Homework 5

Run LDA topic model algorithm, analyze your dataset, and visualize your results. For the dataset, it can be your project dataset, or Harvey Twitter dataset.

➢ Choose a proper topic model numbers.

➢ For the analysis, you may compare the topic movements or popular topics dynamic changes, or anything you feel interesting.

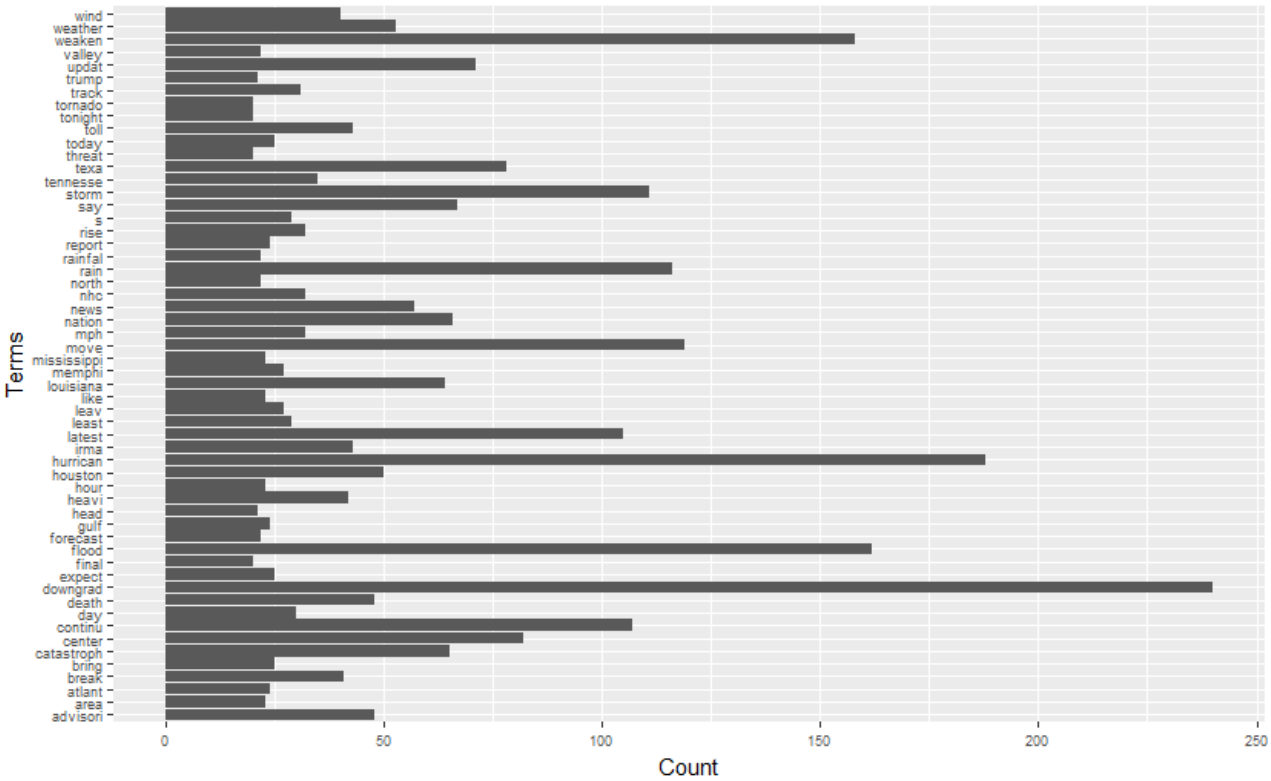
➢ For the visualization, you can use word cloud or any other visualization methods.

1. DATA SET

For this homework, I use Harvey Twitter dataset. Our original dataset contains 5,306 observations. After removing stopwords, url, punctuation, whitespace and duplication. Our final data contains 1.102 records. In addition to the built-in stopwords, I add some more words which I think not have much meaning.

Extra_stopwords = *"just","still","al","ion","gu","kno","amp","ap","ne","w","via","us","near","now","rt","will"*
Keywords in Harvey Twitter are also excluded because we assume that these keywords should appear in every document when we retrieved data from Twitter API. These keywords are *"harvey", "tropical", "depression"*

2. TERM FREQUENCY



From the Term Frequency, it can be seen that the terms downgrad, hurrican, flood, weaken are the most popular terms found in the dataset. These terms appear as we expected since they have close meaning to the search keywords

vinh.nguyen@ttu.edu

3. TOPIC MODELING

Topic modeling with k =8

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**
hurrican	flood	downgrad	hurrican	move	move	hurrican	downgrad
downgrad	downgrad	weaken	storm	texa	hurrican	rain	hurrican
rain	weaken	flood	latest	weaken	downgrad	flood	flood
storm	updat	hurrican	flood	continu	storm	say	rain
continu	louisiana	say	continu	storm	weaken	weaken	catastroph
updat	houston	latest	move	rain	flood	continu	weaken
latest	texa	rain	rain	updat	continu	nation	news

Topic modeling with k=7

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**
latest	downgrad	flood	hurrican	downgrad	flood	hurrican
storm	hurrican	downgrad	rain	weaken	weaken	rain
move	continu	weaken	storm	flood	latest	weaken
hurrican	texa	move	downgrad	rain	hurrican	storm
texa	move	center	move	louisiana	rain	center
weather	flood	continu	weaken	latest	advisori	downgrad
news	louisiana	nation	say	texa	move	continu

Topic modeling with k =6

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
hurrican	downgrad	storm	downgrad	downgrad	rain
weaken	flood	rain	hurrican	move	flood
move	hurrican	downgrad	continu	storm	weaken
center	weaken	move	center	center	updat
continu	latest	updat	texa	flood	catastroph
flood	rain	weaken	weaken	report	move
downgrad	louisiana	latest	latest	updat	death

Topic modeling with k = 5

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
downgrad	weaken	flood	downgrad	hurrican
move	hurrican	weaken	flood	move
weaken	storm	rain	hurrican	flood
center	say	nation	center	texa
storm	continu	downgrad	rain	news
continu	move	storm	latest	downgrad
hurrican	louisiana	hurrican	texa	latest

Playing around the some number k (5,6,7,8), we see that some terms appear together is most topics (weaken, downgrad, hurrican, flood). These terms are also have the highest frequency in the dataset. In practice, setting the initial number of k to look for a desired topic does not have to much meaning, instead we have to go around with several key values and select topic we want. In another word, this is not a fully automated process, there is still a need for human intervention. We found that k =6 provides a good distribution of term across topics.

4. WORDCLOUD

As expected from the frequency table and topic modeling: weaken, flood, downgrad, move, hurrican, rain, latest, storm are most visible in the word cloud.