NLP in Medical Document Retrieval

12/12/2020
Team: Ashwin Pothukuchi, Hongting Zhu, Joel Guo
Thanks to: Brian Noble, Kevin Hastie, and Daniel Dsouza
Background

- Large database with billions of records
- Authoritative sources for students & professionals
- Powerful and efficient search for medical papers
But...

- Difficult search query composition
- Inaccurate results from outdated algorithm
How to improve?

- Similar document suggestion with user-selected result
- Unsupervised learning
Dataset collection

- Corpora.xml from sponsors
- Popular research databases
- Comparison and evaluation
Deep Learning in NLP

- Generally higher performance
- NLP tasks include:
  - Machine translation
  - Document summarization and classification
  - Word prediction
- Popular deep learning models for NLP include:
  - Word2Vec, GloVe, fastText
  - RNN-based: LSTM, ELMo
- Transformer-based: BERT variants, XLNet, OpenAI Transformer
Document to Vector

- Represent document as a numeric vector which best describes it characteristics
- Cluster vectors based on vector cosine similarity
- Measured accuracy
- Compared with the old algorithm in use as baseline
Performance Evaluation

- K-Means Clustering on all documents
- Textual coherence
  - calculate Jensen-Shannon Divergence (JSD) for every cluster
  - Normalize and average the JSD for the entire corpus
- Grant-to-linkage evaluation (G2A)
  - Grant-based similarity
  - How many similar articles have the same grant
  - herfindahl-hirschman index
Result - Textual coherence

- Textual Coherence of new clustering solution: 0.198
- Textual Coherence of old clustering solution: 0.194
- Biased towards old solution, still promising to see new outperform old by 2%
Result - G2A

- New vs Old
- Higher precision, faster convergence
Conclusion

- Better similarity from retrieved documents
- Very low cost of evaluation instead of expensive human rating
- Great scalability for large corpus
Future challenge

- Integrate query information
- Ranker on similarity results
  - Telescoping
  - Learn-to-rank rankers
Thank you!