Novelty Detection: Identificação de Tópicos Emergentes em Corpora de Noticias

- Martín Ávila
- RA:2274183
- martinete.avila@gmail.com,
- Graduando: Eng. Computação, UTFPR

Informações Iniciais

Dados:
Dataset: LREC2018 corpus

Pasta source: Contém 3 notícias iniciais usados como referência.
Pasta target: Contém notícias gos adicionais classificados como "novos" ou "não novos" em relação às notícias da pasta source. Essas notícias são analisadas para determinar se trazem informações novas (novidade) em comparação com as notícias base.

Total : 6104 rows × 16 columns

	category	event_id	news_id	content	is_source	DOP	publisher	title	eventid	eventname	topic	sentence	words
0	SPORTS	SPTE001	SPTE001SRC003	Dangal: Baba Ramdev to wrestle it out with Rus...	True	00/00/0000	www.indiatvnews.com	Dangal-Baba Ramdev to wrestle it out with Russ...	SPTE001	Baba Ramdev wrestling challenge	SPORTS	13	269

Objetivos

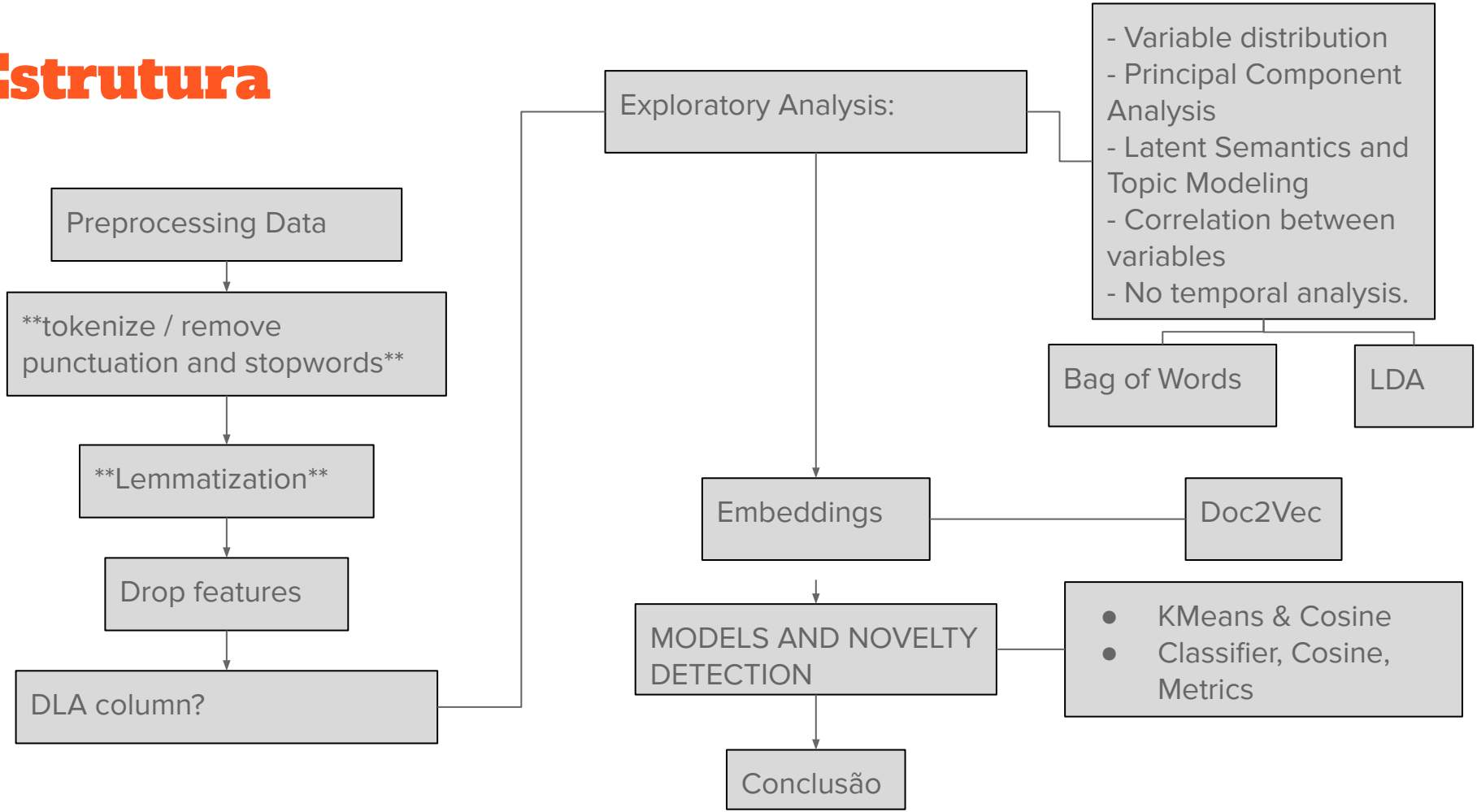
Deteção de Novidades
Identificar artigos como novos ou não novos, comparando as informações dos notícias alvo com os originais.

Desenvolvimento do Sistema
Construir um modelo capaz de distinguir entre informações novas e não novas usando dados da pasta *source* como referência para os artigos da pasta *target*.

Análise de Tópicos
Explorar diferenças temáticas entre artigos de novidade e não novidade.

Avaliação de Precisão
Validar o desempenho do modelo comparando seus resultados com o atributo Document Level Annotation (DLA) utilizando a pontuação F1.

Estrutura


Preprocessing Data
tokenize / remove punctuation and stopwords
Lemmatization
Drop features
DLA column?
Exploratory Analysis:
- Variable distribution
- Principal Component Analysis
- Latent Semantics and Topic Modeling
- Correlation between variables
- No temporal analysis.
Bag of Words
LDA
Embeddings
Doc2Vec
MODELS AND NOVELTY DETECTION
KMeans & Cosine
Classifier, Cosine, Metrics
Conclusão

Preprocessing Data

tokenize / remove punctuation and stopwords

First 10 stopwords: ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]

['Dangal', 'Baba', 'Ramdev', 'to', 'wrestle', 'it', 'out', 'with', 'Russian', 'Olympic' ...]

Lemmatization

Dangal: Baba Ramdev to wrestle it out with Russian

CORPUS CLEANED: dangal baba ramdev wrestle russian

Drop features

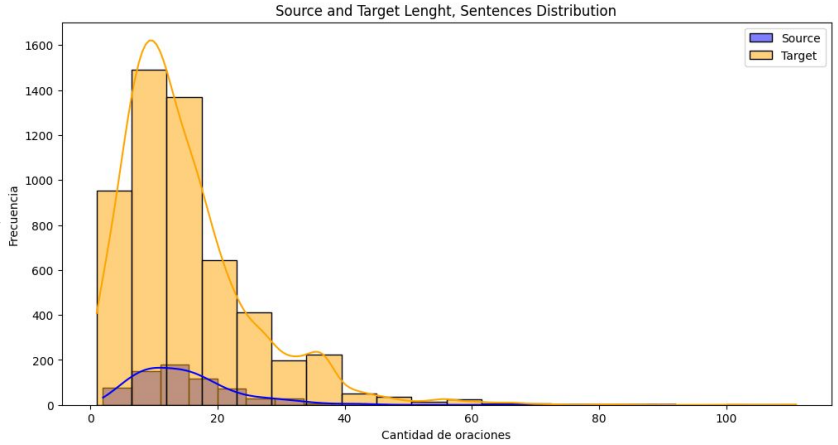
['event_id', 'news_id', 'eventid', 'sourceid', 'title', 'DOP', 'eventname' , 'SLNS'],

DLA column?

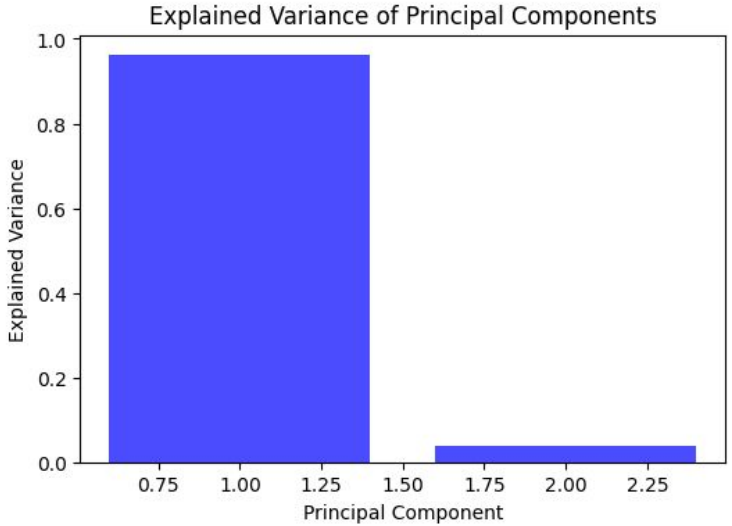
```
Percentage of each category in the 'DLA' column:
Non-Novel: 43.61%
Novel: 42.39%
non-novel: 7.95%
novel: 6.02%
Non-Novelvel: 0.02%
NovNon-Novelel: 0.02%
```

Exploratory Analysis:

- Variable distribution



- Principal Component Analysis



Palavras

Frases

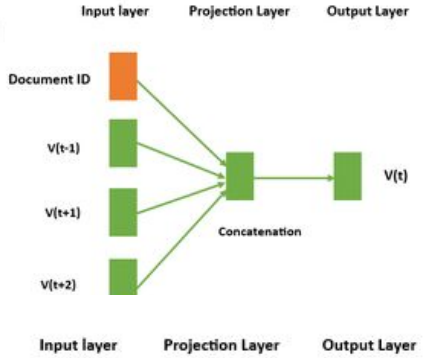
Trabalhos relacionados

- Latent Semantics and Topic Modeling

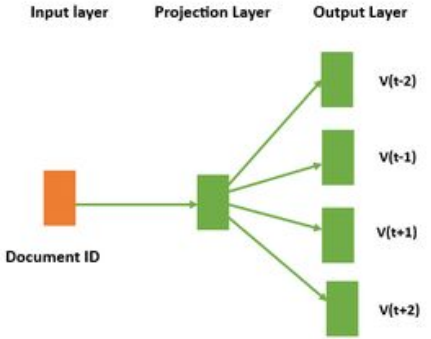
LDA
(Alocação Latente de Dirichlet)
num_topics=11

Embeddings
(Doc2Vec)

Distributed Memory (DM)



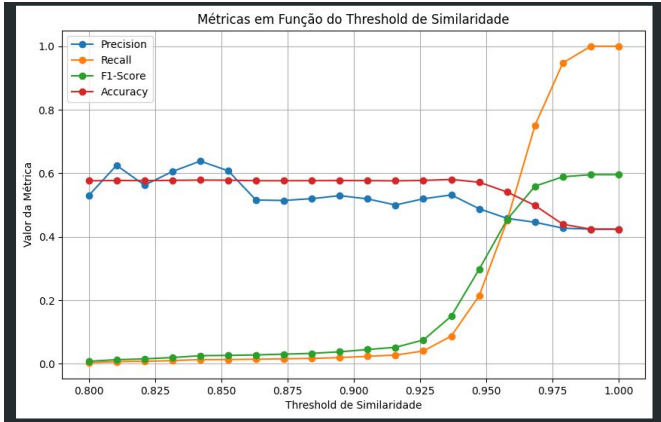
Distributed Bag of Words (DBOW)



- No temporal data. → 😓

Modelos e Resultados

- Clusters
  - KMeans
  - cosine_similarity



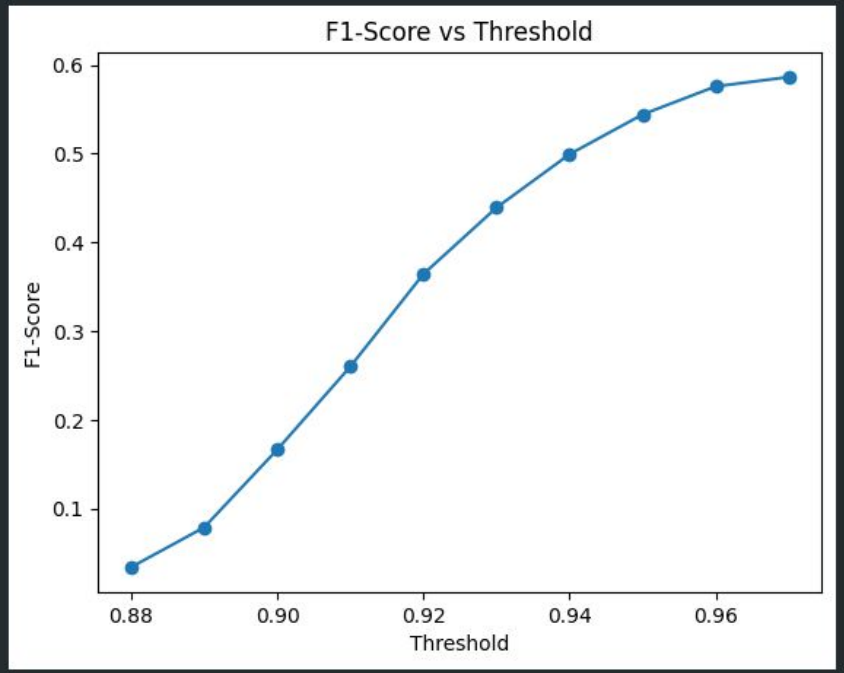

```
Melhor Threshold: 0.99
Métricas no Melhor Threshold: {'threshold': 0.9894736842105263, 'precision': 0.42391904323827045, 'recall': 1.0, 'f1_score': 0.595425765602791, 'accuracy': 0.42391904323827045}
```

- Topics Classifier XGBoost
- Novelty Detection (cosine_similarity)
- Metrics

```
Best Parameters: {'colsample_bytree': 0.7, 'learning_rate': 0.2, 'max_depth': 4, 'n_estimators': 300, 'subsample': 0.8}
XGBoost Classifier Performance (Best Model):
              precision    recall  f1-score   support

           0       0.50      0.05      0.09       503
           1       0.44      0.37      0.41       655
           2       0.47      0.27      0.35       466
           3       0.41      0.09      0.15       411
           4       0.22      0.02      0.04       624
           5       0.26      0.57      0.36       337
           6       0.41      0.92      0.57      1350
           7       0.71      0.12      0.20       277
           8       0.20      0.01      0.02        90
           9       0.67      0.56      0.61       722

    accuracy                           0.43      5435
   macro avg       0.43      0.30      0.28      5435
weighted avg       0.44      0.43      0.36      5435

Accuracy: 0.4269, F1-Score: 0.3581
```

Somente coseno

```
Average Cosine Similarity: 0.928593922649281
Best Threshold: 0.9700000000000001
Best F1-Score: 0.5862523044508823
Classification Report:
              precision    recall  f1-score   support

           0       0.46      0.02      0.04      3131
           1       0.42      0.97      0.59      2304

    accuracy                           0.42      5435
   macro avg       0.44      0.49      0.31      5435
weighted avg       0.44      0.42      0.27      5435
```



Conclusões e Limitações

1. **Resultados e Limitações:**
   - Apesar de alcançar até 0.6 de F1-Score, os resultados ainda estão abaixo das expectativas.
   - O desempenho limitado esteja relacionado à qualidade dos embeddings, indicando a necessidade de explorar técnicas mais avançadas, como o uso de BERT, para gerar representações mais precisas.

2. **Sistema de Classificação em 2 Etapas:**
   - A ideia de um sistema de classificação em duas etapas mostrou-se promissora, mas sua eficácia depende diretamente da qualidade das representações iniciais (embeddings) e da modelagem de tópicos.
   - Ajustes na abordagem LDA e nos embeddings são essenciais para validar essa solução.

3. **Objetivos Parcialmente Cumpridos:**
   - O objetivo de Detecção de Novidades foi parcialmente alcançado, dado que o sistema conseguiu identificar padrões, mas ainda precisa de refinamento para atingir maior precisão.
   - O desenvolvimento do sistema foi concluído, mas ajustes adicionais são necessários para que ele cumpra a tarefa principal com eficiência.
   - A análise de Bag of Words evidenciou diferenças na frequência de palavras por tópico, o que ajudou a calibrar o Doc2Vec e entender melhor a dinâmica dos dados.

Limitações/trabalhos futuros

1. Entendimento do dataset
2. Ajustar LDA e Embeddings
3. Tempo de Processamento e Tests
4. Sistema de classificação

GitLab - > https://gitlab.com/mab0205/introcd2-novelty-detection

Ref

Doc2Vec:
LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. *In*: Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, China, 2014. p. 1188-1196. Disponível em: https://arxiv.org/abs/1405.4053. Acesso em: 3 dez. 2024.

LDA (Latent Dirichlet Allocation):
BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, v. 3, p. 993-1022, 2003. Disponível em: https://jmlr.org/papers/volume3/blei03a/blei03a.pdf. Acesso em: 3 dez. 2024.

KMeans:
MACQUEEN, James. Some methods for classification and analysis of multivariate observations. *In*: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967. p. 281-297. Disponível em: https://projecteuclid.org/euclid.bsmsp/1200512992. Acesso em: 3 dez. 2024.

XGBoost:
CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A scalable tree boosting system. *In*: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016. p. 785-794. Disponível em: https://arxiv.org/abs/1603.02754. Acesso em: 3 dez. 2024.