HOMEWORK - 3

- Do you think the way to quantify concreteness makes sense? Any other thoughts to measure concreteness or abstractness? Share your ideas in pdf.

  Yes, I think the way we are quantifying concreteness in this assignment makes sense to me. We are using articles, adpositions, and quantifiers to find out the concreteness of the text given, the method involved is quite veracious for me.

- Do you think this method is able to generate a good summary? Any pros or cons have you observed?

  1) Yes, I think this method(metric-cosine) is able to generate a good summary to a great extent. I have tried 5 other metrics below the cosine method and some other methods in the bonus question - I can clearly see a similarity in the summary of all other methods to this method and hence we can say that this method generates a good summary.

  2) Cosine similarity works best when there are a great many (and likely sparsely populated) features to choose from. Likewise, cosine similarity is less optimal under spaces of lower dimensions.

- Do these options lemmatized, remove_stopword, remove_punctuation, and use_idf matter? Why do you think these options matter or do not matter? If these options matter, what are the best values for these options?

  Yes, these options matter. Having options like these is an excellent benefit to the function - as it gives the control of the function in the hand of the user. Lemmatizer allows to decrease noise and speed up the user's task. Lemmatization provides better results by performing an analysis that depends on the word's part of speech and producing real, dictionary words. Remove_stopword and remove_punctuation are helpful as the dataset size decreases and the time to train the model also decreases. Removing stopwords and punctuation can potentially help improve the performance as there are fewer and only meaningful tokens left. They could overall increase the classification accuracy. The option of use_idf lets us choose whether we want the function to return tf or smoothed_tf_idf and is quite useful at times.

BONUS QUESTION

- Can you think of a way to improve this extractive summary method? Explain the method you propose for improvement, implement it, use it to generate a new summary, and demonstrate what is improved in the new summary.

  I have tried out 6 other methods for extracting the summary and they are the following

1) Cityblock Similarity

2) Euclidean Similarity
3) L1 Similarity
4) L2 Similarity
5) Manhattan Similarity
6) Jaccard Similarity

I have observed that city block, L1, and Manhattan Similarity extract the same summary. While Euclidean and L2 extract the same summary. Jaccard and cosine give out a slightly unique summary from the rest. I don't see any huge improvement from the cosine summary. I would also like to add that - Jaccard similarity is good for cases where duplication does not matter, and cosine similarity is good for cases where duplication matters while analyzing text similarity. Eg - For two product descriptions - Jaccard similarity will be a better choice.