Project Specification

Project Name: Sentiment Analysis of Comment or Text.

Project Stages:

In order to infer whether the author of a given comment leans a certain way, politically, we use three steps:

1. Preprocess the data, so that we can extract meaningful information, and remove distracting ‘noise’.

2. Extract meaningful information.

3. Train classifiers, given labeled data.

DataSets:

Choose any corpus of the following corpora to applying this project. These corpora should be existing in nltk\_data/corpora folder:

1. movie\_reviews
2. product\_reviews\_1
3. product\_reviews\_2
4. twitter\_samples

**Task1: Preprocessing Stage Requirements:**

1. Remove all newline characters.

2. Replace HTML character codes (i.e., &...;) with whitespace.

3. Remove all URLs (i.e., tokens beginning with http or www).

4. Split each punctuation (using library called string to detect punctuation symbols) into its own token using whitespace except:

1. Apostrophes.

2. Periods in abbreviations (e.g., e.g.) are not split from their tokens. E.g., e.g. stays e.g. (see abbreviation file in the section materials)

3. Multiple punctuation (e.g., !?!, ...) are not split internally. E.g., Hi!!! becomes Hi !!!

4. You can handle single hyphens (-) between words as you please.

5. Split clitics using whitespace (see clitics file in the section materials).

1. Clitics are contracted forms of words, such as n’t, that are concatenated with the previous word.

2. Note: the possessive ’s has its own tag and is distinct from the clitic ’s, but also must be separated by a space; likewise, the possessive on plurals must be separated (e.g., dogs ’).

6. Each token is tagged with its part-of-speech using nltk tagger or StanfordCoreNLP toolkit.

7. Remove stopwords. (See StopWords file in the section materials.)

8. Apply lemmatization using nltk.

9. Convert text to lowercase.