Word Beach

29-06-2021

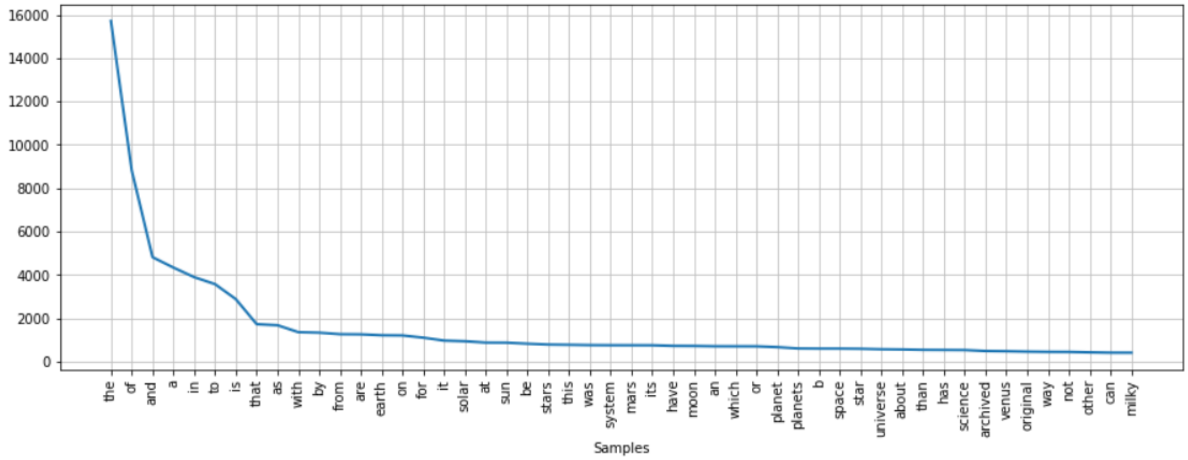
Vamshi Krishna Bonagiri
2020114011

Overview

Word clouds present a low-cost alternative for analyzing text from online surveys, plus it's much faster than coding. Thus the need of such tools is important, but there are not many tools like this available for this in Telugu, which is why this project aims to be able to generate Telugu wordclouds. This word cloud was built using beautifulsoup for extraction of data and nltk libraries for processing the data, and word cloud library for the visualization.

Analysis (Frequency graphs and their analysis)

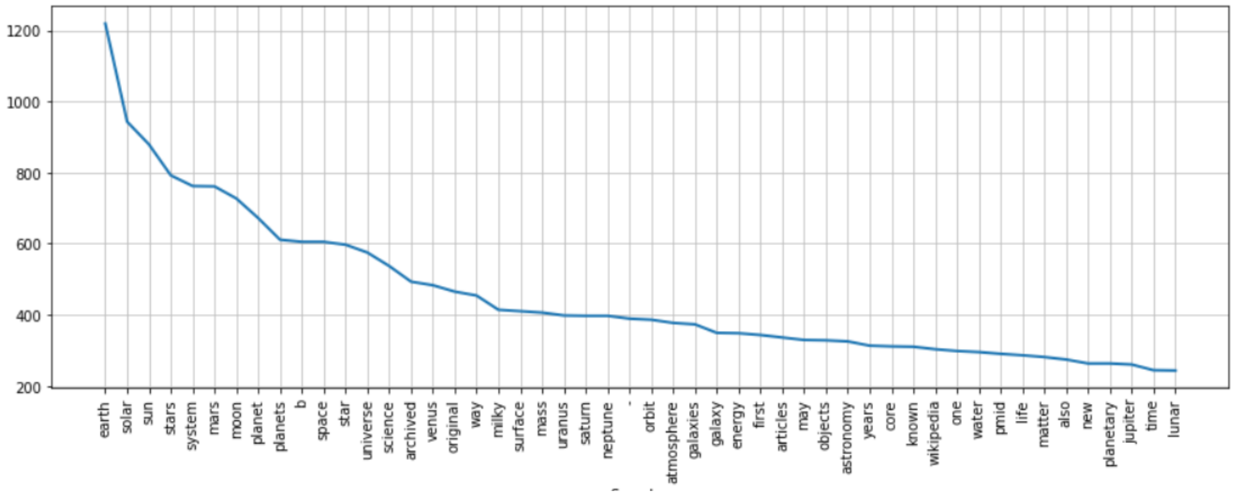
Tokenization



The most frequent words in the data seem to be words like “the” “of” “and” etc. Most of the top 50 words are these, this is true for most of the english text taken since they occur so frequently. These are called stopwords.

The analysis was the same when it was telugu text, which is why I made a manual list of telugu stopwords.

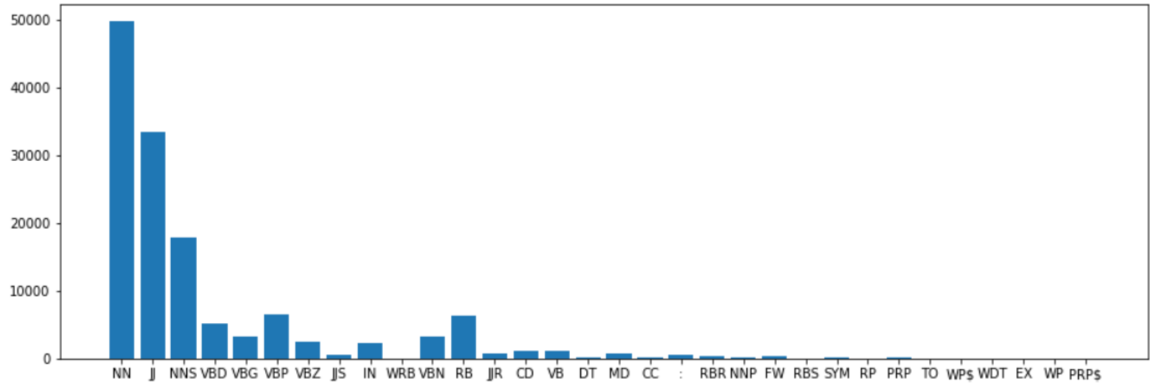
After removing Stopwords



After removing english stopwords, the text seems to resemble the article more now. But we still have some words like “years, one, new, also” which are not really describing the articles chosen(science).

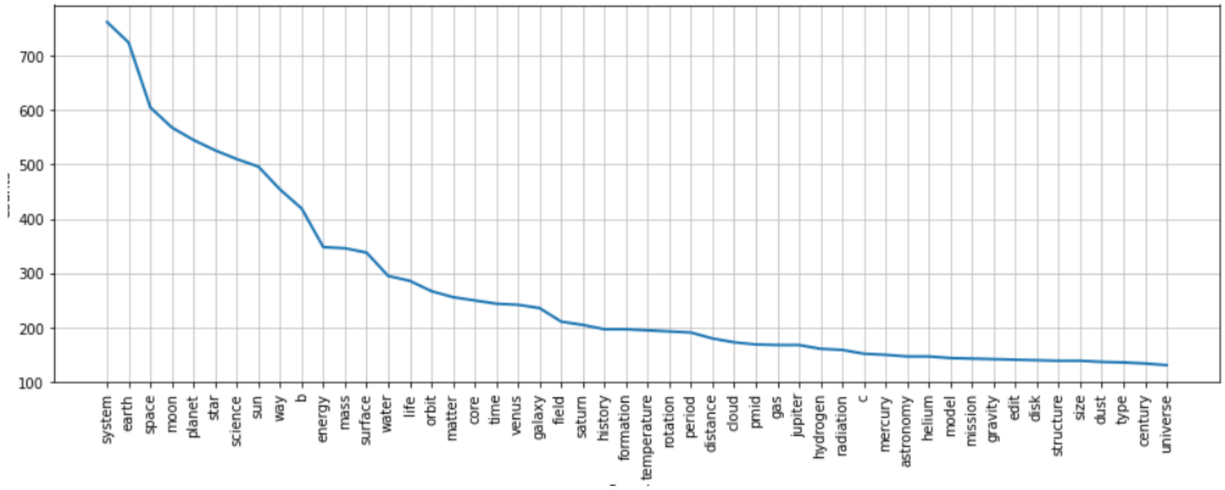
Since stop words are not very helpful in analyzing text, we will remove them in the article.

Parts of Speech



Frequency distribution of parts of speech, most of the data seems to be coming from NN, JJ and NNS, lets focus on those parts.

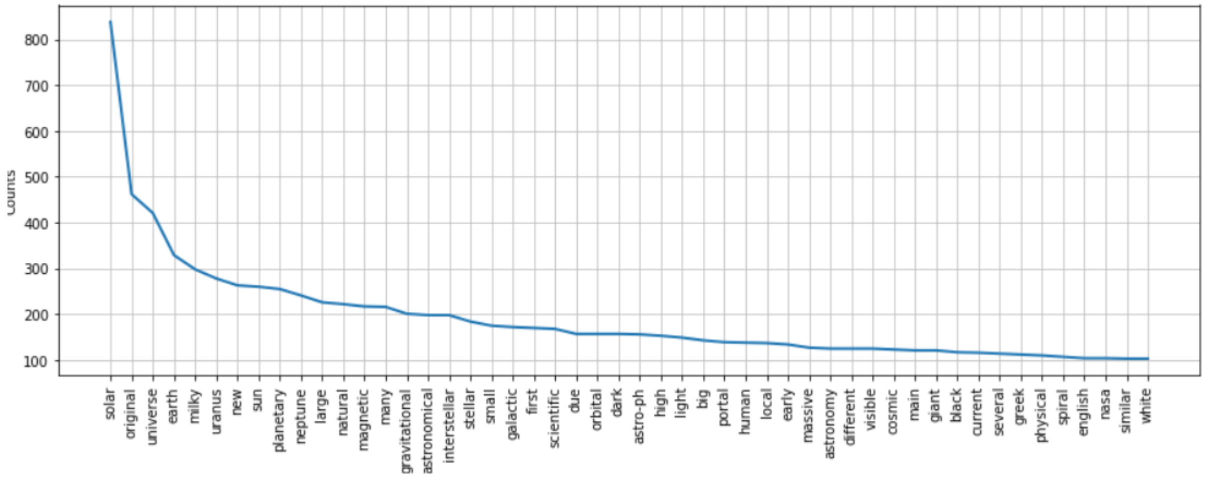
Nouns (NN)



All of the words in this list seem very important, as repeated nouns implies that the articles are talking about these specific entities multiple times. The dataset seems to have mostly nouns, and the more frequent ones look more important

We see a somewhat gradual slope here

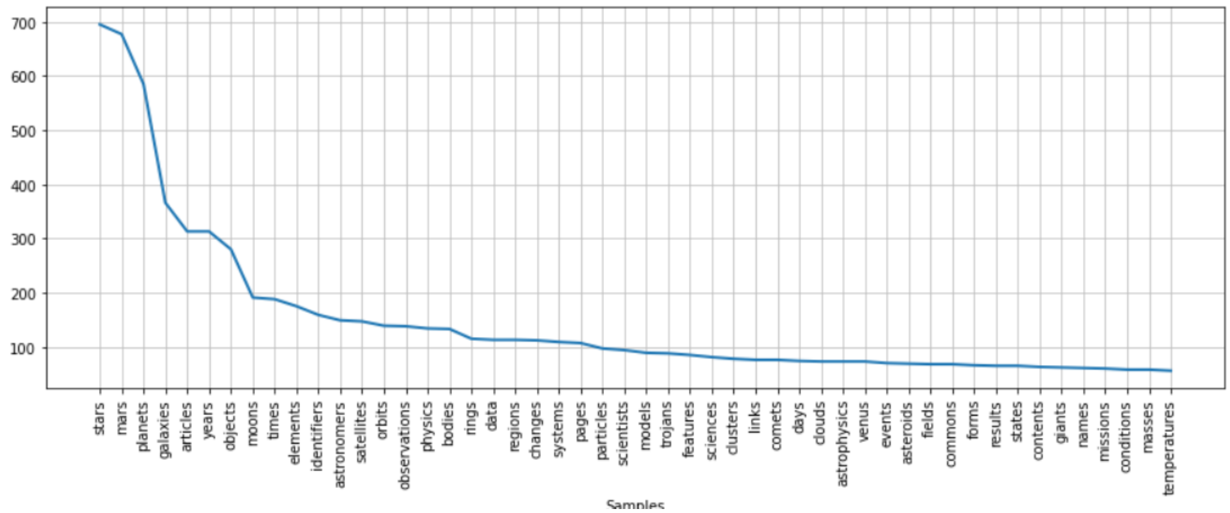
Adjectives (JJ)



Adjectives also look important and talk about what exactly is happening with the concerned topic.

We see an exponential decrease here

Plural Nouns

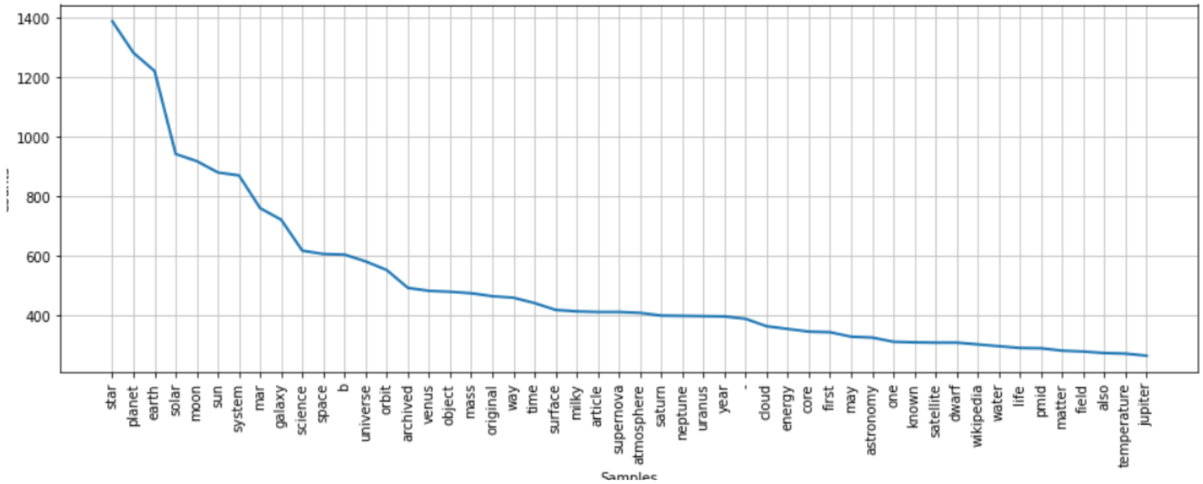


Plural nouns, although not as frequent as singular ones, seem equally important and describe the articles very well.

We see a somewhat exponential slope here

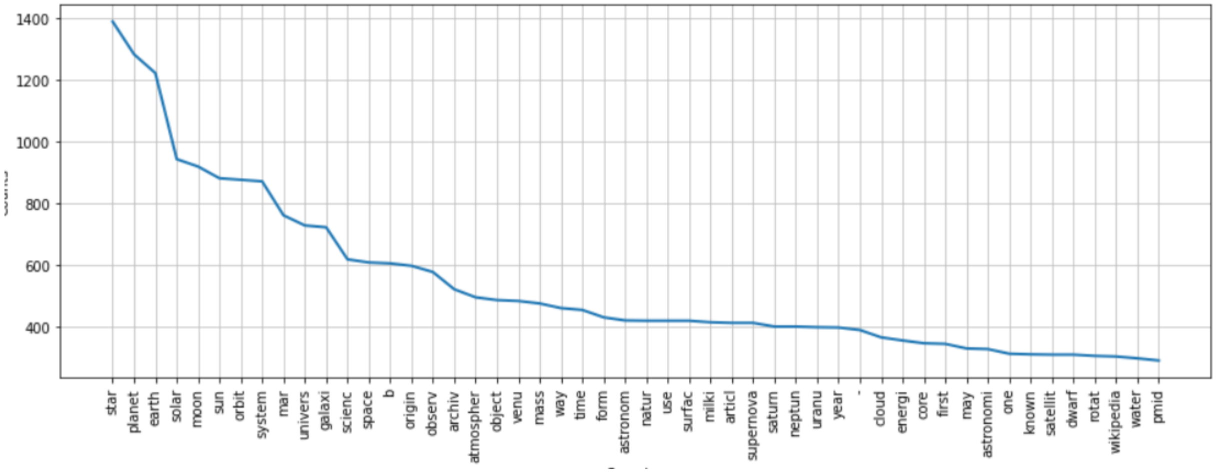
Lemmatization, Stemming

Total text lemmatized



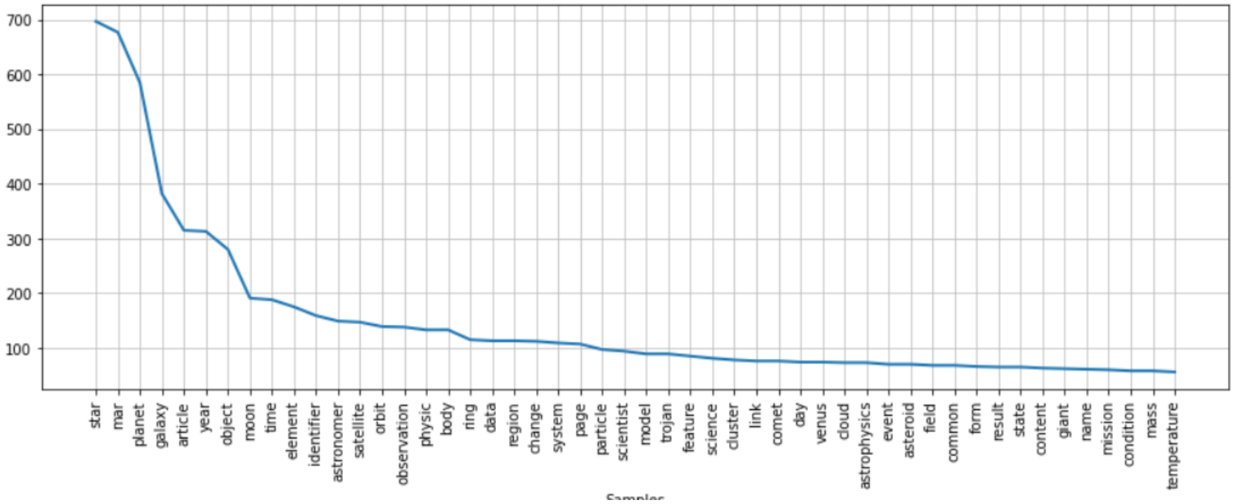
We see a similar graph to the one analysed after removing stopwords, implying that no major changes are taking place. We also see that **mars** has become **mar,** which is messing up with the original data

Total text stemmed



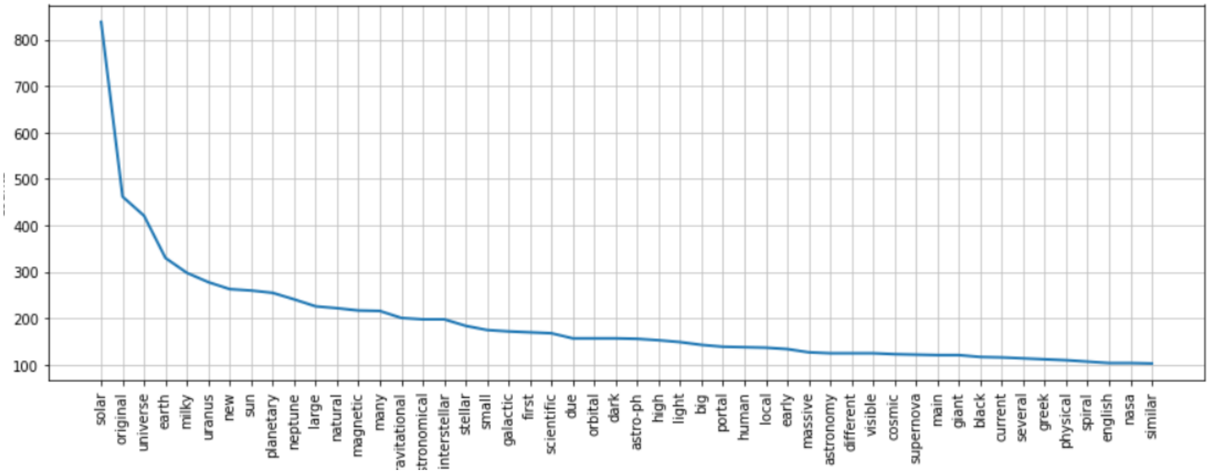
There seems to be a slight change in curve compared to stopwords removal, a frequency increase is visible, meaning some of the common words were in different forms, but this also changes a lot of data and makes it meaningless, and there is not much change in the frequencies.

Lemmatizing Plural nouns



Lemmatizing Plural nouns would be helpful in combining singular and plural forms of the same word in text.

Lemmatizing Adjectives

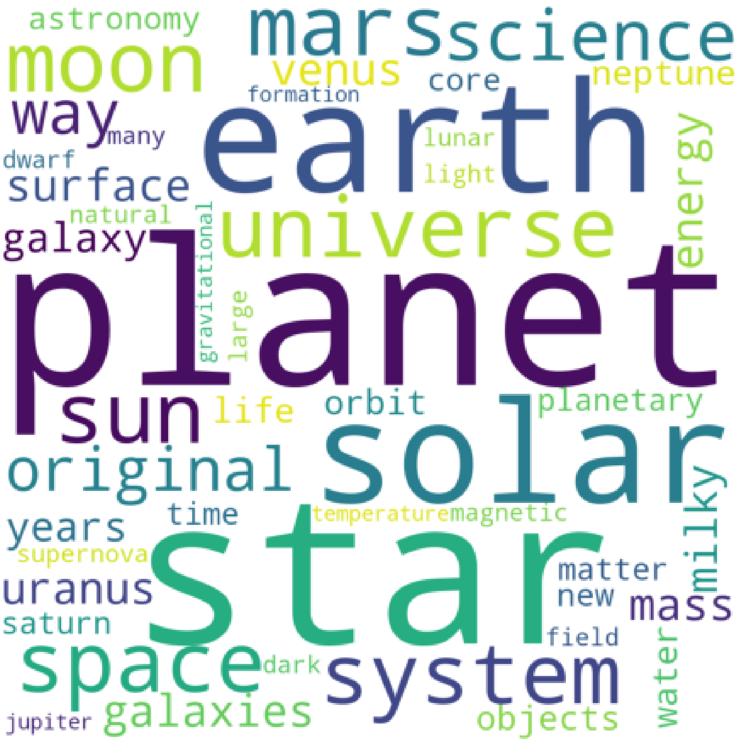


Lemmatizing adjectives would also help in combining NNS, NNP and JJ data to create a more accurate data representing articles, although some words like **gravitational** and **gravity** would imply the same thing but would be considered different parts of speech. Which is why lemmatizing such words is important.

The Algorithm

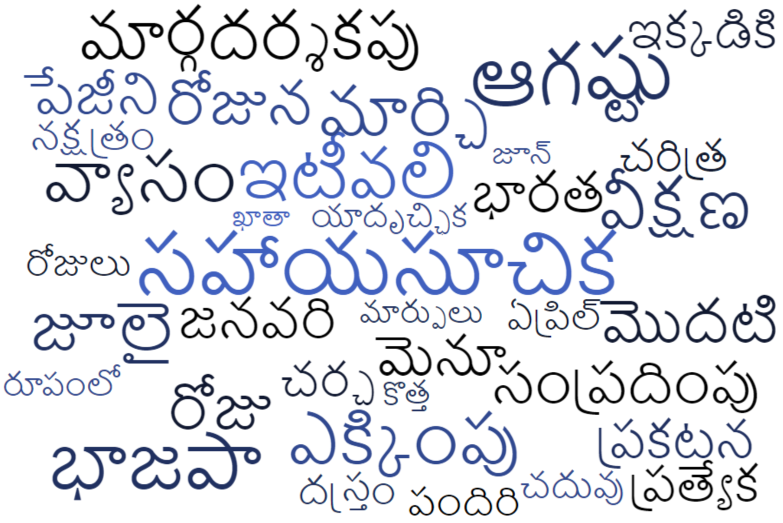
1. Tokenize data
2. Remove stopwords
3. POS tag the data
4. Take a list of data with tags NN, NNP and JJ
5. Lemmatize data with JJ tag (Avoided lemmatizing NNP since it was changing the data a lot, if the lemmatizer was better, NNP lemmatizing would help)
6. Take the top 75 frequent words of the data, this will be the data for our word cloud
7. Visualise it

[nltk Stemmers were not perfect and gave false output for most of the data, so I have avoided using that]



NOTE:

The exact same algorithm with one change -> Lemmatizing/stemming NNP words was supposed to be implemented for telugu, but was not possible due to the lack of resources. By just removing stop words, telugu word cloud was generated, here is the visualization:


మార్గదర్శకపు
ఇక్కడికి
పేజీని
రోజున
మార్చి
ఆగష్టు
నక్షత్రం
జూన్
చరిత్ర
వ్యాసం
ఇటీవలి
భారత
వీక్షణ
ఖాతా
యాదృచ్ఛిక
రోజులు
సహాయసూచిక
జూలై
జనవరి
మార్పులు
ఏప్రిల్
మొదటి
మెనూ
రూపంలో
రోజు
చర్చ
కొత్త
సంప్రదింపు
ఎక్కింపు
ప్రకటన
భాజపా
దస్త్రం
పందిరి
చదువు
ప్రత్యేక