Determining Relatedness of Textual Data

Advisor: Professor Rafail Ostrovsky

Justin Lu
Previous Work

• Developed a system of relatedness that utilizes fast dimension reduction to its advantage (i.e. uses some large vector representation)

• Implemented the actual dimension reduction and examined the preservation of information

• Emphasis on an efficient space representation
Result

- Preliminary results were a little disappointing as it appears information wasn't being preserved very well.
- Trying out other approaches to the problem that may hold promise.
- First implement the common machinery, then experiment with different representations.
New Approach

- Avoid dimension reduction altogether while still trying to maintain compact representation
- Continue to use the common word information and large data set to generate these signatures
- Instead of keeping all information, just keep the indices of the top few hundred results
Implementation

- Randomly grabbed a hundred web pages, this represents what real input will look like.
- Implemented the scheme on a large data set. ~3 million Wikipedia articles.
- Used hash table where a word is the key and the values returned are a list of all the articles indices that contain the word.
Finally, will implement a system where given some search terms, an appropriate certificate can be generated.

This will be done by finding the closest matching articles and using those as a basis for comparison.

This would be a nice demonstration for the utility of such a system.