NLP Homework

Data and code given to you:

- glove.6B.50d.10k.mat: A set of embeddings from GLoVE of 10,000 words.
- word-sim.csv: Human rated scores of word similarity.
- readWordVectors.m: A function to extract words and embeddings.
- word2vec.m: A function to look up the embedding of a single word.
- testAnalogies.m: A function to test your analogy.m code.

Usage:

```
[vectors, words] = readWordVectors('glove.6B.50d.10k.mat');
embedding = word2vec(vectors, words, 'shoe');
```

1 Word Similarity

Implement the function wordSimilarity.m which computes the cosine similarity between two given words.

Test it on a few words of your choice to make sure that the results are reasonable.

For example, "cat" and "pet" should return a high value (close to 1.0), while "jump" and "coast" should return a lower value.

Now, implement the function computeBehaviorCorrelation.m. In this function, you will compare human rated similarities with the word embedding similarity measure you implemented in wordSimilarity. Follow the comments in the code to return the Spearman correlation. You should get a value of 0.6298.

2 Closest Words

Implement the function mostSimilar.m that will return the 10 closest words (in terms of cosine distance), to the given embedding.

Hint: The norm and vecnorm functions may be useful.

When that is working, implement plotMostSimilar.m to plot a word and its 30 closest words. You will need to project the embeddings onto 2D space using

PCA.

3 Analogies

Finally, implement the `analogy.m`. Use `testAnalogies.m` to run your function through a set of analogies. You should get 13 out of 13 correct.

Next, find 2 analogies that don't work and 2 additional ones that work (they must be different than those given in `testAnalogies.m`.

4 Clustering

This is an exploratory analysis. Run KMeans on the vectors with different number of clusters. Write a paragraph or two on the clusters you found, whether they're meaningful of not.

5 Reference/Acknowledgements

This is just for reference.

- GLoVE embeddings: https://nlp.stanford.edu/projects/glove/
- Word similarity: http://alfonseca.org/eng/research/wordsim353.html