Seunghyuk Baek

[Sbaek44@vt.edu](mailto:Sbaek44@vt.edu)

Assignment4

To begin topic analysis for this assignment, I choose 3 topics that related coming U.S. midterm election. Three topics are ‘midtermelection2018’, ‘redwaverising2018’ and ‘bluewavecomming2018’. The reason why I choose above topics was to acquire the general theme of this coming election (using ‘midtermelection2018’) and the theme of each political party.

With topic modeling using genism library, and examining 1000 tweets for each topic, I came up with words that is most related to each topic. However, to eliminate emojis of twitter, I printed 10 words using ldamodel.print\_topics instead of exact 3. Also, extending word parameter gave more meaningful topic modeling result. Following is the result of ldamodel function of genism

Topic 1. Got, people, back, jojoh888, sick, brown, California, known

Topic 2. Gop, dear, you’ve, vile, wagon, married, hitched, traitorous, threetime

Topic 3. Redwaverising2018, redwaverising, remember, jojoh888, midterm2018

It seems that increasing the number of tweets does not change the result much. However, it could be the number of tweets that increased from the original tweet collection is not significant enough.

Since hashtag is to show the topic of a tweet selected by a user, it can be a good measurement to show how differentiate the modeling is from actual topic.

**Here are hashtags from all 3 topics**

[('', 1467), (u'RedWaveRising2018', 739), (u'RedWaveRising', 504), (u'Midterms2018', 477), (u'BlueWaveComing2018', 293), (u'RINO', 92), (u'FlipItBlue', 91), (u'RedtoBlue', 89), (u'Democrats', 67), (u'MAGA', 64), (u'FollowBackResistance', 63)]

**Here are hashtags from topic ‘midtermelection2018’**

[(u'Midterms2018', 474), (u'RedWaveRising2018', 474), (u'RedWaveRising', 474), ('', 473), (u'MidtermElections2018', 14), (u'midtermelections2018', 10), (u'farmers', 8), (u'JamesComey', 8), (u'Midterms', 8), (u'tariffs', 8), (u'ItsMuellerTime', 7)]

**Here are hashtags from topic ‘bluewavecomming2018’**

[('', 466), (u'BlueWaveComing2018', 293), (u'FlipItBlue', 91), (u'RedtoBlue', 89), (u'FollowBackResistance', 63), (u'Democrats', 50), (u'VoteThemOut', 22), (u'VoteBlue', 19), (u'BlueWave2018', 15), (u'FBRParty', 14), (u'bluewave', 14)]

**Here are hashtags from topic ‘redwaverising2018’**

[('', 528), (u'RedWaveRising2018', 265), (u'RINO', 92), (u'MAGA', 63), (u'Beto', 49), (u'RedWaveRising', 30), (u'WednesdayWisdom', 27), (u'AmericaFirst', 20), (u'Trump2020', 19), (u'BlueWave2018', 17), (u'Democrats', 17)]

It seems that only little portion of topic-model is included in hashtag.

For next analysis I ran sentiment analysis for each topic and all topics combined. Among 3000, 2112 tweets were positive about 2018 midterm election. For topic ‘midtermelection2018’, 953 out of 1000 tweets were positive. For ‘bluewavecomming2018’ topic 653 tweets out of 1000 topics were positive. For, ‘redwaverising2018’ topic, only 506 out of 1000 tweets were positive.

If stopping words are not removed in topic-modeling, topics such as ‘to’, ‘who’, ‘very’, ‘have’, ‘are’ ‘rt’ take over existing topics. If punctuations are included in topic-modeling, the result topics are very similar to the original topic-modeling. Topics resulted from not removing punctuation is followed:

Topic 1. Got, back, people, jojoh888, sick, brown, California, known

Topic 2. Gop, dear, you’ve, vile, traitorous, bully, hitched, attacking

Topic 3. Redwaverising, remember, jojoh888, midterms2018

If work stemming is conducted on tokenized sentences, I could see some words are different but some of frequent words are also disappearing.

Topic 1. Lie, people, got, back, sick, California, brown, th, known

Topic 2. Redwaverising2018, redwaveris, state, jojoh888, remeb, midterms2018, border, sendtroop

Topic 3. Redvaverising2018, vote, good