Recommended Resumes using NLP and Unsupervised Learning

Mitchell Seiter

Abstract

The goal of this project was to use natural languages processors (NLPs) to build a topic model based off of resumes. From the topic model a recommendation system was built that would recommend resumes based off of a job description. The recommendation system leverages the topic word matrix that is built from the topic model, and finds the resumes that most closely relate to the job description.

Design

This project stems from the need for recruiters to have to go through a myriad of resumes and try to filter them out manually. This resume filter would drastically reduce the amount of time spent on the preliminary screening process when trying to find suitable applicants for a job.

Data

The dataset consists of a little over 1000 resume text documents with an average of 3100 words per document. These were unlabeled resumes and the text is the only feature available. The dataset is available for view in my github.

Algorithms

Text preprocessing

1. Removing a lot of imported new lines that appeared in the dataframe
2. Lowercase all of the words to make standardization easier
3. Tokenized the words to separate each individual word
4. Lemmatized the text documents to make similar words the same
5. Removed named entities such as places and peoples names that appeared in the resumes
6. Removed standard stop words plus some that were specific to resume such as “skill” , “resume”, and “experience.”

Topic Modeling

Used an LDA topic model as the text documents were a little bit larger and this seemed the most appropriate. The topic model had 2294 unique words and consisted of 25 different topics.

Recommendation System

The recommendation system used sklearn pairwise distance to calculate the cosine distance between the input job description and the resumes in the corpus. It would then return the top 5 most closely associated resumes.

Tools

- Numpy and Pandas for data manipulation
- SPACY and NLTK for test preprocessing
- LDA and SKLEARN for modeling
- Matplotlib for plotting
- pyLDAvis for visualization

Communication

This material will be presented in a 5 minute slide presentation that will be available on my github.