



Abstract-Paper Similarity

-Team ReddyCarry
(Manvith, Saransh, Shahbaz, Shanmukh)




Problem

- To create an accurate similarity matrix given abstracts and research papers as inputs
- Assign a similarity score ranging from 0 to 1, where 0 represents least similar and 1 represents exactly similar
- Ensure that the solution works for specific domain (5G, IoT, Neural Networks) as well as general domains
- Avoid using pre-trained models such as Word2Vec and GloVe



Approach

- Preprocessing - cleaning and tokenizing data using nltk libraries
- Get word frequencies using nltk libraries, generate set of frequent words
 - Use keyword/keyphrase extraction to add to set to improve results (explained later)
 - Size of set chosen based on testing results (too large or too small leads to problems)
- Run LDA (Latent Dirichlet Allocation) to assign topic distributions to documents
 - LDA is a Bag of Words (BoW) model
 - Crucial that stop words are removed so as to not impact frequency
- Model is run on list of research papers
 - Unseen data = abstracts, seen data = research papers
 - Model is **trained in runtime** in less than 20s

- 
- Jensen-Shannon divergence distance is used as a metric for similarity
 - Probabilistic model, 0 represents similar, 1 represents dissimilar
 - Similarity matrix requires the opposite - different methods to mitigate this involve subtracting from 1 (preferred implementation), or scaling the values by a number obtained by linearly mapping the reciprocal of the best match, i.e, the smallest value
 - Each solution has its disadvantages - for instance, the scaling solution fails if there are no matches larger than 50%



Incorporating keyword extraction

- Idea is to improve LDA by adding keywords to set of frequently occurring words
- Number of keywords to choose and add to set (chosen in decreasing order of keyword score) have been determined through testing on various datasets and weighing tradeoffs
- Keywords for input test data are generated by training during run-time on a custom corpus generated by scraping articles pertaining to 5G, IoT, Neural Networks, and other related terms
- For the general case, the custom corpus can be replaced by the Brown Corpus, support for which is available in the NLTK library



Thank you!