

NLP
Assignment -2
Abhinay Natti(150438)

doc/ classif.	Naive Bayes	SVM	Logistic Regres	Feed forward
BOW	84.42	84.78	86.9	85.78
Norm Tf	84.57	85.21	75.92	84.93
Tfidf	84.42	87.61	87.92	85.69
w2vec with wts	68.49	81.52	81.65	81.45
w2vec no wts	70.99	84.04	82.66	83.72
Glove With wts	71.27	81.27	81.24	80.27
Glove No wts	68.96	82.61	82.11	81.52

Accuracy table

➤ Bag of words:

- For BOW I used countvectorizer from feature extraction of sklearn . Fit_transform makes the vocabulary and feature vector for each document.
- Transform makes feature vectors for the corresponding document according to the vocabulary generated by Fit_transform
- Binary=true is used for BBOW

➤ Normalized term frequency and tfidf:

- I used Tfidftransformer from feature extraction.
- For normalized I used hyperparameters use_idf=false and norm = 'l1'

➤ Word2vec And Glove:

- I used the online models for w2vec and glove which had word vectors for each word.
- Then for a word in a document if it is present in the online model it gets the same word vectors otherwise zero vector of the same dimension.
- Then we mean over all the vectors which makes the avg of the wordvectors

- With weights we just multiply the frequencies for each word to the word vector.

➤ Classification Algorithms:

- I used various libraries from sklearn.
- SVM : `svm.LinearSvc()` (LinearSVC gives better accuracy for larger number of samples)
- Naive-Bayes : `BernoulliNb()` (As we have binary data for bow,tf and tfidf)
- Logisticregression(`LogisticRegression`) and `FeedForward`(Mlp classifier)

➤ As we can see naive bayes has the least accuracy for any representation which it should be because of naive bayes's "naive" assumptions.