CS 194 Checkpoint 1

Determining Relatedness of Textual Data

Advisor: Professor Rafail Ostrovsky

Justin Lu

Objectives

- Two main steps
- Develop a system of relatedness that utilizes fast dimension reduction to its advantage (i.e. uses some large vector representation)
- Implement the actual dimension reduction and verify that desired properties are preserved
- Emphasis on an efficient space representation

Systems

- If the system doesn't produce good results even before dimension reduction, it's useless
- To test different approaches used a subset of Wikipedia articles as a data set (2000 articles from a total of 2 million)
- First Approach: Had vectors represent word presence-Produced poor results, something more sophisticated is needed

Systems

- Second Approach: Utilize the number of words each pair of articles have in common-produced good results but scales poorly in time and space
- Third Approach: Utilize the data from the second approach but choose a different representation-produces decent results, is space and time efficient, and has some other advantages

Representation

- Each article was given a n-element vector, the ith element denotes the number of common words with the ith article
- Vectors can be compared to determine relatedness
- If the total number of articles, x, is small then x can equal n, but more generally n can consist of a random subset of articles, this keeps time and space complexity down

Representation

- If using a significance threshold then most of the vectors became sparse-this useful
- Produces good results even with small test n=200, larger sample sizes and data sets should produce even better results
- Top 3 results for article Bay of Bengal: Indus River, Ganges River, Mackenzie River (note: this is out of just 2000 articles)

Goal Assessment

- Met the first goal of finding a good representation
- Implemented Fast Dimension Reduction via type of Johnson-Lindenstrauss transform
- Distances are somewhat preserved, more refinement is needed
- After that will focus on filling in the details: tweaking word choice, much larger data sets, getting a good reference vector