Topic Modeling on Summarized News Articles

Giovanni Noè - 881765

Francesco Volpi Ghirardini - 933735

Text Mining and Search

Introduction

Aim of the project

This project investigates the interaction between topic modeling & extractive text summarization

Dataset

CNN/Daily Mail, news articles (subset of 25,000 documents from each)

Data Cleaning and Exploration

Cleaning and Exploration

Cleaning

- Headers and footers are removed
- Web links are removed
- Escape characters are removed
- Missing full stops are added
- Data is shuffled
- Documents and summaries are separated

Exploration

Both the documents and summaries have a skewed distribution for the sentence count, the extreme values are removed. This distribution was the same in the subset and in the original data

Topic Modeling on Full Articles

Text Processing

Text Processing

- Strip non-alphabetic characters, lowercasing, tokenization, lemmatization
- Remove stop words (both general and domain-specific)
- Build dictionary (types in 0.1-80% of docs, max 20,000 types)

Text Representation

- TF (Term Frequency): used by LDA, normalized for pLSA
- TF-IDF: used by LSA
- Word embeddings (BERT): used by BERTopic

THE MODELS

LSA

- Uses SVD on TF-IDF matrix
- Optimal topics*: 10

pLSA

- Probabilistic model on TF matrix
- Optimal topics*: 30

LDA

- Probabilistic model on normalized TF matrix
- Optimal topics*: 20

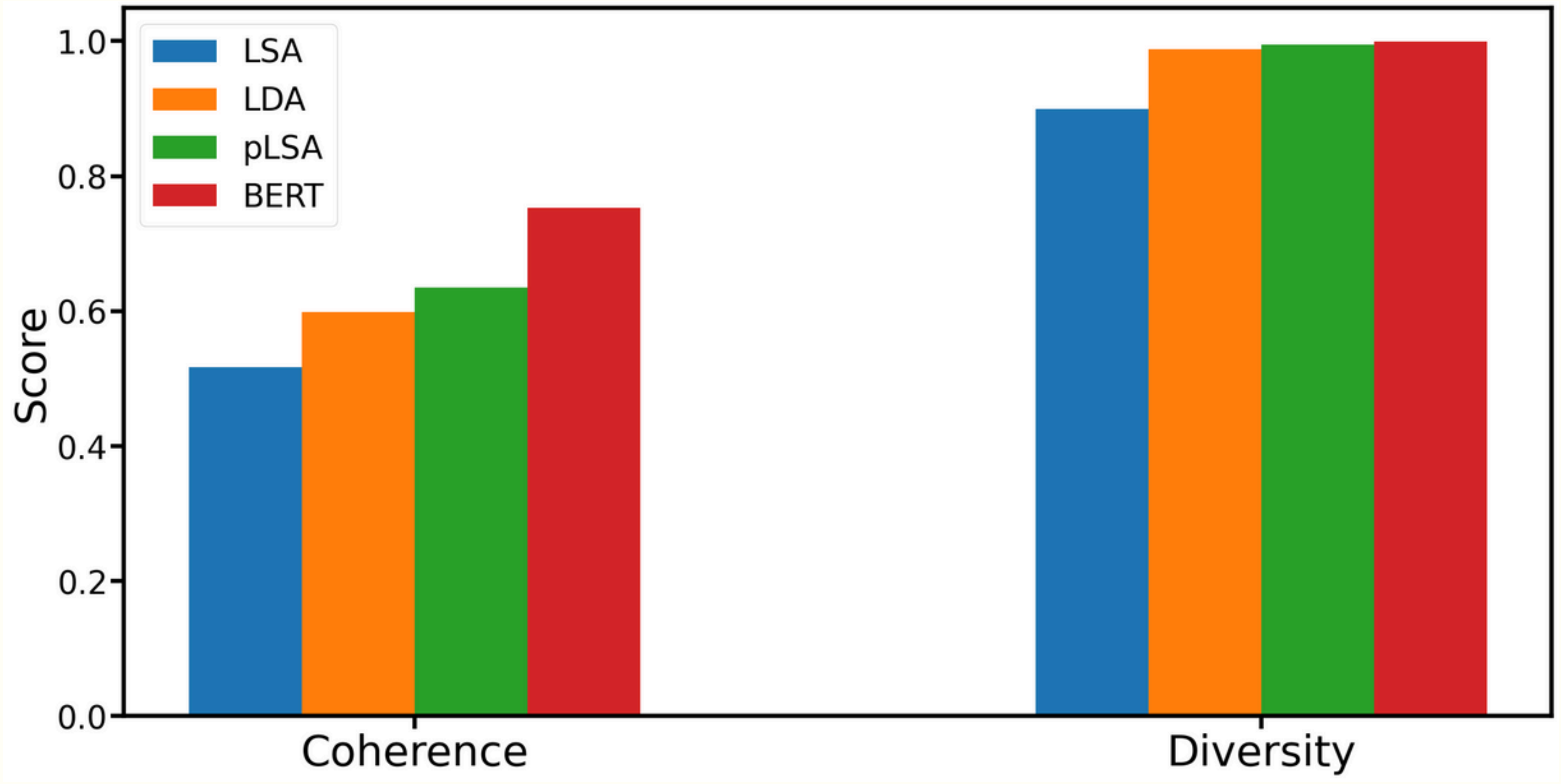
BERTopic

- Uses BERT embeddings
- No fine-tuning required (automatically determines topics)

*Estimated through tuning tests

Evaluation

Metrics



Time Comparison

Model	Tokenization	Tuning	Fitting	Total
LSA	3m	37m	3s	40m
LDA	3m	77m	7m	88m
pLSA	3m	122m	25m	150m
BERTopic	---	---	41m	41m

Visual Inspection

Visual inspection shows that LSA performs the worst, while pLSA, LDA, and BERTopic produce more meaningful topics

Text Summarization

The Method

Key Points

- Extractive summarization
- 5 sentences per document kept
- 85-15% train-test split

Upper Bound Estimate

MMR with ROUGE scores and cosine similarity on BERT embeddings

Baseline

Random summarizer

The Models

First Algorithm

TF-IDF → HITS → MMR

Second Algorithm

GloVe → PageRank → MMR

Third Algorithm

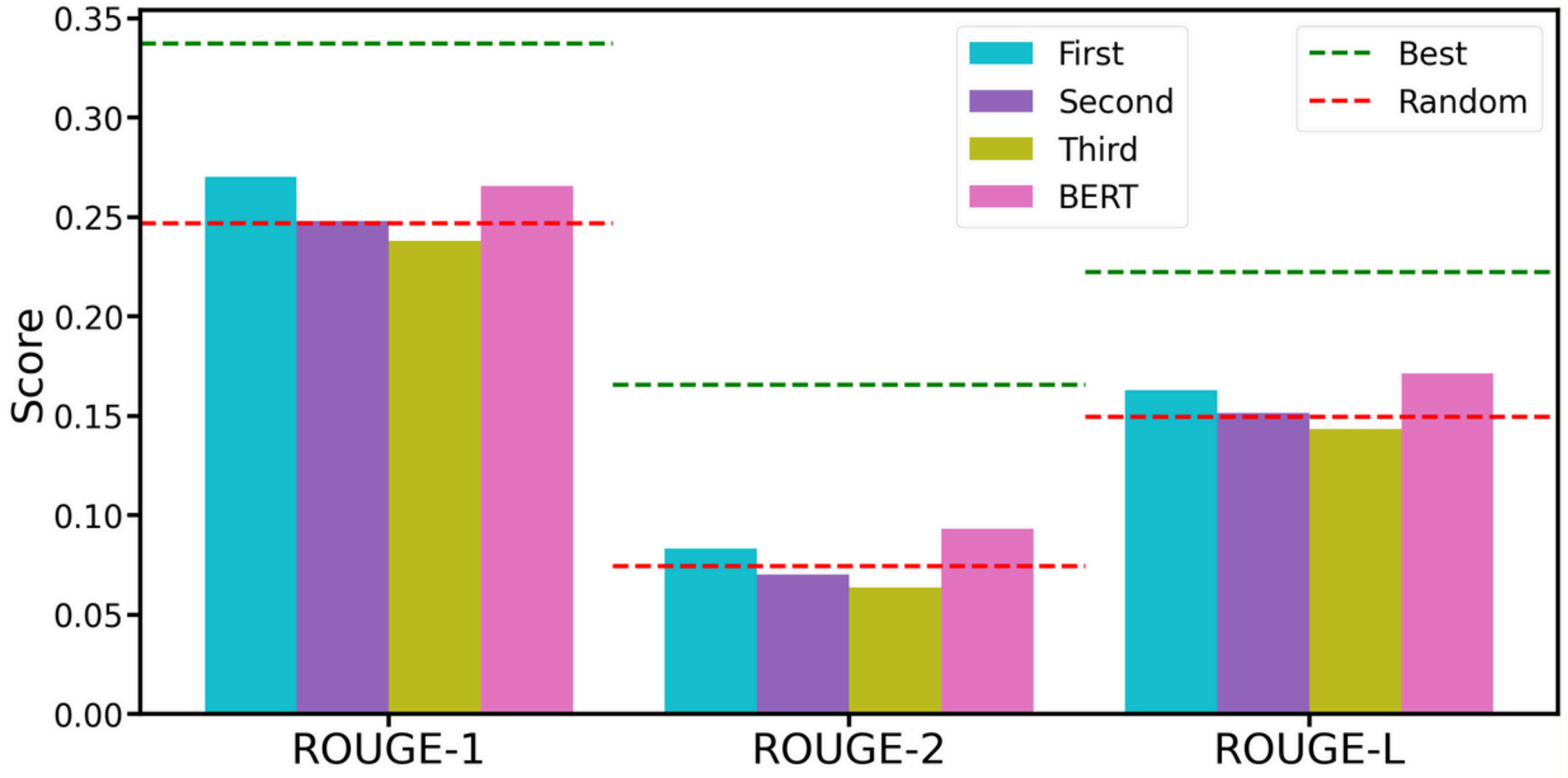
BERT → XGBoost → MMR

BERT Summarizer

BERT → KMeans

Evaluation

Metrics



Time Comparison

Model	Embedding	Training	Test	Total
First	---	---	1m	1m
Second	---	---	2m	2m
Third	16m	31m	12s	47m
BERT	---	---	22m	22m

Topic Modeling on Summarized Text

The Method

Key Points

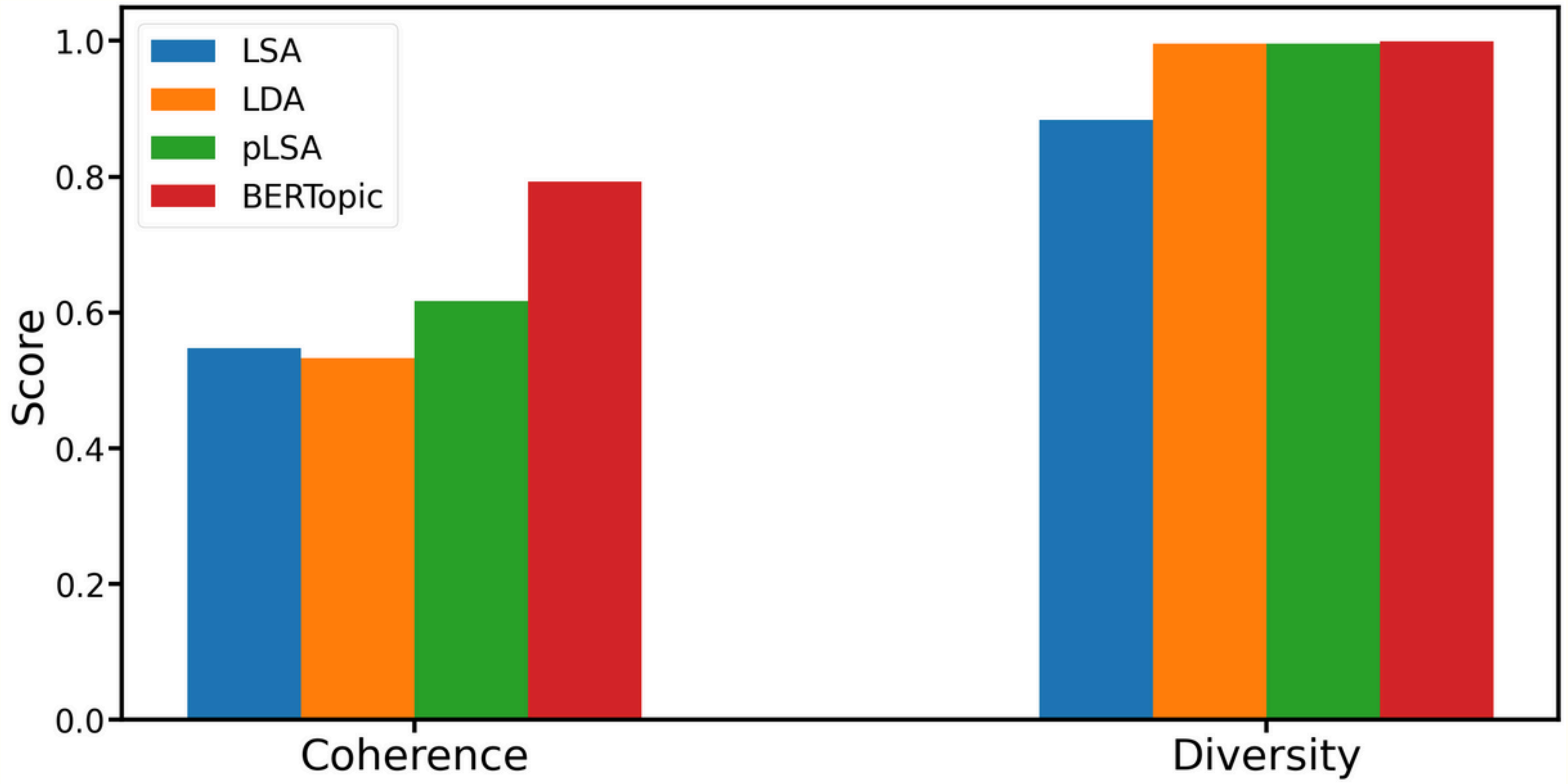
- Apply topic modeling to the summarized corpus
- Use the same pipeline as for original articles for a fair comparison

Pros and Cons

- Potential benefits:
  - Less noise and redundancy
  - Lower computational costs
- Possible drawback:
  - Loss of context

Evaluation

Metrics



Time Comparison

Model	Tokenization	Tuning	Fitting	Total
LSA	44s	47s	1s	2m
LDA	44s	8m	3m	12m
pLSA	44s	7m	3m	11m
BERTopic	---	---	39m	39m

Visual Inspection

By visual inspection BERTopic remains the best-performing model, producing the most coherent and meaningful topics, while LDA performs much worse, pLSA slightly worse, and LSA only slightly better

Conclusions

Conclusions

General Considerations

Both the topic modeling and text summarization experiments worked, with models employing BERT embeddings performing better than traditional algorithms on both tasks

On The Aim of The Project

The topic modeling on the summarized text gave mixed results, with some models achieving better topic quality while others greatly decreased their processing time

Thanks For Your Attention

!!