Analyzing Keyphrase Generation Methods in Scientific Domains

Debopam Sanyal[1], Anna Giannakou[2], Oluwamayowa Amusat[2], Lavanya Ramakrishnan[2]
[1]University of Illinois Urbana-Champaign, [2]Lawrence Berkeley National Laboratory

RESEARCH QUESTION

Which unsupervised algorithm extracts the best quality keywords from scientific data?

BACKGROUND

- Keyphrase extraction deals with the automatic extraction of representative phrases that can summarize a document
- Manually assigning keyphrases to documents is tedious and time-consuming
- Must develop automated keyphrase extraction techniques to enable search over large scientific data
- Part of the science search metadata extraction pipeline
- There is no "best" NLP algorithm across all domains
- Evaluation is done by calculating the similarity between the human-curated and the automated keyphrases
- Unsupervised methods involve selection of candidate lexical units, ranking them and forming keyphrases
- Two categories: statistics-based and graph-based methods

DATASETS

- Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE): (581 documents)
- Joint Genome Institute (JGI): Genomics data (184 documents)

RESEARCH METHODS

- Statistics-based: **KP-Miner** & **YAKE**; Graph-based: **TextRank**, **RAKE**, **TopicRank** & **MultipartiteRank**

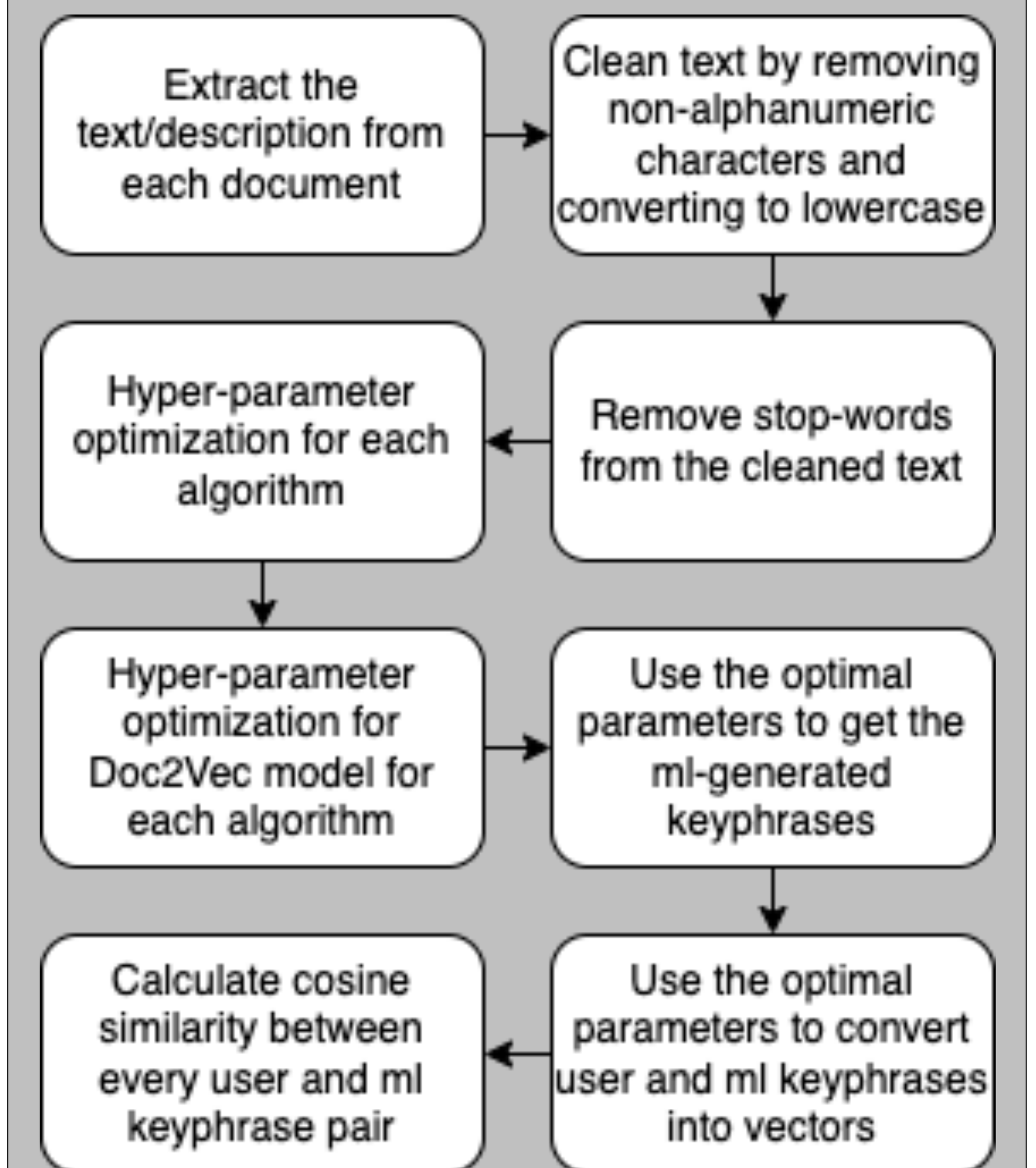

***Figure 1**: Flowchart for finding the best performing algorithm on a dataset*

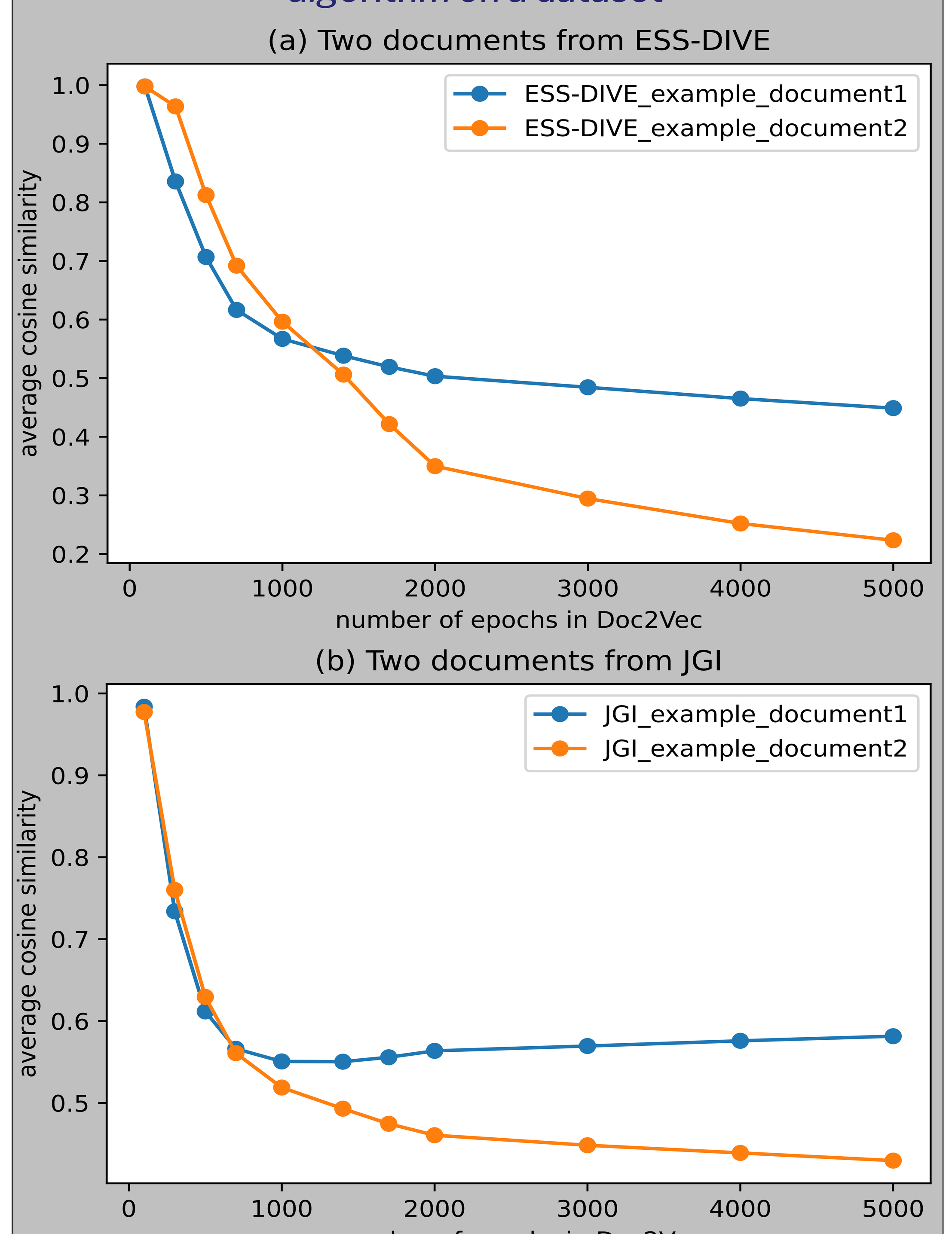

***Figure 2**: Cosine similarity vs number of epochs plot for two documents in: (a) the ESS-DIVE dataset and (b) the JGI dataset, respectively*

RESULTS

- Used the pke module for the algorithms to return top five keyphrases

Human-curated	ML-generated
ammoniaoxidizing bacteria	**ammoniaoxidizing bacteria**
earlydiverging fungi	alga chlamydomonas reinhardtii
climatology	global cloud **climatology**
CO_2 emissions	**CO_2 emissions**
tillage	european **tillage**
soil	potential acid sulfate **soil**
singlecell	thermus aquaticus polymerase
alpine tundra	**alpine tundra**
river corridor	**river** crested butte

***Table 1**: Some examples of best-matched human-curated vs ML-generated keyphrases*

- Returned **top 10** documents based on average cosine similarity

Algorithm	top 10 docs	all docs
KP-Miner	0.84	0.38
YAKE	0.71	0.32
TextRank	0.69	0.32
RAKE	0.71	0.32
TopicRank	0.75	0.31
MultipartiteRank	0.71	0.34

***Table 2**: Average cosine similarity per algorithm using the ESS-DIVE documents*

Algorithm	top 10 docs	all docs
KP-Miner	0.99	0.59
YAKE	0.99	0.59
TextRank	0.99	0.59
RAKE	0.91	0.49
TopicRank	0.99	0.59
MultipartiteRank	0.99	0.59

***Table 3**: Average cosine similarity per algorithm using the JGI documents*

CONCLUSION

- KP-Miner outperformed the next best algorithm by ~5% on the ESS-DIVE data; rest were within ~2% of each other
- RAKE performed ~10% worse than the other algorithms on the JGI data
- Examples show that there is significant semantic similarity between the human-curated and ML-generated keyphrases
- There are many common documents in the top 10 lists across different algorithms, showing the consistency of our evaluation process
- Some dissimilar human-ML keyphrase pairs have high cosine similarity score

FUTURE WORK

- Find out why KP-Miner works the best on ESS-DIVE data and RAKE performs poorly on JGI data
- Preprocess human-curated keyphrases to remove non-alphanumeric characters
- Use exact/partial matching as the method for measuring similarity b/w human & ML keyphrases; compare results with cosine similarity
- Conduct manual inspection of high-scoring human-ML keyphrase pairs
- Try returning more than five keyphrases

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) and uses resources at the National Energy Research Scientific Computing Center, a U.S. Department of Energy Office of Science User Facility under Contract No. DE-AC02-05CH11231.