There are two main components: 1) Generating LDA representation of tweets; and 2) Classification. The

Training

getLDA

LDA model

Training data: Tweets

Training data: Tweets:words

Bot detector

getClassifier

Training data: Tweets:topics

Results

Testing data: Tweets:topics

Testing data: Tweets:words

Testing

classifier

1)

getLda(datafile,row,column=496411,ntopics=200) returns the generated LDA model.

A pickled LDA model, ’lda200.pkl’, is available.

The ‘lda200.pkl’ can be loaded using cPickle.

The input file should be in the Blei’s LDA-C format (<https://radimrehurek.com/gensim/corpora/bleicorpus.html>).

We need a vocabulary file to transform words into IDs, which is also available (vocab.txt).

Using the LDA model, a topic representation of a sentence can be obtained using lda.inference()

2)

getClassifier(filename='inf200.txt',row1=7061,row2=87474): is to train a bot detection classifier based on the training data ‘inf200.txt’, where first 7061 rows are bots and the rest are regular users. I use cross validation here.

You can try different models other than linear SVM through replacing svm.SVC(kernel='linear', C=1.0,max\_iter=500).