NLP: Resume Extraction

Virtual Internship

Agenda

Team member’s details

Problem description

Approach

EDA

Recommendations

Team member’s details

- Group Name: NLP: Resume Extraction

- Name: Tatiana Moteu Ngoli
- Email: mtatiana@aimsammi.org
- Country: Germany

- Github repo link: https://github.com/TatianaMoteuN/Data-Glacier/tree/master/week11

Problem description

Resumes contain surfeit information that is not relevant for the HR/authority, and they have to manually process the resumes to shortlist the promising candidates for them. And, thus making the shortlisting task a herculean task for HR. By making use of the NER(Named Entity Recognition) model of NLP this problem can be solved by finding and classifying the entities that are present in each resume into predefined classes such as person name, college name, academics information, relevant experiences, skill set, etc.

Task

- Problem Understanding
- Data annotation
- Named Entity Recognition (NER)
- Model building & training
- Performance evaluation & reporting
- Model Deployment
- Model Inference

Data Exploration

- 2 columns: content & annotation
- Total data points :160
- 18 NER tags found: ORDINAL, WORK_OF_ART, NORP, GPE, FAC, TIME, ORG, DATE,LANGUAGE, PRODUCT, PERCENT, MONEY, LAW, EVENT, PERSON, QUANTITY

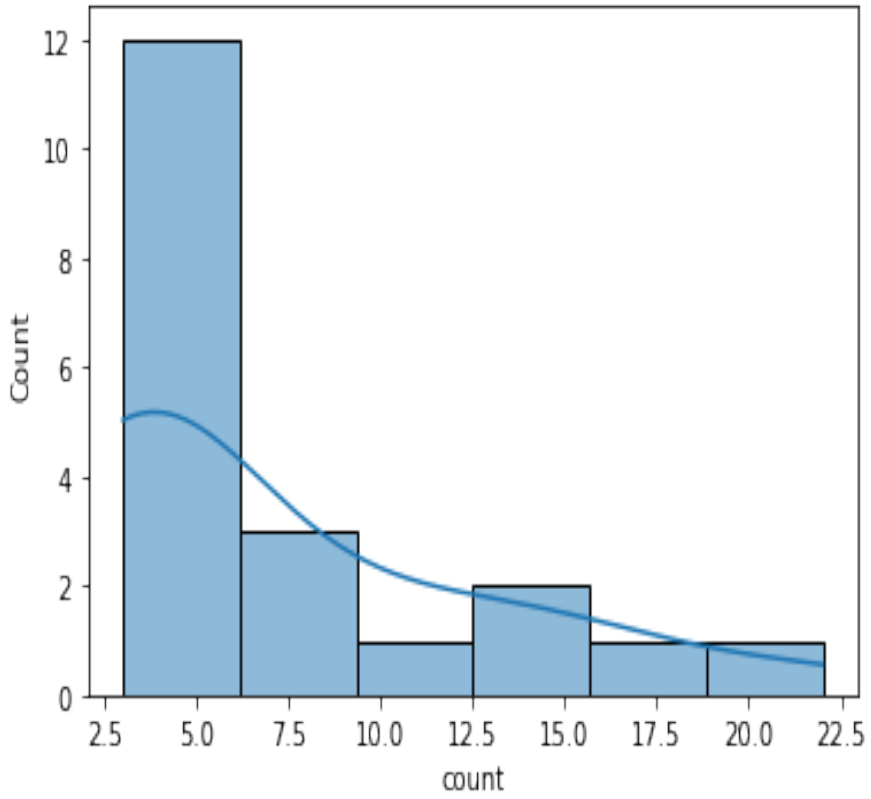
Assumptions:

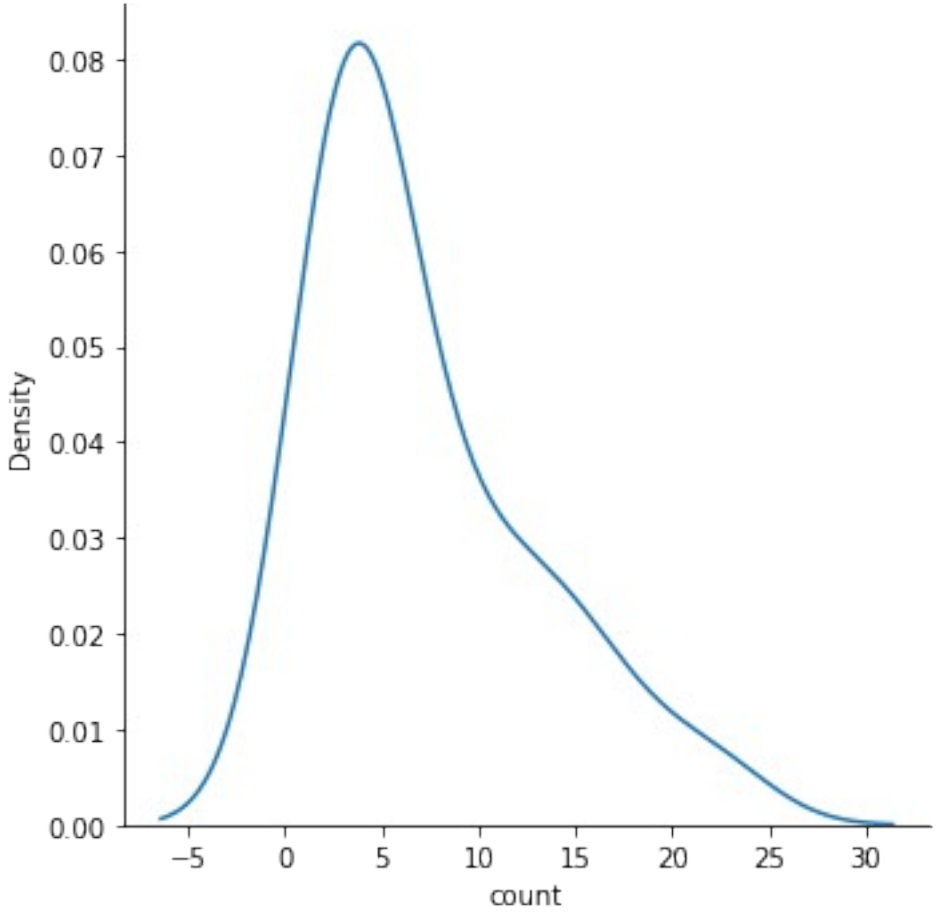
- We need to know what kind of person have applied in that company.

- The language speak by those persons

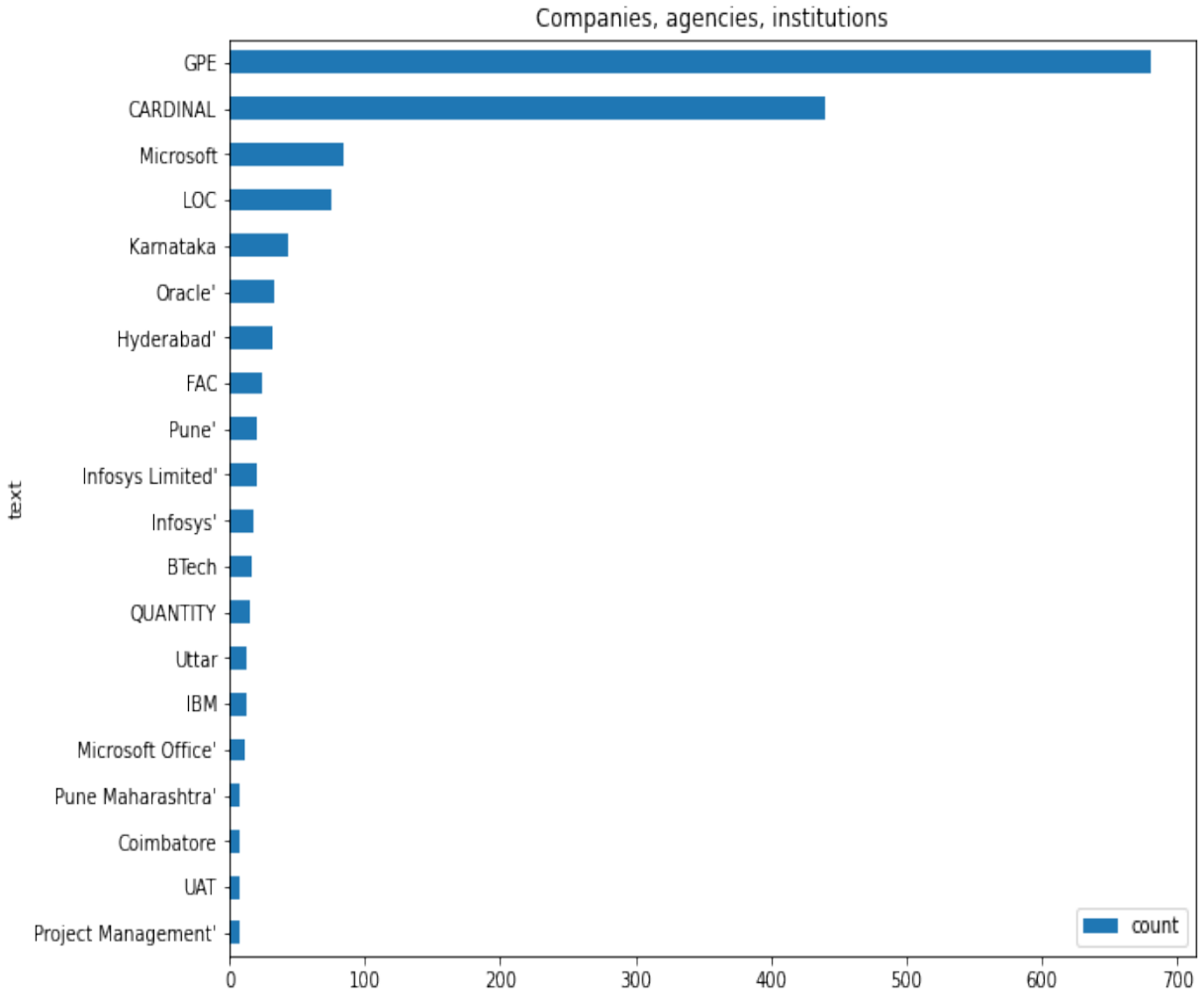
- Their past works
- And where they come from

Density PERSON Analysis



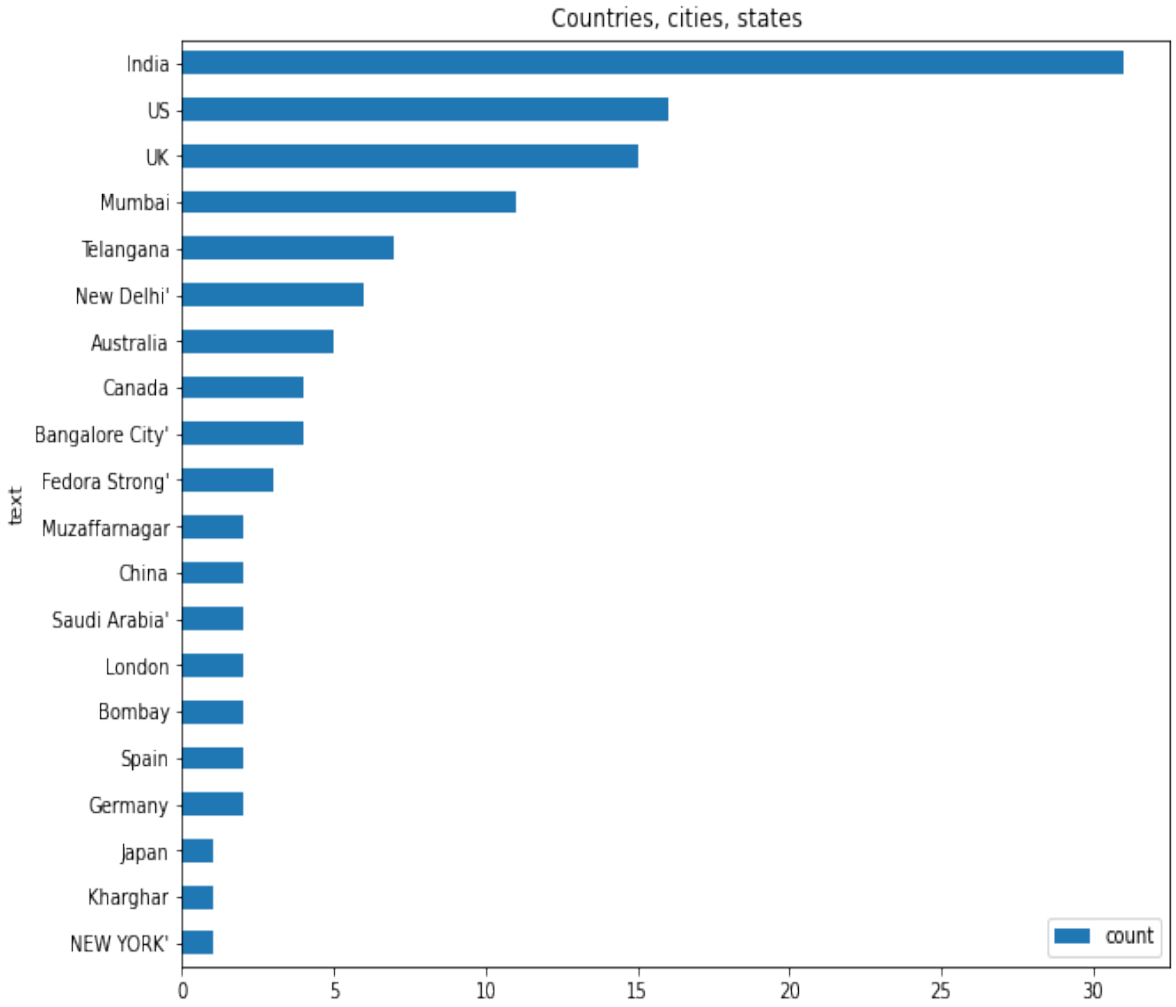


ORG Analysis



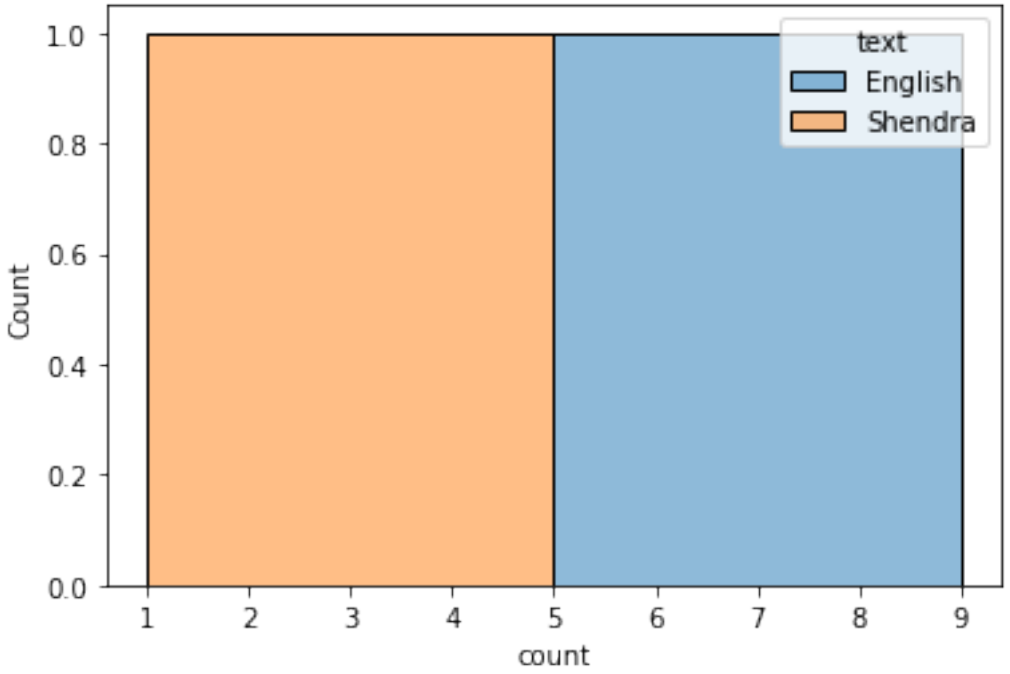
We can observe that most of people worked for GPE which showed a high distribution and next come CARDINAL and MICROSOFT

GPE Analysis



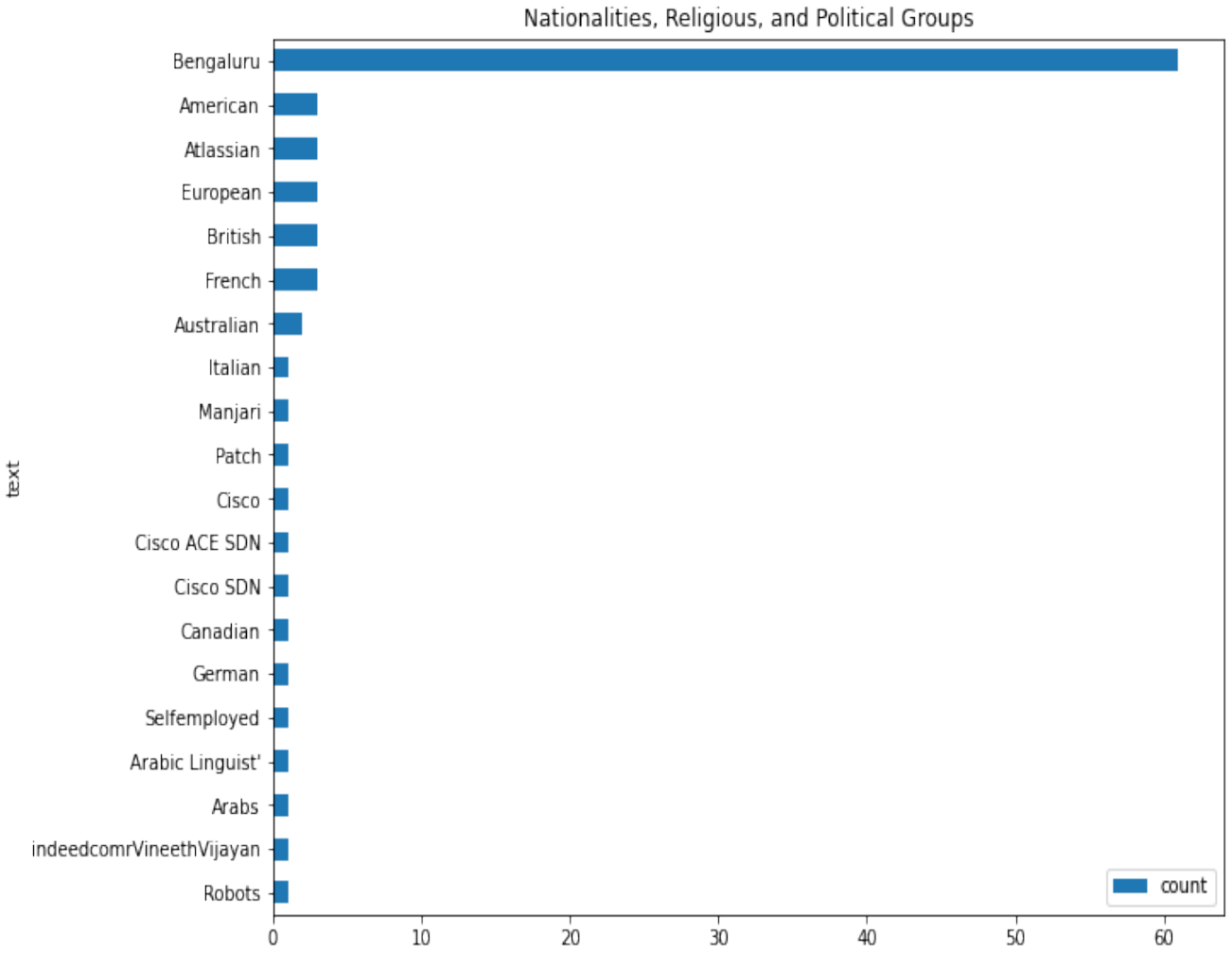
We can observe that the Indian country appear the most in the data which means that the resumes have been mostly submitted by people who leave in Indian. Then come US, UK and Mumbai.

LANGUAGE Analysis



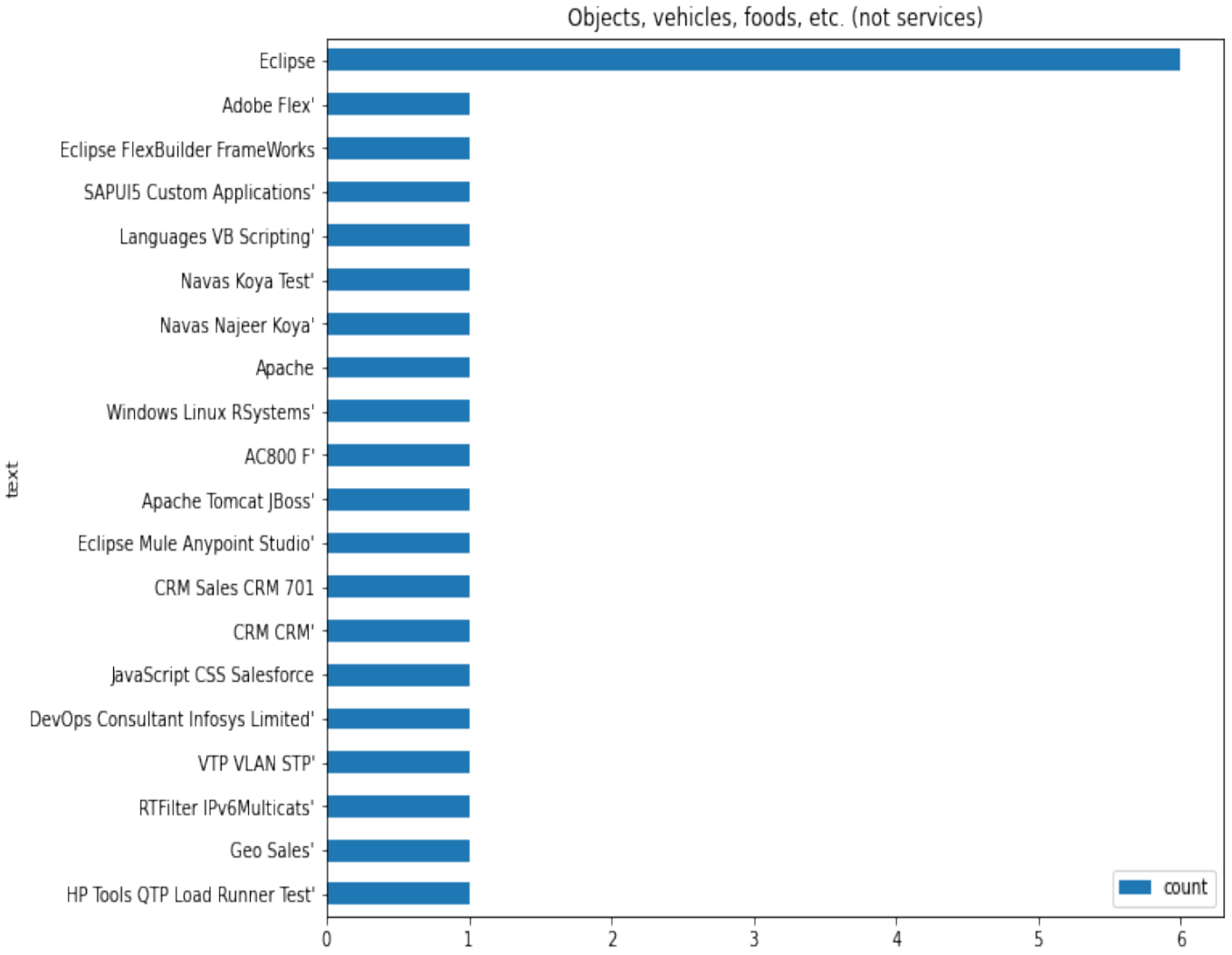
We can observe an equal distribution between the two languages (English and Shendra) found in the data which means that the candidates speak both English and Shendra

NORP Analysis



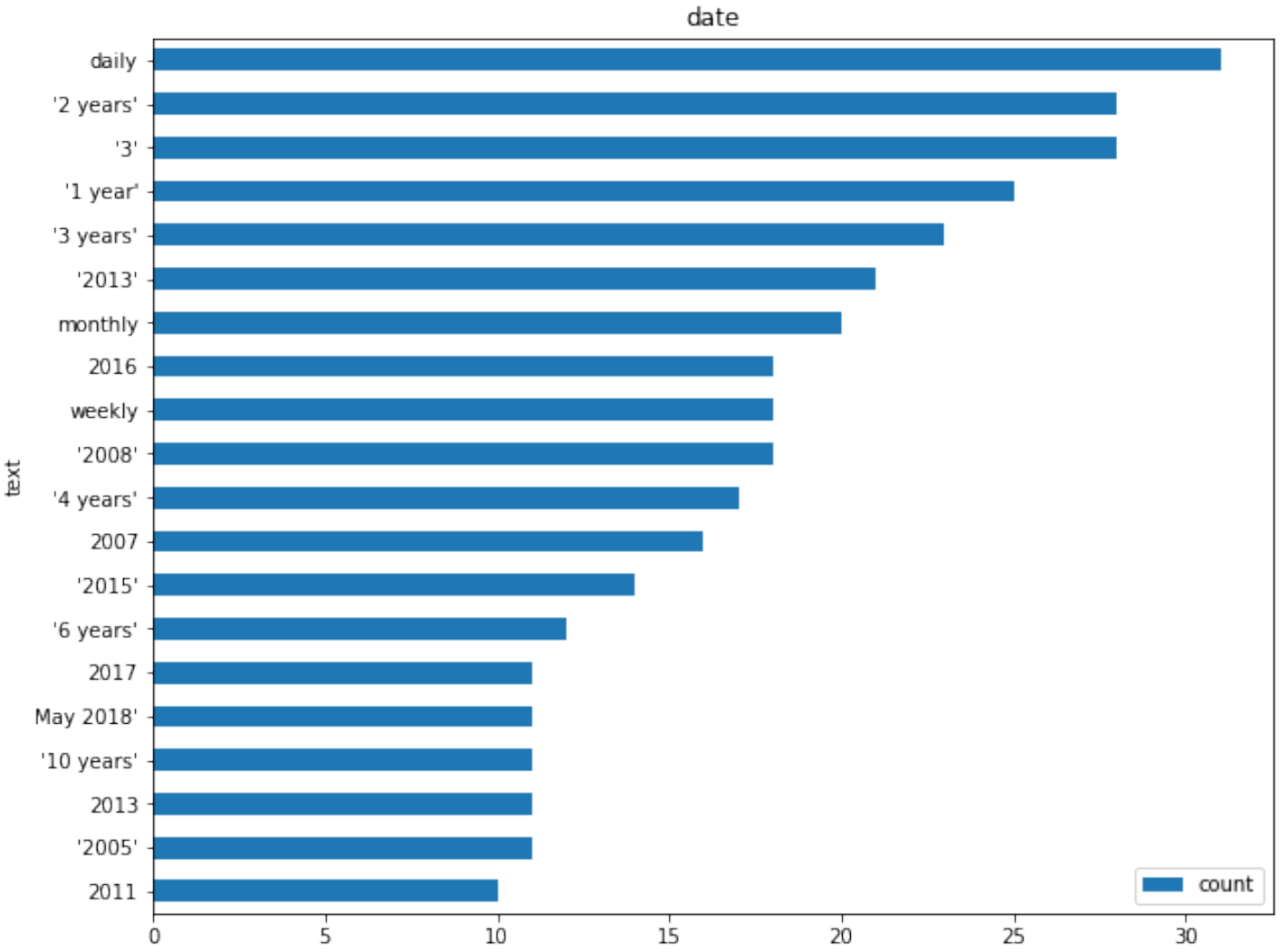
We can observe a higher distribution on Bengaluru which means that most of candidates come from Bengaluru. Then come American, Atlassian, European, British and French people.

PRODUCT Analysis



We can observe that many candidates have used Eclipse as material in their past work

Numbers of experience Analysis



We can observe that most of the candidates have done daily works and then most of them has 2-3 years of experiences in their fields

Recommended models

based on these observations, we recommend to focus on:

Years of experience : the years of experience of a candidate will give a certain idea of his profile and will testify to his skills and aptitudes to be shortlisted.

- **Companies:** The companies where they worked at will also give a considerable advantage to be shortlisted

- **Materials:** the materials employed by each candidate in his past work will determine wether the candidate fit the position or not.

On the analysis of above point and the given datasets , we will recommend to use a custom NER model in Spacy.
To do so we will need to:

- **Install Spacy and Spcay transformers**
- **Have the text with the corresponding annotations**

Thank You