Regex – Split, findall, match, search

Tokenization

Sent\_tokenize

Word\_tokenize

Regexp\_tokenize

TweetTokenizer

Charting word length with NLTK

Bag of Words: Just dictionaries of words and occurrences

Counter from collections

Lemmatization/stemming – shorten words to root stems

Stop words from ntlk.corpus

Word vector

Dictionary from gensim.corpora.dictionary

Doc2bow

Tf-idf Term frequency inverse document frequency

Tfidfmodel from genism.models.tfidfmodel

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Named Entity Recognition – pos\_tag

SpaCy

To check the type of comments or text, we need to use Count Vectorizer along with Bag of Words.

Count Vectorizer is similar to Gensim corpus. It removes stop words.

fit\_transform is used to create mapping of words with IDs and Vectors how many times appeared in plot.

Count vectorizer should be applied on both training data and testing data to identify similar words.

Count Vectorizer and TFIDF are 2 types of word vectorizer