

NLP and Text Mining data in R

Dataset

The initial step to perform NLP was finding the right dataset. For this, initially I used a dataset downloaded from Kaggle but that caused more encoding issues.

I later scrapped the web for hindi data. My current dataset is an article from the Jagran website regarding the reasons behind the decline of INR to USD.

Difficulties

Finding the right dataset and managing the encoding of the file while reading it in R.

Cleaning the Dataset

The following three methods were used. The last method being the best among the three.

Method I:

- In this approach, I remove the stopwords from the lines read from the file. Removing the stopwords resulted in unwanted symbols added to the text.

```
for (sw in stopwords)
  hindi_faulty <- gsub(sw, "", hindi)
```

- This resulted in the symbol '◌' (unwanted symbol) being added to the cleaned data.

```
# Hence cleaning the data by removing the symbol.
hindi_faulty <- gsub("◌", ".", hindi)
```

Method II: Using regex to remove stopwords

- In this approach, I make a pattern of all the stopwords present in the stopwords-iso package and then use the gsub function to remove the stopwords from the file.
- Like the previous method this also causes symbols to be added to the file

```
pat <- paste0("\\b(", paste0(as.list(stopwords), collapse="|"), ")\\b")
# Removing the stopwords from the file.
hindi_without_stopwords <- gsub(pat, "", hindi)
```

Method III: Working method (The most efficient method of the 3)

- In this approach, begin removing the stopwords after annotating the text. It is done by removing the stopword rows from the annotated vector
- This approach doesn't add unwanted symbols on removing the stopwords
- If tokens in the dataframe (x\$token) matches with any of the stopwords from the list we delete that row from the annotated vector.

for (sw in stopwords)

```
x <- x[x$token != sw, ]
```

Analysis:

From this we see the parts of speech that are most frequent. Using this we can say what parts of the speech will helps us to identify the article better and in short.

The screenshot shows the RStudio IDE with a script editor and a console window. The script editor contains a function definition for 'xupos' and a call to the function. The console window shows the output of the function, which is a list of words and their corresponding POS tags.

Script Editor:

```
new_hindi.R x
Source on Save
head Next Prev All Replace Replace All
In selection Match case Whole word Regex Wrap
146:1 (Top Level)
R Script

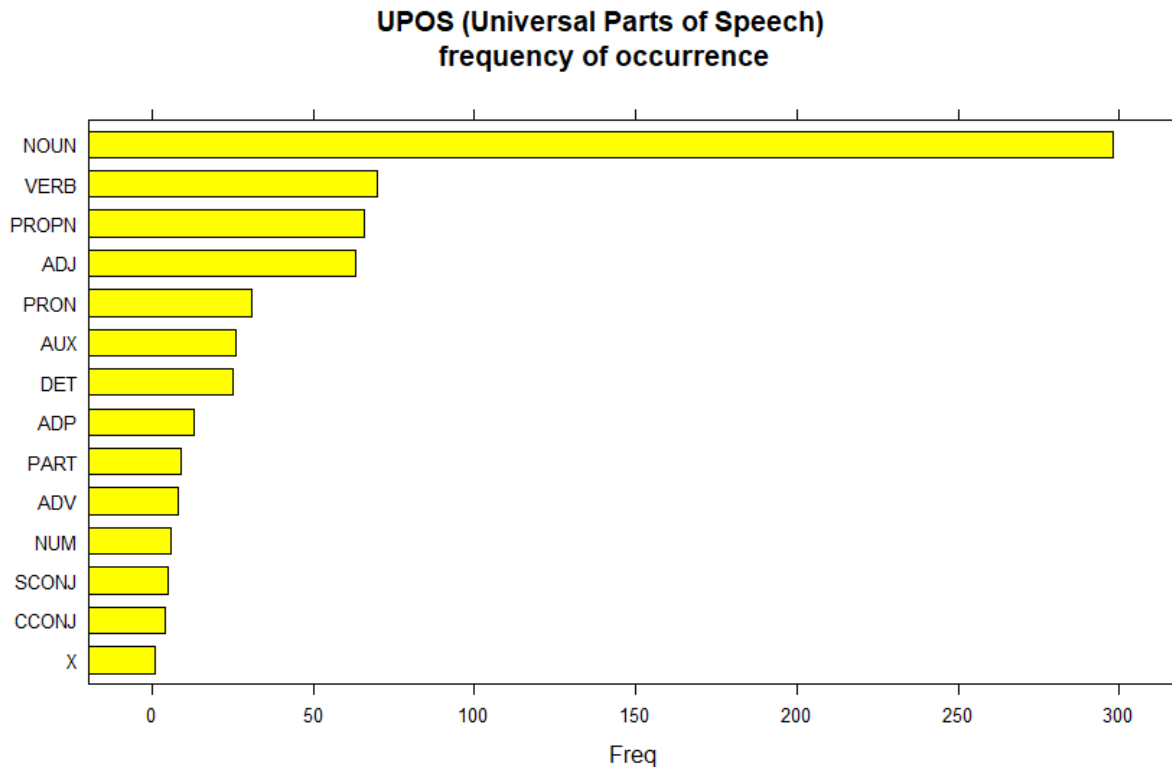
# Function definition
xupos = function(x) {
  # POS tagging
  # ... (omitted code) ...
}

# Call to function
xupos(x)
```

Console Output:

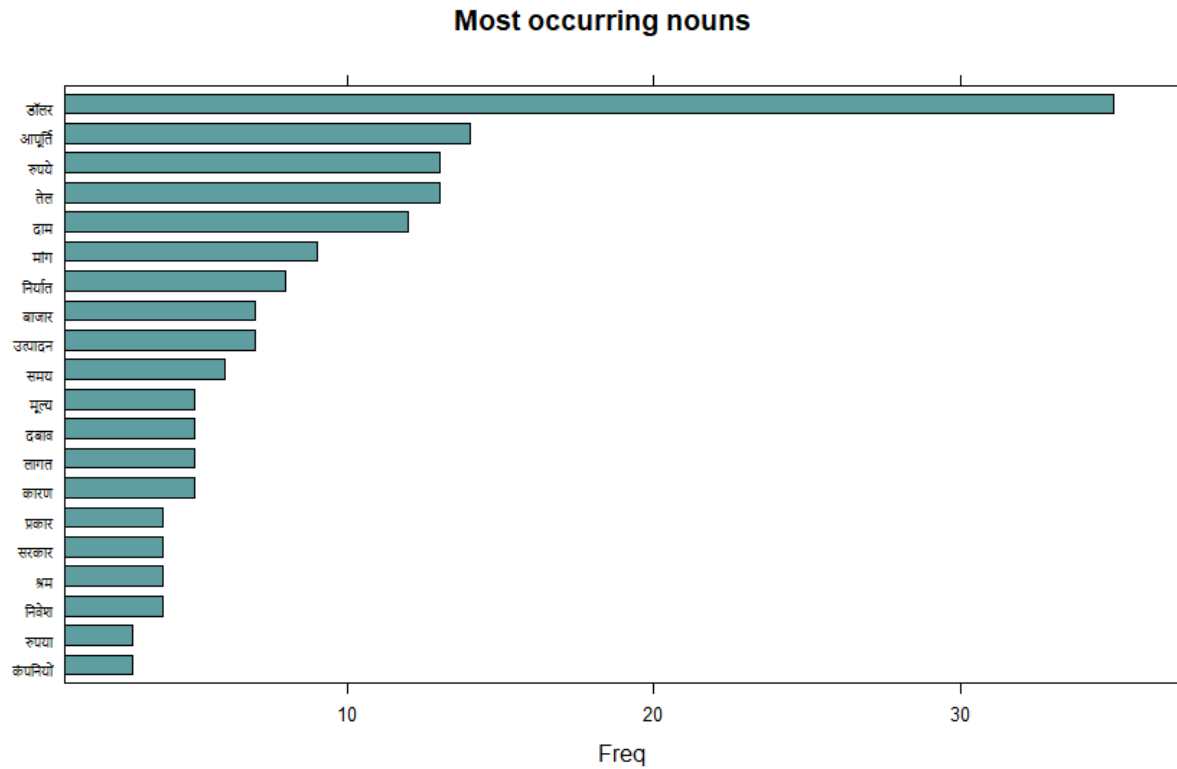
```
1
> table(xupos)
ADJ 63 ADP 13 ADV 8 AUX 26 CCONJ 4 DET 25 NOUN 298 NUM 6 PART 9 PRON 31 PROPN 66 SCONJ 5 VERB 70 X 1
> xupos
[1] "PROPN" "ADJ" "NOUN" "NOUN" "NOUN" "PART" "NOUN" "PART" "NOUN" "ADJ" "ADJ" "NOUN" "VERB" "NOUN"
[15] "NOUN" "VERB" "SCONJ" "NOUN" "ADP" "NOUN" "ADJ" "NOUN" "ADP" "X" "VERB" "AUX" "PRON" "VERB"
[29] "AUX" "NOUN" "NOUN" "NOUN" "DET" "NOUN" "NOUN" "VERB" "NOUN" "NOUN" "NOUN" "PRON" "NOUN"
[43] "NOUN" "ADJ" "NOUN" "NOUN" "NOUN" "NOUN" "NOUN" "NOUN" "ADJ" "NOUN" "ADJ" "PROPN" "NOUN"
[57] "NOUN" "VERB" "VERB" "NOUN" "NOUN" "DET" "NOUN" "NOUN" "DET" "NOUN" "NOUN" "DET" "AUX" "ADP"
[71] "NOUN" "NOUN" "NOUN" "DET" "NOUN" "DET" "NOUN" "NOUN" "VERB" "NOUN" "NOUN" "VERB" "ADV" "PRON"
[85] "ADJ" "NOUN" "NOUN" "NOUN" "NOUN" "DET" "NOUN" "NOUN" "DET" "PRON" "AUX" "NOUN" "NOUN" "VERB" "NOUN"
[99] "NOUN" "ADP" "NOUN" "VERB" "NOUN" "VERB" "PROPN" "ADJ" "NOUN" "NOUN" "NOUN" "NOUN" "NOUN" "PART"
[113] "NOUN" "NOUN" "VERB" "NOUN" "PART" "NOUN" "NOUN" "NOUN" "NOUN" "PART" "NOUN" "VERB" "NOUN" "PART"
[127] "NOUN" "NUM" "NOUN" "ADV" "VERB" "NOUN" "NOUN" "VERB" "PROPN" "NOUN" "NOUN" "NOUN" "NOUN"
[141] "NOUN" "DET" "VERB" "AUX" "PRON" "NOUN" "NOUN" "DET" "NOUN" "NOUN" "VERB" "NOUN" "NOUN" "VERB"
[155] "NOUN" "VERB" "NOUN" "PRON" "PRON" "DET" "NOUN" "NOUN" "VERB" "AUX" "SCONJ" "NOUN" "NOUN" "VERB"
[169] "AUX" "NOUN" "NOUN" "ADJ" "AUX" "SCONJ" "NOUN" "NOUN" "VERB" "NOUN" "DET" "NOUN" "VERB" "PRON"
[183] "ADJ" "NOUN" "DET" "NOUN" "NOUN" "NOUN" "VERB" "DET" "NOUN" "ADJ" "AUX" "NOUN" "NOUN" "VERB"
[197] "AUX" "NOUN" "NOUN" "NOUN" "ADV" "VERB" "AUX" "PRON" "VERB" "NOUN" "NOUN" "NOUN" "NOUN" "ADV"
[211] "VERB" "AUX" "NOUN" "NOUN" "VERB" "NOUN" "NOUN" "NOUN" "NOUN" "NOUN" "NOUN" "AUX" "PRON" "NOUN"
[225] "NOUN" "VERB" "NOUN" "NOUN" "PRON" "PROPN" "NOUN" "VERB" "NOUN" "PRON" "NOUN" "PRON" "ADJ" "NOUN"
[239] "NOUN" "CCONJ" "NOUN" "ADP" "NOUN" "NOUN" "VERB" "NOUN" "ADJ" "NOUN" "ADJ" "NOUN" "NOUN" "VERB"
[253] "PRON" "NOUN" "VERB" "VERB" "NOUN" "NOUN" "NOUN" "PRON" "NOUN" "NOUN" "SCONJ" "ADP" "NOUN"
[267] "PRON" "NOUN" "DET" "ADJ" "ADV" "ADJ" "NOUN" "VERB" "NOUN" "NOUN" "DET" "AUX" "ADJ" "NOUN"
[281] "NOUN" "NOUN" "CCONJ" "PROPN" "NOUN" "ADJ" "NUM" "NOUN" "NOUN" "NOUN" "NOUN" "ADJ" "NOUN" "DET"
[295] "NOUN" "NOUN" "ADP" "PRON" "NOUN" "NOUN" "NOUN" "DET" "NOUN" "ADJ" "PRON" "NOUN" "NOUN" "PRON"
[309] "NOUN" "NOUN" "VERB" "NOUN" "PRON" "DET" "NOUN" "NOUN" "ADJ" "NOUN" "VERB" "NOUN" "ADJ" "AUX"
[323] "PROPN" "PROPN" "PROPN" "PROPN" "PROPN" "PROPN" "ADJ" "NOUN" "NUM" "NOUN" "VERB" "NOUN" "NOUN"
[337] "ADJ" "ADJ" "NOUN" "NOUN" "PRON" "NOUN" "NOUN" "PROPN" "ADJ" "NOUN" "ADJ" "AUX" "ADJ" "ADJ"
[351] "NOUN" "NOUN" "VERB" "AUX" "NOUN" "VERB" "NOUN" "ADJ" "NOUN" "NOUN" "PART" "NOUN" "PROPN"
[365] "PROPN" "PROPN" "PROPN" "PROPN" "PROPN" "NOUN" "PROPN" "NOUN" "ADV" "NOUN" "NOUN" "NOUN" "VERB" "PROPN"
[379] "PROPN" "PROPN" "PROPN" "NOUN" "NOUN" "PROPN" "ADJ" "NOUN" "ADJ" "NOUN" "NOUN" "PRON" "ADJ" "NOUN"
[393] "VERB" "AUX" "NOUN" "NOUN" "ADJ" "ADJ" "NOUN" "NOUN" "VERB" "CCONJ" "NOUN" "NOUN" "NOUN"
```

The same thing can be graphically represented.



Resultant Analysis I:

- Through this we realize that if just plot the words that come under nouns we can a gist of what the entire article is about.
- Hence, we plot the Top 20 frequently occurring Hindi nouns. As you can see, the word dollar is the highest followed by service/import(aapurti) and then rupees. As you continue to the see the words that follow you understand that the article is related to the reasons behind the rise and fall of the rupee.

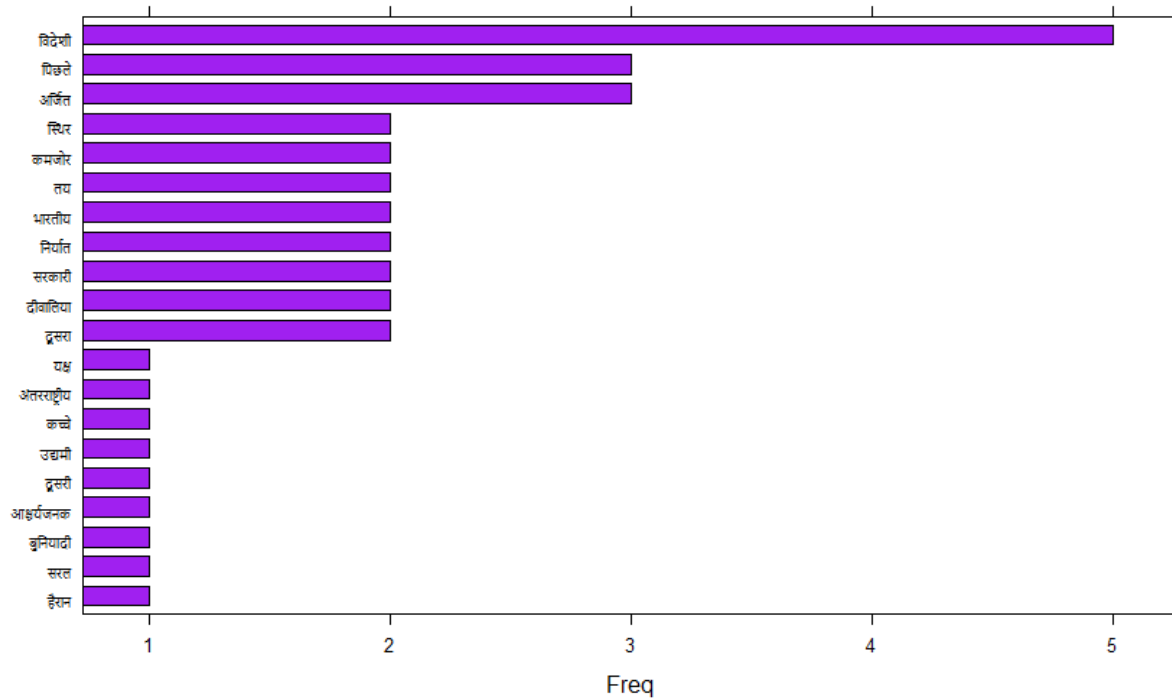


- The same can be plotted in word cloud. This visualization makes it easier to understand the frequency of the words. The words in yellow and along with the word in blue plays the most important part in the document.
- Through this, one can see that the rate of the rupee to dollar is related to the importing of petrol (oil/thel).

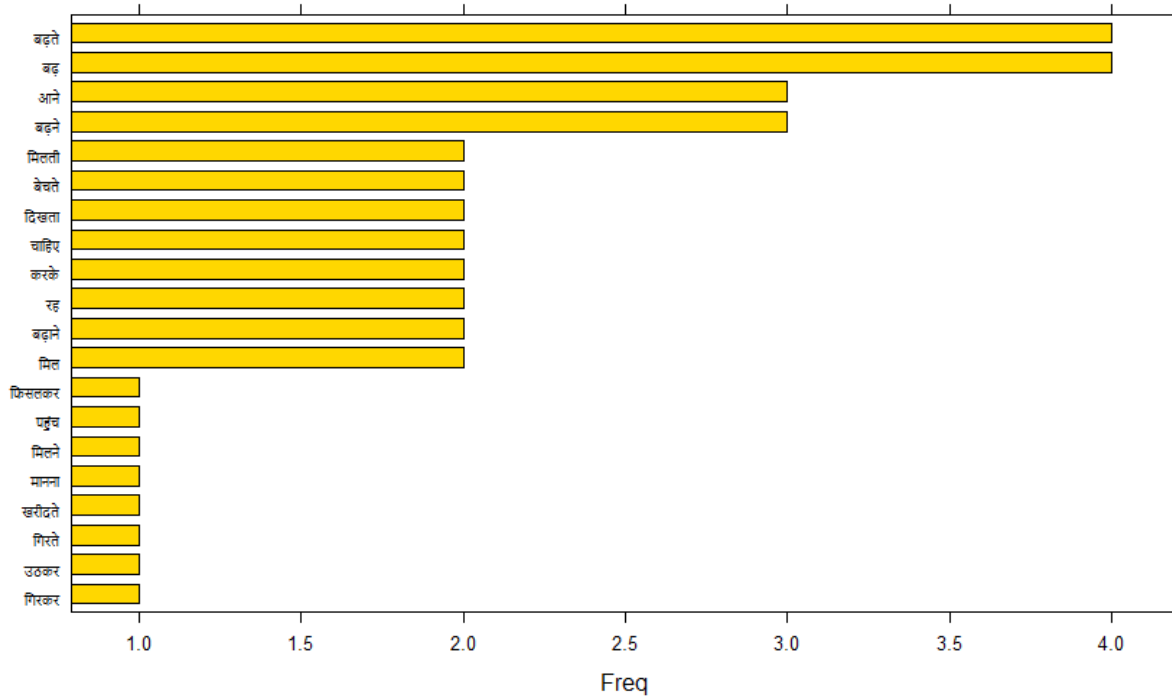


- Plotting adjectives and verbs so that we can get a better description of what the article is about.

Most occurring adjectives



Most occurring Verbs



- The frequency of the nouns, verbs and adjectives for a better understanding.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

new_hindi.R x

Source on Save

head Next Prev All Replace Replace All

☐ In selection ☐ Match case ☐ Whole word ☐ Regex ☒ Wrap

146:1 (Top Level)

R Script

Console Terminal

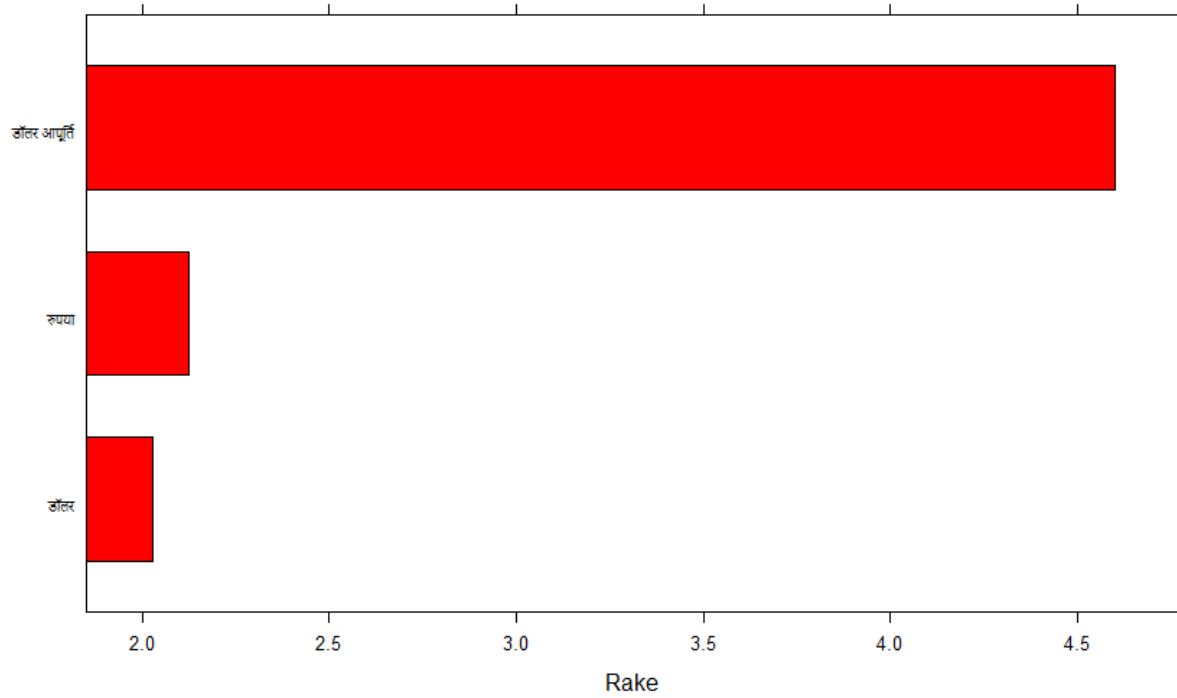
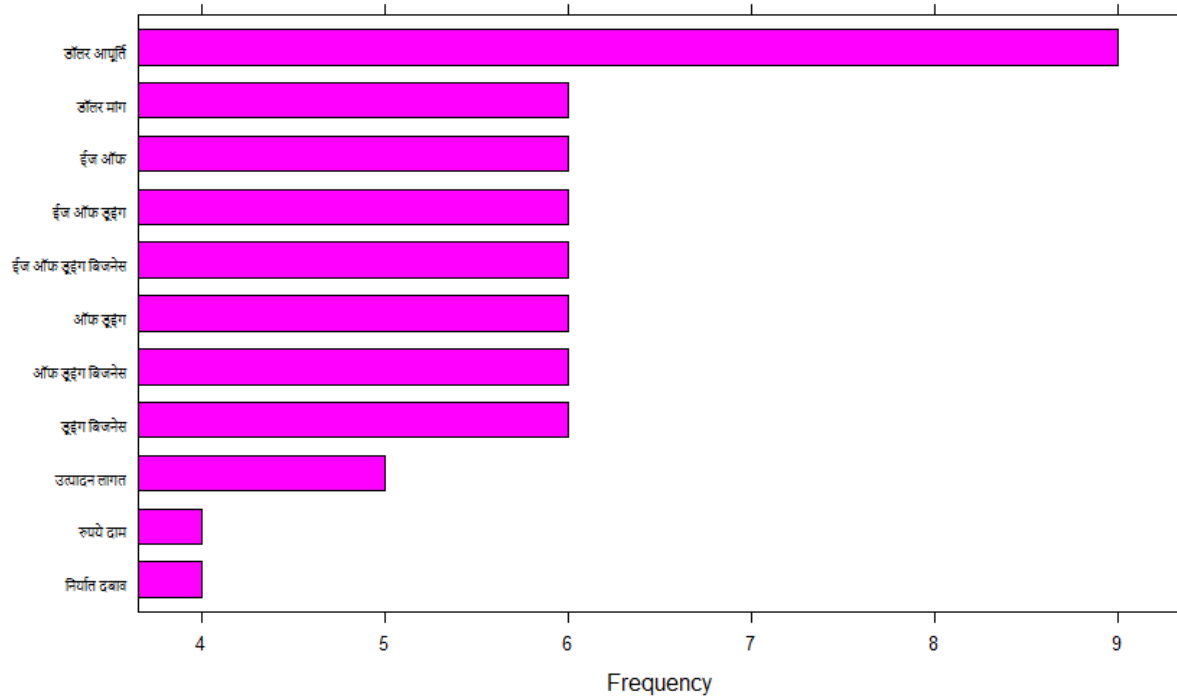
C:/NU/DSP/text/

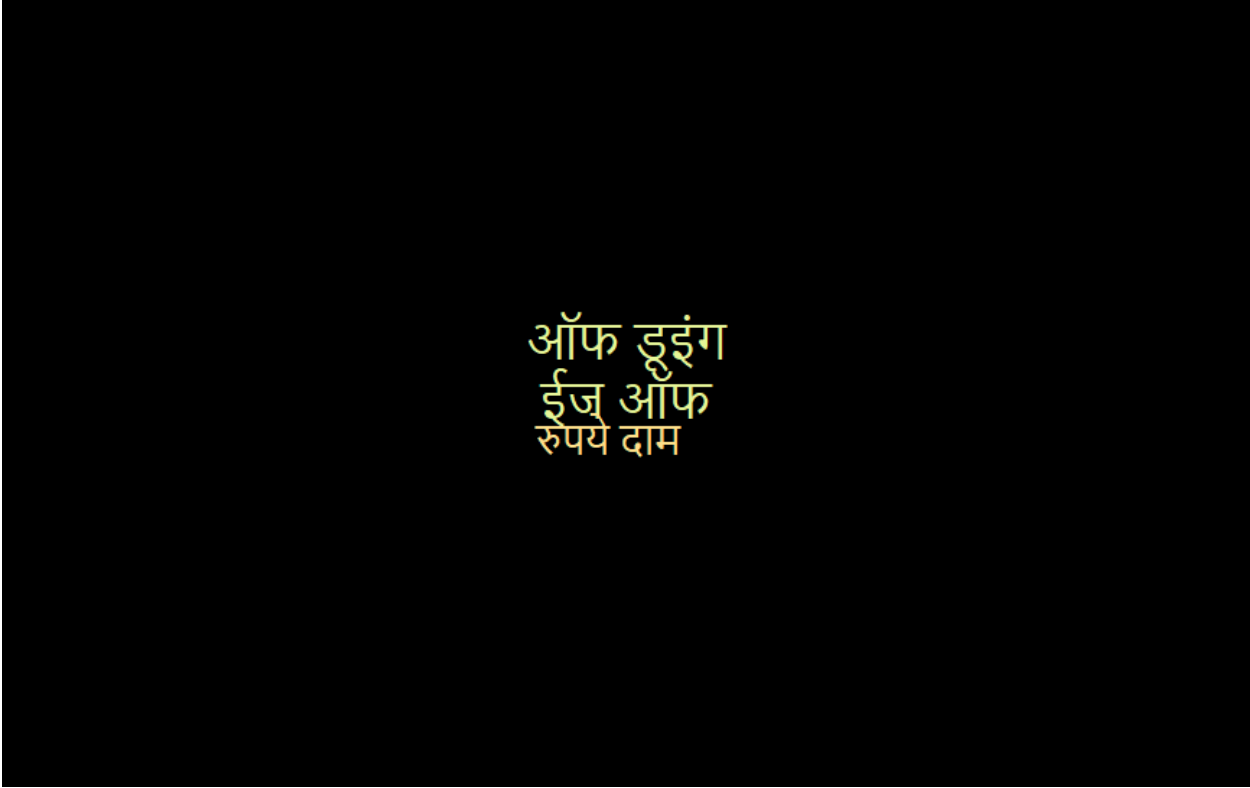
```
> sorted_freq <- sort(table(x$token), decreasing=T)
> sorted_freq
```

डॉलर	आपूर्ति	तेल	रुपये	ज्यादा	दाम	निर्यात	भारत	कम
35	14	13	13	12	12	10	10	9
मांग	उत्पादन	बाजार	ईज	ऑफ	कारण	डूइंग	बिजनेस	समय
9	7	7	6	6	6	6	6	6
इसलिए	दबाव	प्रति	मूल्य	रही	लागत	विदेशी	हम	हमारे
5	5	5	5	5	5	5	5	5
अमेरिका	गई	चीन	तरफ	निवेश	प्रकार	बढ़	बढ़ते	श्रम
4	4	4	4	4	4	4	4	4
सरकार	हमारी	हमें	अथवा	अर्जित	अर्थव्यवस्था	आने	उत्पादकता	कंपनियों
4	4	4	3	3	3	3	3	3
कर्मियों	खरीद	दोनों	पिछले	बढ़ने	मात्रा	मुख्यत	मुद्रा	यानी
3	3	3	3	3	3	3	3	3
रुपया	लगभग	विश्व	शिक्षा	सुधार	अंतरराष्ट्रीय	अधिक	अधिकारी	अप्रैल
3	3	3	3	3	3	2	2	2
अब	अरब	आलू	कदम	कमजोर	करके	चार	चाहिए	जाएगी
2	2	2	2	2	2	2	2	2
तय	तुलना	दामों	दिखता	दीवालि	दूसरा	देशों	दोगुना	परिस्थिति
2	2	2	2	2	2	2	2	2
परिबास	पूंजी	पेंसिल	प्रमुख	बहाव	बिजली	बीएनपी	बेचते	बैंक
2	2	2	2	2	2	2	2	2
बैरल	बढ़ाने	भारतीय	मंडी	मात्र	मिल	मिलती	रकम	रह
2	2	2	2	2	2	2	2	2
रैंक	ले	वर्ष	विपरीत	सरकारी	स्थिर	अदा	अधिकार	अनुपलब्धता
2	2	2	2	2	2	2	2	2
अनुरूप	अन्यथा	अभाव	अमेरिकी	अर्थ	अवधि	अहम	आगे	आती
1	1	1	1	1	1	1	1	1
आधार	आपूर्तिकर्ता	आश्चर्यजनक	आसान	इससे	उठकर	उठाए	उठाने	उद्यमी
1	1	1	1	1	1	1	1	1
उद्योगों	ऊंचा	एनडीए	ऐसा	ऑटो	कंपनी	कच्चे	कनेक्शन	कमा
1	1	1	1	1	1	1	1	1
करनी	कमी	कह	काट	कानून	काफी	काम	कार	कुशल
1	1	1	1	1	1	1	1	1

Resultant Analysis II:

- Through the graphs you can understand that due to the demand for petrol in India the dollar value increases.
- After the Trump government coming in, the flow of dollars from US to developing countries has reduced which affects the ease of doing business.
- Also, the cost of labor in India is disproportional to the goods produced in comparison to China which is noted by the less frequent words in the document but also playing a key factor in the ease of doing business.
- This trend is evident from the type of highly occurring adjectives and verbs plotted in the graph.

Keywords identified by RAKE**Keywords - simple noun phrases**



ऑफ डूइंग
ईज ऑफ
रुपये दाम

Analysis I & II: Justification

Furthermore, through the use of the RAKE library you can see the key noun phrases that are commonly occurring in the data. The key focus is on the ease of doing business which justifies the above key adjectives and verbs plotted.

Conclusion:

The above graphs and wordclouds justify what the article is about and with zero to little effort you can understand that the entire article focuses on the dollar to rupee rate. It also emphasizes about the factors that affect the decline in the rupee rate to dollar.