SLP Assignment F October 8, 2020

**Word Embeddings**

In this assignment you will explore the use of vector representations of lexical identity.

Part 1. For your Information Retrieval assignments, briefly describe how you could use word vectors and what benefit you would expect. Please answer concisely, with a short paragraph for each.

By using word embeddings, in combination with, for example stop lists, we could theoretically combine the vectors of the descriptions of professors’ interests provided to us, and then observe what words are close to the combination of vectors. For example, we could talk about combining the phrase “deep learning for the advancement of stochastic methods”. Here we could prune for, the, of, advancement and combine the vectors of deep, learning, stochastic, methods. Hopefully by making combinations like the previous combination of the words deep, learning, stochastic, and methods we map to the neighborhood of Artificial Intelligence, and hence with this knowledge produce better predictions when we query for “Artificial Intelligence”.

Also, by using similarity scores between documents (e.g. cosine similarity score and tanimoto) we could implement the use of word embeddings to compare the vectors of the query words to the vector of the description document (the addition of the word vectors of the document). Then we could compute the distances between the vector for the description document and each of the query words-vector and that way make an inverse-distance calculation that adjusts the similarity scores, so we can expect lower distance vectors and some similarity to be greater than some similarity and a lot of distance. That way we would eliminate options that seem to be valid because of some mentions of query words but for which our query words were not important in the context.

Part 2. Get familiar with the use of some set of pretrained dimensions. Links to word2vec, glove, and Ratnaparkhi’s bitstrings are on the course homepage, but for this I recommend the glove 50-dimension embeddings. Specifically, determine whether *rabbit* or *bunny* is closest to *food*, and determine whether *texas-el* is closer to *juarez,* to *chicago* or to *rabbit.*

Text

Description automatically generated

In this case by using the embeddings provided by Glove 50d I saw that the vector food is closer to rabbit that to bunny.

Text

Description automatically generated

In this case I observed that the closest word to “texas-el” was “juarez”.