CSE 523: Machine Learning

Group 17 - Hardly Humans
Weekly Project Report - 4

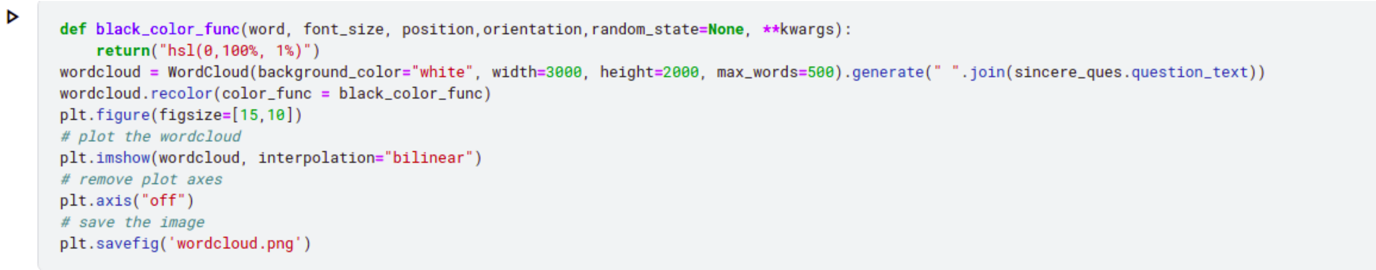
Quora Insincere Questions Classification

Name	Enrolment Number
Malav Doshi	AU1940017
Parth Shah	AU1940065
Sanya Zaveri	AU1920064
Mihir Pathak	AU1920138

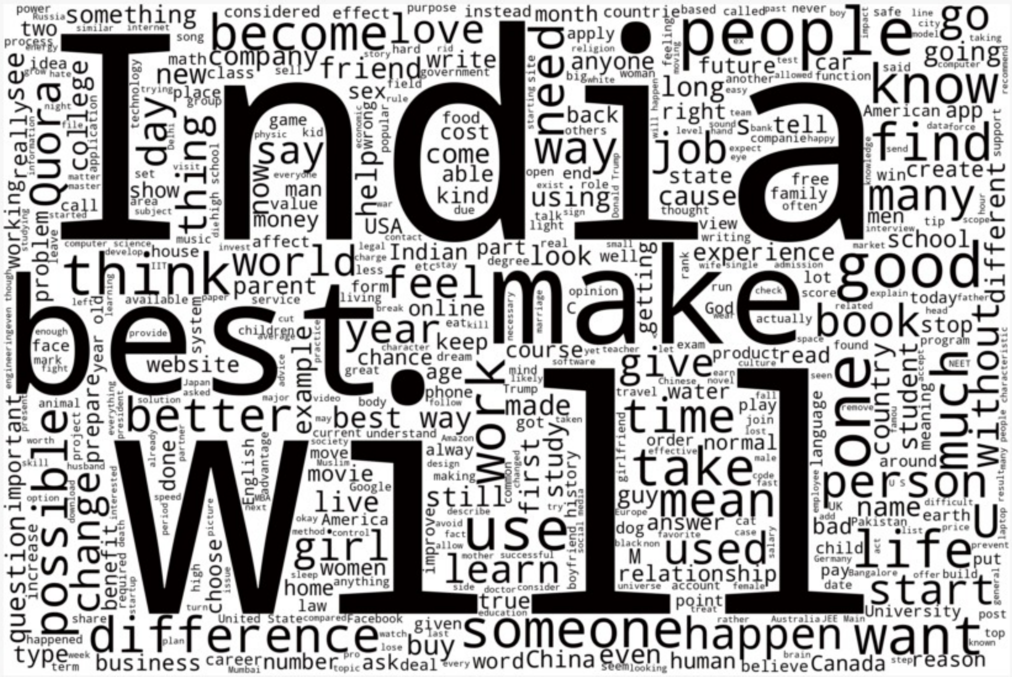
1) Tasks Performed in the week.

- Understanding the Question-text.
- Finding the frequency of Words.
- Importing necessary modules and libraries.
- Creating a word cloud.
- Analyzing the word cloud.

2) Outcomes of the tasks performed.



```
def black_color_func(word, font_size, position,orientation,random_state=None, **kwargs):
    return("hsl(0,100%, 1%)")
wordcloud = WordCloud(background_color="white", width=3000, height=2000, max_words=500).generate(" ".join(sincere_ques.question_text))
wordcloud.recolor(color_func = black_color_func)
plt.figure(figsize=[15,10])
# plot the wordcloud
plt.imshow(wordcloud, interpolation="bilinear")
# remove plot axes
plt.axis("off")
# save the image
plt.savefig('wordcloud.png')
```



3) Tasks to be performed in the upcoming week.

- Data Preparation:
    1. Word Tokenization: Splitting the data into words for text processing.
    2. Removal of Stop Words: Stop words have no value and hence need to be removed for decreasing the redundancy.
    3. Finding Root word: Identifying similar words that co-exist in the data but in different tense or different forms. For instance, India and India.
    4. Text Vectorization: Convert into a series of numeric factors for processing it further.