A Comparative Study of Topic Modeling Techniques

Pedro Castro

Introduction

- Topic Modeling is a crucial technique in natural language processing (NLP) used to uncover latent themes or topics within a collection of documents.
- Exploratory dataset focused in comparing topic modeling pipelines with Gensim, Scikit-learn, PyTorch, and BERTopic.

Dataset Description

- Utilized the all-the-news dataset found in Kaggle which comprises a directory of over 10,000 news articles sourced from various domains, including politics, sports, entertainment, and economics spanning several years.
- All models were tested with the standard LDA topic modeling implementation with slight differences in each implementations. Included BERTopic model at the end as a contrast to the other implementations.

Gensim Approach

Strengths:

- Gensim’s preprocessing simplified the data handling.
- Simplicity of Gensim procedures (the dictionary and the LDA models built-in) facilitated topic modeling pipelines massively.
- Had a more diverse set of topics out of the other implementations.
- Easily testable for coherence

Weaknesses:

- Was not as flexible as other implementations attempted for LDA when it came to customizing the code outside its libraries.

Scikit-Learn Approach

Strengths:

- Allowed more in-depth utilization of TF-IDF vectorization and cosine similarity for a more robust topic modeling
- Utilized parallel processing to improve efficiency of preprocessing tasks

Weaknesses:

- Topics seemed slightly more focused on specific domains compared to Gensim implementation, which suggests limited versatility of the model.

PyTorch Approach

Strengths:

- Allowed for an in-depth customization of the dataset creation and training.
- Utilized parallel processing to improve efficiency of preprocessing tasks

Weaknesses:

- Coherence similarity was more difficult to compute than for other LDA models.

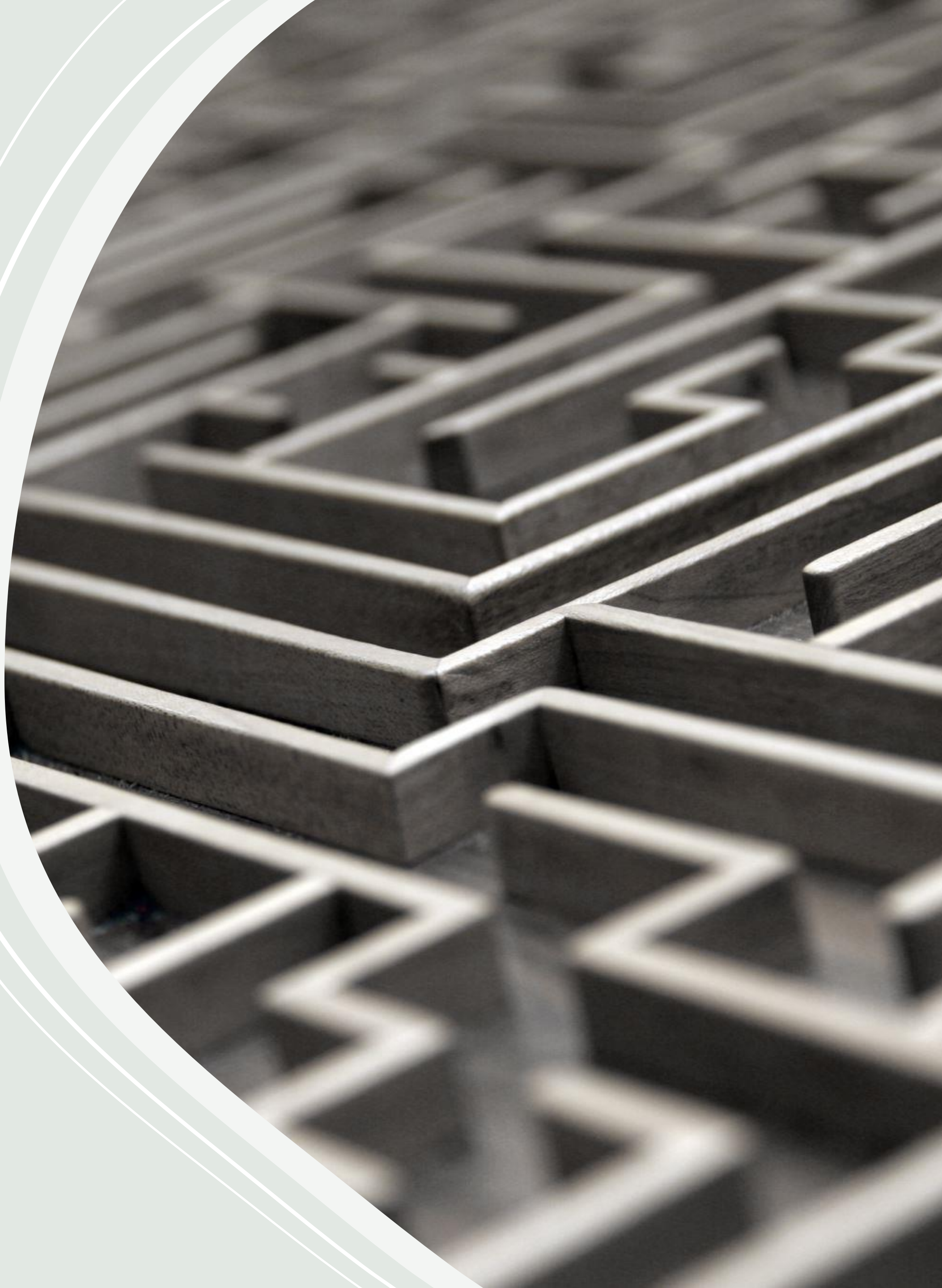
BERTopic Approach

Strengths:

- Instead of using statistical patterns and matrix factorization, this model utilizes BERT embeddings which give a contextual understanding of the data.

Weaknesses:

- Computationally very expensive to run and scalability is increasingly difficult to achieve as a cause of its computationally intensive procedures.



Results

GENSIM	SCIKIT-LEARN	PYTORCH	BERTOPIC
• Topic 0: International Relations • Topic 1: Media Influence • Topic 2: Sports • Topic 3: Economy and Policy • Topic 4: Urban Development • Topic 5: Law and Crime • Topic 6: Political Campaigns • Topic 7: Health and Medicine • Topic 8: Military and Conflict • Topic 9: General Life and Society	• Topic 0: General Life • Topic 1: Business and Finance • Topic 2: Law Enforcement and Crime • Topic 3: Health and Research • Topic 4: Politics - Elections • Topic 5: International Trade and Politics • Topic 6: Politics - Legislation • Topic 7: Law and Education • Topic 8: International Conflict and Military • Topic 9: Politics - Government and Intelligence	• Topic 0: Diseases • Topic 1: Brazilian Politics • Topic 2: Sports • Topic 3: Entertainment • Topic 4: Sports Doping • Topic 5: Economy • Topic 6: Crime • Topic 7: Sports Events • Topic 8: Politics • Topic 9: Economy & Business	Topic Count ... Representation Representative_Docs 0 -1 20599 ... [and, the, of, in, to, was, he, that, it, is] [Policing in America today is a rib dinner pai... 1 0 663 ... [obamacare, insurance, health, care, repeal, b... [It looks like the beginning of the end for Ob... 2 1 453 ... [clinton, hillary, she, her, mrs, campaign, tr... [Should she win the presidency, Hillary Clinto... 3 2 444 ... [sanders, bernie, clinton, democratic, says, h... [Hillary Clinton and Bernie Sanders will debat... 4 3 406 ... [gun, guns, awrhawkins, hawkins, awr, backgrou... [On January 11 The Atlantic pointed out that P... .. .. 606 605 10 ... [debt, cbo, trillion, ceiling, 2046, spending,... [The Congressional Budget Office's (CBO) dire ... 607 606 10 ... [bacteria, gut, microbiota, microbiome, flossi... [If you're making resolutions for a healthier ... 608 607 10 ... [valeant, ackman, lachs, pershing, allergan, r... [SAN FRANCISCO — Kirsten Green had only dab... 609 608 10 ... [xi, china, chinese, li, hong, kong, beijing, ... [President Xi Jinping of China plans to stride... 610 609 10 ... [chemical, syria, syrian, sheikhoun, strike, w... [(CNN) Survivors of a deadly airstrike in Syr...

Conclusion and BerTopic

When comparing these four models, these can be seen as the clear takeaways:

- Gensim offerse a straightforward and efficient implementation for LDA with robust preprocessing capabilities. While it is not as flexible as the other two implementations, it does indeed work well for simple topic modeling pipelines.
- Allows a comprehensive implementation of LDA with TF-IDF vectorization and cosine similarity coherence calculation.
- The most flexible and scalable library for LDA because of its implementation of tensors and its dataset creation and DataLoader usage. However, the coherence implementation solely on this library was shown to be overly complicated in comparison to the other two.
- When it comes to text understanding and NLP utilization, BERTopic is the most comprehensive way of doing topic modeling. However, its computational requirements are to be considered.

Sources

- https://zilliz.com/learn/explore-bertopic-novel-neural-topic-modeling-technique
- https://www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/
- https://umu.diva-portal.org/smash/get/diva2:1763637/FULLTEXT01.pdf
- https://www.youtube.com/watch?v=sZcGuYHWN_w