Machine Learning Engineer Nano-Degree

Capstone Proposal

Khachatur Mirijanyan

Domain Background

Currently, access to peer reviewed, academic research is both abundant and unprecedented. There has never been as much easy to access, high quality information as there is today. However, much of this information is unstructured and disorganized, making it more difficult to obtain the necessary information desired in a timely manner. This is especially important in times of crises, when experts need the most reliable and relevant information as quickly as possible. A popular solution to better organizing unstructured text data is through topic modeling.

Topic modeling has been used for a long time to determine the topics for a group of documents and decide the topical distribution of single documents. There are many different types of topic models such as LSA, LDA, and NMF. They all boil down to using the relationships between words in documents to decide what they mean and how it relates to the documents collectively.

I chosen to work on this due to my personal interest in NLP as well as the state of the current COVID-19 pandemic. The main data used will be elaborated on later, but it is basically a collection of research papers both directly and indirectly related to COVID-19. I believe that this project would be a good exercise for me, as well as potentially provide real benefit to researches

Datasets and Inputs

The primary dataset I will be working from is the CORD-19 dataset. It is a series of academic papers that either directly deals with COVID-19 or is potentially related to disease. The data was provided by many different institutions, including the American government, and subsequently provided in a usable format by major tech companies for exploration my data scientists.

I will be using a specific subset of the data that comprises about 25,000 academic papers. The papers in this particular subset have the full text of the papers, which would enable me to make a model on entire research papers, instead of just snippets.

The text in the data will be preprocessed and converted into the appropriate tokens which will signify the entire document. A corpus will be established from these tokenized documents and both will be used as inputs for various topic models.

Problem Statement

The primary question: How can all the accumulated knowledge and research obtained by experts be properly organized so that there is less time spent looking for the correct information?

One potential way of solving this problem is using a Latent Dirichlet Allocation (LDA) topic model. LDA is a generative probabilistic model, a description of which can be found in *Latent Dirichlet Allocation, Blei et. al* [1]. The model uses the latent relationships in the text to build topics, and then determines what the topic distribution of document is. Each of the topics has words associated with them that best represent the topic. These related words of each topic, as well as the topic distribution assigned to a document can determine what the document is about as well as how it relates to other documents in an unsupervised manner.

Solution Statement

The solution to the problem is to create the best topic model with the most coherent topics that best describe the documents. This will be done by create multiple topic models using different methods to find the best one. This will require significant text processing to ensure that the right amount and type of tokens are used. There will also need to be a lot of hyperparameter tuning to achieve find the best model. The main obstacle will be the number of tokens used to create the models, since too many tokens will overload my compute capacity.

[1] http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Benchmark Model

The benchmark used for this problem is a simple LDA I created using the abstracts and titles of the papers in my subset of CORD-19. The full implementation of the benchmark can be found within this notebook on my GitHub page[2].

The benchmark first establishes the specific subset of CORD-19 that will be used. Some of the information in the metadata does not match up with the actual files provided, along with issues like missing information and duplicate entries. Once these differences were examined and reconciled, only the papers with available abstracts and full texts were chosen. This resulted in a collection of over 25,000 documents.

For the benchmark model, only the title of the paper as well as the text in the abstracts were used. This reduced the number of tokens and made the model easier to compute, but most likely reduced its efficacy. Very simple text processing was done to prepare the text. Each document was first lowercased and tokenized. Then, all stop words were removed, which in this case was the default nltk stop words set. All completely numeric text was removed as well as any tokens that were one character long. A dictionary of the tokens was created, words that appeared too little or too often were removed, then the dictionary was prepared as a bag of words corpus, which is the input used in gensim's implementation of LDA. This resulted in 9691 unique tokens for the 25,339 documents.

I then used mostly the default parameters in gensim's LDA implementation. All hyperparameters were set to their default, including the number of topics which was set to 100. The only change made the number of passes of the model was increased to 10. I then calculated average *c_v* topic coherence as well as the relative standard deviation of the average topic coherence which came out to .4986 and .3056 respectively.

I believe this to be a simple implementation that can be improved upon, but still adequate for a benchmark model.

[2] https://github.com/kmirijan/ML-COVID-CORD/blob/master/CORD-metadata-explore.ipynb

Evaluation Metrics

The evaluation metrics used will be the average c_v topic coherence value as well as the relative standard deviation of the coherence value. A topic coherence value estimates how coherent the extracted words of a topic are. For example, the top 20 words of an identified topic can be used to determine how coherent a topic is. A further explanation of using coherence values in many different models can be found in *Exploring Topic Coherence Over Many Models and Many Topics, Stevens et. al*[3].

The coherence value I am using in particular is c_v, which is shown to have the best performance based on *Exploring the Space of Topic Coherence Measures, Roder et. al*[4]. Through these papers we can also see that there is a connection between effective models and good coherence measures. In the case of c_v, the measure is between 0 and 1, with 0 being non-coherent and 1 being overly coherent.

The topic coherent value will determine how good a model is. An average value around 0.7 mean that the topic is coherent, but it is not filled with words that are either overly related or are too similar. A very high topic coherence could signify too many topics or overfitting.

Relative standard deviation of topic coherence looks at the spread of topic coherence values across all topics. If this number is high, that means there are topics with wither too high or too low coherence, which could be a sign of too many topics.

Project Design

The entire workflow of the project will be like the benchmark, but with more data, more detailed data processing, and a few extra steps.

The project will start with data collection. Ideally, this will involve extracting the text from the papers, and using the full text of the papers as documents. This step will mostly depend on the computational feasibility of creating a series of models with potentially an order of magnitude more unique tokens.

[3] https://www.aclweb.org/anthology/D12-1087.pdf
[4] http://www.cse.chalmers.se/~richajo/dit862/L13/LDA%20with%20gensim%20(small%20example).html

Next, I will perform the data processing. This will involve many of the same steps as in the benchmark, but with more detail. I will increase the number of tokens by including the bigrams and trigrams of the documents into the corpus. I will also try to remove more words that are not related to the medical field or that could potentially aid researchers.

During the model creation I step, I will be doing a lot of hyperparameter tuning to create the best model. The most important hyperparameter I will be tuning for will be the number of topics. I will also be using at least one other model type outside of LDA. I hope to use NMF for topic modelling as well as continuing to do research and see what state-of-the-art topic models are being used today.

For the evaluation step I will continue to use the topic coherence values, but I will also try to look for other ways to measure topic separation and see if documents with similar topic distributions are indeed similar. This could result in checking my model's document similarity queries with a doc2vec similarity check. I am still researching this. I will also provide some visualizations to help interpret my model and its efficacy.