0.730000
CountVectorizer	Com	0.898571
TF-IDF	Com	0.892857
Word2Vec	Com	0.748571

maior valor da tabela.

No geral, o Pré-processamento não melhorou os resultados.
Isso por que:

- Remover stopwords pode tirar palavras importantes de sentimento
- Lematização pode reduzir palavras que carregam carga emocional

Resultados

item b

	Técnica	Preprocessamento	Acuracia
4	Analise Estatística	Com	0.730000
5	CountVectorizer	Com	0.898571
6	TF-IDF	Com	0.892857
7	Word2Vec	Com	0.748571

maior valor da tabela.

Esses resultados são esperados, pois:

- Word2Vec gera vetores contínuos densos
- MultinomialNB não trabalha bem com valores contínuos
- NB é ideal para vetores de contagem, por isso BoW funciona melhor

Resultados

item c

```
Melhor técnica encontrada no item B: CountVectorizer
```

```
Melhor técnica encontrada no item B: CountVectorizer

--- Resultados item C ---
Comparação entre LEMMA x STEMMING usando a melhor técnica:

  Técnica Método  Acuracia
0     bow  lemma  0.898571
1     bow   stem  0.892857
```

ligeiramente maior

Lemmatização preserva mais significado
Bag-of-Words depende muito da qualidade dos tokens
No português, stemming é mais agressivo

Conclusões

UNIVERSIDADE FEDERAL DO CEARÁ

A acurácia sem pré-processar foi ligeiramente superior. Isso mostra que remover elementos do texto pode retirar informações relevantes ao classificador.

A técnica Bag-of-Words apresentou o melhor desempenho geral

Em relação ao CountVectorizer a lematização produziu resultados superiores ao stemming.

Word2Vec foi a técnica menos eficaz neste cenário

Trabalhos Futuros:

Universidade Federal do Ceará

Experimentar classificadores mais robustos

Aplicar outras métricas

Explorar métodos modernos de representação de texto

Avaliar outras/estratégias de pré-processamento

Obrigado!

Arielly Gonçalves