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.

- Tokenization
  - It is the process of splitting the sentence or the string
- Removing Stop Words
  - Stopwords are those words which does not add much meaning to a sentence
  - We need to remove them to focus on value adding words
  - *stopwords* from *nltk.corpus* is used to download the list of stopwords in the English language
  - Then, we iterate over the list of words in our dataset and discard if it exists in the stop words list.

- Lemmatization
  - It is the process of grouping words to its root word
  - For example, eat is the root word for eating, to eat, eat, etc.
  - WordNetLemmatizer() is used to lemmatize

2) Outcomes of the tasks performed.

- Each question is split into a series of words
- Stopwords are removed from the entire dataset
- Words are grouped into their root word

3) Tasks to be performed in the upcoming week.

- Feature engineering
- Modeling
- Inferences