**2017-05-31**

**Relevent xkcd ->** [**https://xkcd.com/1838/**](https://xkcd.com/1838/)

* **Action items**
  + Kenny: Create list of 50-100 most used hashtags for each topic
    - Done, see my project notes
  + Kenny: Further work on project codebase

**2017-05-22**

* **Action items**
  + Misha D: Research ways of embedding alignments
  + Misha D: Ping Lesha about the new server access and start UK tweets collection
  + Misha D: Ping Lesha about the other project which needs help with word2vec
  + Kenny & Misha D: figure out how projects overlap
  + Kenny: Plan database/organization of code
  + Think of ways to include subculture detection

**2017-05-08**

* **Some notes on Twitter Streaming API**
  + Can follow up to 5,000 users per token
  + For each user specified, the stream **will** contain:
    - Tweets created by the user.
    - Tweets which are retweeted by the user.
    - Replies to any Tweet created by the user.
    - Retweets of any Tweet created by the user.
    - Manual replies, created without pressing a reply button (e.g. “@twitterapi I agree”).
  + The stream **will not** contain:
    - Tweets mentioning the user (e.g. “Hello @twitterapi!”).
    - Manual Retweets created without pressing a Retweet button (e.g. “RT @twitterapi The API is great”).
    - Tweets by protected users.
* **Link to shared folder for relevant work**
  + <https://drive.google.com/open?id=0B2kF35wMVD1fcTh0emRzSGoxOVE>
* **Action items**
  + Misha D: Start the UK twitter collection (need servers from Lesha)
  + Misha D: Look for more research on our topic
  + Misha D: Try out embedding and clustering hashtags (e.g. with word2vec) using the US data from 2016
  + Lesha: Create a git repo
  + Lesha: Create two virtual servers for data collection / storage
  + Kenny: Read through lexical drift papers
    - Subcultures and changes in language over time
    - Divergence of languages between two subcultures

**2017-05-01**

Participants: Kenny, Misha D, Misha G, Anna, Alexey

* **Available data**
  + In our DB: **18.3** million tweets from US politicians (including retweets/replies)
    - from 15 July to 11 September 2016
  + In our DB: **11.8** million tweets collected using keywords
    - from 24 June to 11 September 2016
    - Tweets by issue:

issue | count

-----------+---------

abortion | 2114926

taxes | 849175

political | 8874205

* + Can collect: tweets from UK politicians and political journalists
    - <https://twitter.com/tweetminster/lists/ukmps/members>
    - <http://uk.businessinsider.com/uk-politics-twitter-accounts-2016-8>
* **Action items**
  + Misha D: Check for Twitter API limitations (# of users to collect the data from)
  + Misha D: Figure out how to select top N retweeters (based on “popularity”)
  + Figure out how to snowball
  + Anna + Misha G: revisit firehosing with Twitter!
  + Everyone: Find and review more relevant papers (use the papers below + Google scholar as the starting point “hashtag” “polarization” “hashtag hijacking”)
* **References:**
  + <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2847/3275.pdf>
  + <http://journals.sagepub.com.libproxy.uml.edu/doi/pdf/10.1177/0270467610380011>
  + <http://ieeexplore.ieee.org.libproxy.uml.edu/stamp/stamp.jsp?tp=&arnumber=5701826&isnumber=5701812&tag=1>
  + Hashtag embeddings & prediction:
    - FB AI Research paper: <http://emnlp2014.org/papers/pdf/EMNLP2014194.pdf>
    - Datasift keyword explorer (2015): <http://dev.datasift.com/blog/exploring-keyword-relationships-social-posts-word2vec> (seems like it hasn’t been updated since then, e.g. it doesn’t know about #brexit)
    - This looks like a similar project too: <https://github.com/Data4Democracy/assemble/issues/26>

**2017-04-21**

Participants: Kenny, Misha D, Misha G, Anna, Alexey

* **Action items**
  + Verify that keyword search data is not biased by twitter
  + Look at Peter’s work on community detection / graph-based controversy measures (Peter - [pjpotash@gmail.com](mailto:pjpotash@gmail.com), see Anna’s emails with attached documents, and relevant sections)
  + Compile an inventory of available data -- assemble the info about the size, the source, and get access to the data (Kenny - see the table in the interim report doc)
    - US Political tweets - politicians’ accounts
    - US Elections - keyword-based collection
      * clinton, trump
      * abortion, guns
    - Brexit -- look into the data set recently released (see forwarded email)
    - US political blogs data (alexey / nikita)
    - Look into collecting data for UK elections (Misha D) keywords and/or accounts??
      * Keywords extraction?
  + Get access to dumuzi (Kenny) - Alexey Romanov [jgc128](mailto:jgc128ra@gmail.com)@outlook.com