CIS530 HW4: Submission Example

5. Free-response Questions (10 points)

Results

- Task: Top Five Most Similar Character Vectors to “Macbeth”
- Results
    - GLOUCESTER 0.9545800025616389
    - HAMLET 0.9527282098881199
    - KING HENRY V 0.9518306505244132
      YORK 0.9494973495443843
    - MARK ANTONY 0.9463861448664543
- Discussion
    - When comparing different vector space models and similarity functions, my model found that the five most similar characters to Macbeth were Gloucester, Hamlet, King Henry, York, and Mark Antony. We used tf-idf vectors to eliminate the similarities created by common words and placed more emphasis on less common vocabulary. This model interpretation is producing accurate results. After some research to verify our answers, we are confident that all of these characters share similar themes which are expressed through their dialogue.

Extra Credit (15 points)

Results

- Explanation
    - The purpose of this experiment was to quantify the goodness of one vector space representation over another. In this experiment, we were tasked with creating a code pipeline that would measure the correlation and compare that with human estimates. The dataset displayed showcases a pipeline that created results from a positive pointwise mutual information matrix (PPMI) which was then fed into the cosine similarity function and dice similarity function. The cosine value is useful because it maps words with readable mathematical values - through 1 representing the relative direction of both semantic vectors: 1 indicates the vectors are similar, and -1 indicates they are opposite. After creating a list of similarity scores between the word vectors, we compared that with our SimLex data and calculated the tau coefficient value. During the experiment, we encountered practical resistance in the space complexity of the available servers.

Utilizing list comprehensions to reduce time spent searching was vital to solving the limited size of the servers.

- Results
  -

Similarity Function	Window Size	Tau	Function Score
Cosine Similarity	3	.0142	.806
Cosine Similarity	4	.028	.622
Cosine Similarity	5	.054	.352
Dice Similarity	3	-.0147	.801
Dice Similarity	4	-.004	.935
Dice Similarity	5	.020	.730

  -
- Discussion
  - We experimented with modifying the context window size from 3 to 5. Although the Tau values showed a slight upward trend in one run t = 0.01 → 0.05, the correlations remained very close to zero with large p-values and the function scores decreased approaching zero. This might indicate that the accuracy of our function increased as the window size increased. The setting that best appears to improve performance would be a window size of 5 utilizing cosine similarity function to evaluate the similarity between the words.