IRE Assignment 1

- **Adithya Jain (20161109)**

Method of data extraction:

I extracted the articles from the Dainik Bhaskar E-paper site. I simply copied and pasted articles into a text file and used that file as my corpus.

Stop Words:

के
में
की
है
से
को
और
ने
का
पर
हैं
कि
भी
नहीं
था
लिए
एक
ही
कर
किया
हो
इस
तो
गया

करने
साथ
यह
बाद
कहा
थे
दिया
तक
थी
रहे
उन्होंने
रुपए
गए
साल
गई
वे
लेकिन
पहले
उनके
लोगों
हुए
रहा
जो
कुछ
वह
वाले
हुई
अपने
ज्यादा
कोई
बात
देश
होने

कई
रही
लाख
बार
दो
जा
सबसे
कम
अपनी
बीच
काम
दी
इसके
हुआ
उनकी
अब
जब
महिला
न
शुरू
पहली
किसी
उन्हें
दोनों
व
नाम
तीन
बहुत
करोड़
रूप
करते
ये
लोग

जाता
नई
दिन
चाहिए
हजार
लेकर
आ
बताया
लिया
करना
देने
तरह
मैं
पास
यात्रा
जैसे
उसे
सकता
वाली
फिर
हम
अभी
होता
क्या
क्योंकि
होगा
या
अन्य
सकते

Observations:

I bolded the stop words that took me by surprise.

एक, लाख, दो, पहली, तीन, करोड़, हजार are quite commonly used words to signify quantity and thus are in high frequency.
रुपए, देश are words that frequently occur in the news, thus the high freq.
Words like महिला, रूप, यात्रा, पानी appearing in the list really surprised me as although these words are commonly used, I couldn't think of a solid reason as to why they occur so frequently.

Method for extracting stop words:
I first removed special characters like punctuation marks and quotes from the corpus.
Then, I ran a loop through the text and made a dict with the count of each word. I then sorted this dict and took the first 100 words from the sorted dict.