LDA

What we did

Why we chose it

What we expected

What we actually got out

# LDA

## What we did:

One method that is common for text analysis is Latent Dirichlet Allocation (LDA). This method groups documents into topics. The results of an LDA is a probabilistic grouping of documents based the terms in the entire set of documents.

The jobs data was cleaned to remove extraneous characters and to remove terms which we did not want to be used in the category. Specifically, our business purpose is to rank jobs based on job similarity, so the name of the company was removed from the text of the document. Otherwise it was likely that the name of the company repeated multiple times in a document would become a strong influence in the document breakdown in this analysis. Additionally, the standard text cleaning methods of converting to lowercase, removing punctuation, and cleaning whitespace were performed. The corpus was then converted to a document term matrix (DTM).

Next, an LDA was run on the DTM. The first model run was with the number of topics set at k=5. The next model was at k=50. Due to the time consuming nature of running the LDA, these were the only allocations that were run. The top 10 terms for each lda was produced to get a representation of the major topics in each allocation. Next the topic probabilities were calculated to identify the increased likelihood of a topic based on each term. Additionally, term probabilities were calculated to identify the probability of each term occurring in each of the topics.

The resumes were loaded into a document term process using the same clean-up and process that was used to prepare the jobs data for the LDA. Next, the topic likelihoods and term likelihoods were calculated for the resumes using the posterior() function to calculate based on the previous LDA. The most likely topic for each resume was also calculated.

The next analysis that was planned was to use the LDA to identify the most relevant words (based on a not-yet identified cut-off) for that document. Then we anticipated that we would be able to search resumes for those particular terms only and perform a distance measure from each topic. Additionally, it was anticipated that an analysis would be performed to identify how much of the data was covered by the LDA. Currently it is unknown whether the model is underfit or overfit. Further, more analysis can be performed to identify the characteristics of each topic more than just getting the top 10 characteristics.

## Why we chose it