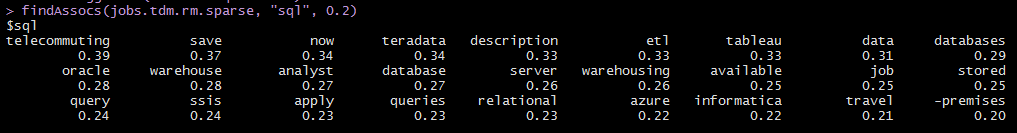
# Evaluation

Further evaluation is required in the future, as only minor evaluation of resumes has been accomplished (per the business purpose).

The TF-IDF has successfully identified both the term frequency and inverse document frequency. Additionally, the model is capable of doing word associations for any input word to determine how well it correlates with the other terms in the document set. The next steps required for this is to expand this are to find the most important words in a document and then run the association for each of those terms. Alternatively, it would be useful to take words from resumes and run those through the association function to determine which related topics are useful for candidates.



The LDA model is successfully able to identify categories of documents. Additionally, the model is successfully able to take an input resume (in text format) and identify which topic(s) the resume best matches. This is possible for both 5 topics and 50 topics. An example output for the 5 topic model is in the figure below. This model suffers from having too many terms which are not relevant to real world HR review of resumes.



The kmeans clustering is likewise capable of identifying categories of job postings, and identifying the most probable category for a resume. However, even more than the LDA, this model suffers from very frequent words which are not critical to HR reviewers.

Further analysis is necessary to evolve these models. The team had intended to develop a dictionary of critical terms and processes using text analysis. However, due to time contraints this was not possible. This is the most important next step because it will enable the evolution of the categorization models and will enable simple word-find checks of resumes.