Team ZeRoS

Reeka Hazarika (UAE) – Free Lancer

hazarika.reeka@gmail.com

Swapnil Vishwakarma (India) - SIES Graduate School of Technology

swapnilvishwakarma7@gmail.com

Zyad Al-Azazi (Lebanon) – Lebanese American University

zyadazazi@gmail.com

Specialization: Natural Language Processing

Week 10 Deliverables:

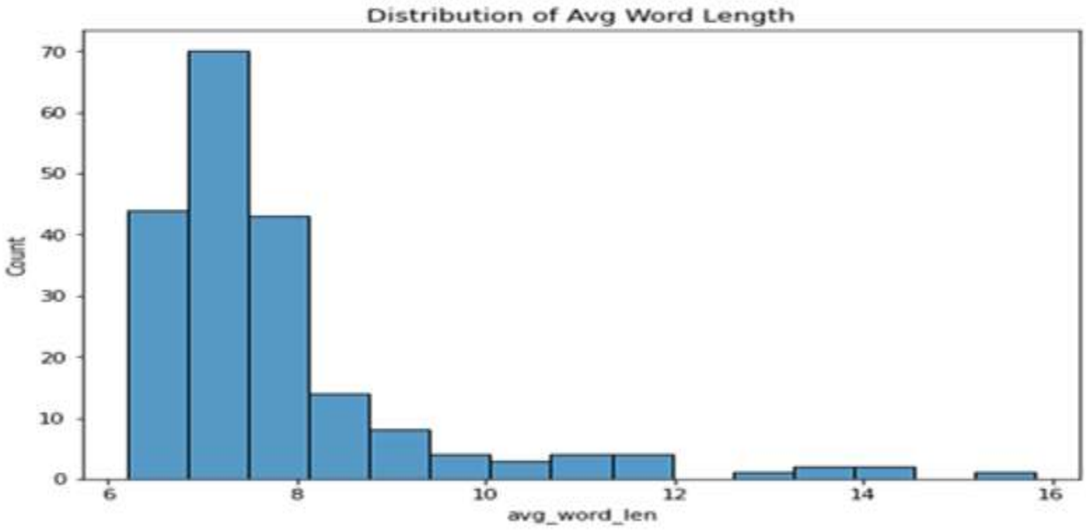
Problem Description:

When companies recruit for any position, they usually end up receiving thousands, if not millions, of resumes. Such a huge number of resumes makes the task of going over all these resumes an extremely difficult and tedious job for HR employees. This made a lot of companies opt for systems that take the necessary information from the candidate after they fill an application with all the required fields. The solution worked greatly for employers; nevertheless, candidates have always found it very illogical to spend tens of hours sharpening their CVs and cover letters only to find out that they must spend another hour or so re-entering all the information they have on their CVs in the designated fields.

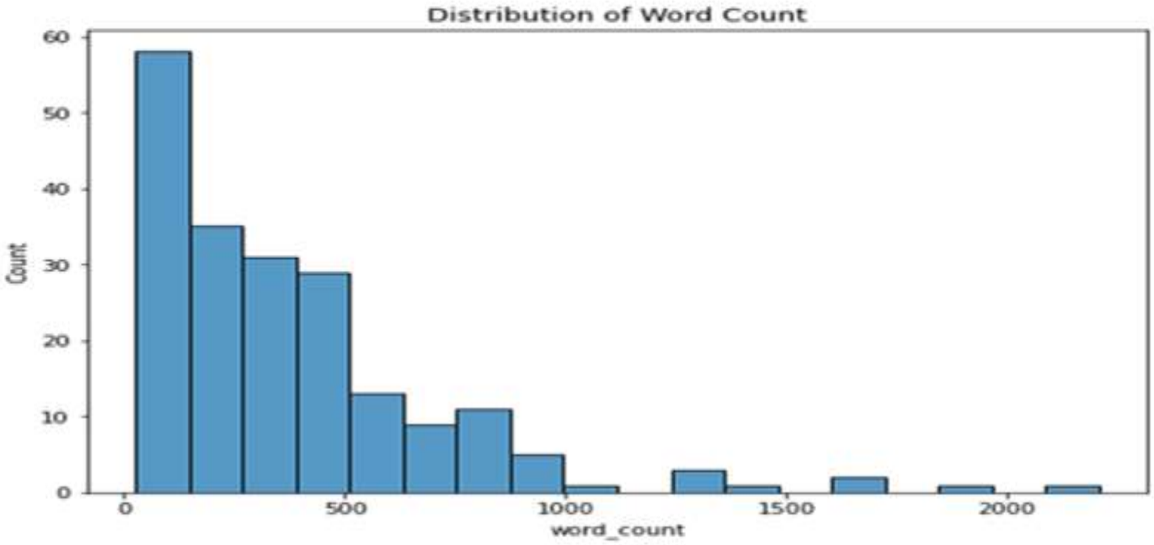
Github Repo link:

https://github.com/zyadalazazi/resume_extraction_team_zeros

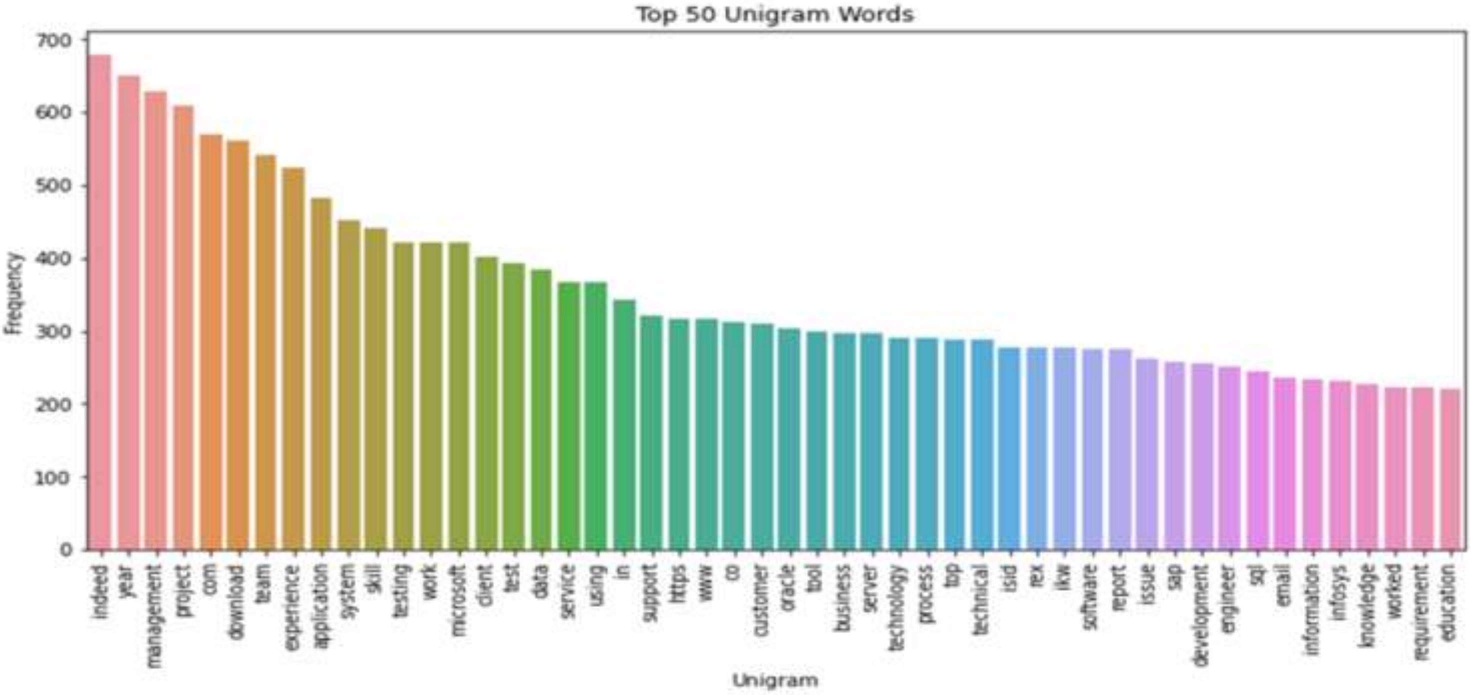
EDA performed on the data:



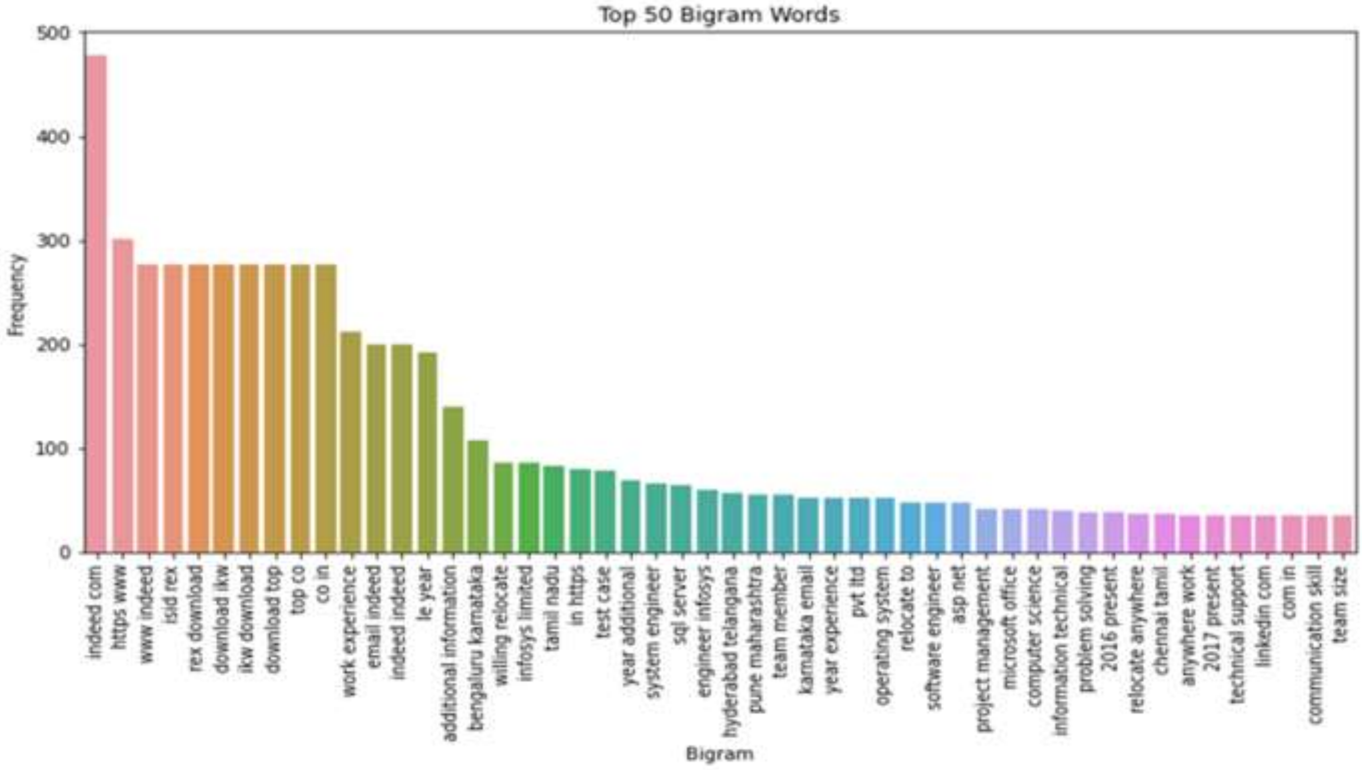
- **It is very clear from the plot that the average length of the word is 7 words.**



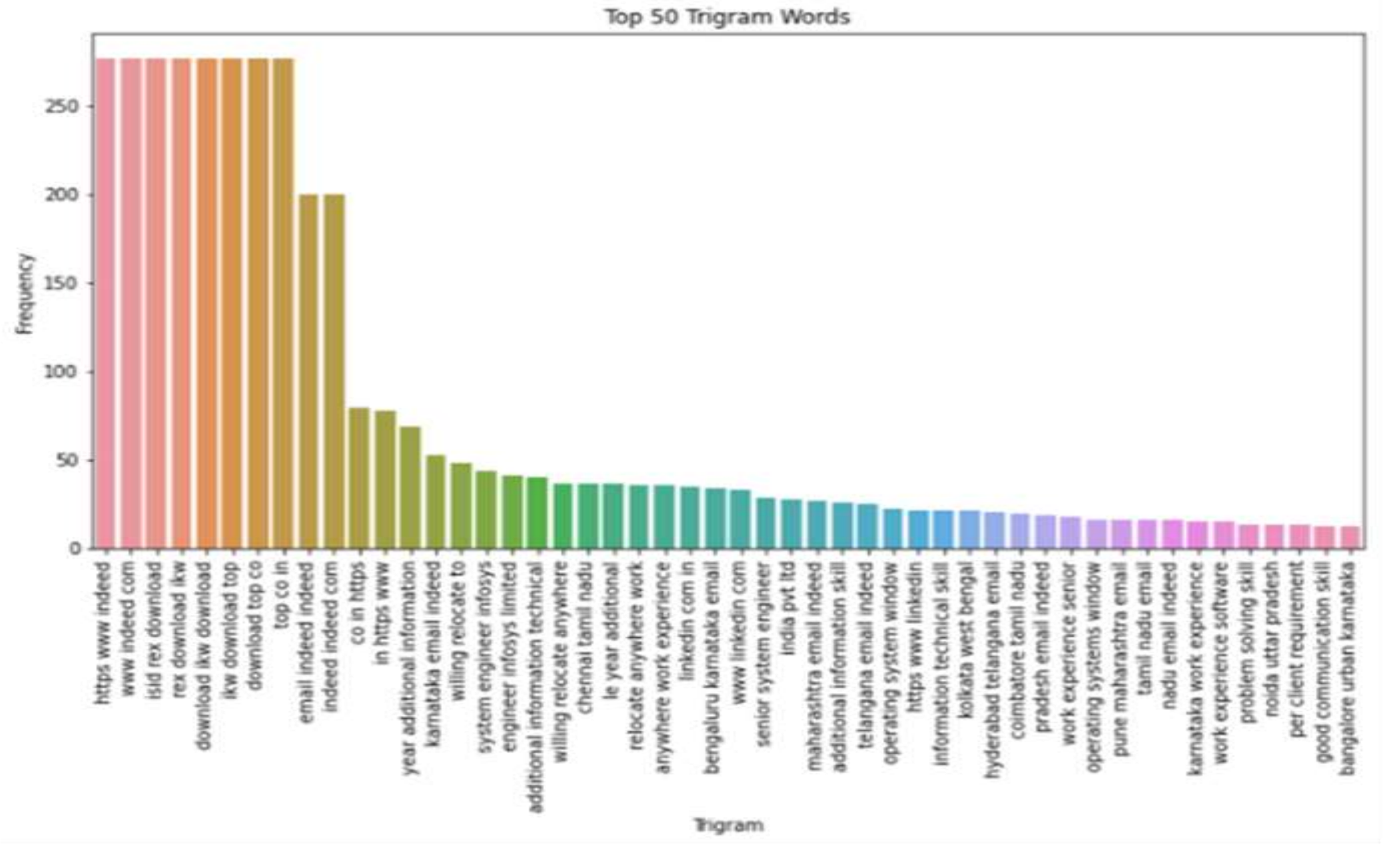
- **From the following graph it can be concluded that most of the Resumes contain less than 500 words.**



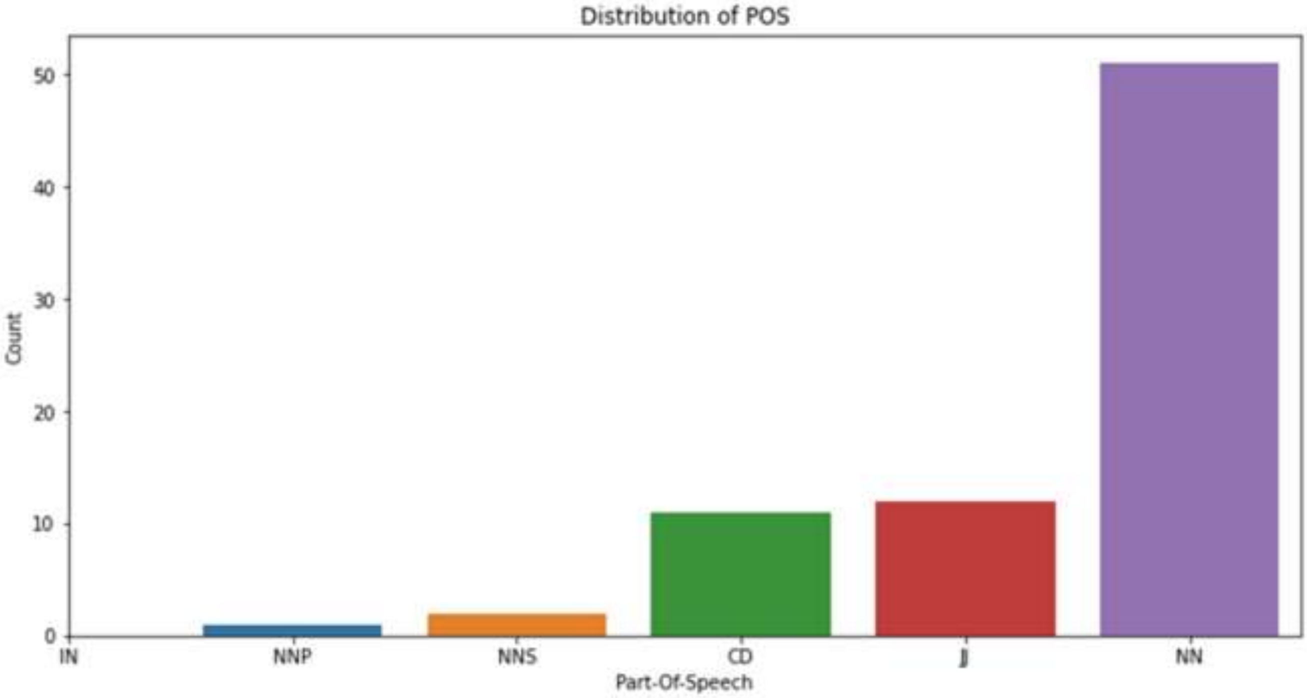
- **Top 3 frequent words are ‘Indeed’, ‘Year’ & ‘Management’.**



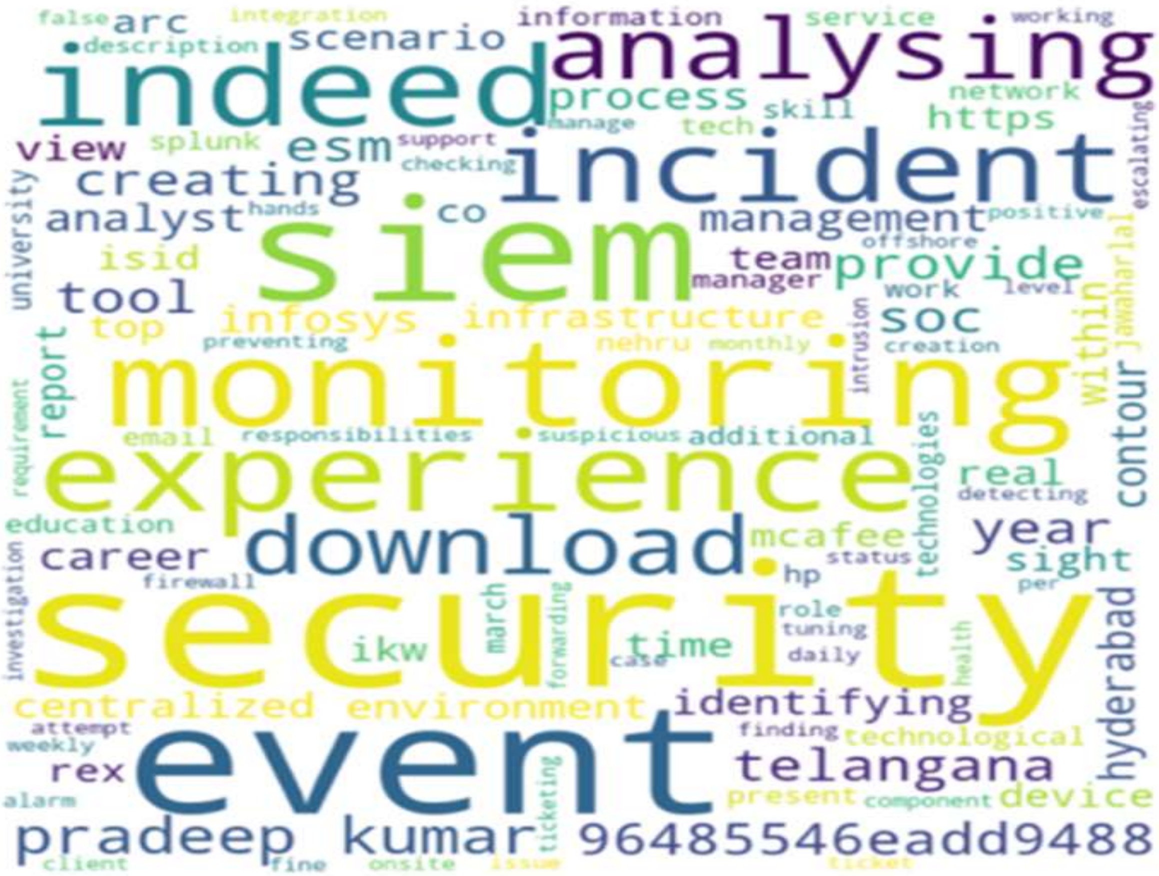
- **Top frequent words occurring in a pair are ‘Word experience’, ‘Email indeed’ & ‘Additional information’ that could be useful.**



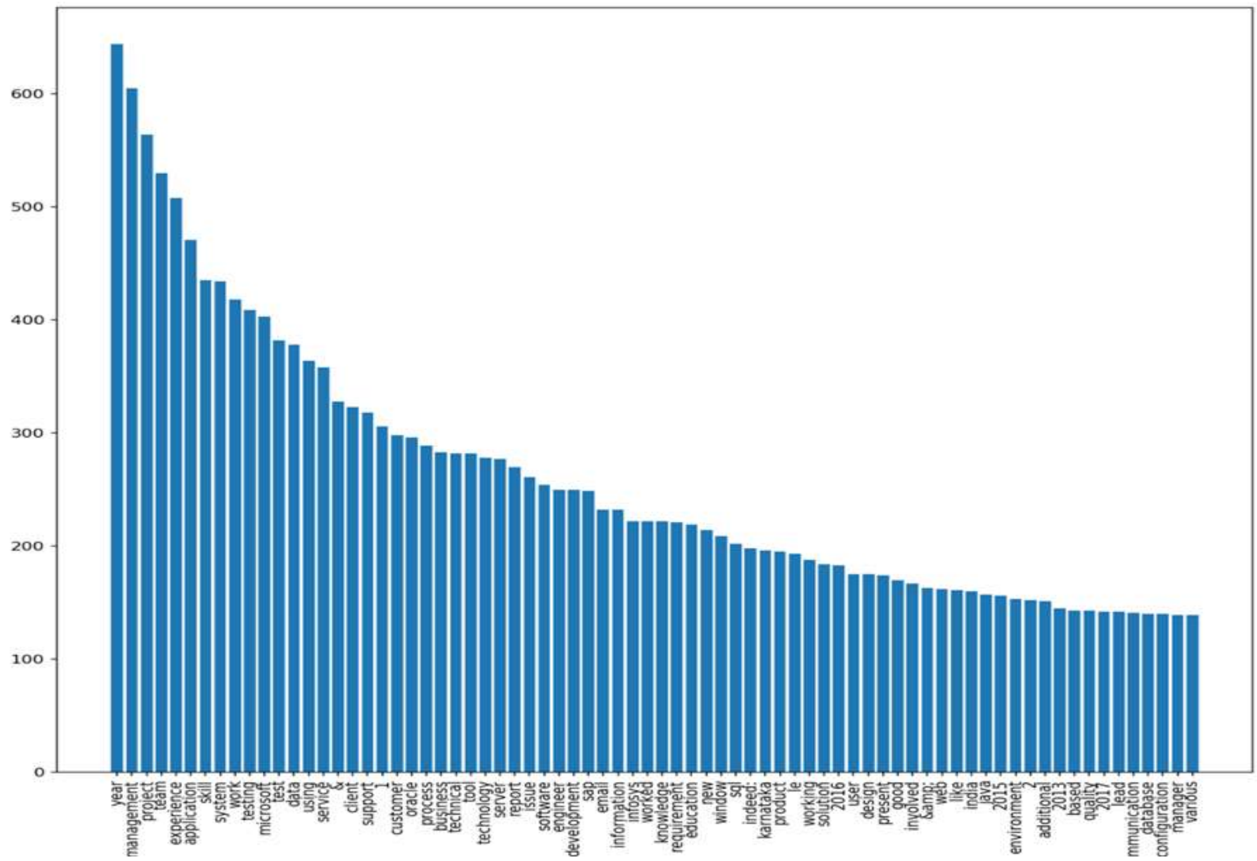
- **Top frequent 3 words occurring together are ‘Year additional information’, ‘Karnataka email indeed’ & ‘Willing relocate to’ that could be useful.**



- **‘Singular noun’ is most commonly used, then comes ‘Adjective or numeral’, ‘Numeral (cardinal)’, ‘Proper plural noun’, ‘Singular proper noun’ and ‘Preposition or conjunction, subordinating’ respectively.**



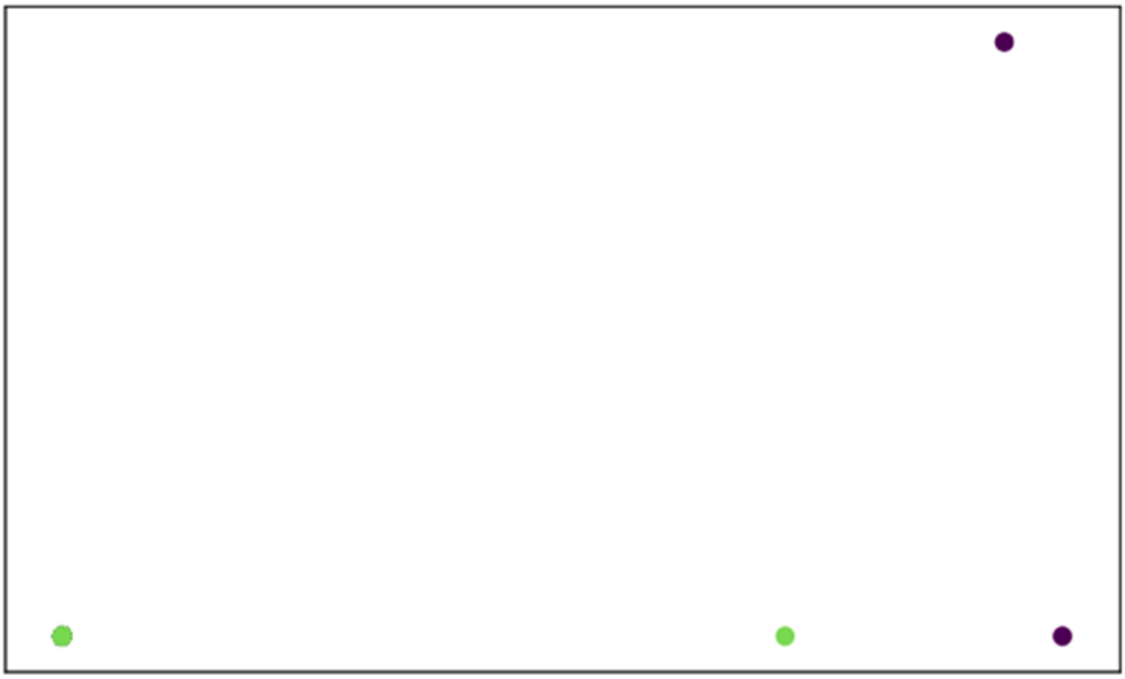
- **Graphical representations of word frequency. The larger the word in the visual the more common the word was in the resumes.**



- **This bar graph shows the frequency of the 75 most-frequent in all the resumes.**
- **Insights from the Bar Graph:**

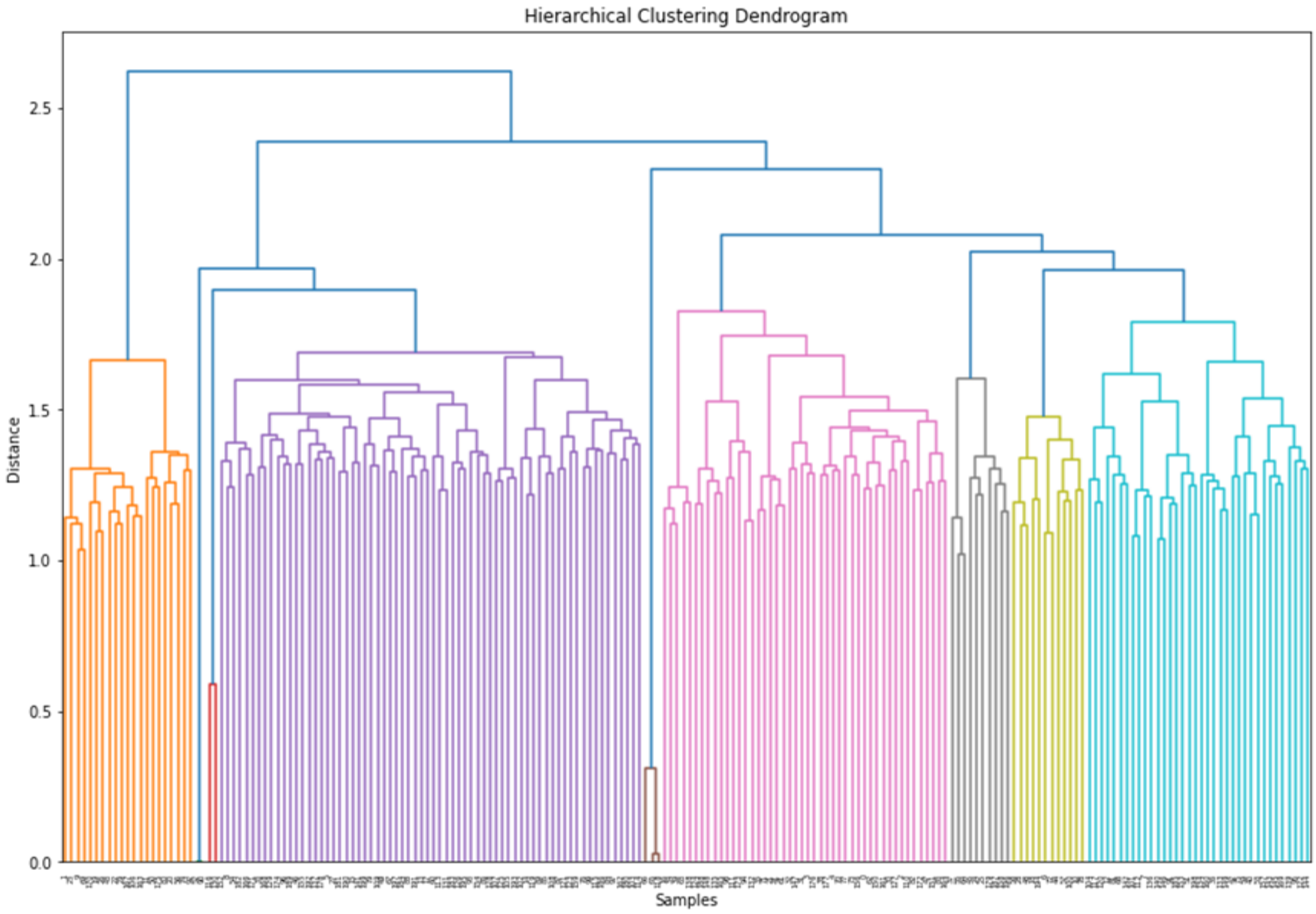
- **The most mentioned companies on the applicants resumes were Microsoft and Oracle.**
- **Some of the keywords that the applicants emphasized on, in order, were management, data, testing, customer, business, technical and software.**
- **Under the assumption that all the applicants are applying to the same vacancy, we can hypothesize that this job is a leadership role that requires business and customer-communication skills accompanied with technical skills.**
- **The word “experience” was the 5th most frequent word, which demonstrates the high importance of experience over education in the job market.**

K-means clustering:



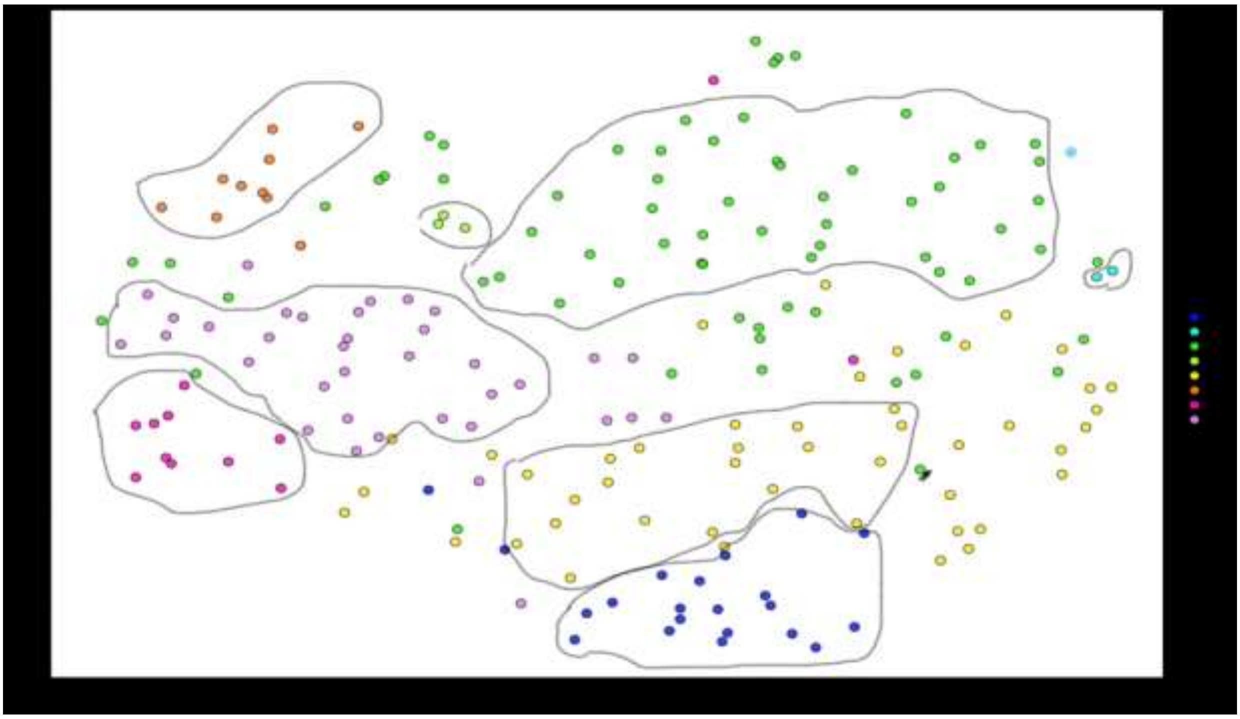
- When attempting to represent the 198 resumes using K-means clustering, it is clear that the graph fails to correctly cluster all the points. The reason behind this is the high number of dimensions that each vector possesses since we are using the Tf-Idf vectors to the resumes.

Hierarchical Clustering Analysis:



- **Graphical Analysis:**
  The tree shows that we can split our data into 8 clusters.
  To identify each clusters which group all features' values of resumes we added colours to each clusters.

- **We applied a method of how unstructured text data for a specific field, namely, recruiting can be organized. With the right feature engineering (TF-iDF transformation) it's possible to split resumes into different groups (here we see 8 groups).**



Final Recommendation:

- We plotted word count distribution to find out that most of the resumes contained at most 500 words.
- Next, we wanted to see the frequency of each word occurring in the dataset, so we created the distribution of Unigram words, Bigram words, and even Trigram words to analyse the most frequent words solely, in a pair and three words at-a-time, respectively.
- The distribution of Part of Speech was done to find out that most of the words were a singular noun.
- The next step was to plot the Word Cloud which is a visual representation of the frequency of different words present in a document. It gives importance to the more frequent words which are bigger in size compared to other less frequent words. We were also able to come up with some interesting insights when looking at the 75 most frequent words. The most mentioned companies on the applicants' resumes were Microsoft and Oracle.
- Some of the keywords that the applicants emphasized on, in order, were "management," "data," "testing," "customer," "business," "technical," and "software." If we assume that all the applicants were applying to the same vacancy, we can hypothesize that this job is a leadership role that requires business and customer-communication skills accompanied with technical skills.
- After this, we tried exploring the dataset using the unsupervised machine learning technique using the technique called "K-means clustering." When attempting to represent the 198 resumes using K-means clustering, it was clear that the graph fails to correctly represent and cluster all the points– resumes in our case. The

reason behind this is the high number of dimensions that each vector possesses since we were using the Tf-Idf vectors to represent each resume.

- **After the failure of K-means clustering, we used Hierarchical clustering which groups data over a variety of scales by creating a clustering tree or a Dendrogram. The tree is not a single set of clusters, but rather a multi-level hierarchy, where clusters at one level are joined as clusters at the next level. The tree shows that we can split our data into 8 separate clusters. We applied the method of how unstructured text data for a specific field, namely, recruiting can be organized. With the right feature engineering (TF-iDF transformation) it is possible to split resumes into different groups (here we have 8 groups).**

……………………………………………………………