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 paragraph 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 these text files into vector representations, we might be able to annotate the files 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 factors that make the companies look alike or distinguish them from one another (like the industries they play in, or the profitability, etc), in view of the 10-k wording.
- Additionally, we expect to see the wording of 10-k files will evolve over this time period, by studying the changes associated with the principle features.

Algorithms:

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

- KMeans: we expect the algorithm to help separate the 10-k files into different clusters. As we mentioned, each file will be embedded into a vector representation, with the help of Word2Vec. The files can be clustered using the KMeans, based on their vector values.
- Single Vector Decomposition: after the clustering analysis, each cluster can produce a matrix by staking up all the vectors. We can compute the loadings of the words on the factors, which help us to identify the underlying topics associated with the clusters.