Resume extraction

Sukriti Macker, Nishchay Vaid, Anish Mitra and Zeynep Nilsu Bozan

Problem description

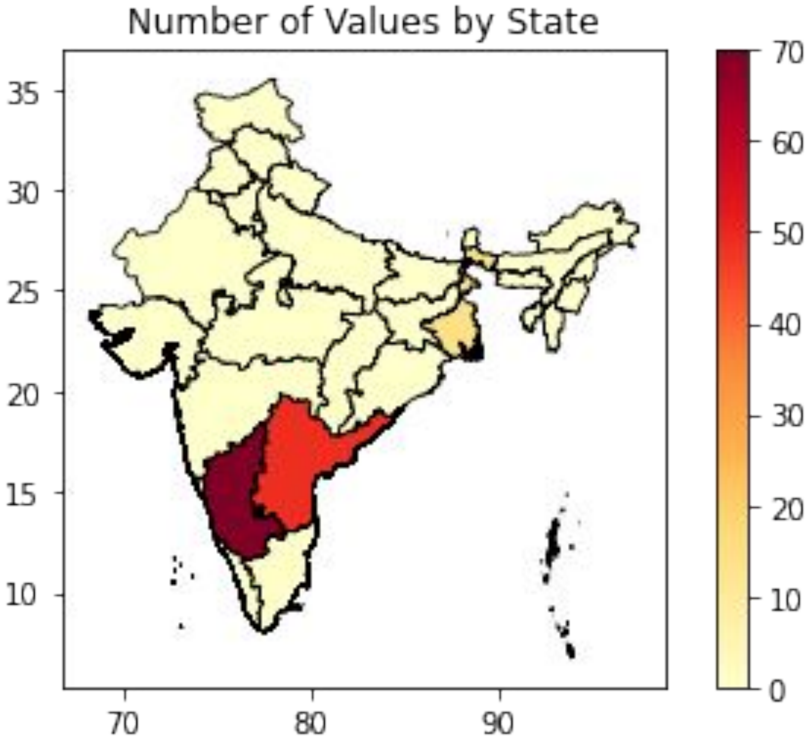
- Parsing resumes with very different formats and inputting the important information into the company database is something that sites with widespread usage such as workday do not do accurately.
- This is frustrating to both job seekers and HR professionals.
- We are trying to ameliorate this problem through this project.
- Implementing NLP techniques, we hope to reduce the difficulties involved in applying for a job and thus streamline the process in the human resources industry.

Proposed model

- We are proposing a Named Entity Recognition model for further use
- Reasons:-
  a. It allows us to classify data according to field(Location, Name, Companies Worked At, College etc.)
  b. It is an NLP technique.
  c. NER can deal with unstructured text.
  d. It is the most appropriate model for parsing resumes.
  e. It can extract and identify important information from the text while discarding superfluous information.

EDA

Location



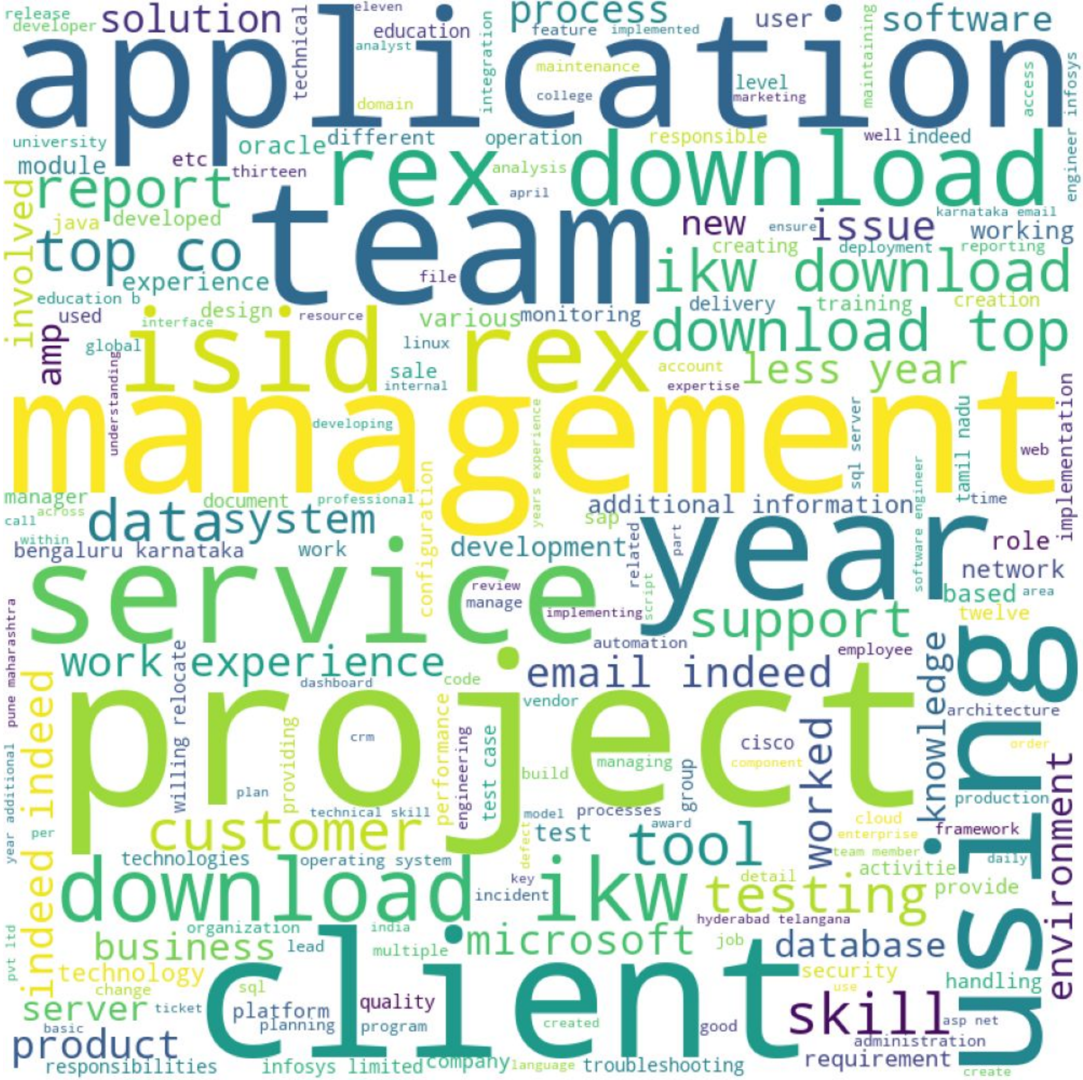
State	Workers in state
Karnataka	70
Andhra Pradesh	50
Maharashtra	49
Delhi	14
West Bengal	14
Chandigarh	6
Tamil Nadu	1

This slide shows the number of workers located in each state and a corresponding visualization on the Indian map.

Word cloud

A word cloud of the most common words appearing in the resumes.

Our observations emphasized that 'application', 'team', 'management', 'year', 'project', 'client', 'service' are most common words used in the resumes.

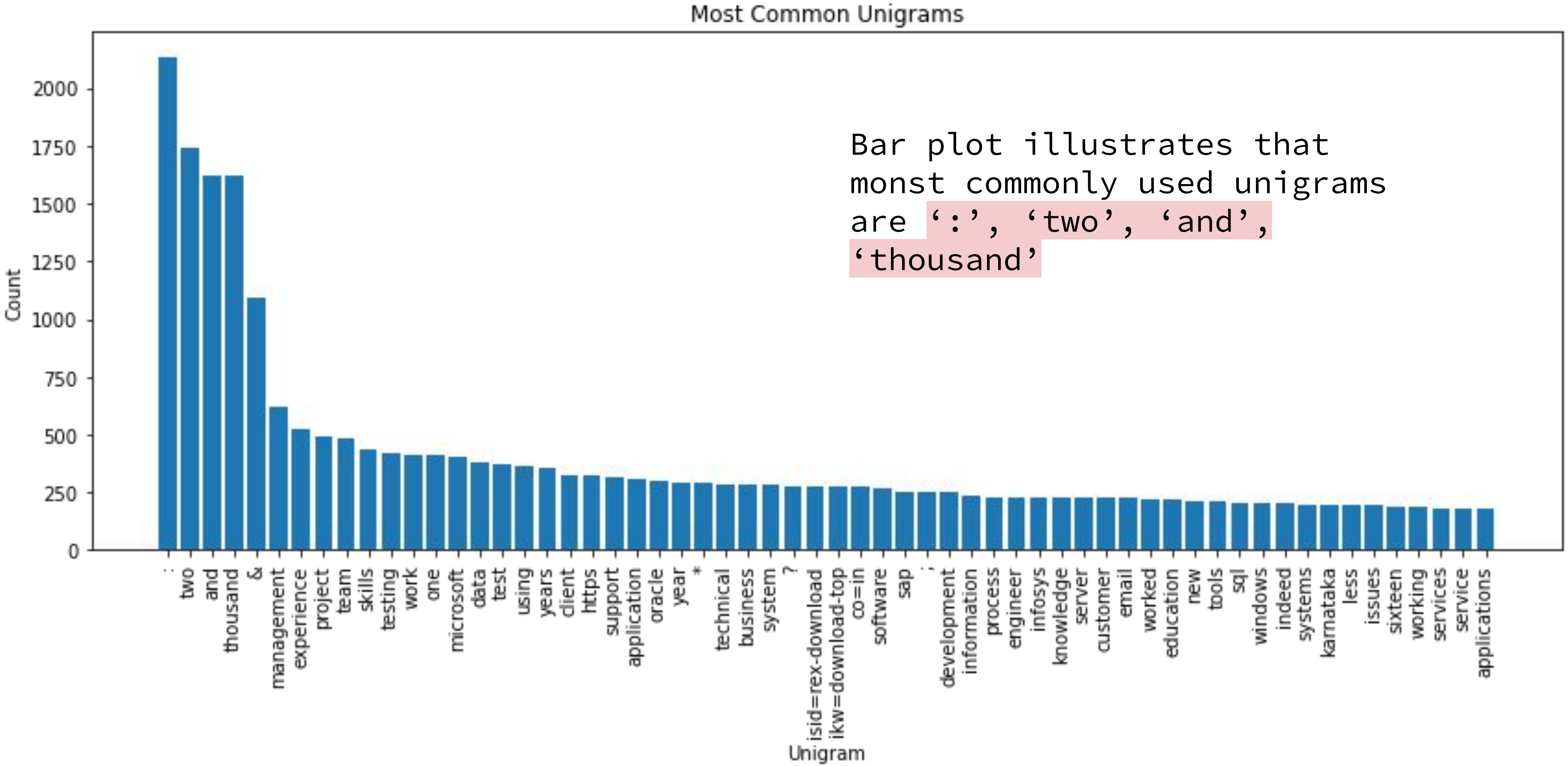


N-GRAM ANALYSIS

N-gram analysis is a technique used in natural language processing to extract contiguous sequences of n items (typically words) from a text. It helps identify patterns, relationships, and frequencies of word sequences, enabling insights into language usage and context.

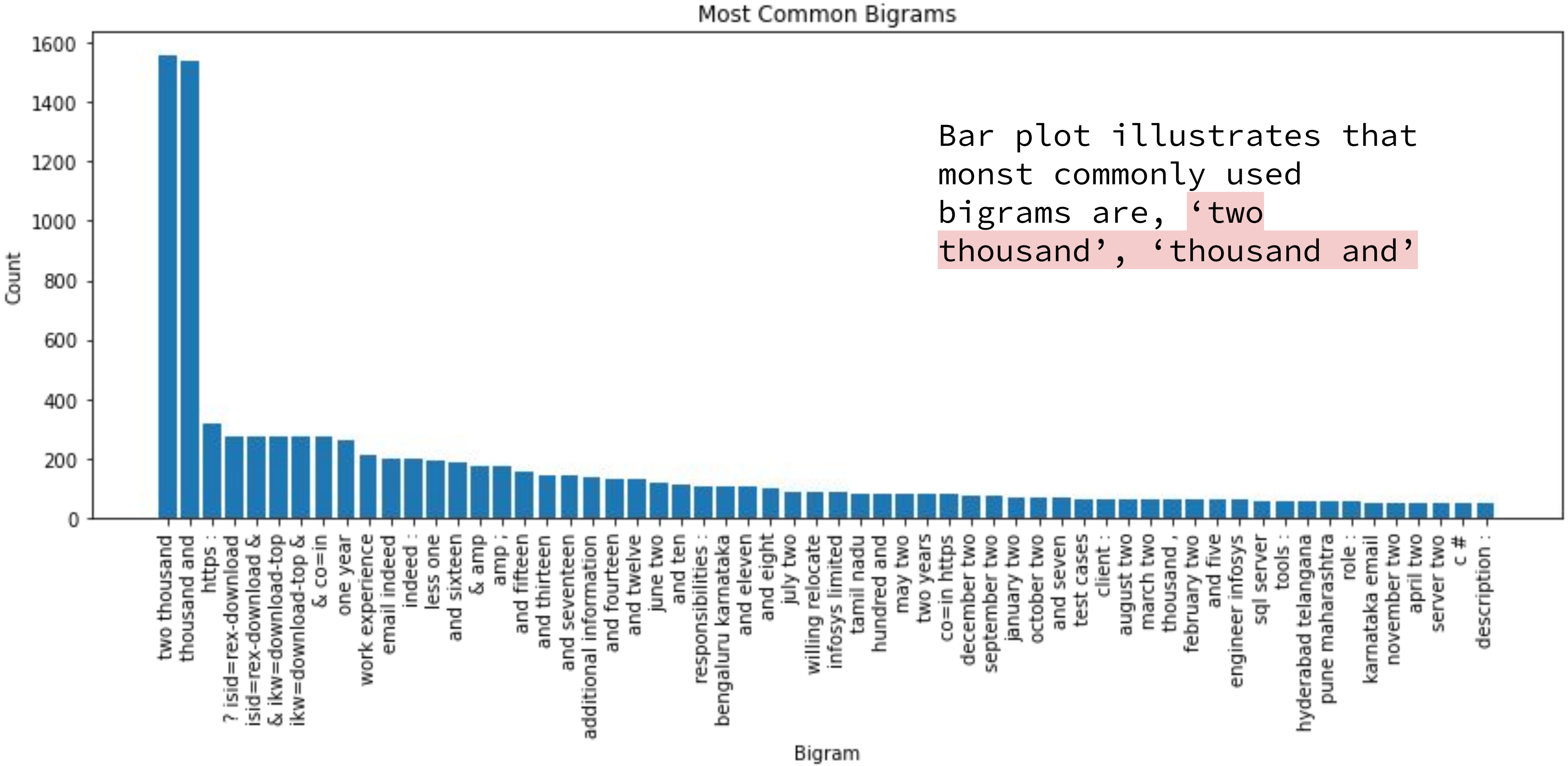
We computed unigrams, bigrams and trigrams to successfully detect most commonly used word groups.

Unigrams



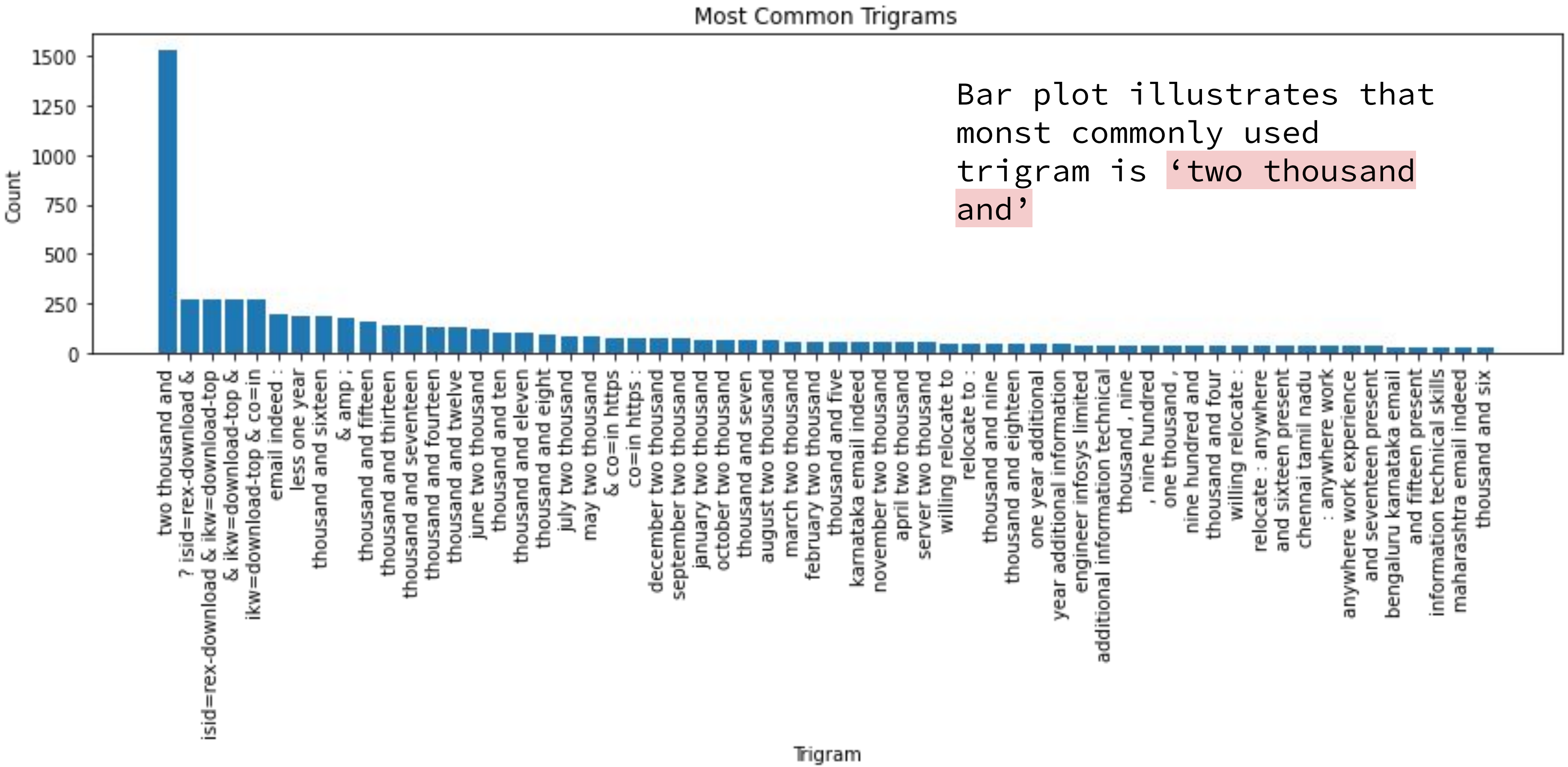
Bar plot illustrates that monst commonly used unigrams are ‘:’, ‘two’, ‘and’, ‘thousand’

BIGRAMS



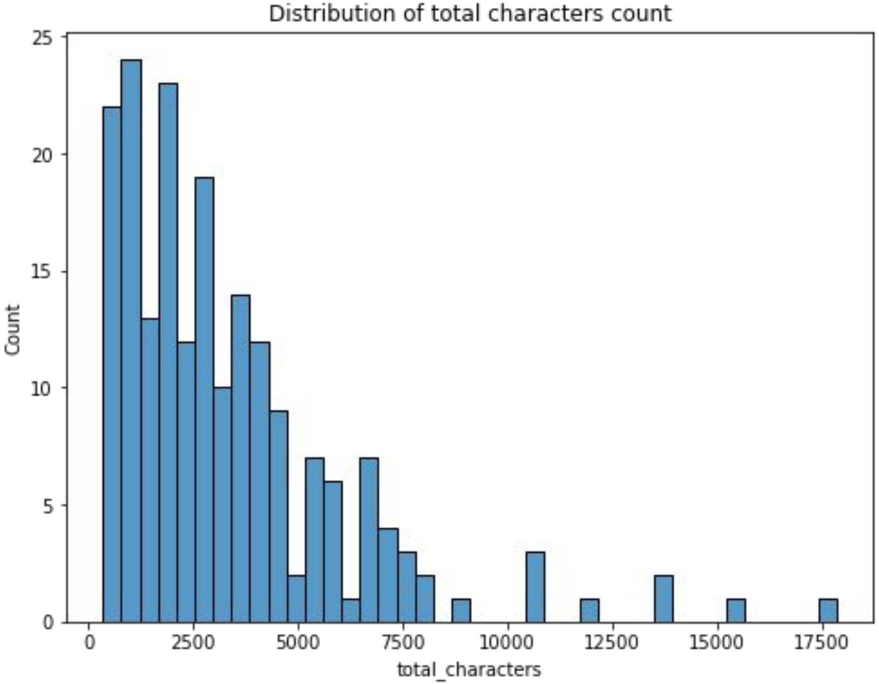
Bar plot illustrates that monst commonly used bigrams are, ‘two thousand’, ‘thousand and’

TRIGRAMS



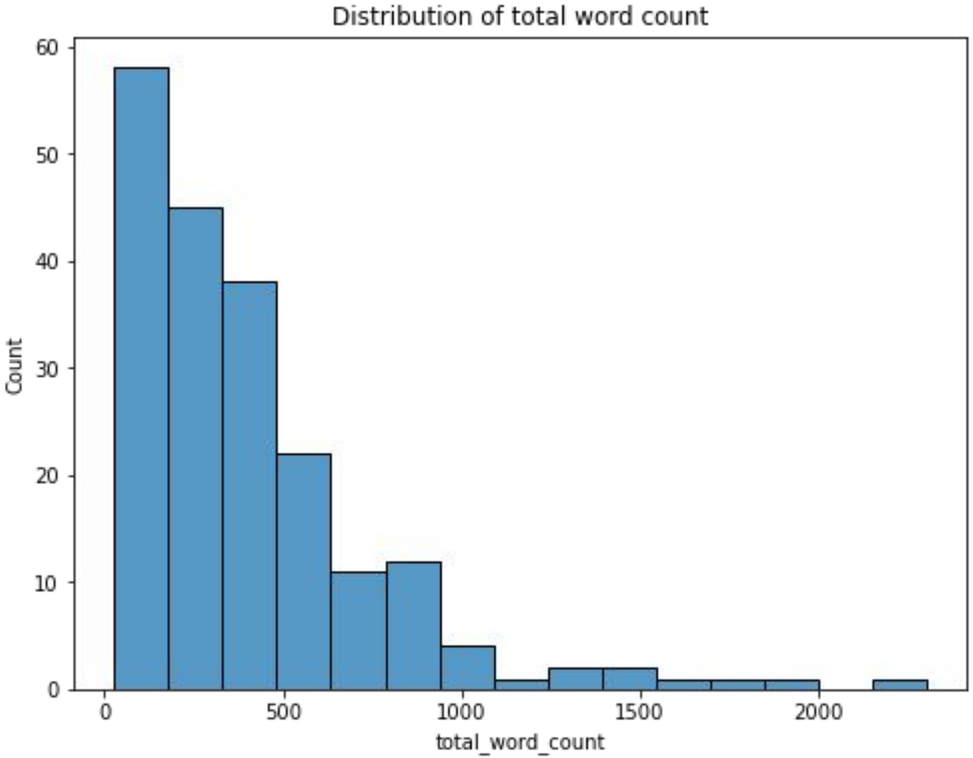
Bar plot illustrates that monst commonly used trigram is ‘two thousand and’

CHARACTER COUNT



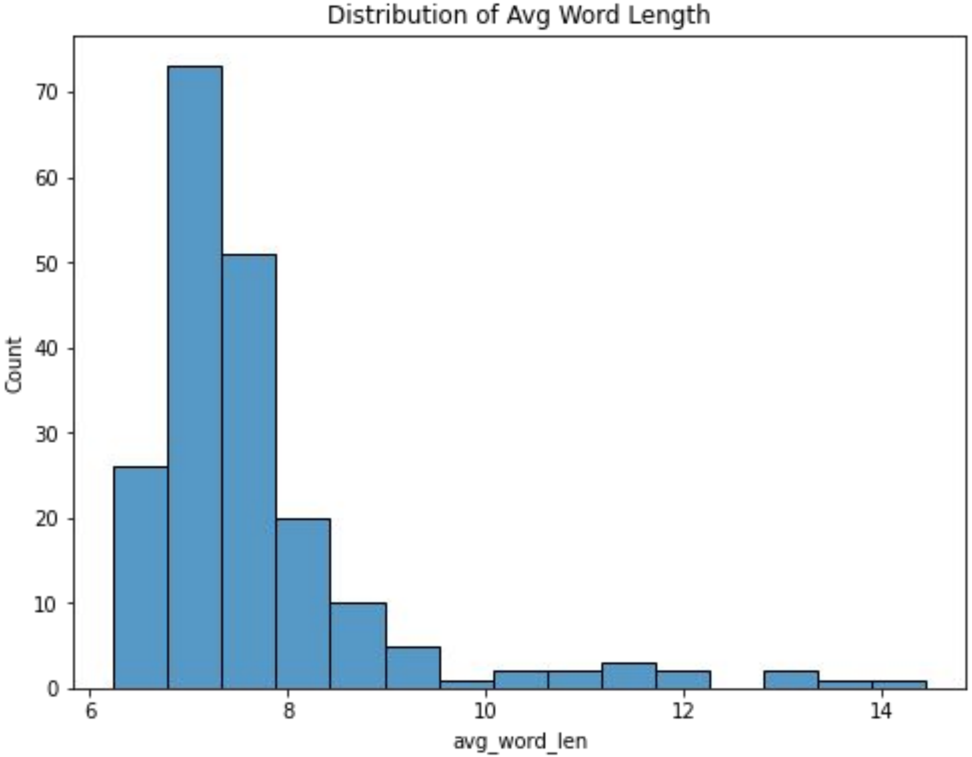
The average number of characters in the resumes is 3337 ± 2850 words. The resume with the fewest words is 337 while the highest is 17866 words.

Word count



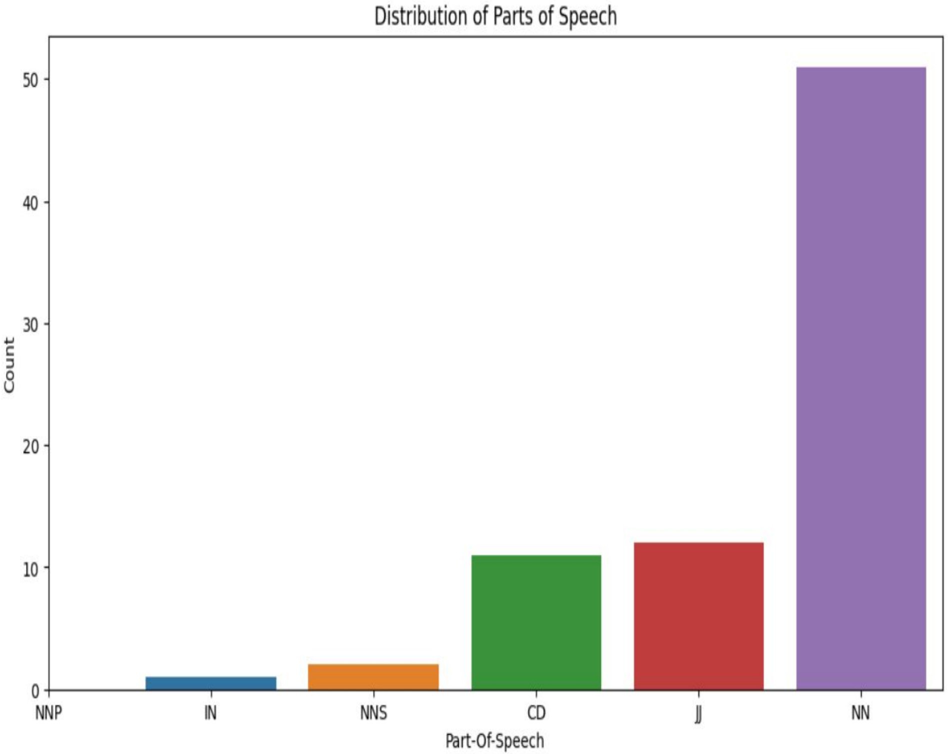
The average number of words in the resumes is 406 ± 362 words. The resume with the fewest words is 28 while the highest is 2801 words.

Average word length in resumes



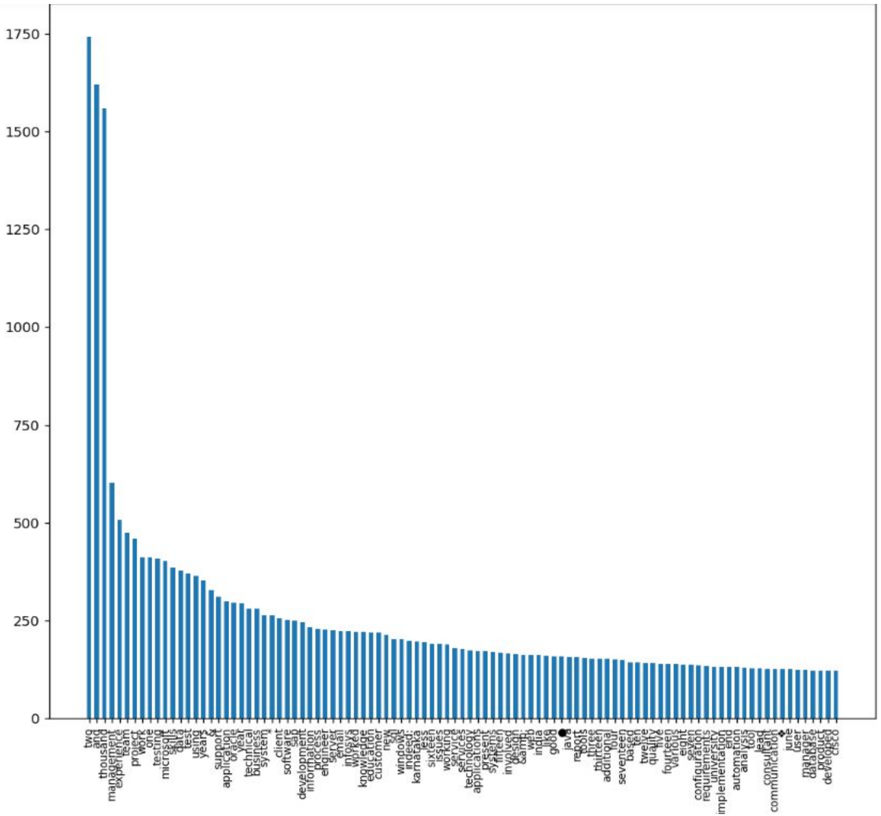
The average length of the words in the resumes is 7.70 ± 1.32 characters. The shortest average word length is 6.24 while the highest is 14.4 words.

Parts of speech analysis



When specifying the resume details, singular nouns dominate the description of each candidate. This showcases. This in turn can highlight the use of concrete details in most resumes to satisfy ATS format.

Frequency Distribution of 100 Most common words



1. "Experience" ranked as the fifth most commonly used word, emphasizing its importance in the job market over education.
2. Management experience is listed as the fourth most sought-after skill by employers.
3. Microsoft and Oracle are the most frequently mentioned companies on applicants' resumes.

THANK YOU!