Week 14: Code Clustering Part I

Published Edit ⋮

In this assignment, you will perform a basic topic modeling analysis of tweets after a fatal Tesla crash. You should use the dataset posted on this module of the course.

The steps to complete the topic model are:

1. If necessary, install the topicmodels, tidytext, and ldatuning libraries
2. Load our standard library set along with topicmodels, tidytext, and ldatuning
3. Read in the crash tweets dataset
4. Use the unnest_tokens function to convert the tokens column to words
5. Use the anti_join(stop_words) function to remove stopwords from the dataset
6. Use the count function to count the word frequencies by tweet id
7. Use the cast_dtm function to cast the word fruency result from #6 to a document term matrix
8. Use the LDA function to fit a topic model. You can select the number of topics. More topics will take longer but will give a more informative model. You should select a number between 5 and 25.
9. [Optional] You can alternatively use the FindTopicsNumber function (see https://quantdev.ssri.psu.edu/sites/qdev/files/topic_modeling_tutorial-Gutenberg-chapter_as_document.html (https://quantdev.ssri.psu.edu/sites/qdev/files/topic_modeling_tutorial-Gutenberg-chapter_as_document.html)) to optimize the number of topics. Beware that this will take a very long time.
10. Use the following code to generate a dataset of the top 10 terms by each topic:

    1.
       ```
       [your lda model] %>% tidy(matrix = "beta") %>%

         group_by(topic) %>%

         top_n(10, beta) %>%

         ungroup() %>%

         arrange(topic, -beta)
       ```
11. Plot a bar chart of the beta's of the top ten terms in each topic faceted by topic.

Note the following helpful resources:

https://www.tidytextmining.com/topicmodeling.html (https://www.tidytextmining.com/topicmodeling.html)

https://quantdev.ssri.psu.edu/sites/qdev/files/topic_modeling_tutorial-Gutenberg-chapter_as_document.html (https://quantdev.ssri.psu.edu/sites/qdev/files/topic_modeling_tutorial-Gutenberg-chapter_as_document.html)

Points	10
Submitting	a file upload
Allowed Attempts	2

Due	For	Available from	Until
Apr 23	Everyone	-	-

R coding assignment						
Criteria	**Ratings**					**Pts**
R code functionality Does the R code run as expected and produce the expected result.	**5 pts** **Yes**	**4 pts** **Yes, but with one minor issue**	**3 pts** **Yes, but with several minor issues**	**2 pts** **Mostly, but there is one major issue**	**0 pts** **No it does not**	5 pts
Code reproducibility Does the code include elements to make it reproducible (e.g., comments/annotations, random seeds, documented data manipulation)?	**5 pts** **Yes**	**4 pts** **Yes, but with one minor issue**	**3 pts** **Yes, but with several minor issues**	**2 pts** **Mostly, but there is one major issue**	**0 pts** **No it does not**	5 pts
						Total Points: 10