**S&P 500 Company Job Description –Text Analytics Report**

**Student name: Bojian Zhang, Minghan Wang**

**Submission date: 09/09/2018**

## Introduction

About 120,000 rows of job descriptions of S&P 500 companies were obtained from Glassdoor website. After cleaning and normalizing the text data, we extracted features for each industry and calculated the similarities between each industry by cosine similarity score. We also generated trigrams that are TFIDF weighted. The purpose of this project is to find out key features of job descriptions of each industry and compare among them.

## Data Description

1. Crawling process
   1. The period of crawling is from June to July in 2017, (job description is also around similar time, April 2016 earliest time, mostly May, June, July 2017)
   2. Used the module of urllib2 in Python and regular expression to design the crawling program.
   3. The job information in every row includes id, url, title, industry, company name, company score, review number, job type, city, state, posted date and job description.
2. Data statistics
   1. Data includes 446 companies from S&P 500.
   2. Crawled 118,677 rows data.

With stop words

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Average Len | Min | Q1 | Median | Q3 | Max | Std | Mode |
| 618 | 3 | 402 | 574 | 787 | 4,712 | 309 | 452 |

Without stop words

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Average Len | Min | Q1 | Median | Q3 | Max | Std | Mode |
| 343 | 0 | 227 | 321 | 435 | 2,377 | 167 | 245 |

1. Industry sector

GICS (<https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard>)

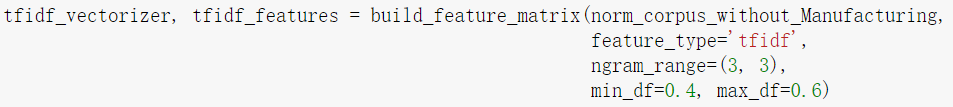
## Normalization

First step is stopwords removal and tokenization. Several words such as ‘gt’, ‘br’, ‘li’ are added to the stop word list to make the results better. We also remove punctuations and numbers, only text corpus is kept.



## Feature Extraction

In this project, we implement TFIDF model to build feature matrix which will be used to calculate similarity score among industries. We set minimum document frequency as 0.4 and maximum frequency as 0.6. We only extract trigram phrases as features.



## Document Similarity and Network

After extracting features for each industry, we calculated the similarities between each industry by cosine similarity value.

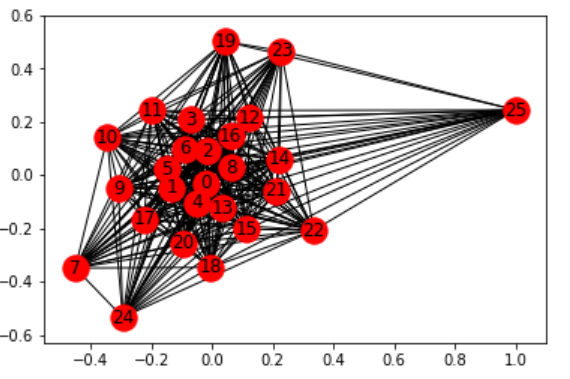


Figure1. Industry network



Table1. Industry indices

Using the indices defined in Table 1, Figure 1 shows a very clear representation of the similarities between each industry. From above network, we can see that manufacturing is in the center of all industries, which means job descriptions of manufacturing include some skills or experiences referred by most of other industries. Meanwhile, biotech, pharmaceuticals, information technology, oil, gas, energy, utilities and business service industries have highest similarity score with manufacturing, it is perhaps that these industries are closely related to manufacturing. In contrast, government, arts, entertainment, recreation and utilities industries have some specific job requirement that they have least similarity with other industries.

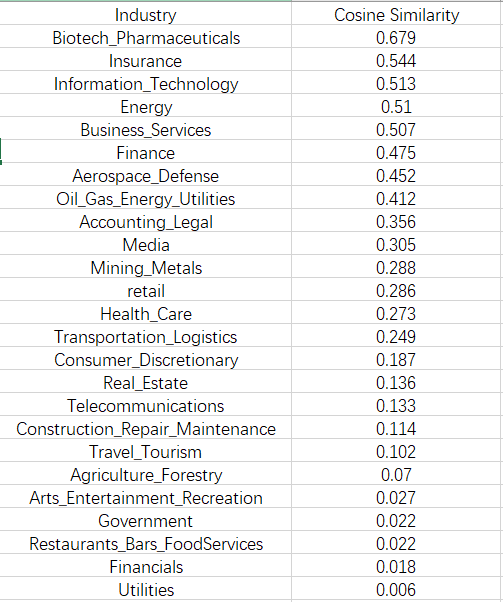


Table 2. Cosine Similarity of 25 Industries Comparing to Manufacturing

From the above table, we can see that biotech and pharmaceuticals, insurance and information technology are the three industries that are most similar to manufacturing. It infers that the job descriptions are rather similar within these four industries and maybe job positions in these industries requires similar skills.

## TFIDF Weighted Trigram Key Phrase Extraction

For each industry, 30 features were generated. Within each industry, we first extract features for each job description by term frequency and then used TF-IDF to get features for the whole industry.

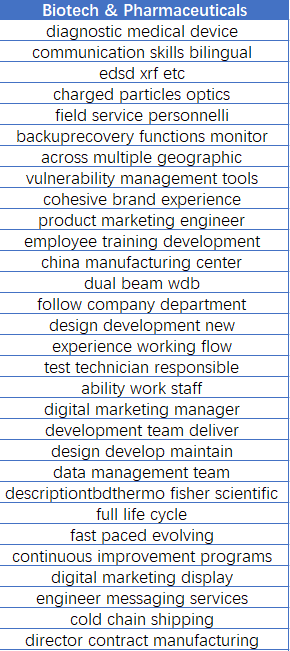


Table 3. Key phrases of Biotech and pharmaceuticals industry

For more details of key phrase of each industry, please refer to appendix. Table 3 displays key phrases of biotech and pharmaceuticals industry. Some phases indicate characters of this industry, such as ‘diagnostic medical device’, ‘communication skills bilingual’, ‘vulnerability management tools’ and most of companies require candidate to work ‘across multiple geographic’.

## Appendix

