Homework Assignment 4: Clustering Word Embeddings

Assignment

Word embeddings are high-dimensional vector representations of words that are designed to capture some aspect of semantics, such that words that are more similar are closer together in vector space. One famous and early example of pre-trained word embeddings is Global Vectors for Word Representation, known as commonly as GloVe.

For this assignment, I'd like to you do use the different clustering techniques and dimensionality reduction methods we've covered to explore three different newsgroups from the 20 Newsgroups data set in Scikit-Learn. Please process the three newsgroups that have to do with religion.

Your goal is to obtain the "best" clustering you can. Use different dimensionality reduction methods (e.g., PCA, tSNE, UMAP) to project the data, and then different clustering methods (e.g., K-Means, HAC, HDBScan). Try at least 2 dimensionality reduction methods and 2 clustering methods. You can define “best” however you want — this could be subjective (explain your criterion) or data driven (explain your choice).

1) What happens if you project the data before you cluster? What happens if you don't? Explain which you prefer and why.

2) What are the major clusterings in the different newsgroups? Please summarize them for each of the three newsgroups. *Note: A summary goes well beyond a list of clustering elements, and tries to capture the underlying themes or characteristics that define each cluster within the newsgroups.*

3) How do the three newsgroups differ? How are they the same?

A Jupyter notebook has been provided to get you started. Answer the above questions in this notebook, in cells you develop after the cells that have already been provided.

Note that even though we are using stop words and TF-IDF to filter the term list, a number of terms sneak in that aren't really relevant (e.g., com, edu). Feel free to remove the high-frequency terms that you don't think contribute to your understanding of the data.

Please submit everything in the provided Jupyter notebook. There is no need to write a separate markdown file, but be sure to include text in your notebook that answers each of the above questions, along with visualizations. The text in your notebook should be nicely formatted, and free from grammatical and spelling errors. Remember to use markdown cells with headings to clearly delineate the sections of your notebook.

Rubric

1. Did you explore and identify an effective approach for clustering/dimensionality reduction?
2. Did you explore and respond to the 3 questions above?
3. Is your Jupyter notebook well formatted and free from errors. Does it include well-written answers, free of grammatical and spelling errors?