Project Proposal

Group Name:
No Errors No Warnings
Members:
Ken Chen
Shanglun Li
Shaojie Ma
Shuyan Huang

Dataset:

- Data size: 2.25GB
- Records: 61,424 separate files
- Time range: 1996 - 2007
- Details: Our data is scraped from the online annual financial reports of the US listed companies. Every file is the essentially the first Item (business) of the company's 10-k report in that year.

Hypotheses to test:

Regularly, companies' financial reports are structured in very stable patterns, which highlights the information regarding the company's financial performance and business outlook. The salient readability of these files makes it possible for our algorithms to learn their latent topics and the linkage across 10-k reports. We will test the following hypotheses:

- After converting all the words that appeared in these files into vector representations, we might be able to annotate the words into groups of clusters and devise the principal features of each cluster. By studying these features, we may be able to identify what are the underlying topics and compute the loadings of each file on these topics.
- Additionally, we expect to see the wording of 10-k files will evolve over this time period, by studying the how topic loadings change over time.

Algorithms:

Here we propose two primary algorithms that we plan to employ:

- KMeans: we expect the algorithm to help unveil the latent topics associated with these files. With the help of Word2Vec, all the words in these 10-k files will be embedded into vector representations. The Kmeans algorithm is them utilized to cluster the words and then helps to induce the latent topics.
- Single Vector Decomposition: after the clustering analysis, we will employ the SVD algorithm to compute the loadings of each document on these topics. We might be able to see the distribution of the document topics and how that changes during the time span between 1996 to 2007.