DSCI-560: Data Science Practicum Laboratory Assignment 8

Representing Document Concepts with Embeddings
Instructor: Young H. Cho

1) Accuracy Comparison Doc2Vec Embeddings

Return to the data and the results of the previous lab on data extraction, analysis, and clustering of Reddit postings. If you used doc2vec to represent your data into vectors, these vectors are called embeddings.

Doc2vec allows the user to configure the dimensions and values of the vectors including vector_size, min_count, epochs, etc. (https://radimrehurek.com/gensim/models/doc2vec.html) For this task, take all of the collected data from the previous lab and generate their embeddings/vectors using THREE different doc2vec configurations, especially with different vector size. Then, cluster the data based on the vectors and cosine distance metric.

Once the clusters are formed, examine the data in each cluster to argue which embedding configuration is better than others. There is not a right or wrong answer to your answer because of the subjective nature of this evaluation process. However, you are expected to do your utmost to assess and draw conclusions based on your qualitative and/or quantitative results.

2) Embeddings based on Word2Vec and Bag-of-Words

There are other ways to vectorize documents. One notable way to generate embedding for a document is to make a vector of word frequencies. To convert a body of words into a variable-sized vector of word frequencies, you will need to group/cluster the words into a specific number of bins first. For this task, one can manually bin all the words that are found in his/her data set based on the similarities and differences of the word definitions. Alternatively, you may accelerate this task by using another program called word2vec to vectorize all words in your data set and cluster them into specific bins based on the vector distances.

Therefore, go through the tutorial on the following link to get familiar with word2vec and its usage.

https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/

Vectorize then cluster the words (all or selected) found in the same data extracted from Reddit used above. The dimension may be adjusted to optimize the accuracy of the representation but keep in mind that it is only used for binning/clustering purposes.

Once the words are binned based on their distance metric, you can use it to create a vector for each document based on the normalized frequency of the words in each bin.

In other words, to generate a vector of N-dimension for a single Reddit post, you can use word2vec to cluster all words in all your data into K-bins. Then, for each post, you can partition all its words into K-bins and create a vector with K entries, each representing the number of words in each bin. To normalize the vector, you can divide all numbers by the number of words in the post.

Using this method, embed all the data with the vectors of the same THREE dimensions you experimented with using doc2vec. Compare and contrast the quality of the embedding methodology. Which of the two methods is superior for a specific dimension?

3) **Team Discussions**

Your team is expected to meet in person / virtually each day of the week and discuss the assignment progress & next steps. Document and compile minutes of all meetings in a separate file called **'meeting_notes_L8_<team_name>.pdf'**

4) **Submission**

Make one submission per team. Each team must submit all the code files for the working solution, a readme document containing information for running the code in PDF format, and a document that outlines the minutes of all team meetings in PDF format.

You must include a detailed GitHub history describing what each team member submitted.

Provide a video per team that demonstrates the entire working solution and explains how the data tables were loaded, demonstrates query results, and talks about the design decisions made along with reasoning for the same. Also, include details about how your team preprocessed the data. Please include the team's name and the names of all members in the video.

There will be a 50% penalty for all late submissions.