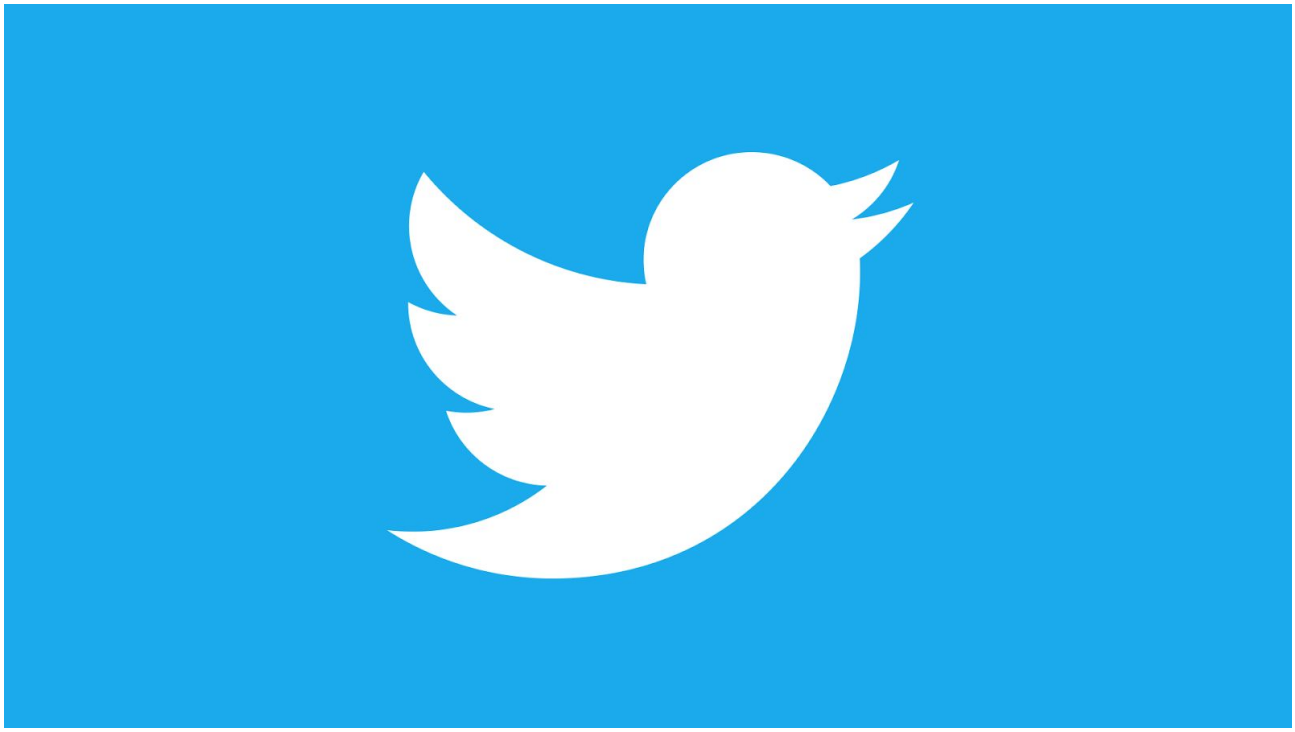
Lab Twitter Analytics



Avishek (Joey) Bose

29.04.2016

Architech Labs

INTRODUCTION

The goal of this project is to retrieve useful topics from large unsupervised data that consist of short text documents pertaining to an initial filtering query. Furthermore, from the resulting list of topics we retrieve a list of top documents that are strongly correlated to the bag of words for each topic given to us by the model.

PROCEDURE OVERVIEW

1. Collect Data set, specifically stream tweets and save them to disk. The code for this can be found in **Lab-twitter-analytics / twitter_downloader / twitter_stream.py**
2. Process tweet dump for stop words and query tokens. Concatenate replies to tweets as one document and other basic NLP. the code for this can be found in **Lab-twitter-analytics / twitter_downloader / process_big.py**
3. Create word vectors for the Data set using word to vec model. The code for this can be found in **Lab-twitter-analytics / models / train_gensim_word2vec.py**
4. Use Diffusion Map algorithm to reduce the dimensionality of the word vectors and embed them in the diffusion space. Note the code for this will have to be repurposed with your input files but it can be found here. **Lab-twitter-analytics / models / w2v_diffusion.py**
5. Cluster the word vectors in the new embedded space and remap the Dataset to cluster centres, effectively collapsing the vocabulary. Using the remapped Dataset we then create a Gensim Dictionary and Gensim Corpus. This also saves a lookup table of sorts to go from cluster centres to words in the cluster. The code for this can be found here **Lab-twitter-analytics / models / diff_remap_words.py**
6. Train an LDA model using the Gensim Dictionary and Corpus. The code for this can be found in **Lab-twitter-analytics / models / train_gensim_lda.py**
7. To view the results of the LDA model one can use the following code to get a bag of clusters. Furthermore, we can get top tweets per topic based on filtering the tweets on a per cluster basis and then scoring them based on the probability of a word from a tweet appearing in any of the top 3 clusters for a topic and then ranking them based on score. The code for this can be found in **Lab-twitter-analytics / models / cluster_to_words.py**

RESULTS

There is a visualization of the of the LDA model without the word Layer in the following ipython notebook **Lab-twitter-analytics / models / LDA_visual.ipynb.** The top tweets for the middle east dataset can be found here: **Lab-twitter-analytics / models / top_middle_east_tweets /**

The results are presented by first listing out the top 3 topic clusters and its respective words, followed by the top 20 tweets for that topic. Here is an example from topic 1 which is about ISIS.

[[u'isis', u'innocent', u'aiding', u'despots', u'defenceless', u'terorist', u'maiming', u'mullas', u'freethinker'], [u'country', u'fight', u'crimes', u'accuse', u'strictly', u'arresting', u'terrorised', u'colluding', u'publically', u'mny', u'victi', u'crusading', u'staunchest', u'incarcerating', u'indisputably', u'rationalist', u'highjacking'], [u'propaganda', u'becoming', u'lied', u'equal', u'legitimate', u'surrender', u'succeed', u'correctly', u'interfere', u'populace', u'interfering', u'confuses', u'campaigners', u'outsiders', u'ppls', u'masterminds', u'obliged', u'relied', u'progressed', u'infestation', u'dissenting', u'fea', u'evildoers', u'materially', u'weaponizing', u'countenance', u'deconstruct', u'bogeymen', u'shibboleth', u'mildest', u'fasci', u'infests', u'critisism']]

groups like isis want this to happen bombing areas in which there are innocent civilians is just aiding them dontbombsyria

so f’ed up how we choose to bomb innocent people it s just not right all bombing achieves is aiding isis conscription syriavote

i understand that things with isis are difficult but the decision to bomb thousands of defenceless innocent people is

the mps did not vote to bomb innocent civilians thats for sure they are aiding the forces already attacking apparent isis targets ffs

how can i call rte a muslim as he sold sarin gas to isis killing thousands of innocent a muslim cant be a terorist

how can i call rte a muslim as he sold sarin gas to isis killing thousands of innocent a muslim cant be a terorist https

how can i call rte a muslim as he sold sarin gas to isis killing thousands of innocent a muslim cant be a terorist i şi d

britain and isis are just a bad as each other both killing innocent defenceless people disgusting

isis is terorist and for all those who target innocent people youaintnomuslimbruv

if you are too cowardice to send your troops to fight against isis don t send your jets to kill innocent syrian destroy the country

absolutely the right decision to bomb isis they ve barbarically murdered innocent victims some of our country it s time to fight back

i agree with fighting back against isis not killing innocent people you can t fight terrorism theyre not a fkin country
how are you any different than isis if you fight back by bombing an entire country full of innocent people
over years at the hands of isis watching innocent people die we finally stand fight the country turns it s back
i hate this government i hate this country they are cowards they re scared to fight isis so instead they want to hurt innocent people
use these weapons to kill those innocent rather than fight isis wakeup injustice crimes
uses these weapons to kill those innocent rather than fight isis wakeup injustice crimes
killing innocent people in a defenceless country for the sake of what isis aren t rooted to one country one city
innocent people don t talk big like the us if you want to fight isis than uou must be on the ground don t be scared
our military men women want to go and destroy isis it s the weak politicians in this country that are getting innocent americans killed

CONCLUSION

We created a multi layered approach to topic modelling large unsupervised data that consist of short documents. We do this through creating a word level model based on extracting word vectors and then clustering them in a lower dimensional space that is achieved through diffusion map. After remapping the Dataset based on cluster centres we then use a standard LDA model to extract interesting topics. The effectiveness of this approach stems from the fact that short documents have word co-occurrence matrices that are very sparse and consequently regular topic modelling yields worse results. By clustering in the diffusion space we seek to alleviate this problem. However, this approach when applied to longer text documents does not yield any performance improvements as co-occurrence matrices are dense enough that clustering does not add much to the actual representation.