# RESUME EXTRACTION

By: Team ZeRoS

12-05-2021

# TABLE OF CONTENTS

# EXECUTIVE SUMMARY

It is often observed by HR that the manual process of evaluation of Resumes in bulk which are populated with excess information often becomes tedious and hectic. Therefore, we could automate this process by reading several formats of files (CV). Then using some basic techniques of Natural Language Processing like word parsing, chunking, regex parser and/or Named Entity Recognition to easily capture information like name, email id, address, educational qualification, experience in seconds from a large number of documents.

# PROBLEM STATEMENT

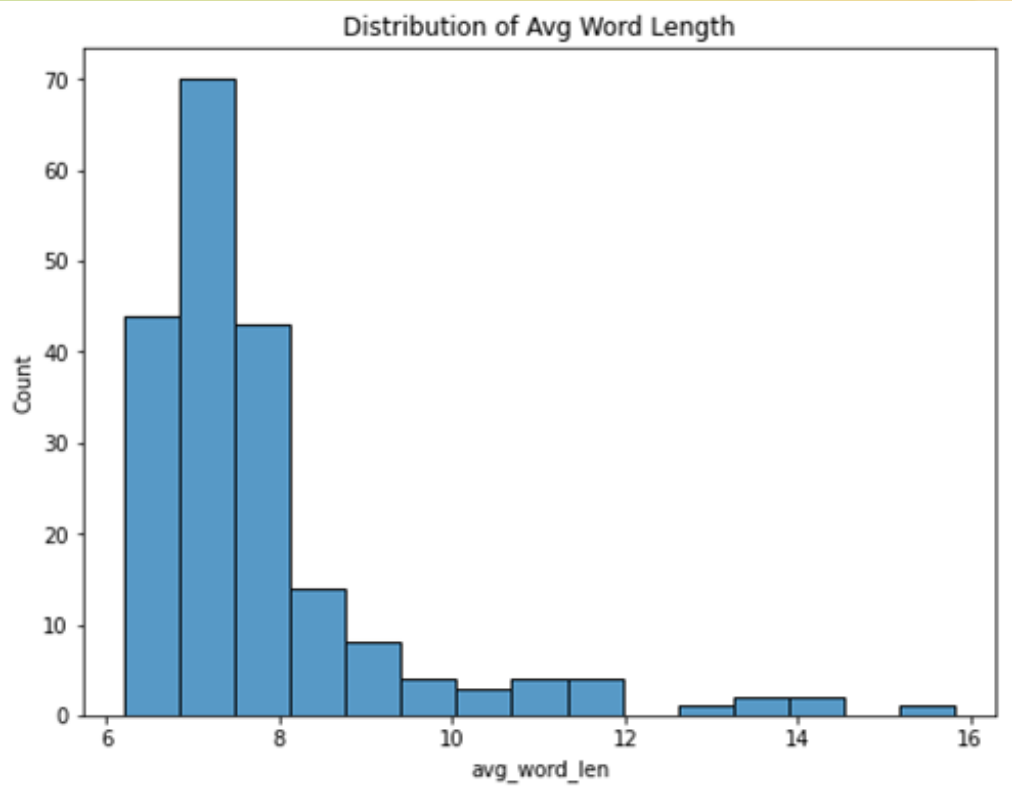
When companies recruit for any position, they usually end up receiving thousands, if not millions, of resumes. Such a large number of resumes makes the task of going over all these resumes an extremely difficult and tedious job for HR employees. This made a lot of companies opt for systems that take 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.

# SAMPLE SIZE

| TYPE | VALUE |
| --- | --- |
| Total number of Observations | 200 |
| Total number of files | 01 |
| Total number of features | 02 |
| Base format of the file | .txt |
| Dataset size | 1.1 MB |

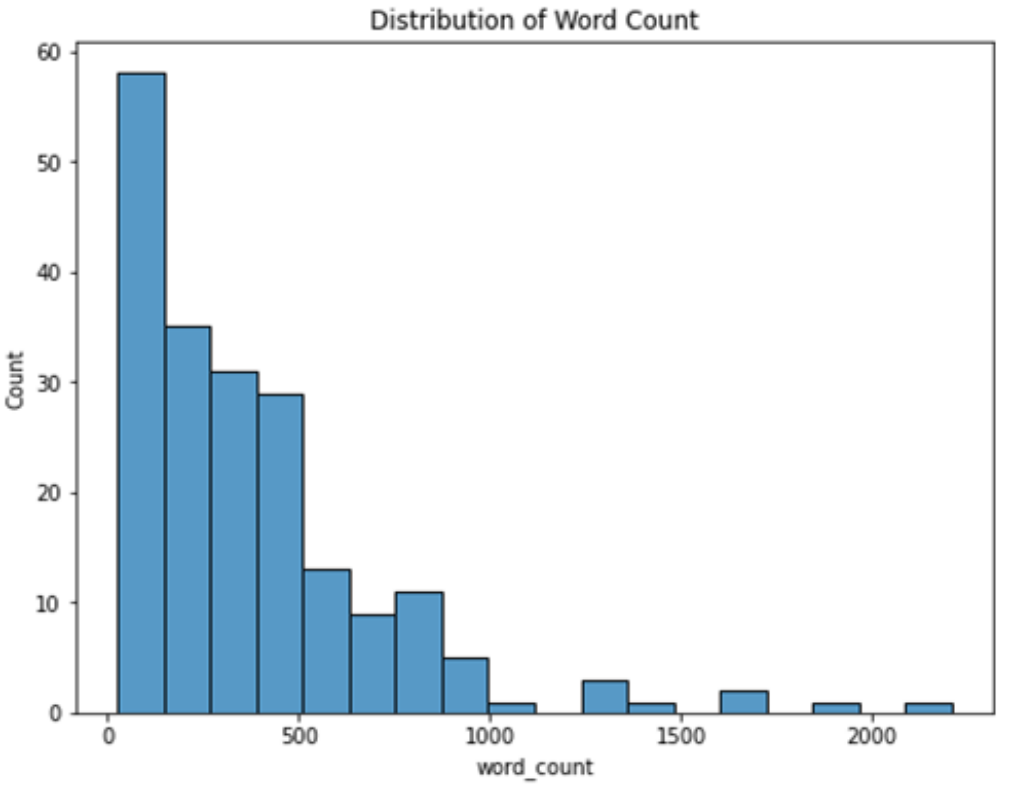
# EXPLORATORY DATA ANALYSIS

# EDA



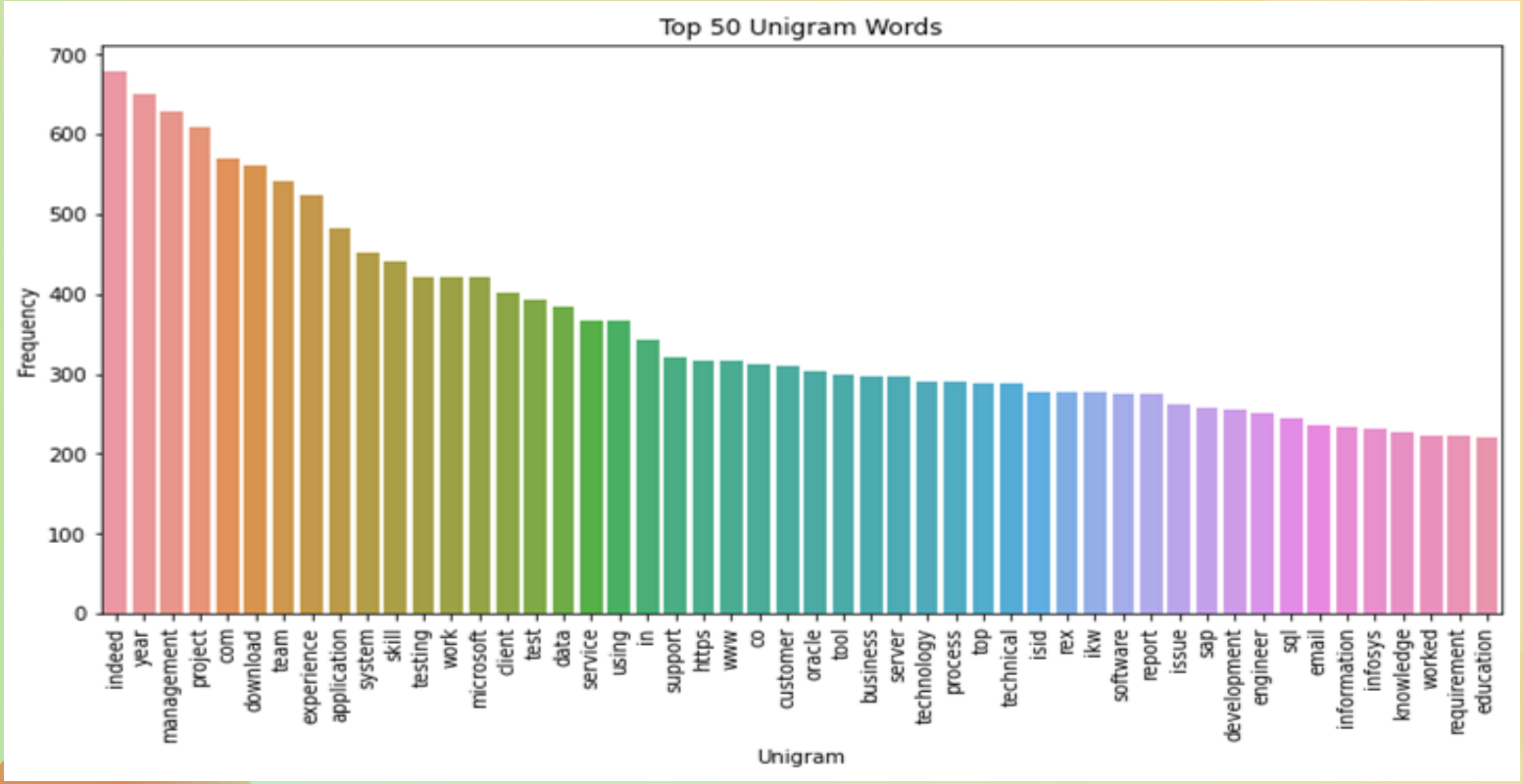

It is very clear that the average length of the word is 7 words.

# EDA



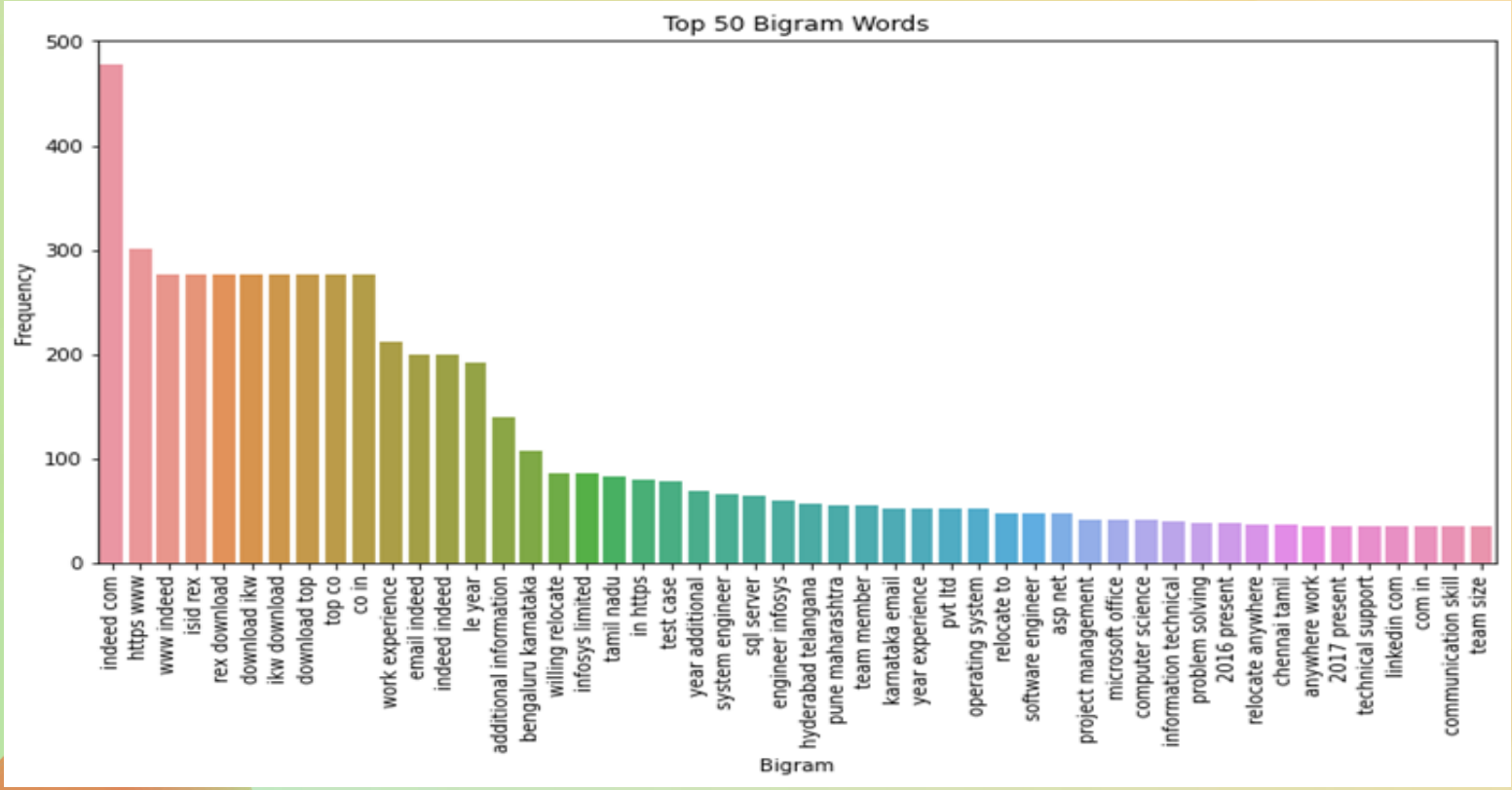

From the following graph it can be concluded that most of the Resumes contain less than 500 words.

# EDA



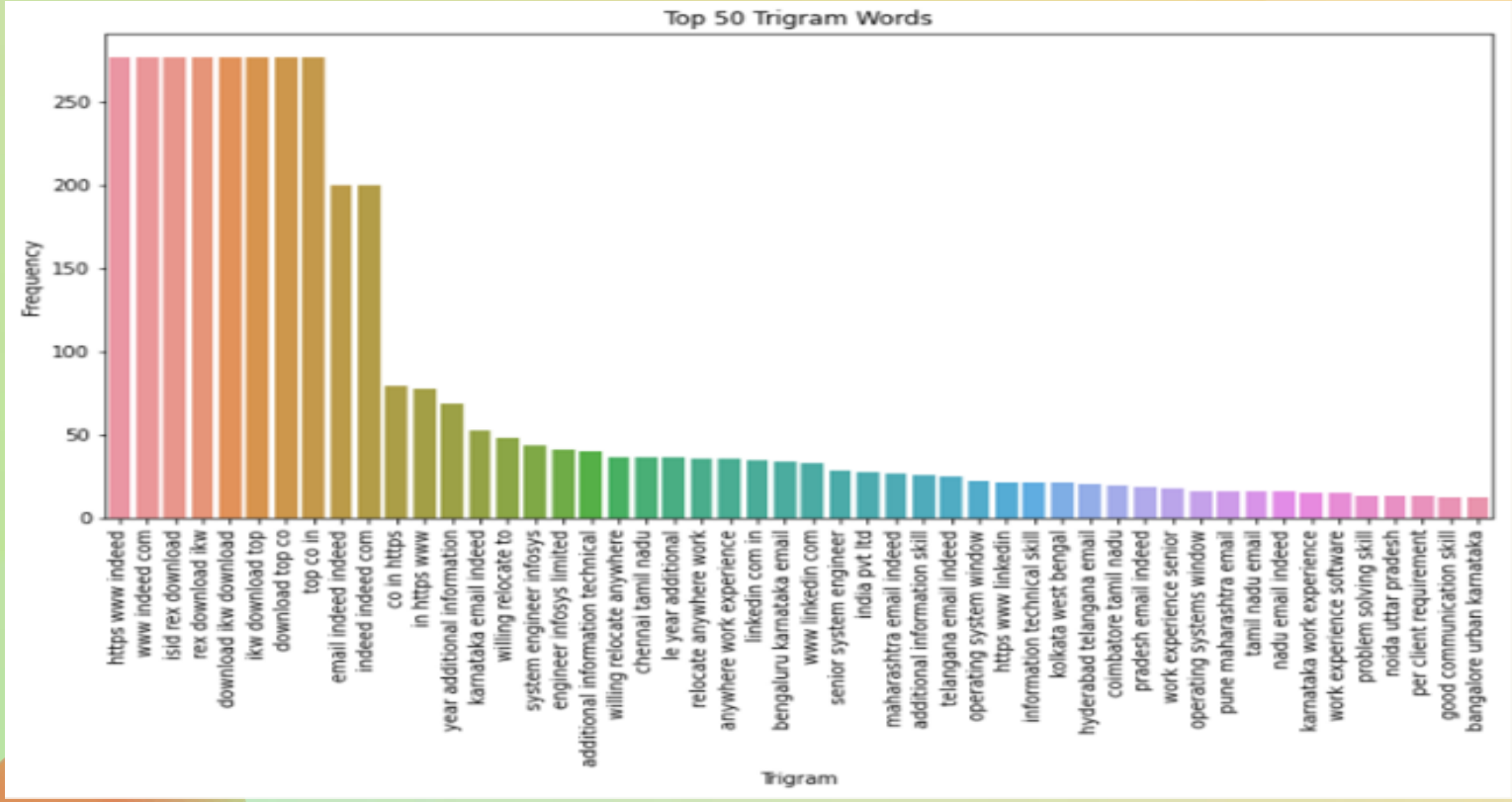

Top 3 frequent words are ‘Indeed’, ‘Year’ & ‘Management’.

# EDA



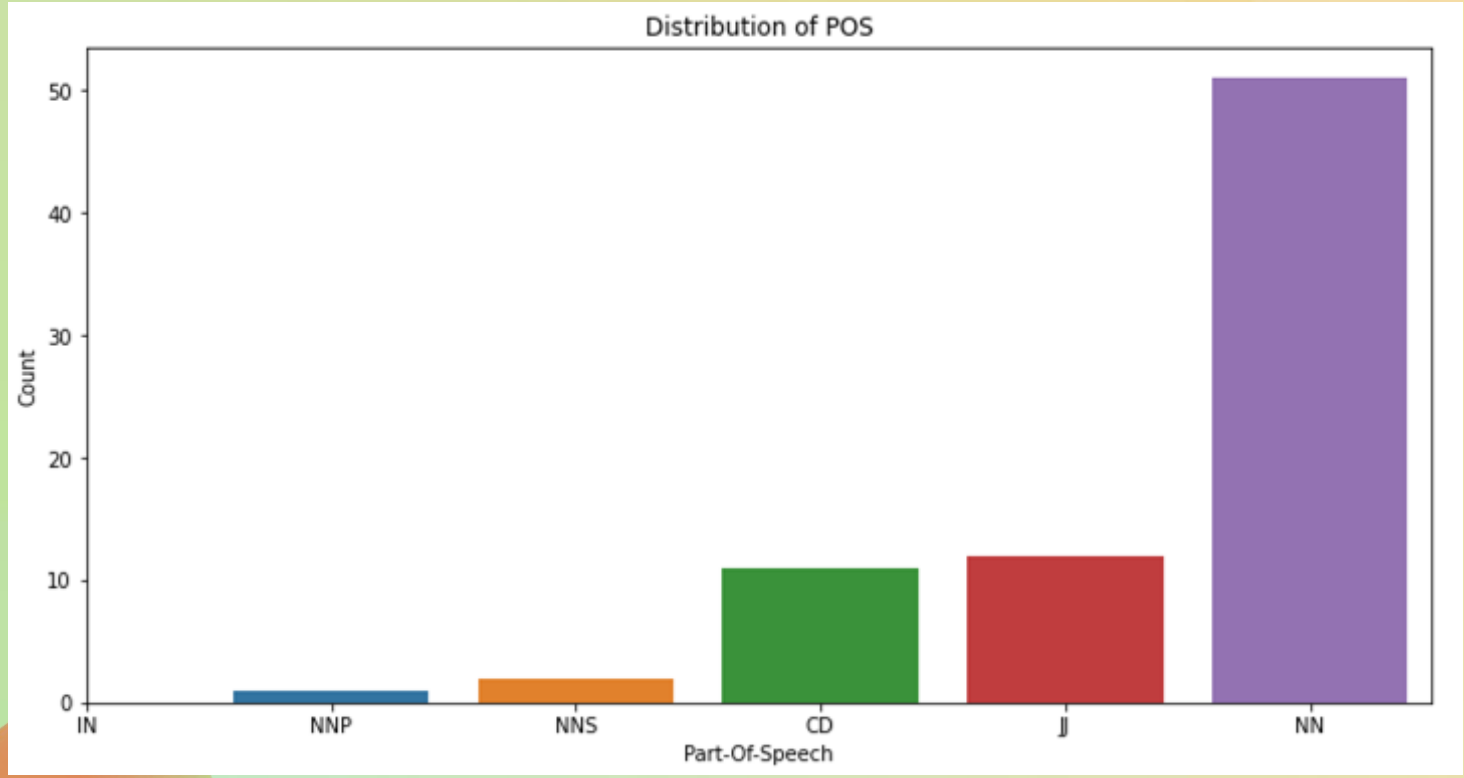

Top frequent words occurring in a pair are ‘Word experience’, ‘Email indeed’ & ‘Additional information’ that could be useful.

# EDA




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

# EDA



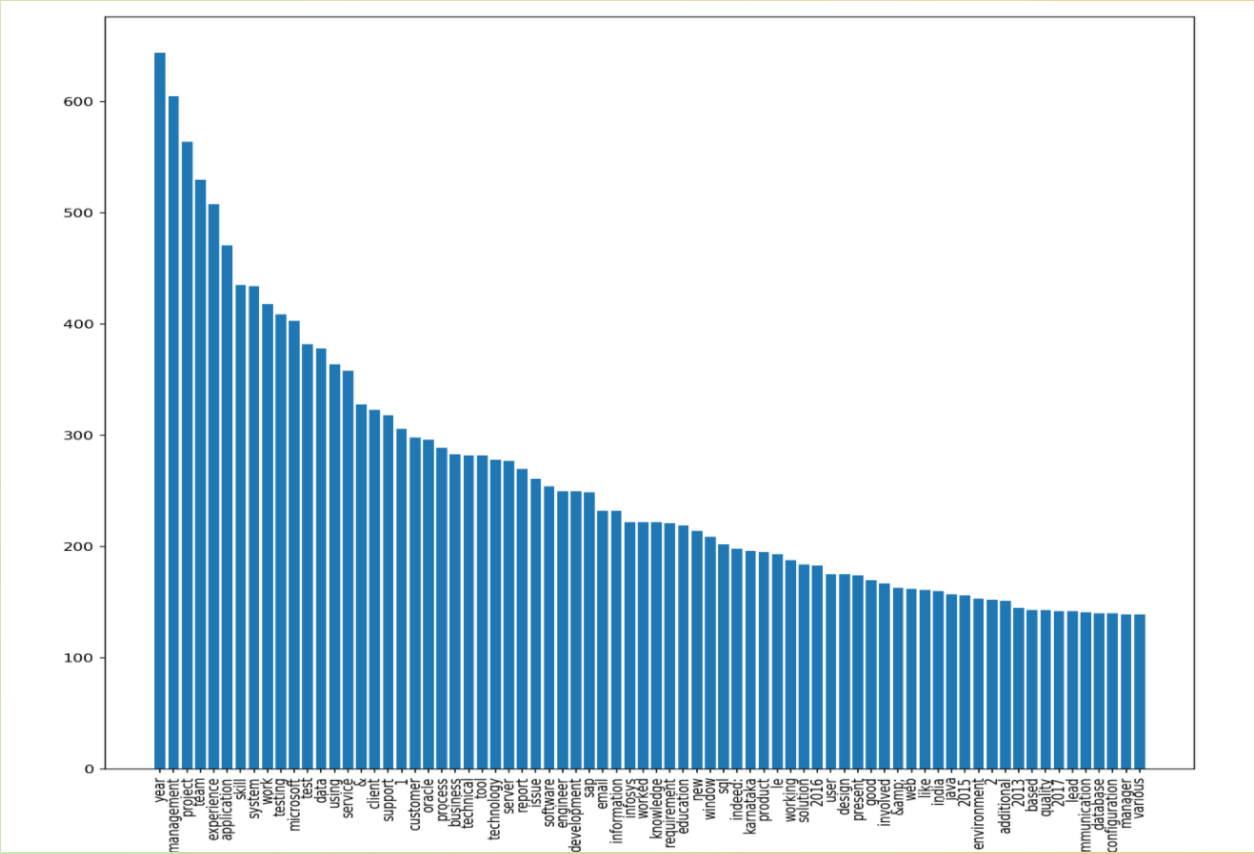

'Singular noun' is most commonly used, then comes 'Adjective or numeral', 'Numeral (cardinal)', 'Proper plural noun', 'Singular proper noun' and 'Preposition or conjunction, subordinating' respectively.

# EDA


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

# EDA

- This bar graph shows the frequency of the 75 most-frequent in all the resumes.

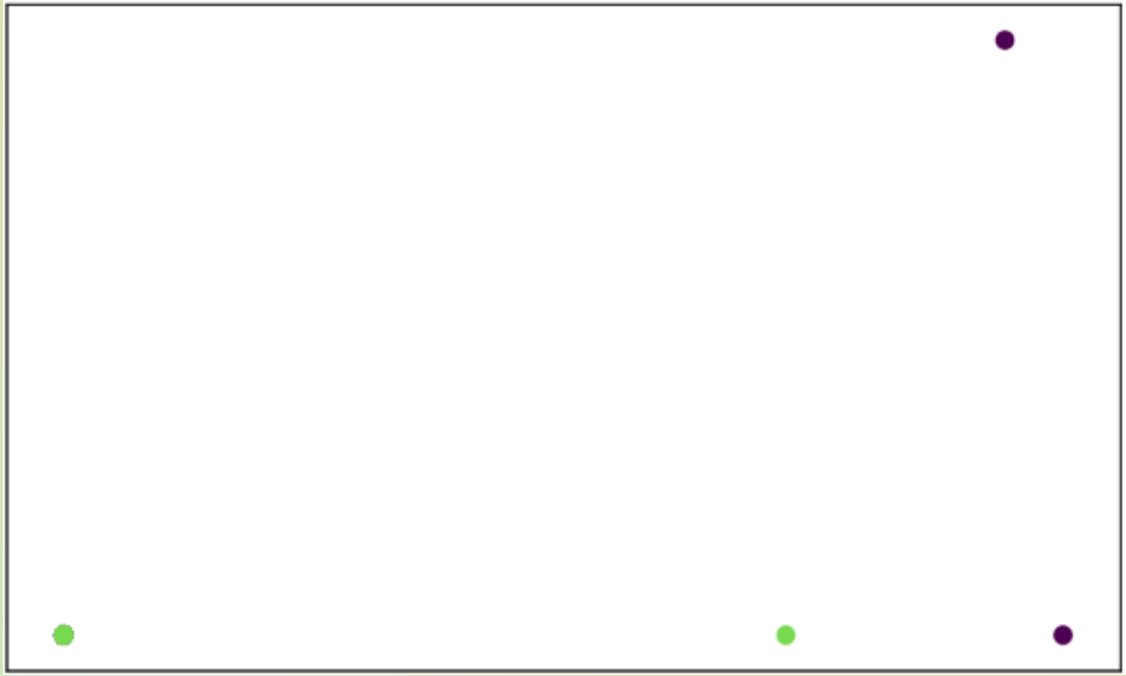


# Insights from the Bar Graph

- 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 5$^{th}$ most frequent word, which demonstrates the high importance of experience over education in the job market.

# Attempting K-Means Clustering

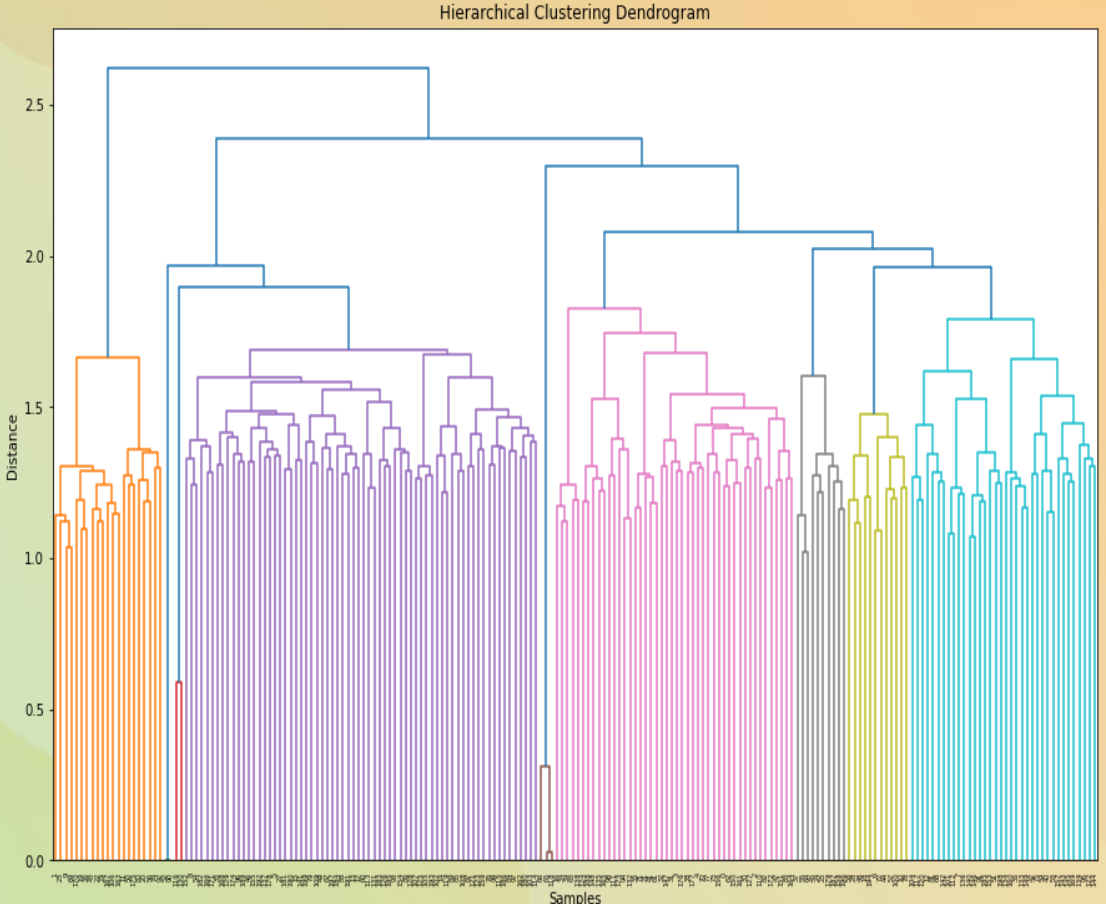
- When attempting to represent the 198 resumes using K-means clustering, it is clear that the graph fails to correctly cluster all the points. The reason behind this is the high number of dimensions that each vector possesses since we are using the Tf-Idf vectors to the resumes.



# Hierarchical Clustering Analysis:

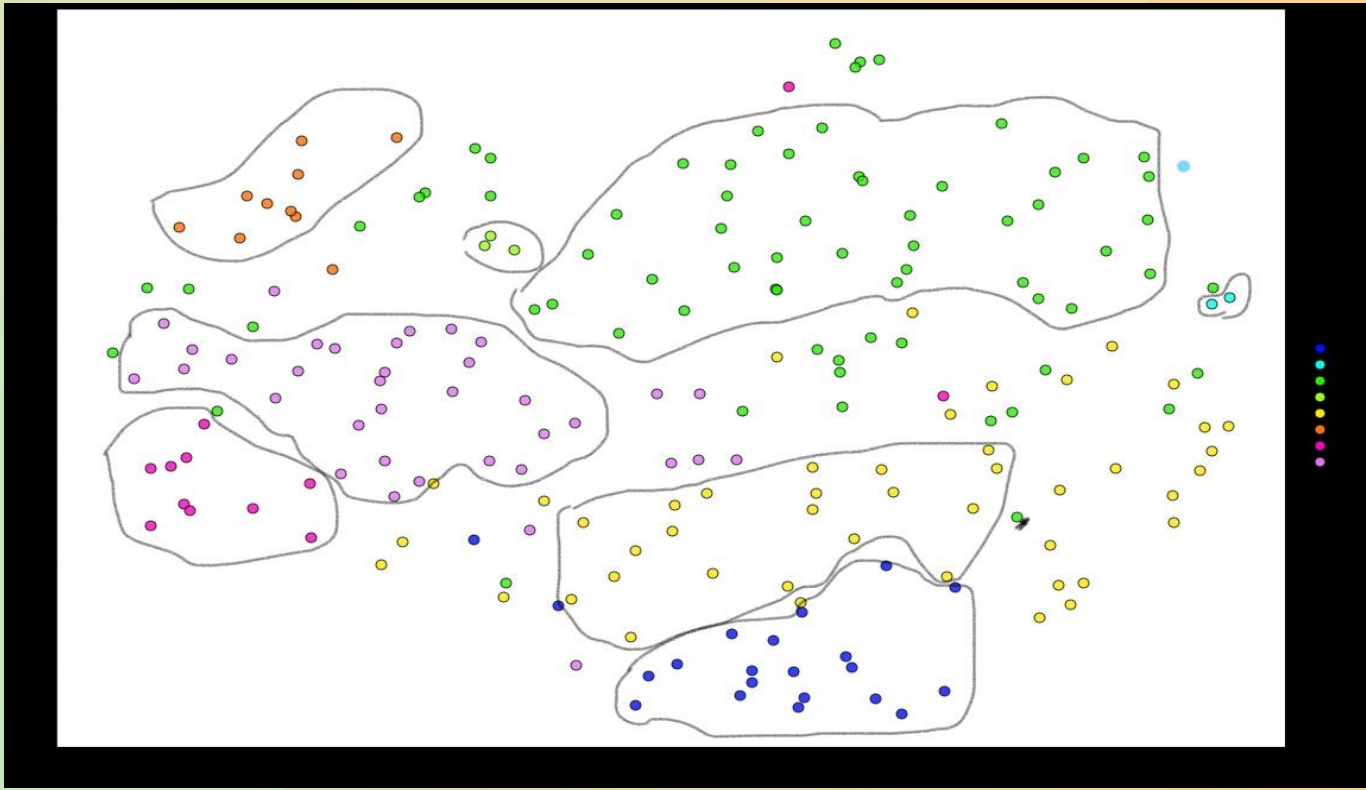
Hierarchical clustering groups data over a variety of scales by creating a cluster tree or Dendrogram.
The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level.

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.



**Taking 8 clusters, we visualize the texts, using UMAP along with Seaborn and Matplotlib. It allowed us to get projections in 2D.**

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).



# Model Recommendation and Selection

- The model we selected is the **SpaCy NER Model**.
- Through our analysis and search, we concluded that Linear models are not an option.
- Due to time constrains, we were not able to delve deep into Deep Learning approaches that could have made of a great alternative for our case.
- Additionally, other ensemble models such as Bagging and Bootstrapping were not serving our purpose.

# THANK YOU!