Resume Matching System

**Introduction:**

Text similarity (hence referred to as "text similarity") is a widely used concept in a variety of tasks, including question answering (Lin and Pantel, 2001), automatic essay grading (Attali and Burstein, 2006), and paraphrase recognition (Dolan et al., 2004). In contrast to the notion of similarity in psychology, which is thoroughly explored and contained in formal models such as the set-theoretic model (Tversky, 1977) or the geometric model, text similarity is frequently used as an umbrella term spanning a wide range of phenomena (Widdows, 2004). We show that the seemingly straightforward question of "How similar are two texts?"

One of the use can be Resume matching, for each job ad, companies often receive thousands of resumes and hire dedicated screening officers to select qualified candidates. For all firms, finding the proper talent is a struggle. If the business is labor-intensive, growing, and experiencing high attrition rates, this difficulty is amplified by the large number of candidates.

Growing markets are scarce in IT organizations. Professionals with a variety of technical talents and business domain expertise are employed and assigned to projects to fix customer difficulties in a typical service organization. Resume screening is the process of identifying the top talent from a large pool of candidates.

Large firms typically do not have the time to examine each CV, so they utilize machine learning algorithms to screen resumes. Companies often receive thousands of resumes for each job posting and employ dedicated employees to screen qualified candidates.

Hiring the right talent is a challenge for all businesses. This challenge is magnified by the high volume of applicants. Automating the screening of resumes help reduce hiring cost for companies.

**Motivation:**

My aim of the project is to build the Resume Matching system with the help of cousin similarity and tf-idf embedding. Job description is used as one the documents and pair-wise cosine similarity is calculated with each resume document. Following key questions are investigated.

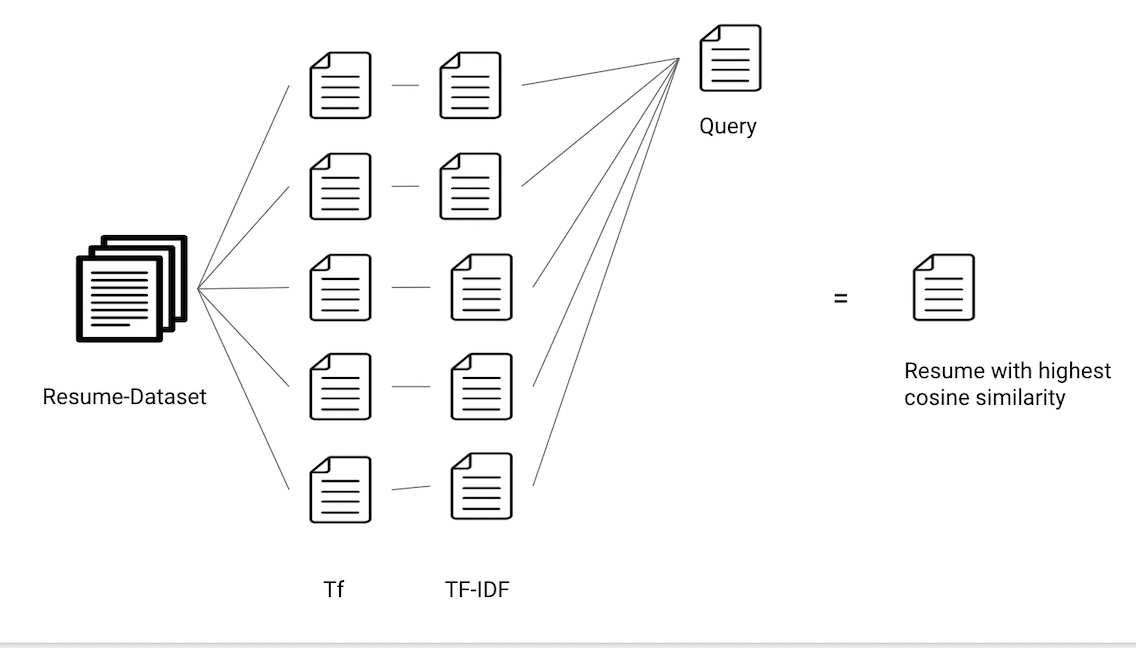
● What threshold for cosine similarity score can be used for resume screening?

● What are most used words in resume dataset?

● What is the variance in the resume dataset? - How far cosine similarity scores are spread out from their average value.

How to solve the problem:

* Converting the data in text format most of the resume were is either pdf or word format.
* Preprocessing and Tokenization.
* Term document frequency of each resume.
* TF-IDF for each resume.
* We have one document for the job description.
* Then with the help of these values we will find the similarity of the query document with each resume. The resume with the highest cosine similarity will be considered as the most matched profile.



* TF-IDF:

The Term Frequency-Inverse Document Frequency (TF-IDF) technique is a well-known text processing technique. This method allows you to assign a weight to each term in a document. The weights of terms that appear frequently in a document are high. Furthermore, phrases that appear often throughout the document corpus have lower weights.

* Cosine similarity:

Cosine similarity is a measure of similarity that quantifies the cosine of the angle between two non-zero vectors in an inner product space. The cosine of 0° is 1, and the cosine of any other angle in the interval [0,2] is less than 1. It is thus a judgement of orientation rather than magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90 degrees have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, regardless of magnitude.

* Dataset:

Dataset link - https://github.com/arefinnomi/curriculum\_vitae\_data

● 2500 dataset of resumes

● The most of the files are collected from random websites in this dataset

● Most of the resumes are from India

● Language of the dataset is English

Parallel computation: In the project parallel computation is done while calculating the tf values and idf ,tf-idf, and cosine similarity for documents.

**Solution:**

1. What threshold for cosine similarity score can be used for resume screening?

Removing stopwords:



Spark command to find cosine similarity:

spark-submit TextAnalyzer.py -m=SIM -i="job.txt" -allFiles="word/\*" -o=simjob1.sim

Tf values:



idf values-



1. The similarity for the first job description which is in text file job.txt:

“In-depth knowledge of Solar power generation systems.”

Match with the resume number 464 with the cosine similarity =0.3623300

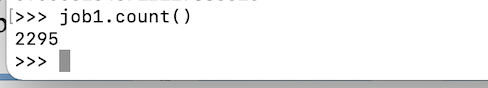


1. The similarity for the first job description which is in text file job.txt:

“Manufacturing Line Manager

Match with the resume number 339 with the cosine similarity = 0.25657

* We can see file count 2295



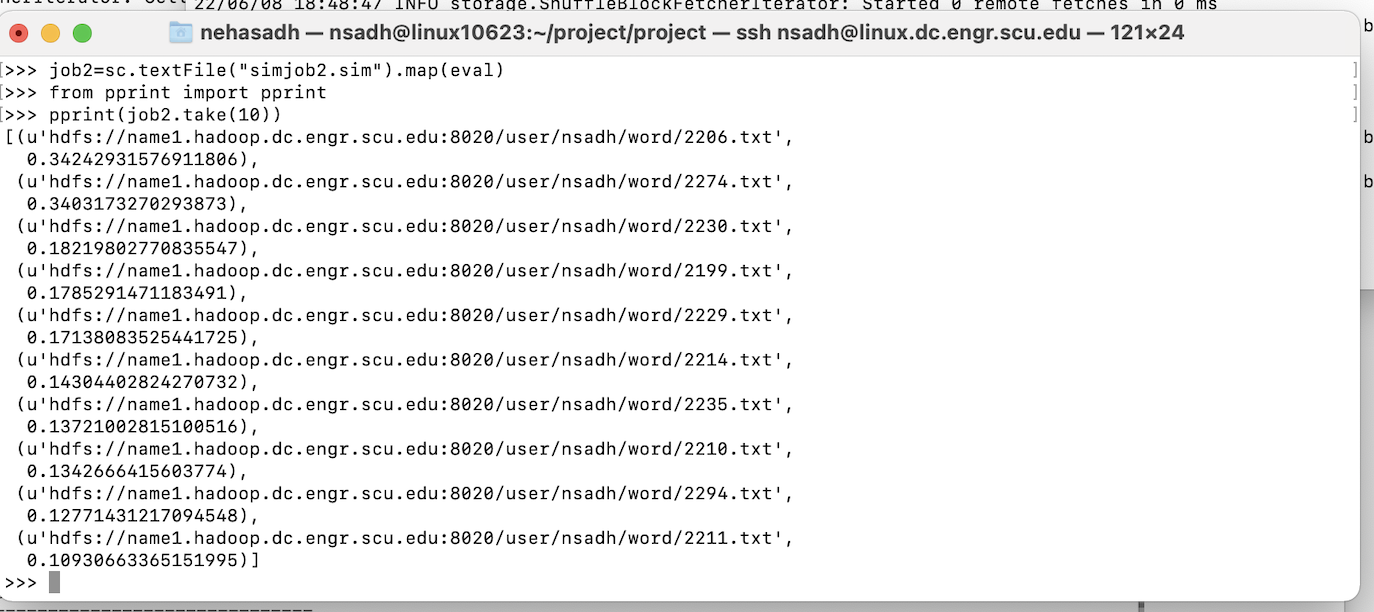
This results the similarity with the resume which have work experience in manufacturing company.



1. The similarity for the first job description which is in text file job2.txt:

“orthopaedics”

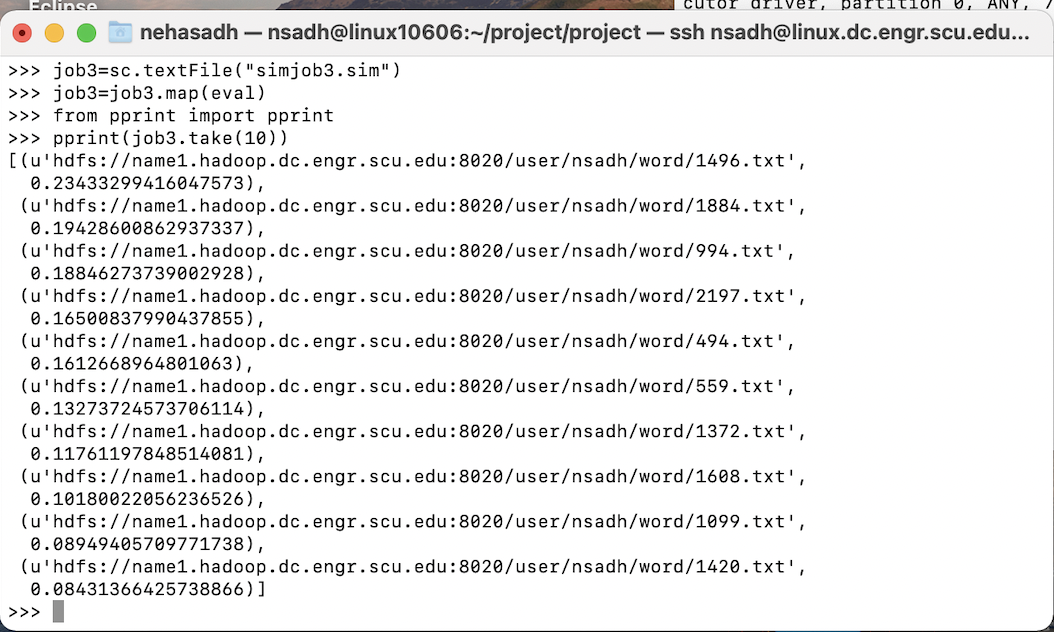
Match with the resume number 2206 with the cosine similarity = 0.34242



1. The similarity for the first job description which is in text file job3.txt:

“Interior Designer”

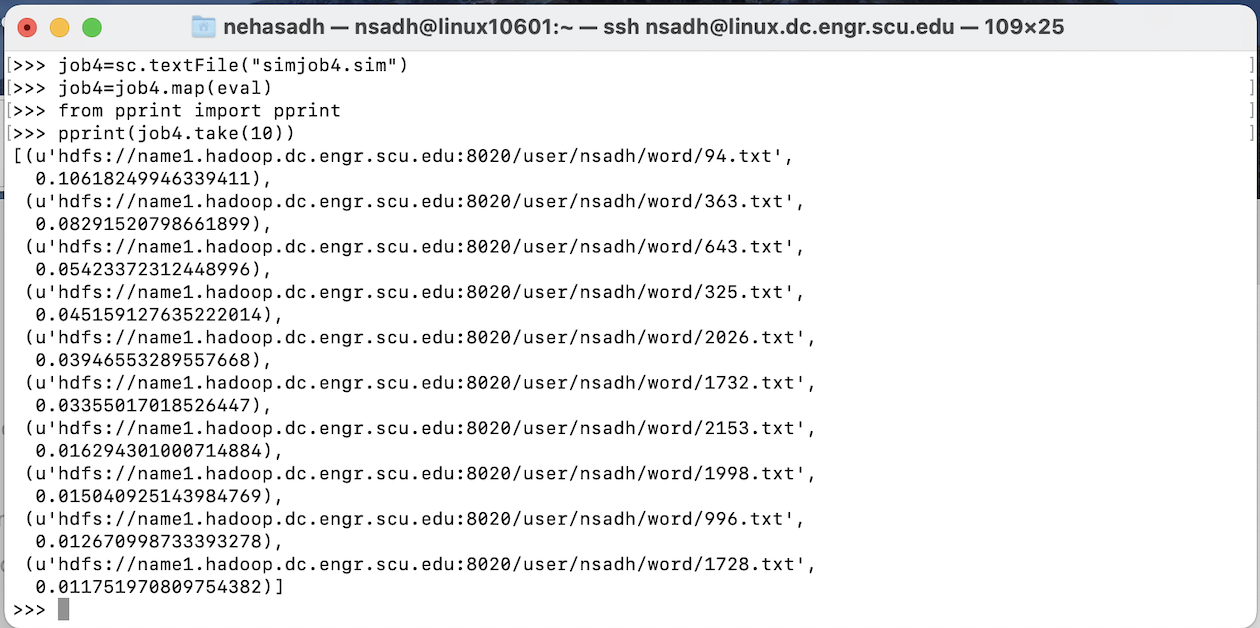
Match with the resume number 1496 with the cosine similarity = 0.2343



1. The similarity for the first job description which is in text file job4.txt:

“Information Communication Technologist”

Match with the resume number 1496 with the cosine similarity = .10618



1. What are most used words in resume dataset?

Data analysis for the words used in resume:

After removing the stop words, tf values of each resume analysis:

|  |  |
| --- | --- |
| Word | TF value |
| Project | 6456 |
| Work | 5632 |
| date | 4432 |
| experience | 4317 |
| management | 4245 |

We can see that still the words with frequency 2000+ are common words of resume.

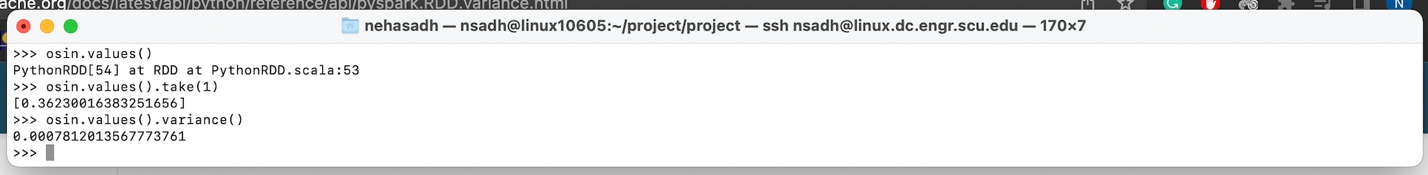
|  |  |
| --- | --- |
| Word | TF value |
| Construction | 1646 |
| business | 1709 |
| design | 1669 |
| technology | 1303 |
| Ability | 1155 |

Words with 1000+ frequency are still some good words.

|  |  |
| --- | --- |
| Word | TF value |
| Complex | 416 |
| Piping | 414 |
| orthopedic | 329 |
| contractor | 339 |
| stock | 351 |

Result which are under 500 have important words. And can remove the most common words.

1. What is the variance in the resume dataset? - How far cosine similarity scores are spread out from their average value.



Variance for cosine similarity of first job.txt document- .000781

Variance of the data is 0.000781235 is very low.

It shows that the data points are very close to each other, the cosine similarity of the documents are not very different from the job description documents.

**Conclusion:**

We found three key observations in the consine similarity algorithm used for resume matching. The average threshold for the cosine similarity in these documents for the five-job description is 0.26. The score is not very great but still we can find the most similar document according to job description. With the term frequency we can find the common words in the resume which we can be used as stop words. We used the threshold frequency of 1000 to identify the stopwords for resume dataset. We found that average variance of consine similarity in the dataset was .000781.

**Reference:**

1. Lin, D., & Pantel, P. (2002, February 15). *Discovery of inference rules for question-answering: Natural language engineering*. Cambridge Core. Retrieved June 8, 2022, from https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/discovery-of-inference-rules-for-questionanswering/31861E00D18EBDC77DC75A4237616207
2. *(PDF) Attali, Y., & Burstein, J. (2006). automated essay scoring with E ...* (n.d.). Retrieved June 9, 2022, from https://www.researchgate.net/publication/274309476\_Attali\_Y\_Burstein\_J\_2006\_Automated\_essay\_scoring\_with\_e-raterR\_V2\_Journal\_of\_Technology\_Learning\_and\_Assessment\_43
3. American Psychological Association. (n.d.). *Apa PsycNet*. American Psychological Association. Retrieved June 8, 2022, from https://psycnet.apa.org/record/1978-09287-001