Homework 1

Due: October 20, 11:59pm CDT

Instructions

This homework consists of two parts: Written Questions and Coding Tasks. For the written questions, please respond to each of the questions and upload a .pdf file to Canvas. Remember that some questions may be fairly open-ended, so do the best you can; there is (usually) no formulaic right or wrong answer (except no answer).

For the Coding Tasks, I am providing a partially filled-in Jupyter Notebook which you will have to complete. Simply fill in the missing functions and upload your completed file, along with the dataframe you'll be constructing.

Written Questions

Question 1

1. Briefly describe the geometric metaphor of meaning. If you need to, consult the second chapter of Magnus Sahlgren's dissertation (which is listed in the syllabus and freely available online).
2. What strengths does the geometric metaphor of meaning have?
3. What weaknesses might the geometric metaphor of meaning have?

Question 2

1. Briefly describe the distributional hypothesis and the related distributional methodology.
2. How do these relate to the geometric metaphor of meaning?

Question 3

1. Formally define a distance metric. Take the time to understand the intuition behind each of the requirements a function must satisfy to qualify as a distance metric.

2. How does a distance metric differ from a similarity metric?

3. We've discussed two metrics thus far, Euclidean distance and cosine similarity. Formally define each of these (i.e. provide an equation).

4. Given a hypothetical embedding of documents in term space,[1] consider two hypothetical functions.

   - **Closest**$_{Euc}$ takes as an argument doc_i and returns $\underset{i \neq j}{argmin}(dist_{Euc}(doc_i, doc_j))$
   - **Closest**$_{Cos}$ takes as an argument doc_i and returns $\underset{i \neq j}{argmax}(sim_{cos}(doc_i, doc_j))$

   Will the two functions always agree? Why or why not?

Coding Tasks

Question 4

1. For the *preprocess()* function, describe and justify each of the steps you took in preprocessing the text. For example, if you lowercased all the characters, explain the rationale behind the decision.

Bonus

Given your embeddings, try to find interesting patterns in your corpus and word embeddings. Without using a *for*-loop, what is the highest cosine similarity score (and, lowest Euclidean distance score) you can find for any pair of movies? For any pair of words?

[1] In other words, given an $N \times D$ matrix, where we have N terms and D documents, consider the document embeddings, where each document is represented by a vector in $\mathbb{R}^N$, each term serving as an axis.