WikiHow Articles: a study on Topic Modeling and Text Summarization

Text Mining & Search project

Tariq Baghrous - 904027   Luca Iarocci - 894066   Roberto Ferrari - 852220

Contents

Introduction

- This project is dedicated to a comprehensive text mining procedure aimed at extracting valuable information from a vast collection of articles.
- The articles are extracted from the famous online platform **wikiHow**, an extensive database of instructional content.

230.000+ articles

2 datasets

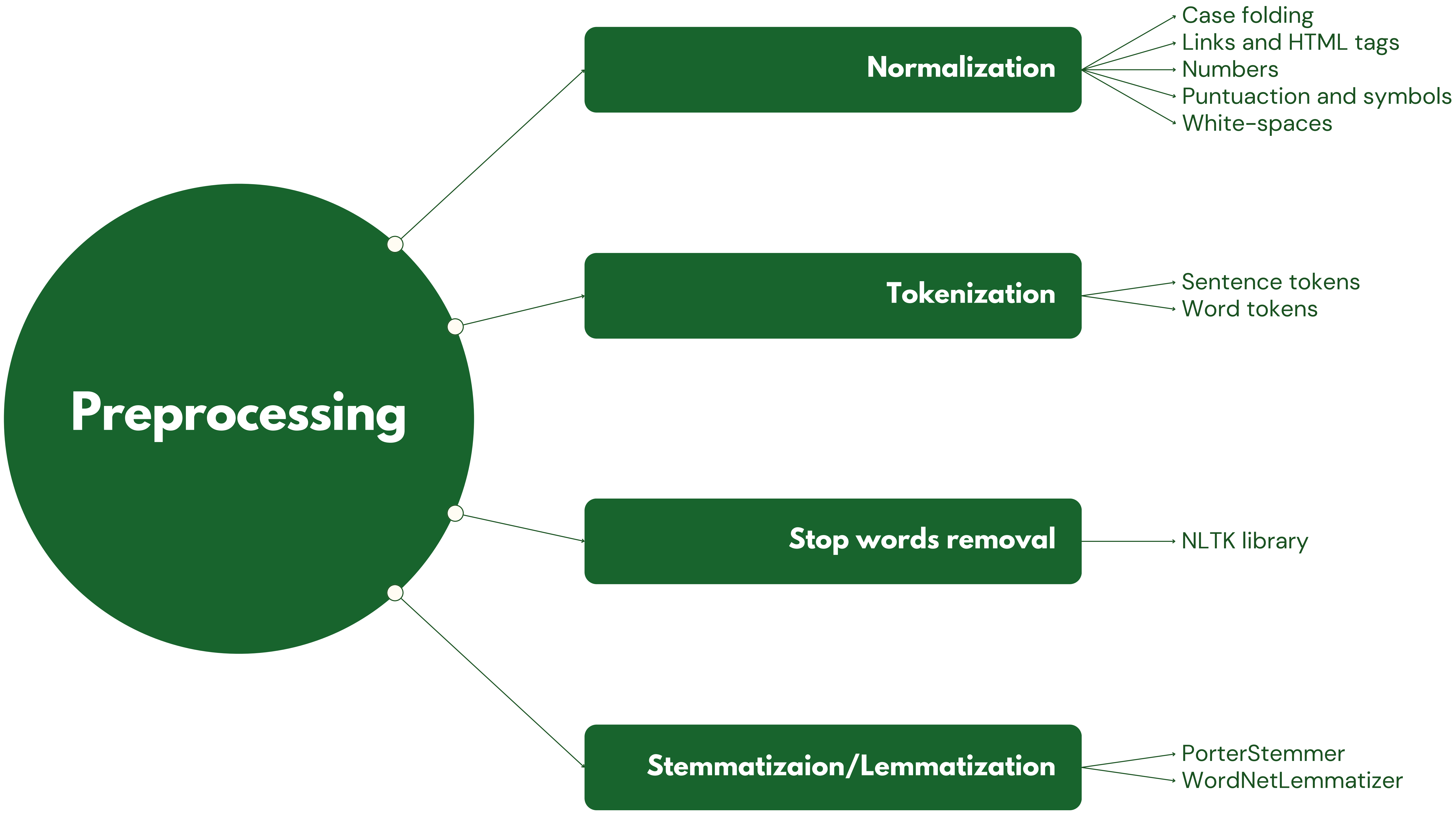
Data Description

wikihowAll

- Concatenation of all paragraphs as the articles and the bold lines as the reference summaries
- **Topic Modeling** task

wikihowSep

- Separate paragraphs as the articles and the bold lines corresponding to each paragraph as the reference summary.
- **Text Summarization** task


Preprocessing
Normalization
Case folding
Links and HTML tags
Numbers
Puntuaction and symbols
White-spaces
Tokenization
Sentence tokens
Word tokens
Stop words removal
NLTK library
Stemmatizaion/Lemmatization
PorterStemmer
WordNetLemmatizer

Topic Modeling

01 Text Representation

02 Latent Semantic Analysis (LSA)

03 Latent Dirichlet Allocation (LDA)

4.1 Text Representation

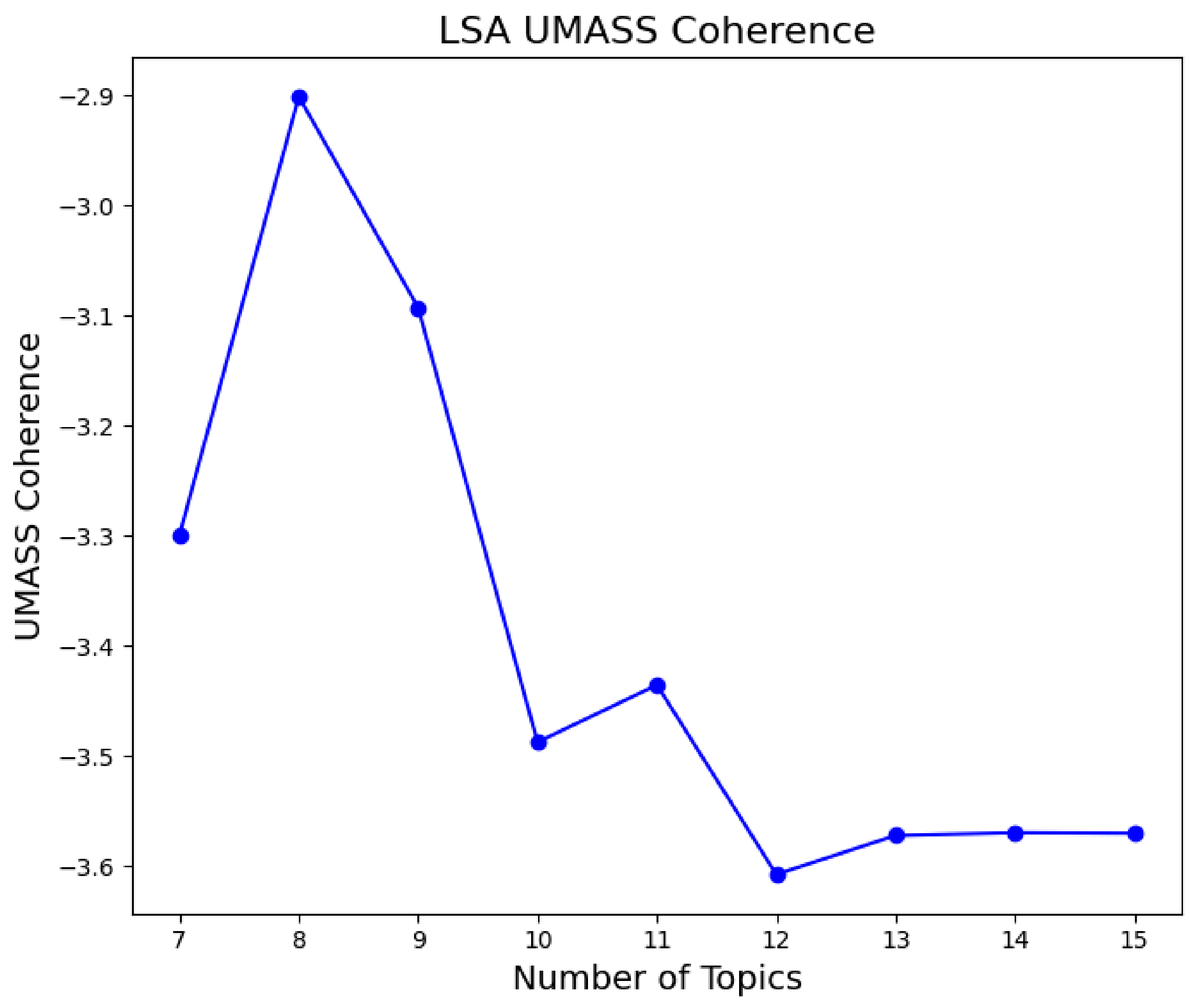
In order to obtain a better performance for both LSA and LDA models, it's good practice to prepare the text corpus with some preliminary operations – starting from the lemmatized text – that will help uncover latent structures between words in the corpus, as well as filter the final dictionary:

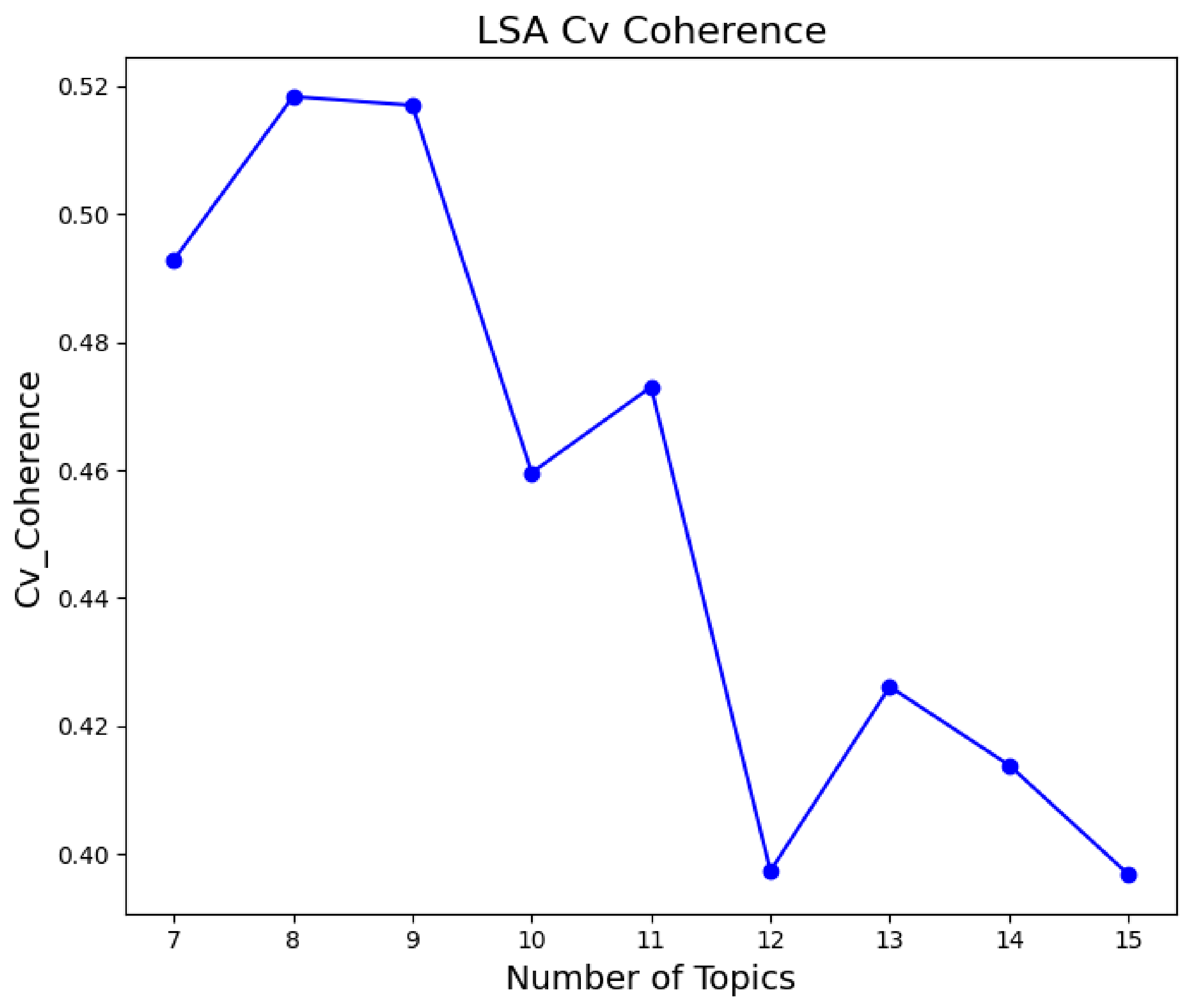
- **Words listing**
- **Bigrams and Trigrams identification** (12434 and 514 respectively)
- **Removing less and most frequent words** (bottom 0.05% and top 20%)
- **Dictionary definition** (cut below 0.1% and above 5% and words with a character length < or equal to 3, for a total of 19413 unique terms)

Finally, two different **text representation** techniques have been applied to the resulting dictionary: **TF-IDF matrix** and **BoW model** for LSA and LDA, respectively.

4.2 Latent Semantic Analysis (LSA)

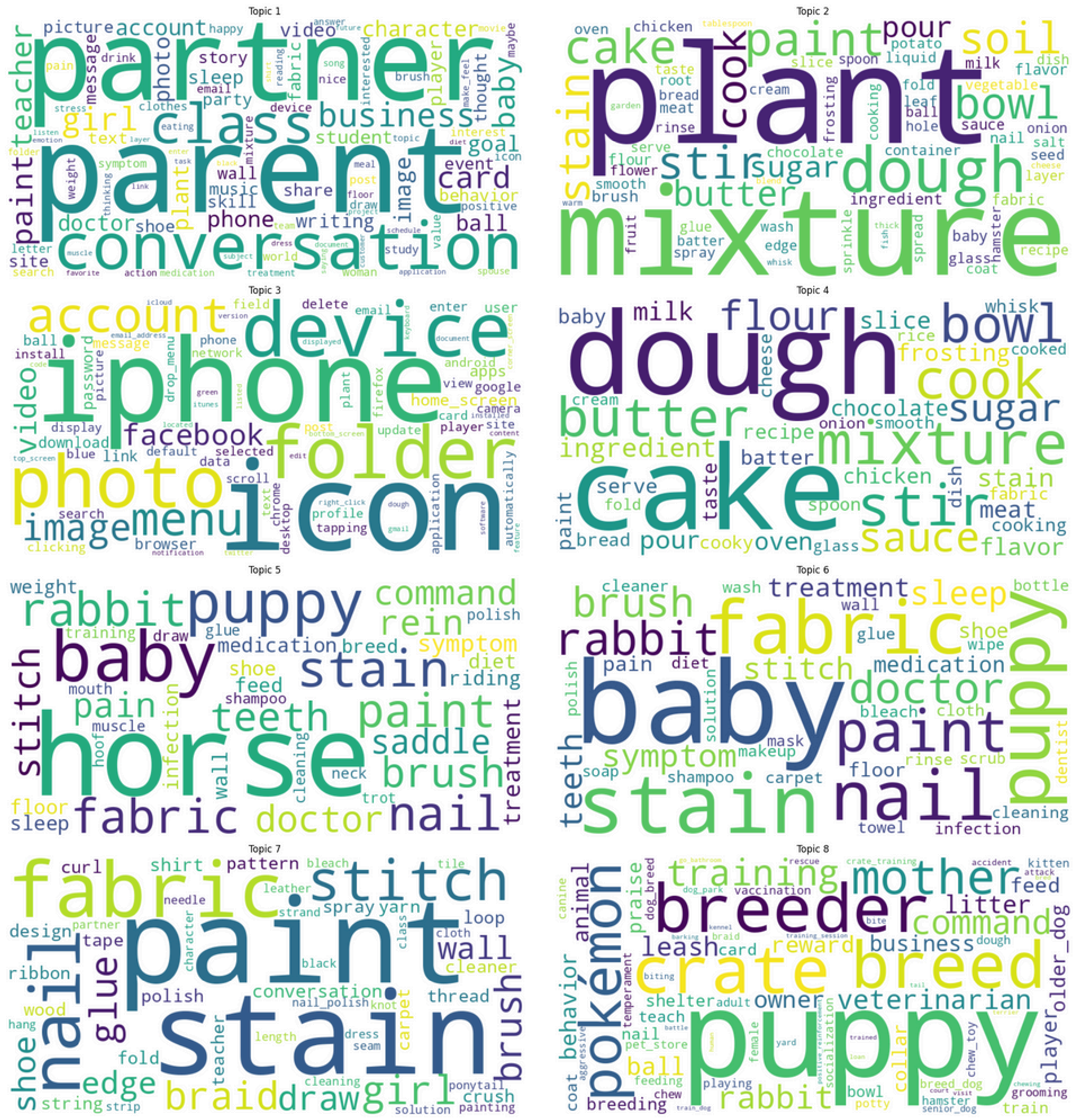
- NLP model used to identify underlying **topics** by grouping together words that frequently appear in similar contexts
- Number of topics evaluated through the comparison between **Umass** and **Cv** coherence scores on a pre-established range





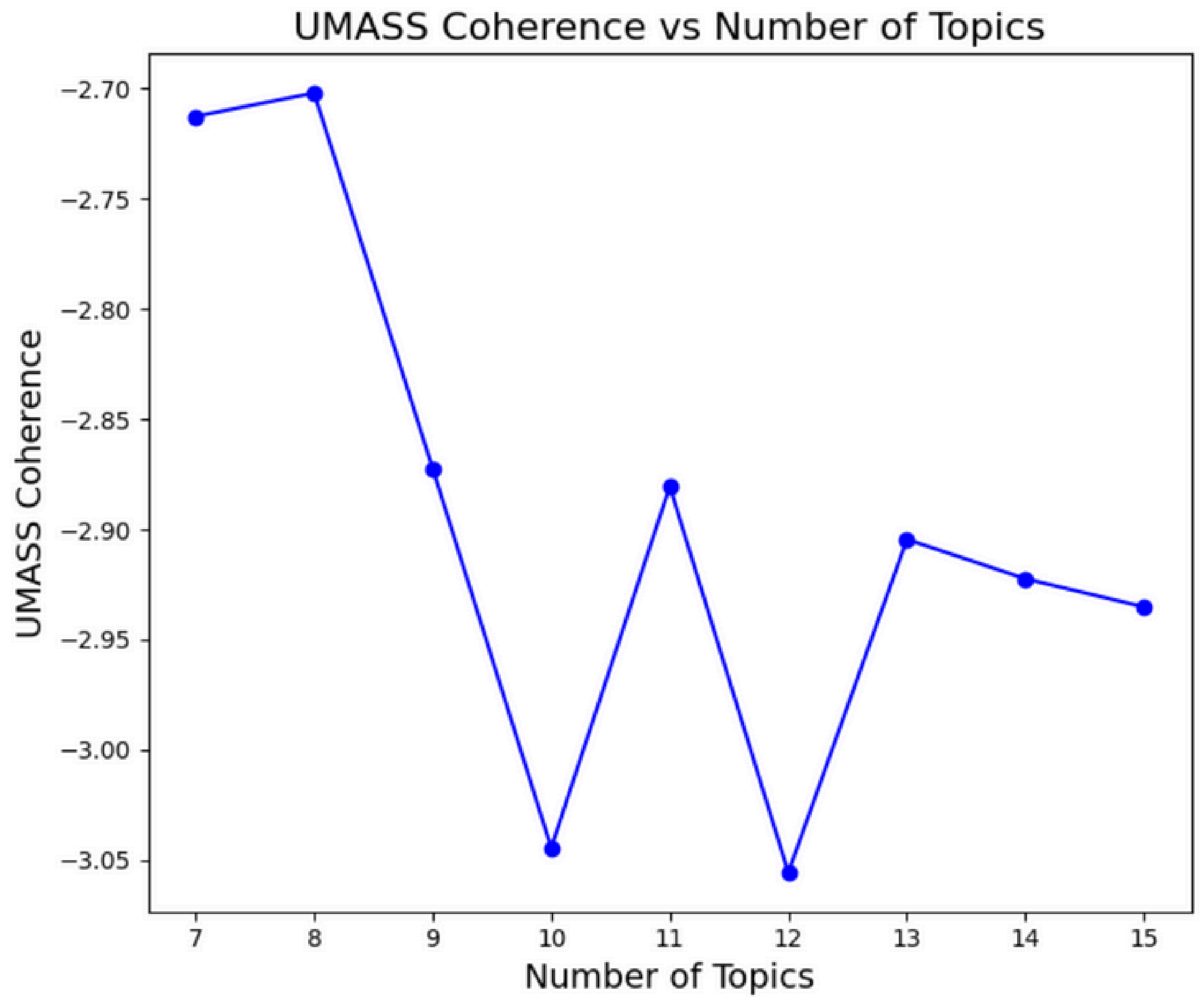
4.2 Latent Semantic Analysis (LSA)

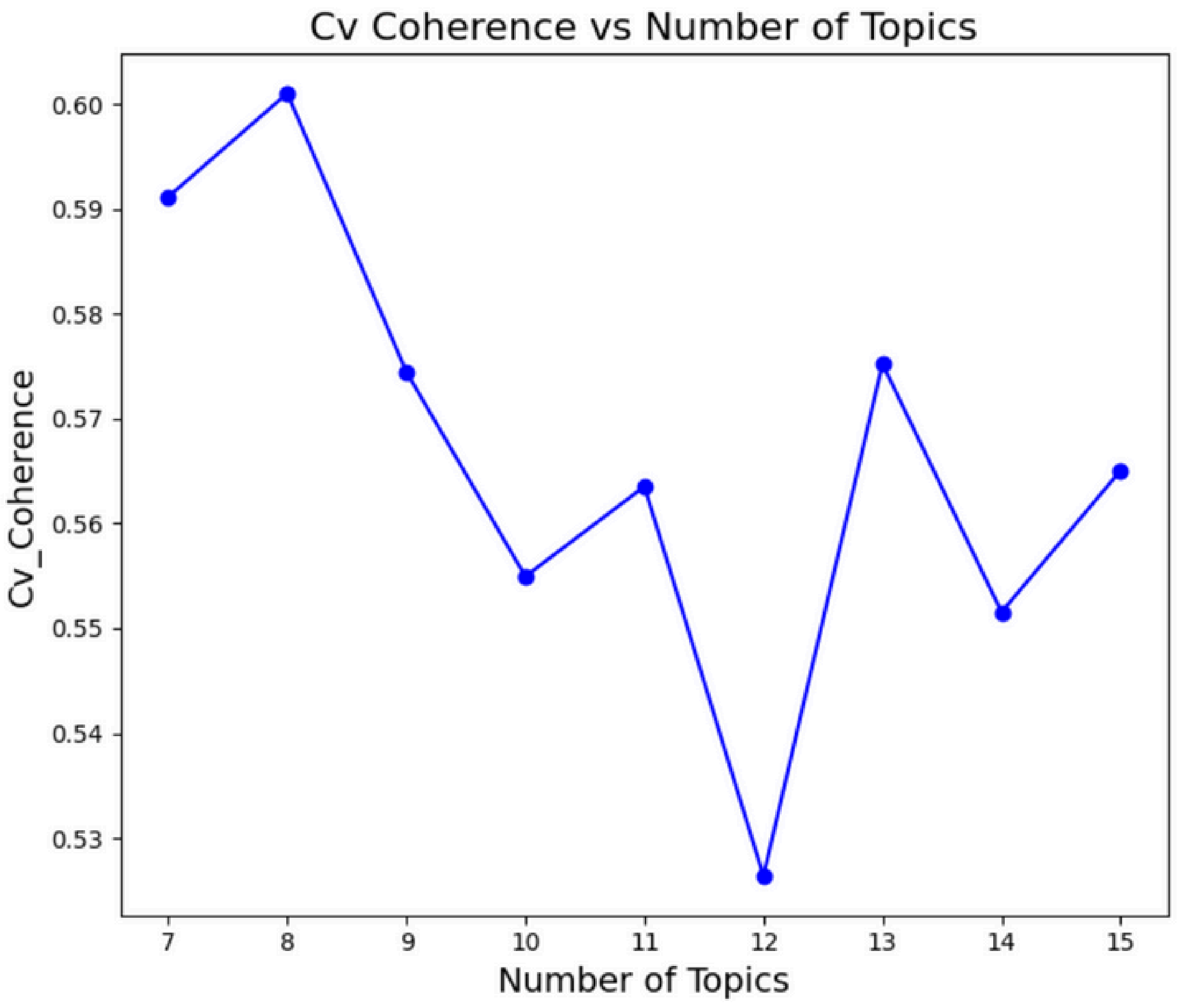
1. **Family and Education:** parent, partner, conversation, class, teacher
2. **Baking/Gardening:** plant, mixture, dough, cake, stir, bowl, stain, soil
3. **Smartphone Usage:** icon, iPhone, device, folder, photo, account, menu
4. **Baking:** dough, cake, mixture, butter, cook, flour
5. **Animal Care:** horse, baby, puppy, rabbit, stain
6. **Household Care:** baby, fabric, paint, nail, puppy
7. **Household Activities:** paint, stain, fabric, nail, stitch
8. **Dog Care:** puppy, crate, breeder, breed, training



4.3 Latent Dirichlet Allocation (LDA)

LDA model is a **generative probabilistic model** used to **classify text** contained in a corpus into a specified number of **topics**, which were obtained, after some testing on a 5% sample, through the comparison between Umass and Cv coherence scores, reaching their best scores in correspondence of **8 topics**. The most relevant topic words will be displayed in the next slide.





Model parameters:

- **Alfa**: 'symmetric'
- **Beta**: 'symmetric'
- **chunksize**: 800
- **passes**: 10

4.3 Latent Dirichlet Allocation (LDA)

1. **Stitching** stitch, string, length, needle, hook, loop.
2. **Gardening**: plant, garden, tree, soil, root, ground, battery, cable, install.
3. **Farming/Gardening**: stem, seed, growth, fertilizer, compost, fruit, vegetable, watering.
4. **Music/Art**: chord, drum, card, image, artist, letter, picture.
5. **Music/Entertainment**: song, music, character, guitar, beat, audience.
6. **Drawing/Creative**: fabric, paint, engine, design.
7. **Investing**: business, sale, sell, price, vehicle, payment.
8. **DIY Activities/Creative**: fabric, wire, paint, metal, glue, wood, tape, hole.

Text Summarization

01 Summarization Approaches

02 Data Exploration and Selection

03 Summaries Evaluation

5.1 Summarization Approaches

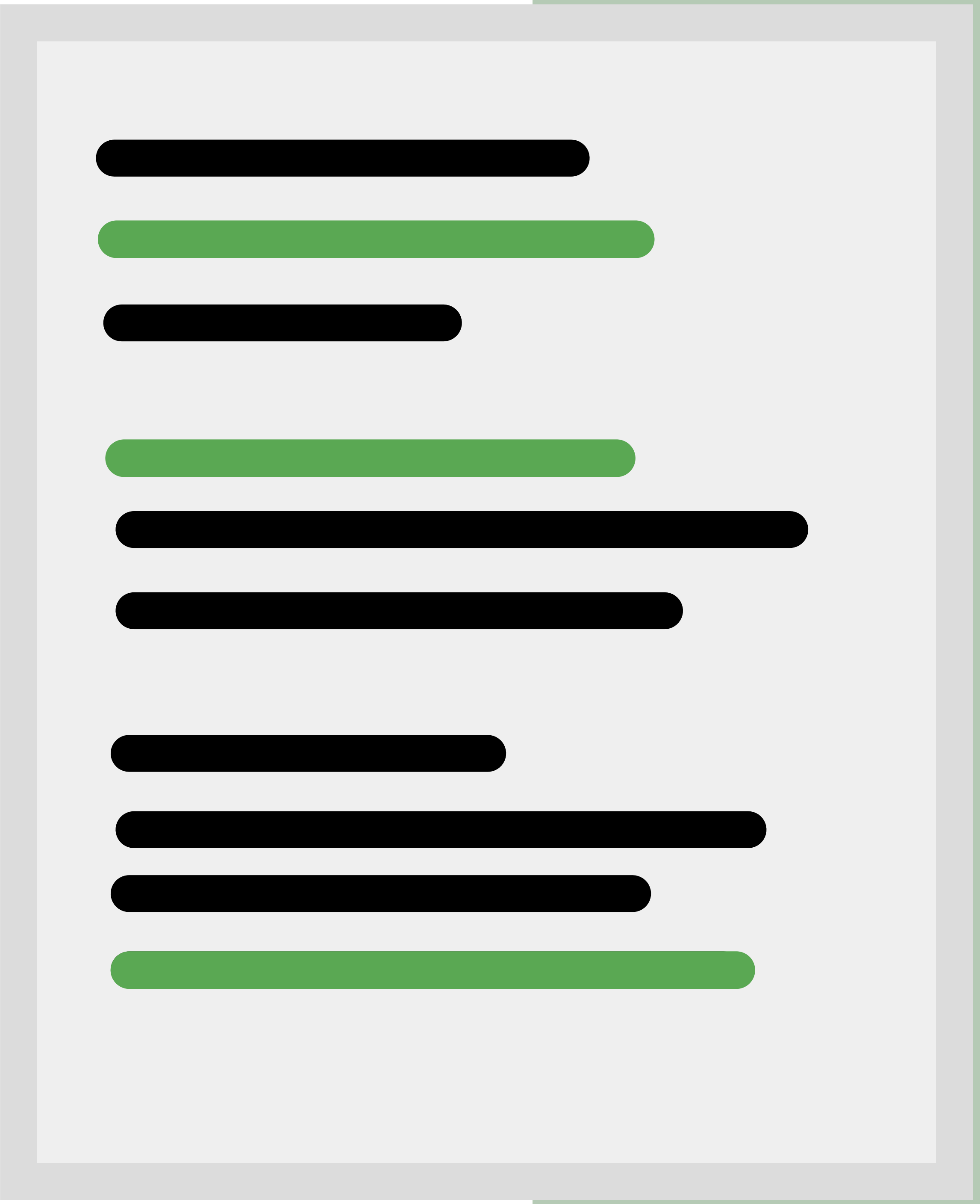
We focused mainly on **2 extractive summarization** technique:

- Latent Semantic Analisis (LSA)
- TextRank (TR)

Inspired by the WikiHow articles bullet-point/step-by-step style we implemented **2 variations**:

- LSA by paragraph (LSAp)
- TR by paragraph (TRp)

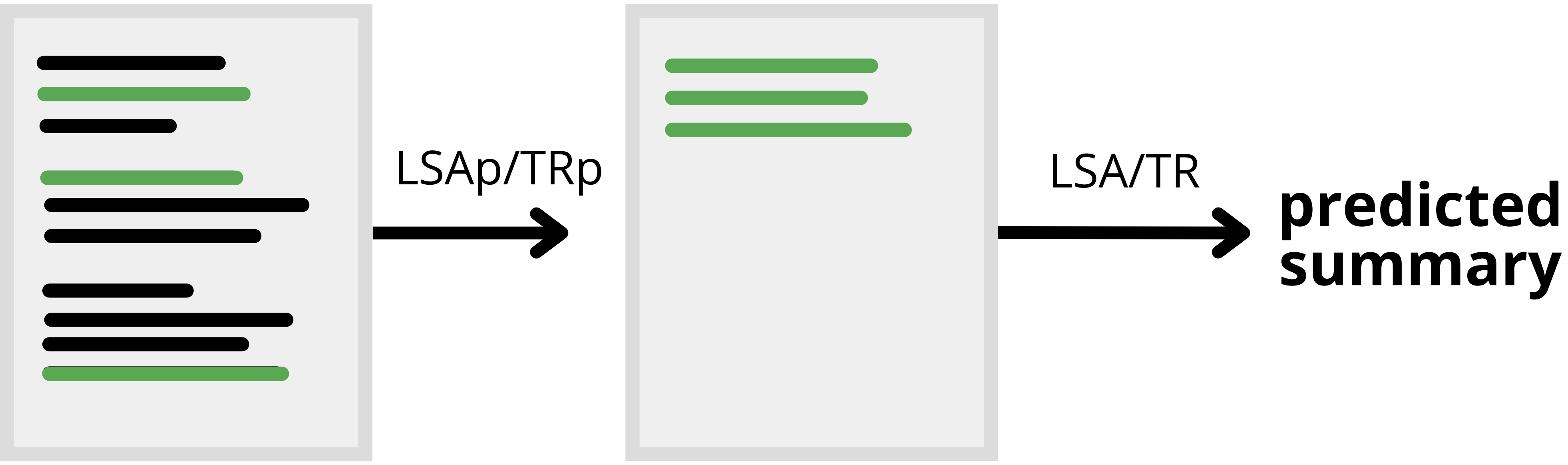
that summarize text paragraph by paragraph.



5.1 Summarization Approaches

We experimented with **3 different target length:**

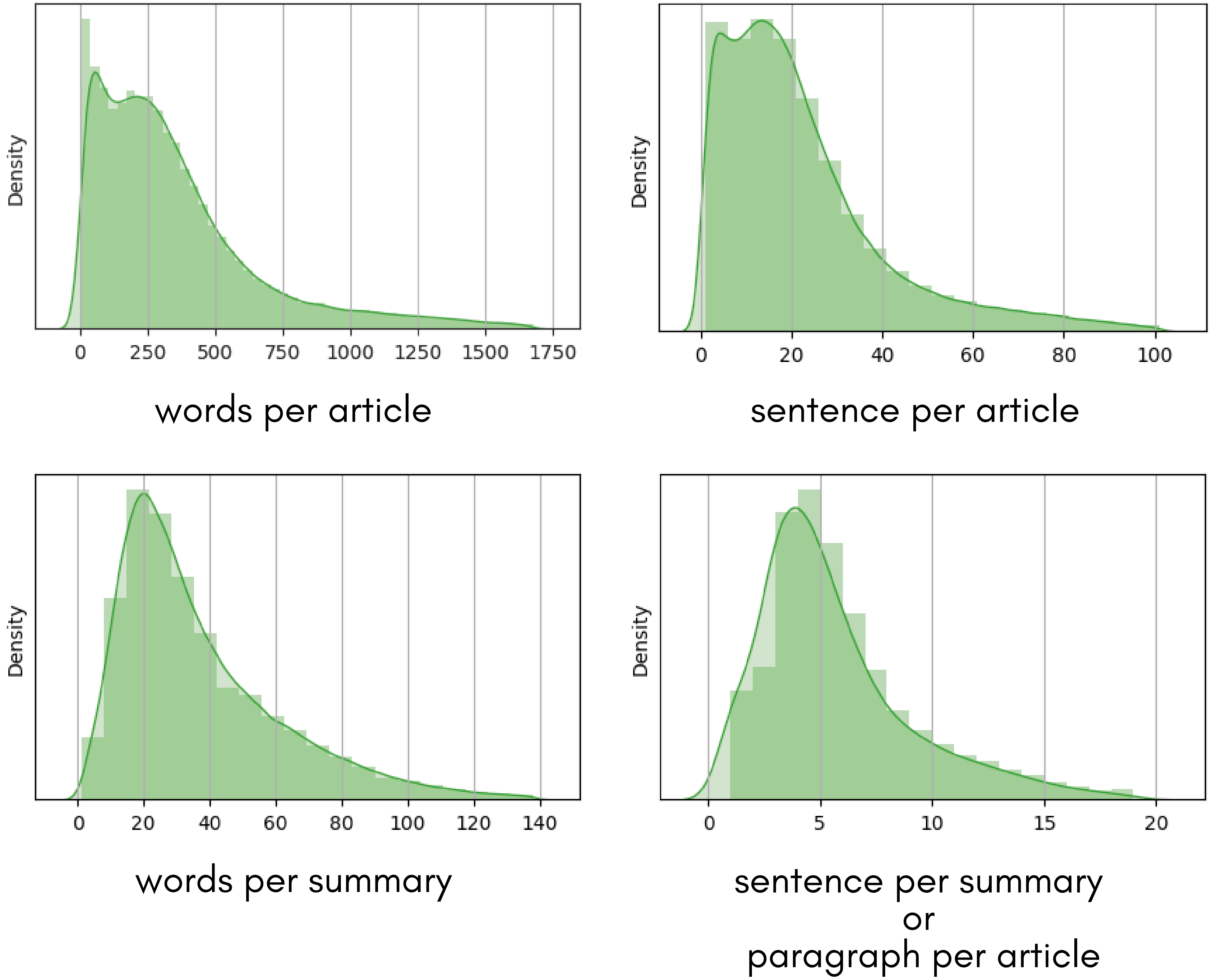
- base length = text paragraph number
- further reductions:
  - 1/2 of base length (rounded)
  - 1/3 of base length (rounded)



5.2 Data Exploration and Selection

Before applying our summarization techniques, we decided to explore the corpus characteristics in order to **select a subset** of suitable articles to be the object of our evaluation.

This choise is in part justified by the fact that our summarization techniques are **unsupervised,** thus no point in applying them to the whole dataset.



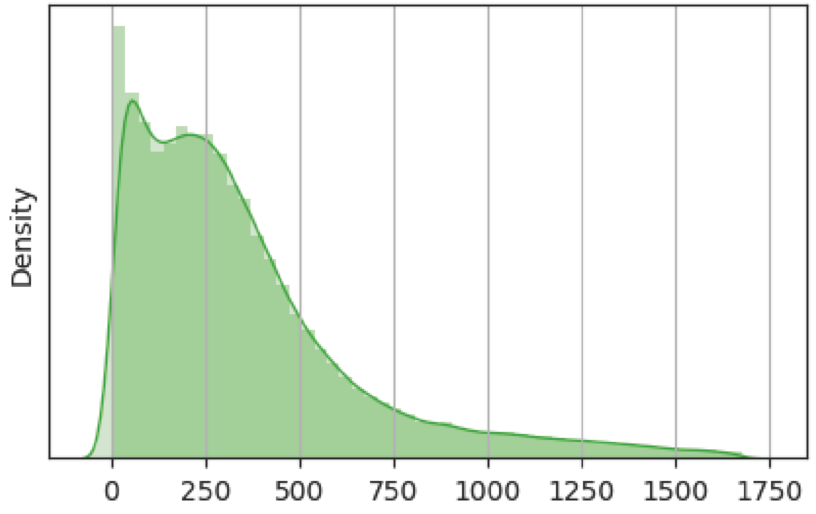
5.2 Data Exploration and Selection

Detect and remove outliers

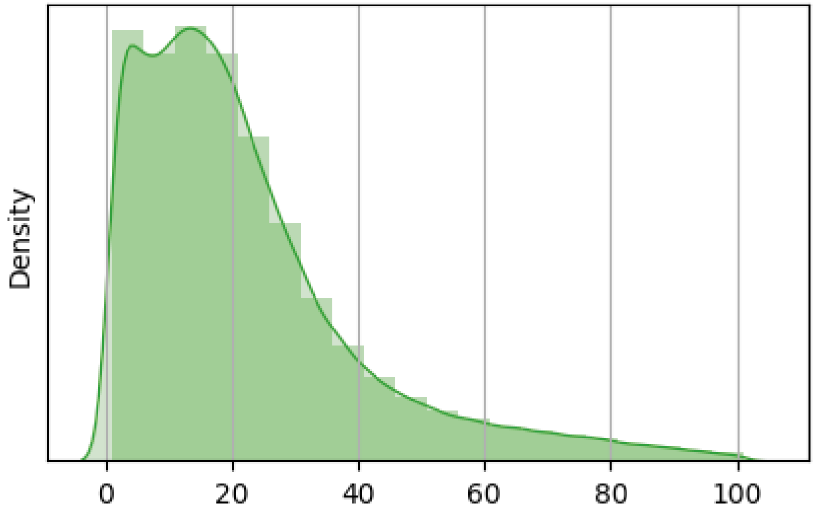
Statistic based selection, records that exceeded the mean value by more than 3 times the standard deviation were excluded.

This was done both for word and sentence length.

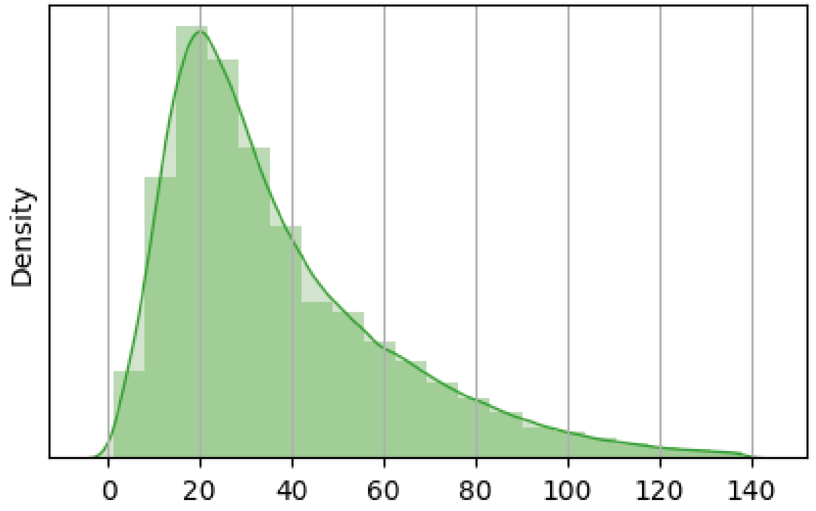
		Text		Summary	
		words	sentences	words	sentences
Paragraph	median	118	7	5	1
	mean	125.03	8.21	5.68	1.00
	std	43.88	2.49	2.84	0.00
Article	median	274	18	30	5
	mean	353.02	22.38	37.16	5.61
	std	311.85	18.64	24.96	3.45



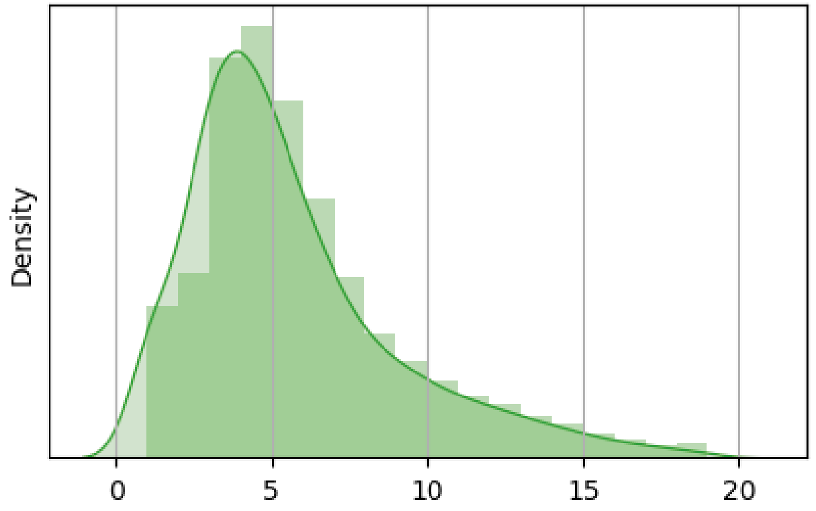
words per article



sentence per article



words per summary



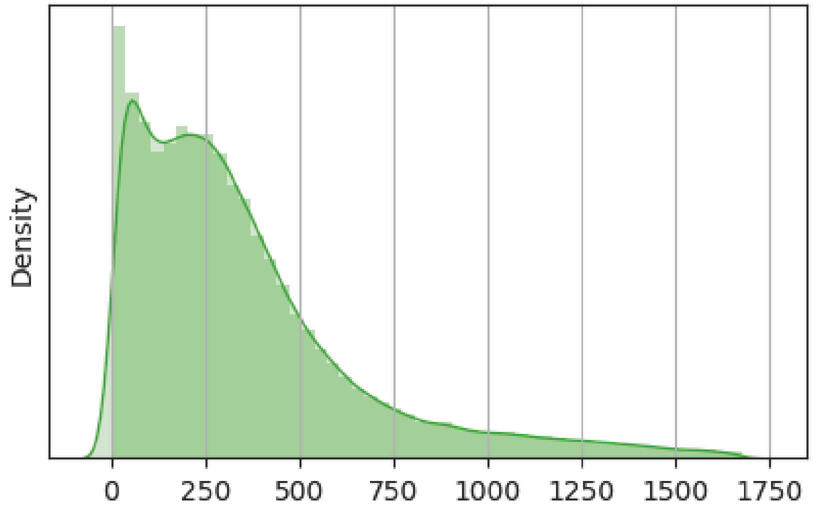
sentence per summary
or
paragraph per article

5.2 Data Exploration and Selection

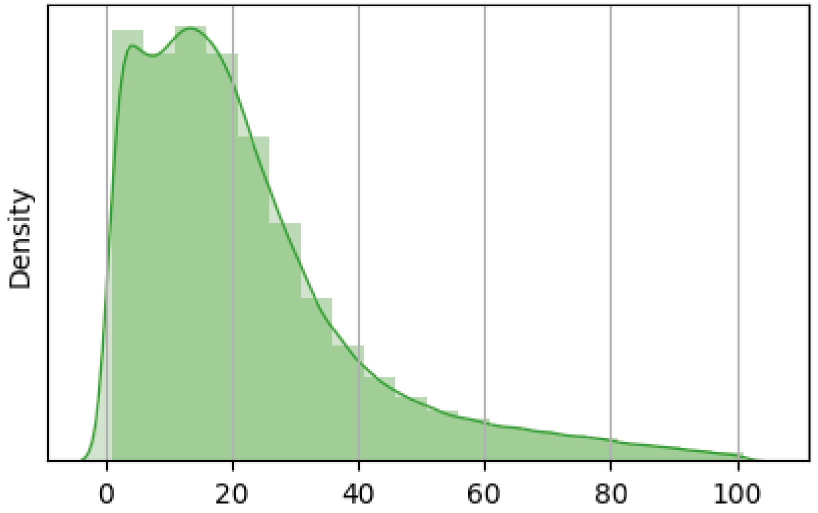
Complexity and length

Records' characteristic based selection, filter for reasonably long and structured article:

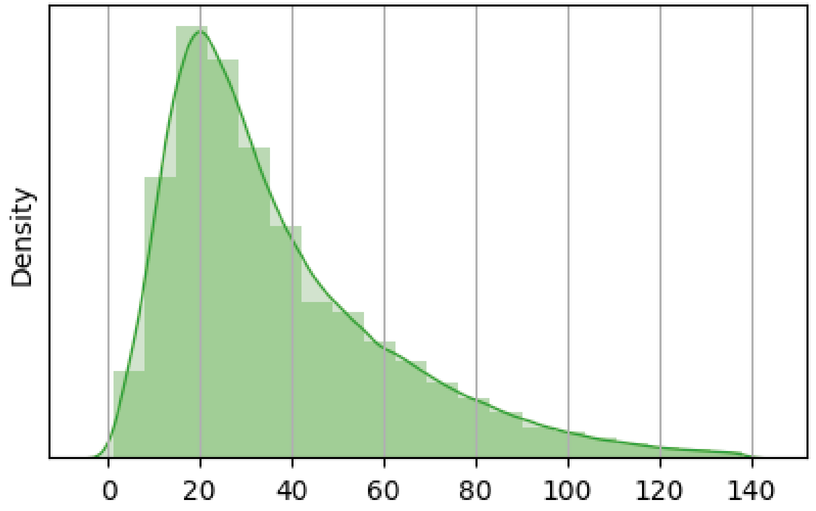
- more than 3 sentences per paragraph
- (more than 10 words per paragraph)
- more than 6 paragraphs/steps per article
- (more than 100 word per article)



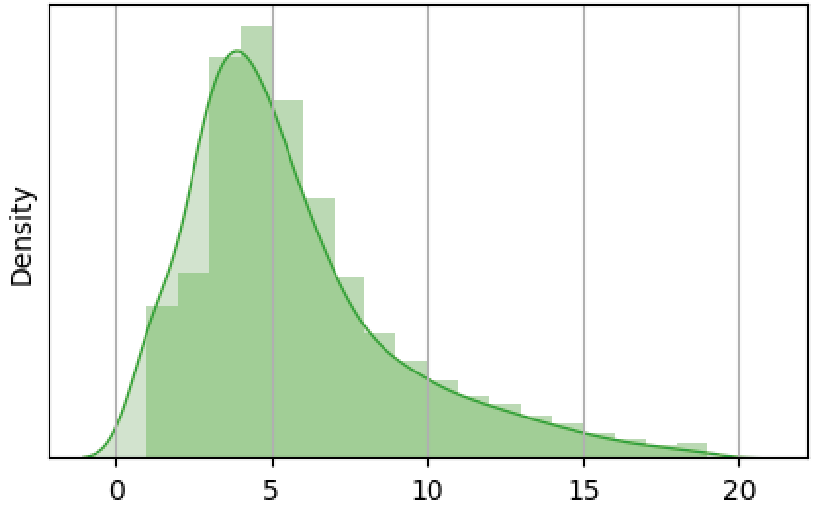
words per article



sentence per article



words per summary



sentence per summary
or
paragraph per article

5.3 Summaries Evaluation

Sampling

Following the described selection crieteria we randomly extracted a subset of 1000 articles.

Benchmark

To evaluete the results of our summarization technique methods we set a benchmark:

- Random (RND)
- Random by paragraph (RNDp)

which resectivly select random sentences from an article and from each paragraph.

Metrics

Each summary was evalueted by computing and averaging the following metrics:

- ROUGE-1
- ROUGE-2
- ROUGE-L

5.3 Summaries Evaluation

Results

For each method and reduction level combination we computed the average ROUGE mean, standard deviation and 0.95 t-test confidence intervial.

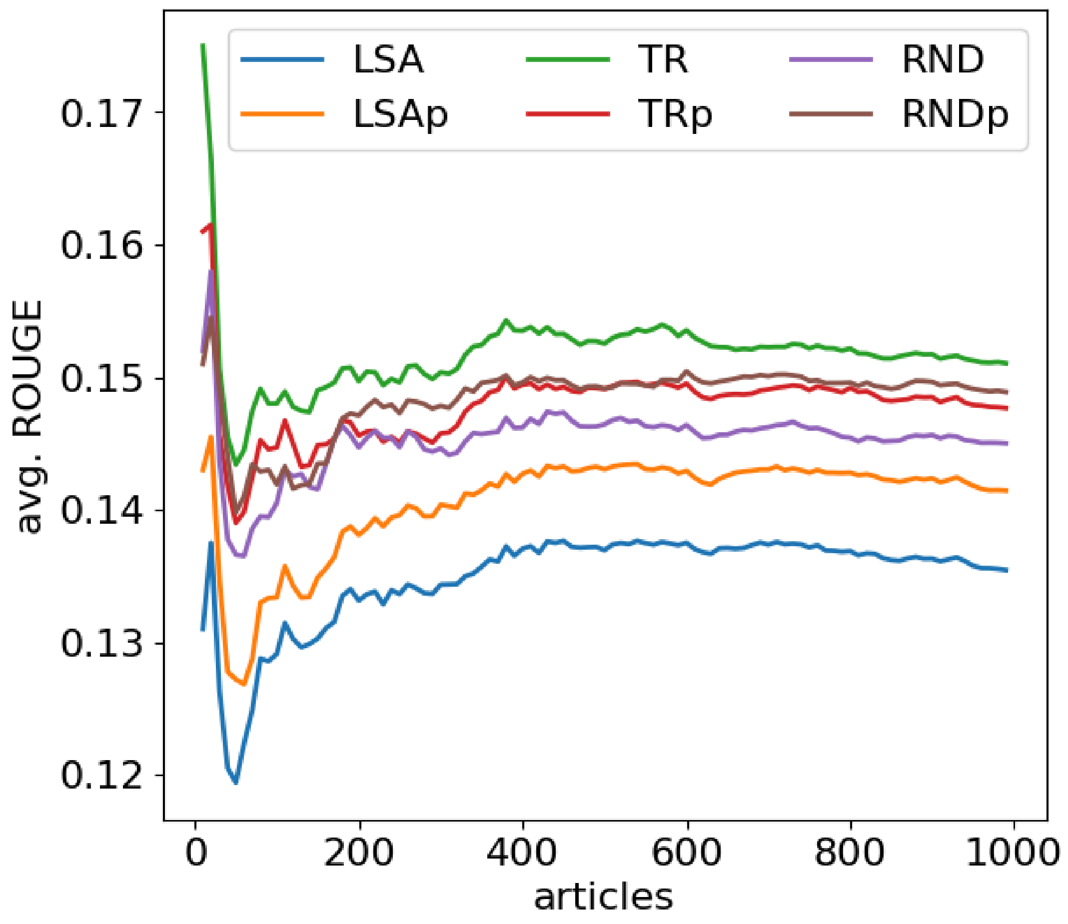
		LSA	LSAp	TR	TRp	RND	RNDp
reduction=1	mean	0.136	0.142	**0.151**	0.148	0.145	0.149
	95% CI	±0.003	±0.003	±0.003	±0.003	±0.003	±0.003
	std	0.047	0.0490	0.0525	0.051	0.048	0.051
reduction=2	mean	0.139	0.142	**0.158**	0.152	0.140	0.145
	95% CI	±0.003	±0.003	±0.003	±0.004	±0.003	±0.004
	std	0.051	0.053	0.056	0.057	0.053	0.056
reduction=3	mean	0.136	0.137	**0.156**	0.150	0.132	0.132
	95% CI	±0.003	±0.003	±0.004	±0.004	±0.004	±0.004
	std	0.054	0.054	0.058	0.059	0.057	0.059

Table 2: *average ROUGE statistics by reduction level*

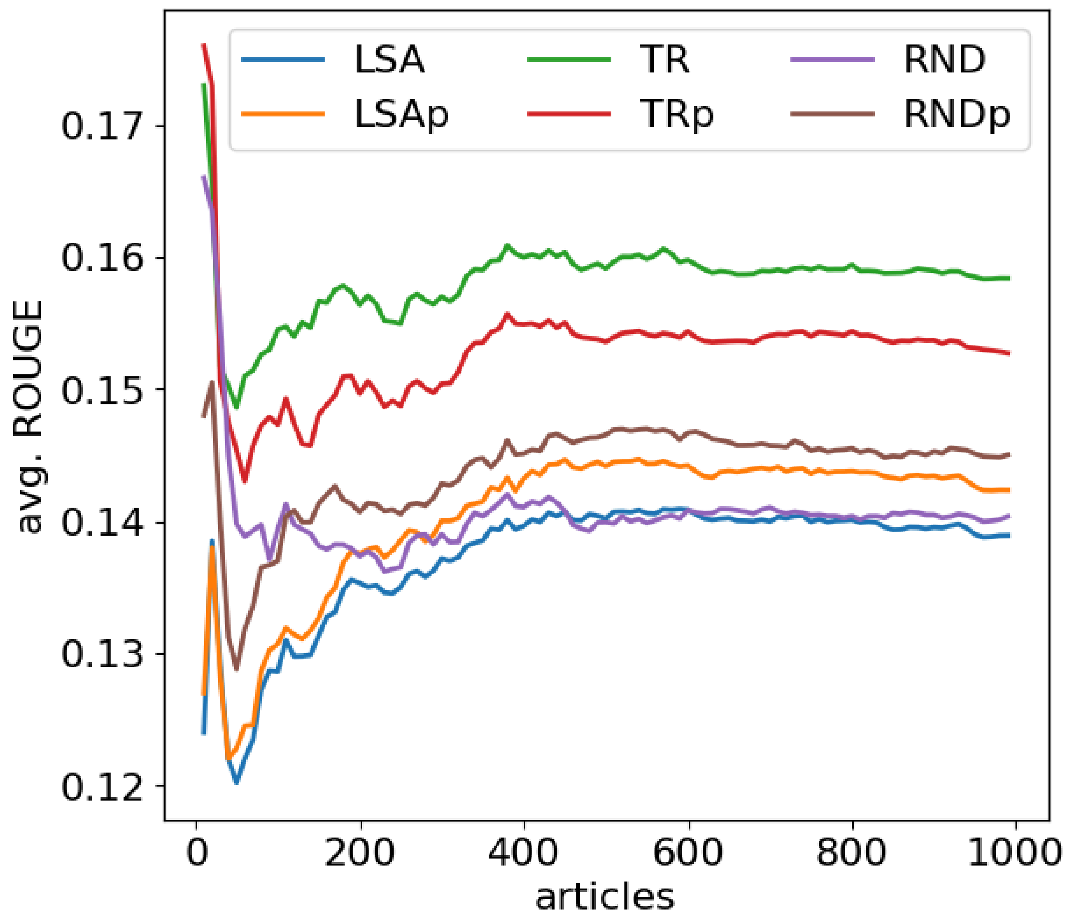
5.3 Summaries Evaluation

Results consistency and visualization

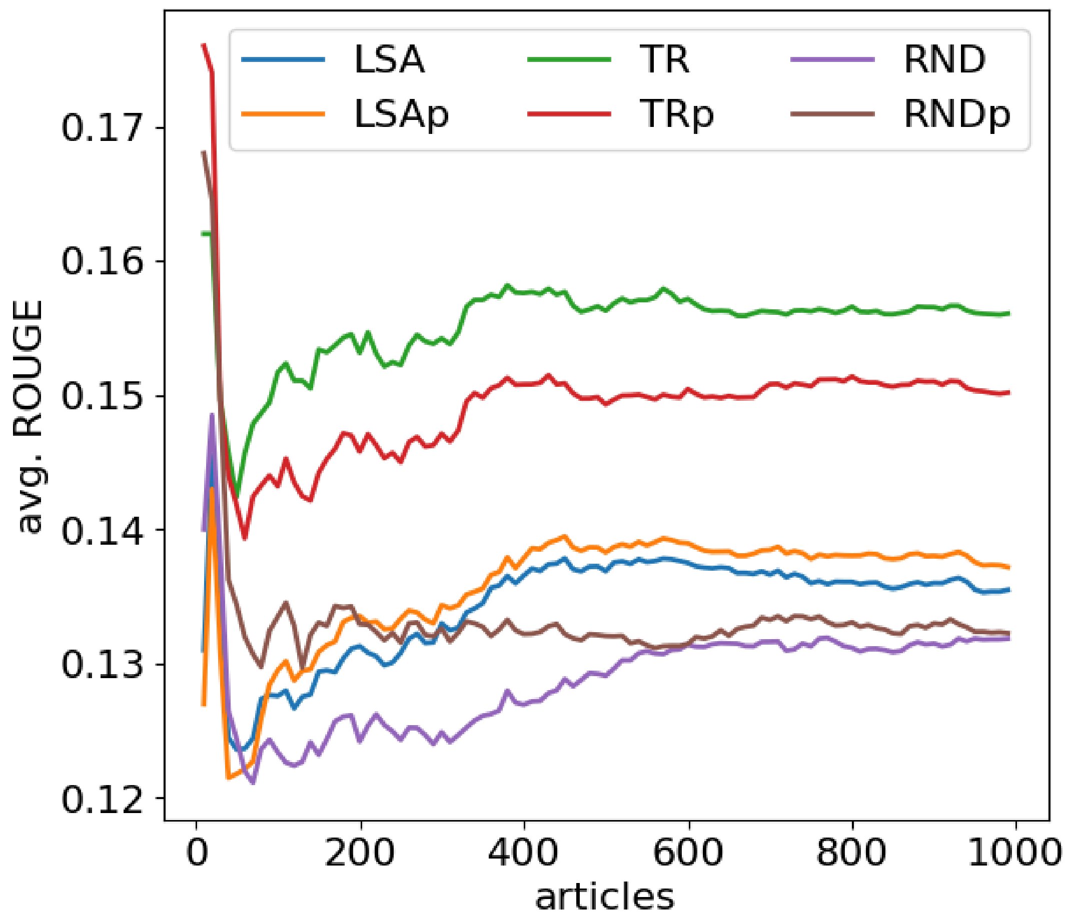
To asses the consistency of our results we studied their trend and ranking as the number of article grows.



reduction=1



reduction=2



reduction=3

Conclusions

Topic Modeling

LSA better efficiency

LDA better coherence

Text Summarization

TR overall best

LSAp > LSA
TRp < TR

Thanks for your attention