Homework 4

11-791

Fall 2013

Callie Vaughn (cvaughn)

Task 1.

I first found tokens by splitting on punctuation and spaces to isolate words, just as in Assignment 2. The vocabulary consists of the set of all tokens found in any document. Each token's frequency (within document) was then found and stored in its Frequency field. These raw frequencies were stored in document weight vectors. Each document vector had one field for each word in the vocabulary. The field for a given token was set to its document-specific frequency.

Once I had the document vectors, I separated the documents into sets based on their QueryID. I then found the cosine similarity between each query and each document that was associated by it via QueryID. Then I ranked the potential answer documents by their cosine similarity (so that the document with the highest cosine similarity to the query was given the rank 1, and so on.)

Once I had the rankings, I calculated the Mean Reciprocal Ranking of my system. For my baseline system (as described above), I got MRR=0.6666666667 on the extended data set (including the two additional queries that were added from Piazza).

Task 2.

I examined the errors that I made, and noticed that capitalization might have been hiding useful information for my system. (For example, the query "It takes a long time to grow an old friend" has a lowercase "old", whereas the correct answer "Old friends are best" has an uppercase "Old" and therefore a cosine similarity of 0.) This also led me to suspect that my frequency weights might not be sufficiently sophisticated, so I changed the raw document frequencies to the more complicated tf-idf measure. (Specific information on the calculation of tf-idf can be found here: http://en.wikipedia.org/wiki/Vector_space_model.)

These changes improved my MRR to 0.7666666.