### **Schedule**

|  |  |  |
| --- | --- | --- |
| Week | Task | Results |
| 5/15/17 | Literature analysis | In project proposal |
| 5/22/17 | Plan procedure and code outline | Procedure planned |
| 5/29/17 | Build code structure – objects and main | In github |
| 6/5/17 | Finalize code | Somewhat done |
| 6/12/17 | Apply code to topics and analyze accuracy | Currently working on it, got switched with below |
| 6/19/17 | Revisit program and procedure for improvements | So far so good... |
| 6/26/17 | Finalize procedure and program |  |
| 7/3/17 | Draft paper & manuscript |  |
| 7/10/17 | Finish paper & manuscript |  |

### 

6/12/17

List of 15 notable hashtags for each issue

-----abortion-----

defundpp

praytoendabortion

prolife

ppsellsbabyparts

abortion

pregnant

healthcare

medical

health

prochoice

antichoice

womensrights

mybodymychoice

fundpp

-----political-----

hillary2016

nevertrump

imwithher

hillary

hillaryclinton

immigration

decision2016

military

usa

americans

makeamericagreatagain

maga

trump2016

trumptrain

votetrump

-----taxes-----

Not really sure how to split, since hashtags aren't as obviously biased on this issue.

bogusclimatechange

middleclassworkers

oligarchy

economics

majorityyoucantrust

america

budget2016

profitsoverpeople

panamapapers

boycottmadeinchina

taxontax

unfairtax

boycott

occupy

accountant

inequality

5/16/17

**Hashtag2vec thoughs**

1. Format and tokenize tweets - lexical analysis
   1. Good starting tokenizer: <http://www.nltk.org/api/nltk.tokenize.html>
   2. Integrate other syntax filters and stop word filters (when counting words only)
   3. Tokens converted to integers for word2vec
2. Use word2vec to create a vectors representing relations of words in our tweets
3. Graph hashtags based on their simultaneous use in a tweet or a user using both hashtags in differenttweets (other ways?)
4. Create hashtag vector spaces by summing over the vector spaces of the word vectors and normalizing
   1. This might be improved by taking into account specificity and volatility of words, which we could then obtain distinctiveness and dynamicity from for each hashtag ( another weight in the graph?) see <https://cs.stanford.edu/people/jure/pubs/identity-icwsm17.pdf>
5. Find cosine of vector spaces between two hashtags to find the weight of an edge in graph from 3.)
6. analyze the now weighted graph(s) of hashtags and look for clusters and signs of conflict, possibly through <http://www.mapequation.org/publications.html#Rosvall-Axelsson-Bergstrom-2009-Map-equation>
7. Compare this to conventional graphing by user interest: ie, looking at vocabulary of users and finding a relationship between them with word2vec profiles

Begin previous work research

Look for gaps in current understanding of topic, and try to formulate topic

Goal - Create a way to compare hashtag similarity, which will then produce another way of modeling conflict

Interface between psql database and tensors

Create tf.variable from multidim array returned from psql

Psycopg2

Cursors

Allow execution of psql queries

Can use as iter in for loop - allows for stream of rows

Tensorflow

Variables - input of tensors

Constant - input of ndarray

Psql database design

Single table or multi table?

Multitable

Tweet id to hashtag id

Tweet id to user

Tweet id to text

Tweet id to int array

Int array is an array of vocabulary integers

Dictionary

Word to integer - also include count of each word in vocab

Integer values greater than 0 is vocabulary

Reverse Dictionary

Integer to word

Word embeddings

Hashtag embeddings

Few tables

Table 1

Tweet id

Text

User id

Hashtags contained

Array of integers ( vocabulary)

Table 2

Dictionary

Table 3

Word embeddings

Table 4

Hashtag embedding

Process

Format Tweets

Tokenize

Remove syntax

Urls

Unicode

Words beyond certain length

Build dictionary

Convert tweets to array of ints

Word2vec

Create NN

Provide batches to NN

Hash2vec

Sum words in