11-791 HW3 Report
Nikolas Wolfe

Analysis

Using a simple space-delimiting tokenizer and cosine similarity as our metric, the example retrieval exercise returns only 1 correct answer out of 20, i.e. it was only able to rank the correct answer first in one instance. Analyzing the mistakes, we find some recurring issues:

1. **Answer Length**: Cosine similarity measures are sensitive to the length of a document. If a correct answer is too wordy, it may be ranked lower than it ideally should be. For instance, in answer to the question "*What has been the largest crowd to ever come see Michael Jordan?"*, the correct answer from our document corpus is as follows:

   "When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last March, an NBA record 62,046 fans turned out to see him and the Bulls."

   Using a cosine similarity measure, the following was deemed a more likely answer:

   "*A supposedly last play of Michael Jordan gathered some of the largest crowd in history of NBA."*

   The correct answer was obscured by superfluous information. If the document had been shorter, it would have been ranked higher.

2. **Tokenization**: Splitting a document only on spaces and not stripping punctuation can lead to identical words being tokenized as different words because they are appended by punctuation mark(s). For example, given the question *"Give us the name of the volcano that destroyed the ancient city of Pompeii",* we see that in every answer the word *Pompeii* is appended by a punctuation mark, e.g. "*Pompeii."* and *"Pompeii;"* An obvious problem here is that Pompeii is being tokenized as three different words, rather than just one.

3. **Stop Words**: Words like "the," "an," "a" and other commonly used words carry little semantic information. Counting these in our word vectors leads to some obvious distortions.
4.