IRE Assignment 1 Report

Vineeth Chelur
201564080
Github Link

August 20, 2018

1 Creating the corpus

Udayavani is an online Kannada newspaper (Link). Scrapy (Python tool) was used to scrap the website for articles in particular categories. The categories scraped were:

- Bollywood news
- Sandalwood news
- Interviews
- State news
- Sports news
- National news
- World news
- Tech news

The articles that had more than 500 words were filtered and the title, content and URL of the article were stored in JSON format. There were a total of 155 articles and were stored in ”udayavani.json” file. The ”crawler” folder contains the settings required to scrape the website. The ”crawler/spiders” folder contains the spider for following the links on the website.

To get the corpus, the following command was run

```
$ python main.py &> main.out
```

2 Getting stop words

”udayavani.json” file was read and blobs were created from the ”content” part of the JSON line using textblob module. The words in each blob was read and a dictionary which computes the frequency of each word was stored. The dictionary was sorted in descending order of frequency and the top 102 words were printed(102 words because 2 of the words in the stop list were numbers). The list of words is given in the next page.

To get the list of stop words, the following command was run

```
$ python stopwords_find.py
```

The list of stop words contained a pretty standard set of stop words, i.e. most of the words were common words that occur in the language. Due to the limited number of articles (155), a coupole of numbers appeared in the list as well. A few underlined words which mean Cinema, Day, Kannada, Work, People, Answer and Situation were a bit surprising because they are not stop words. But, because of the fewer articles, it is possible for these words to have crept in.

ಈ
ಆದರೆ
ಎಂದು
ಅವರ
ಮತ್ತು
ಎಂಬ
ಅವರು
ಒಂದು
ಬಗ್ಗೆ
ಆ
ಇದೆ
ಇದು
ನಾನು
ಮೂಲಕ
ನನ್ನ
ಅದು
ಮೇಲೆ
ಈಗ
ಹಾಗೂ
ಇಲ್ಲ
ಮೊದಲ
ನನಗೆ
ಹೆಚ್ಚು
ಅವರಿಗೆ
ತಮ್ಮ
ಮಾಡಿ
ನಮ್ಮ
ಮಾತ್ರ
ದೊಡ್ಡ
ಅದೇ
ಕೂಡ
ಸಿನಿಮಾ
ಯಾವುದೇ
ಯಾವ

ಆಗ
ತುಂಬಾ
ನಾವು
ದಿನ
ಬೇರೆ
ಅವರನ್ನು
ಎಲ್ಲಾ
ನೀವು
ಸಾಕಷ್ಟು
ಕನ್ನಡ
ಹೊಸ
ಮುಂದೆ
ಹೇಗೆ
ನಂತರ
ಇಲ್ಲಿ
ಕೆಲಸ
ಅಲ್ಲ
ಬಳಿಕ
ಒಳ್ಳೆಯ
ಹಾಗಾಗಿ
ಒಂದೇ
ಜನ
ಅದನ್ನು
ಬಂದ
ಕಾರಣ
ಅವಕಾಶ
ವರ್ಷ
ನಿಮ್ಮ
ಇತ್ತು
ಚಿತ್ರ
ಹೇಳಿ
ಮಾಡಿದ
ಅದಕ್ಕೆ
ಆಗಿ

ಎಂಬುದು
ಅಂತ
2
ಕೆಲವು
ಮೊದಲು
ಬಂದು
ಇದೇ
ನೋಡಿ
ಕೇವಲ
ಎರಡು
ಇನ್ನು
ಅಷ್ಟೇ
ಎಷ್ಟು
ಚಿತ್ರದ
ಮಾಡಬೇಕು
ಹೀಗೆ
ಕುರಿತು
5
ಉತ್ತರ
ಎಂದರೆ
ಇನ್ನೂ
ಮತ್ತೆ
ಏನು
ಪಾತ್ರ
ಮುಂದಿನ
ಸಂದರ್ಭದಲ್ಲಿ
ಮಾಡುವ
ವೇಳೆ
ನನ್ನನ್ನು
ಮೂರು
ಅಥವಾ
ಜೊತೆಗೆ
ಹೆಸರು
ಚಿತ್ರದಲ್ಲಿ