Results

Determining Relatedness of Textual Data:
An Implementation utilizing Dimension Reduction

Advisor: Professor Rafail Ostrovsky

Justin Lu

The Problem

- How, given large sets of arbitrary data (e.g. text from webpages), to determine their relevance to one another
- Has applications in things like better searching and data mining
- How to deliver a good level of quality while still remaining computationally feasible

Work

- Developed a system of relatedness that utilizes fast dimension reduction to its advantage
- Implemented the actual dimension reduction and examined the preservation of information
- Emphasis on an efficient space representation

Methodology

- Use a bit vector to model each piece of data
- Used Wikipedia as a basis for data
- Each bit is the number of common words with that article

Implementation

- Implemented using hash tables to obtain word count
- Used about 3 million Wikipedia articles
- Filtered out common words

Testing

- Randomly picked a hundred pieces of data and obtained representations
- Picked a specific one and compared the vector distance to it
- Ranked the above, this is the canonical” version

Testing

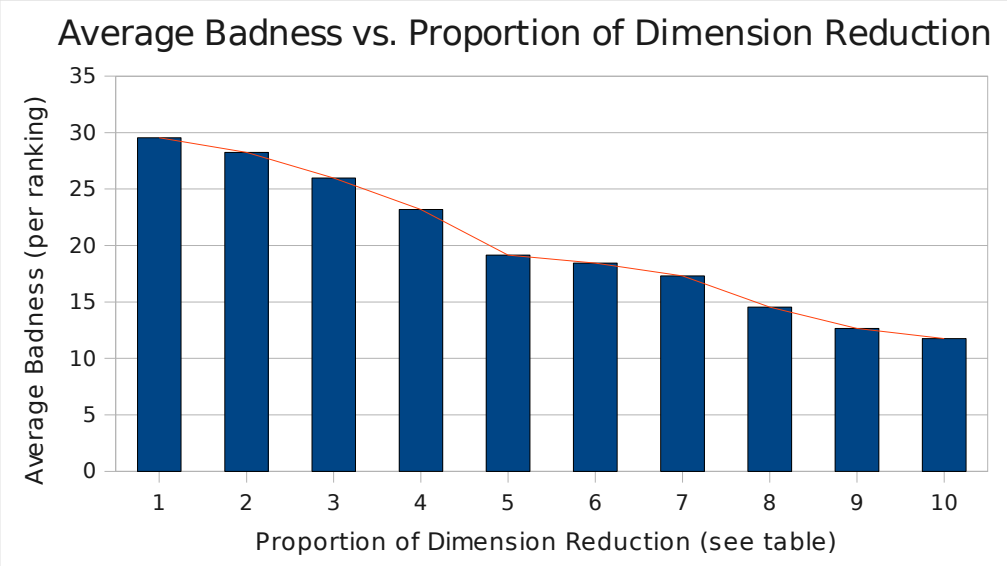
- Implemented dimension reduction via Johnson Lindenstrauss Transform (Ailon, Liberty)
- Reduced the representation to 128 points of data, ranked the same articles using the new representation
- Determined badness from canonical version

Testing

- Compared vs. random basis reduction
- Reduction from d (original basis) to k (target basis)
- First reduced by random basis subset selection to an intermediary representation with a basis i
- Then this intermediary representation was further reduced via JLT to k, the target size.

Data

- d = 65536, k = 128



Number	Intermediary Basis (i)	Average Badness
1	128 (Full Random Basis Reduction)	29.56
2	256	28.25
3	512	25.98
4	1024	23.19
5	2048	19.16
6	4096	18.43
7	8192	17.31
8	16384	14.55
9	32768	12.65
10	65536 (Full JLT Dimension Reduction)	11.76

Conclusion

- Data shows a smooth decrease in badness as more and more JLT dimension reduction is used
- This demonstrates the superiority of using JLT dimension reduction vs. picking a subset for the basis