“Resume Summarization and Matching”

Submitted To:
DR. Sajjad Haider
Submitted By:
Muhammad Asad Baig

Table of Contents

Problem Statement

Selecting a candidate for an available position in your office is quite hectic task for recruiters and HR managers. Big firms can receive 10's and 100's of resume per day and they have to go through all of them one by one. This hectic task can be simplified by implementing text summarization or pattern matching techniques in Text Analytics and Natural language processing which can make this process of finding the relevant candidates easier.

In this project I tried to solve this problem with two different approaches. The first approach is text summarization in which I tried to summarize the candidates resume into basic required information like what is his skills, what is his education and similarly his name, email and phone number. My second approach is focused on the similarity measure of candidates resume with job description posted by company and tried to calculate answers in terms of percentages. So we can have a better look at the candidates profiles that best matches with the job description.

Text Analytics Approaches

There are lot of ways and methods to find similarity between documents and to summarize the documents in text analysis. text summarization is the method of making long bits of text short with the aim to make a reasonable and familiar outline having just the primary concerns illustrated in the archive. Also, continue summarization and matching resumes with the job description is the shorting of a full resume or CV posting the individual data, abilities and encounters. While text outline can be from unstructured content, continue outline can be from unstructured, semi-organized or organized content.

Automated text summarization is accomplishing text summarization with common language handling in Text Analytics.

A portion of the methodologies used to fetch relevant data from resumes incorporate Named Entity Recognition (NER) utilizing Natural Language Tool Kit (NLTK), where text blocks are arranged into sections which are thus cleaved into words and these words are named utilizing the Natural Language Tool Kit which names by appointing labels to each word.

An exploration by talked about some new advances in Deep Learning and how the strategies is advancing and being applied in each circle of life including archive examination portrayed how Artificial Intelligences (AI), and the uses of Natural Language Processing (NLP) are applied in the enrollment industry and how AI is impelling huge mechanization across the employing measures along these lines streamlining the enrollment of value competitors. While a weight-based content outline was model was proposed by where sentences in a record are scored by connecting loads to the diverse language components (thing, phrases, and so forth), applied the synopsis heuristics to produce variable length synopsis separate from a solitary archive. To assess the characteristics of the produced synopses, the exploration further analyzed the first reports to the rundowns and the analysis uncovered that 65% of the reports showed under 10% change in scores.

Using a combination of linguistics and machine learning based approaches, developed a parsing-specifications separation framework for extracting structured information from unstructured resume. The author demonstrated the process using one example and opined that it may not work for other formats.

Creating a summarization and pattern matching system can be very time consuming, as there are no standard algorithms for doing this and unlike other Text Analytics fields, Natural Language processing does not follow a regular pattern. Existing resume summarization and resume matching systems are largely in use in industries and companies that can afford to build such systems as they work entirely based on hard-coded programming.

Algorithm for Summarization

1. Read pdf/doc file from data set

2. Convert file into document list

3. If fail, return error

4. Else, turn the file into text document

5. Remove stop words

6. Perform Lemmatizations

7. convert text into lower case

8. Tokenize the words in document

9. Read Lines in document

10. Read line

11. Use POS Tags to determine name

12. Read Phone with Regex

13. Read Skills that match with Skill set

14. Read Education matching tokens with existing token set

15. Print name, phone, skills, education

16. Go to Next Doc

17. Repeat till doclist.size()

18. End

Algorithm for Matching

1. Read pdf/doc file from data set

2. Convert file into resume list

3. If fail, return error

4. Else, turn the resume file into text document

5. Remove stop words

6. Perform Lemmatizations

7. convert text into lower case

8. Read Job Description Document

9. Repeat Step 5 to Step 7

10. Create list of documents with job description at each entry

11. Calculate similarity between the resume and document

12. Calculate similarity in percentage

13. Repeat till doclist.size()

14. Print candidate name with similarity measure

15. End

Resume Data Set

To find the appropriate data set is quite a handy task itself. After a lot of research I was able to find relevant dataset from Kaggle.com. The author link is also provided below.

https://www.kaggle.com/palaksood97/resume-dataset

To download and unzip this data into my project and convert it into readable form below commands were used:

```
! pip install -q Kaggle

from google.colab import files

files.upload()

! mkdir ~/.kaggle

! cp kaggle.json ~/.kaggle/

! chmod 600 ~/.kaggle/kaggle.json

! kaggle datasets list

! kaggle datasets download -d palaksood97/resume-dataset

! mkdir resume-dataset

! unzip resume-dataset.zip -d resume-datasetdf
```

The dataset for resume consists of 224 candidates with background related to the field of computer science and java. So I carefully choose the relevant job description accordingly to analyze algorithm results in more depth.

NLTK

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Spacy

Spacy library is designed specifically for production use and helps you build applications that process and “understand” large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning.

Genism

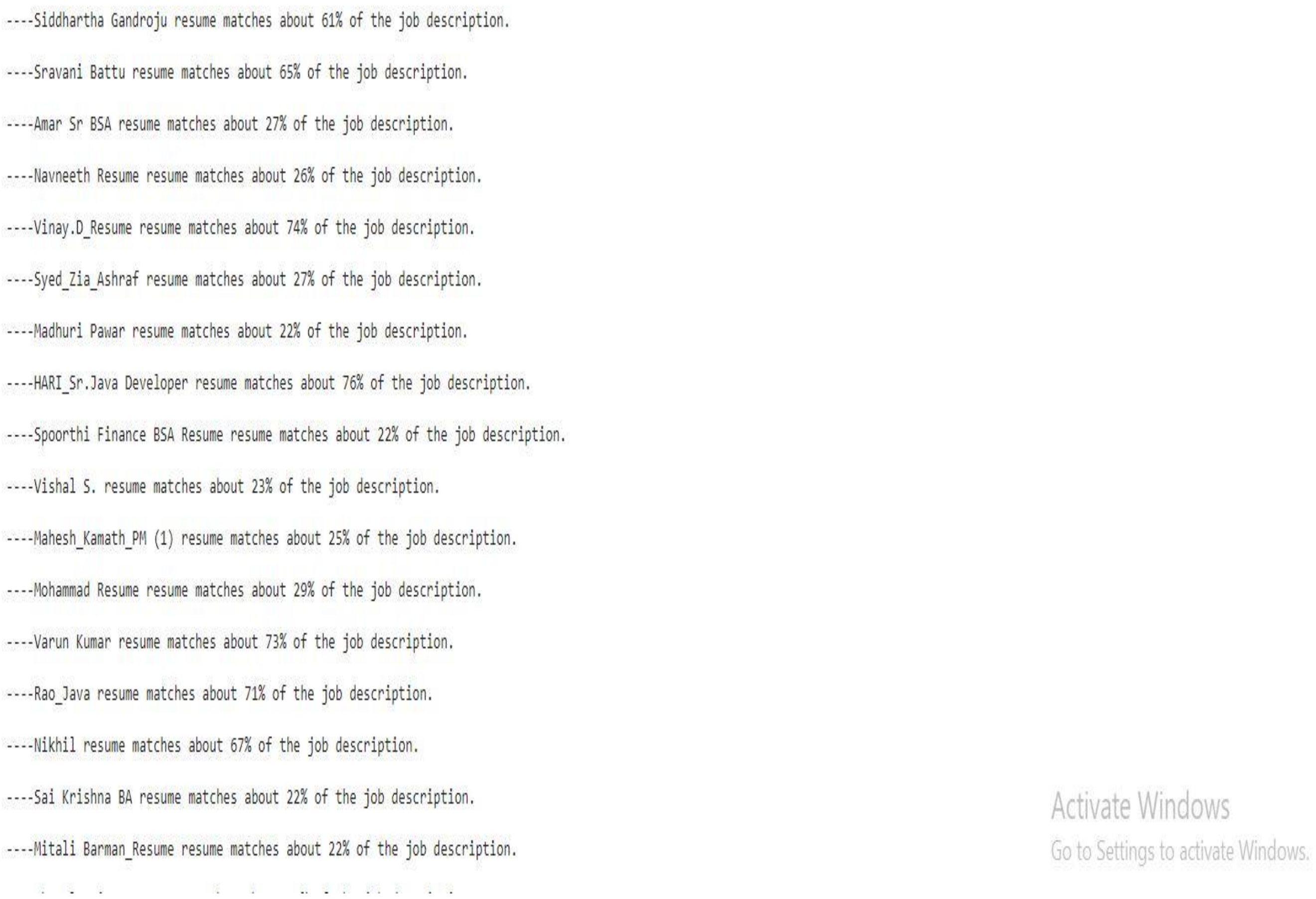
Gensim is billed as a Natural Language Processing package that does 'Topic Modeling for Humans'. ... It is a leading and a state-of-the-art package for processing texts, working with word vector models (such as Word2Vec, FastText etc) and for building topic models.

Doc2Text

Doc2text: detect text blocks and OCR poorly scanned PDFs in bulk. Python module available via pip.

Results

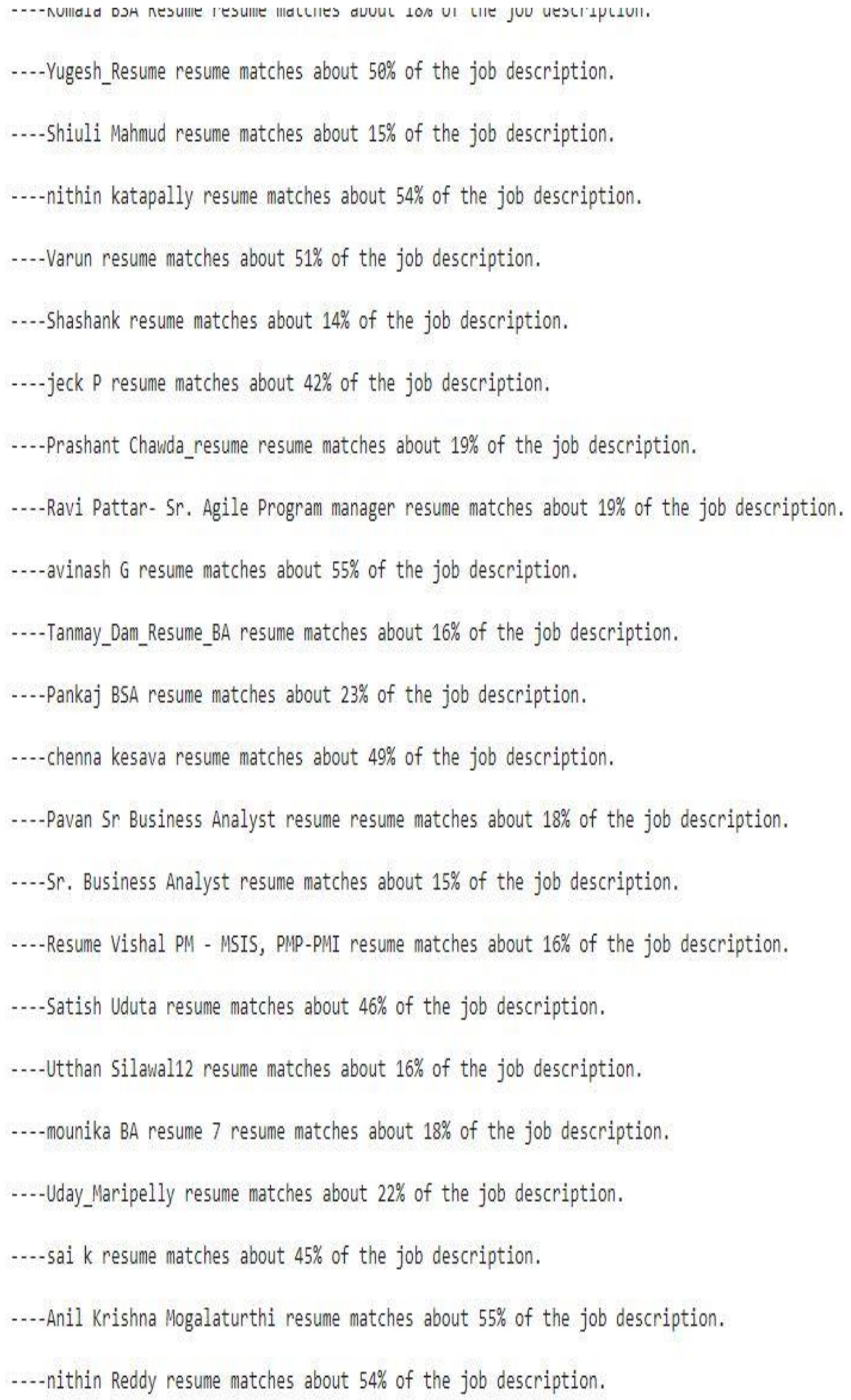
Results after similarity measure between job description and 228 candidates resume are shown below. I have used count vectorizer fit transform method and used it for measuring cosine similarity between job description and resume of candidate. After that I used that cosine similarity and extract percentage from it. The higher the percentage the better candidates match with the job description.



```
----Siddhartha Gandroju resume matches about 61% of the job description.

----Sravani Battu resume matches about 65% of the job description.

----Amar Sr BSA resume matches about 27% of the job description.

----Navneeth Resume resume matches about 26% of the job description.

----Vinay.D_Resume resume matches about 74% of the job description.

----Syed_Zia_Ashraf resume matches about 27% of the job description.

----Madhuri Pawar resume matches about 22% of the job description.

----HARI_Sr.Java Developer resume matches about 76% of the job description.

----Spoorthi Finance BSA Resume resume matches about 22% of the job description.

----Vishal S. resume matches about 23% of the job description.

----Mahesh_Kamath_PM (1) resume matches about 25% of the job description.

----Mohammad Resume resume matches about 29% of the job description.

----Varun Kumar resume matches about 73% of the job description.

----Rao_Java resume matches about 71% of the job description.

----Nikhil resume matches about 67% of the job description.

----Sai Krishna BA resume matches about 22% of the job description.

----Mitali Barman_Resume resume matches about 22% of the job description.
```

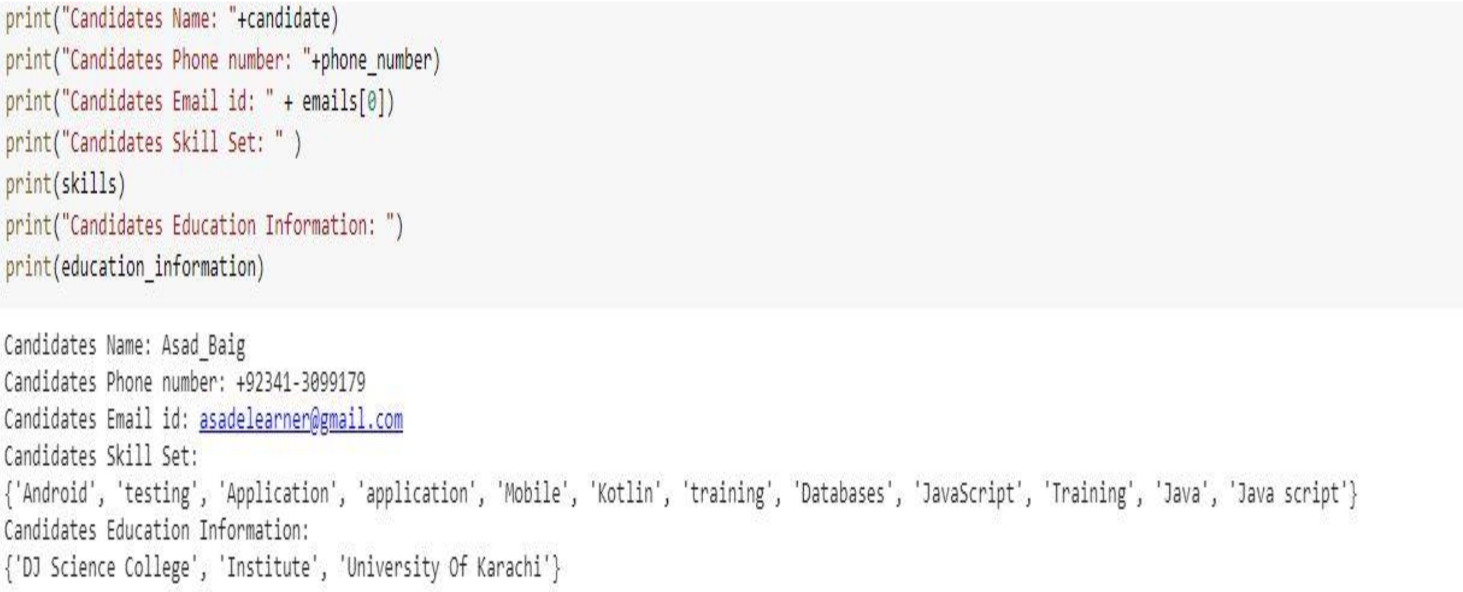
A sample cosine similarity score of my cv with java backend developer cv. I have few skills in java but not java backend and J2EE. So when I compared the skill set manually it shows almost similar score near 30% when I compare my skill set with job description.

```
Simlarity Scores:
[[1.         0.32500793]
 [0.32500793 1.        ]]
```



```
----Komala BSA Resume resume matches about 18% of the job description.

----Yugesh_Resume resume matches about 50% of the job description.

----Shiuli Mahmud resume matches about 15% of the job description.

----nithin katapally resume matches about 54% of the job description.

----Varun resume matches about 51% of the job description.

----Shashank resume matches about 14% of the job description.

----jeck P resume matches about 42% of the job description.

----Prashant Chawda_resume resume matches about 19% of the job description.

----Ravi Pattar- Sr. Agile Program manager resume matches about 19% of the job description.

----avinash G resume matches about 55% of the job description.

----Tanmay_Dam_Resume_BA resume matches about 16% of the job description.

----Pankaj BSA resume matches about 23% of the job description.

----chenna kesava resume matches about 49% of the job description.

----Pavan Sr Business Analyst resume resume matches about 18% of the job description.

----Sr. Business Analyst resume matches about 15% of the job description.

----Resume Vishal PM - MSIS, PMP-PMI resume matches about 16% of the job description.

----Satish Uduta resume matches about 46% of the job description.

----Utthan Silawal12 resume matches about 16% of the job description.

----mounika BA resume 7 resume matches about 18% of the job description.

----Uday_Maripelly resume matches about 22% of the job description.

----sai k resume matches about 45% of the job description.

----Anil Krishna Mogalaturthi resume matches about 55% of the job description.

----nithin Reddy resume matches about 54% of the job description.
```

The results after resume summarization are shown below:



```
print("Candidates Name: "+candidate)
print("Candidates Phone number: "+phone_number)
print("Candidates Email id: " + emails[0])
print("Candidates Skill Set: " )
print(skills)
print("Candidates Education Information: ")
print(education_information)
```

```
Candidates Name: Asad_Baig
Candidates Phone number: +92341-3099179
Candidates Email id: asadelearner@gmail.com
Candidates Skill Set:
{'Android', 'testing', 'Application', 'application', 'Mobile', 'Kotlin', 'training', 'Databases', 'JavaScript', 'Training', 'Java', 'Java script'}
Candidates Education Information:
{'DJ Science College', 'Institute', 'University Of Karachi'}
```

I have used NLTK, spacy and gensim library to summarize the resume’s. Used nltk tokenize methods to extract the skills and education from the data set, also use regex for phone number, POS tags for extracting name of the candidate and email of the candidate is also extracted with the help of the above three libraries. The resume summarization almost give perfect results when I check it on similar resume’s that have same format as my cv. Although when I applied it on 228 candidates cv this thing not work well as expected. Few information went missing because of different cv formats. Over all I think if further work on this thing and improve the algorithm according to the resumes format than this summarization algorithm can work well on all kinds of formats.

References

- https://www.kaggle.com/palaksood97/resume-dataset
- https://spacy.io/usage/spacy-101
- https://towardsdatascience.com/resume-summarizing-and-matching-165840cf9f75