Search and Extract Text Data

Mandatory Exercise

Prof. Dr. Nicolas Meseth

OVERVIEW

Intuitive first steps to text analysis are simple **keyword and pattern searches**, **extractions**, or **replacements** of certain elements from text data. Learning how to do this will improve your understanding of the challenges involved in NLP.

Here are the relevant lessons from the online script. Make sure you study them carefully, including the provided literature, in order to solve this exercise:

- 7 - Searching Text
- 8 - Imposing Structure

Good Luck!

YOUR TASKS

To pass this exercise, you must complete the following tasks and submit your results via ILIAS. You find the submission details below.

Task 1: Search, Extract and Transform Tweets

1. Extract the most common emoticons from the tweets and create a top 10 list by frequency.
2. Extract the common hashtags for the different political parties, such as `#spd`, `#gruene`, or `#fdp` (see if you can find more) from the tweets. Which users include their party's hashtag(s) most often in their tweets?
3. Extract all hashtags in tweets from the current year. Count how often each hashtag was used and create a top 20 list.
4. Whom does Christian Lindner mention most often in his tweets?

5. Create a list with (at least) ten positive words and ten negative words. Search for both sets of words in the tweets and calculate a sentiment score by adding 1 for a match of a positive and -1 for a negative word. Find the tweets with the highest and lowest scores.
6. From the `source` column, extract the app the user cretaed the tweet with. Create a new column called `twitter_app` that contains the values "Twitter for iPhone" or "Twitter for Mac" and so on. How are the values distributed?
7. Search for tweets with URLs in the text and calculate the percentage of tweets containing a URL.
8. Remove the "RT" text from the beginning of all tweets! What does it mean and what column does it relate to?
9. Remove all non-alphanumeric characters from the tweets!

Task 2: Word Count in Tweets

1. Create a new tibble that contains only tweets from the current year and that are written in German!
2. Using the new tibble, prepare the tweet's `text` column for tokenization and convert them to lowercase and remove all unnecessary characters, user mentions, urls, hashtags, and punctuation. Replace all occurrences of more than one space with a single space.
3. After cleaning the text data, split the `text` column into individual words and then count the frequency of words to identify the most common words in the dataset. What is the top 10 of words across all tweets?
4. Remove common German stop words from the list. Which words are now the most frequently used in all tweets?
5. See if you can assign topics to each one of the most frequently used word. Focus especially on the nouns. Try to identify the tweets belonging to each topic you found.

SUBMISSION

For your first exercise, please submit:

- Two R-scripts with your solution to the tasks 1 & 2:
  1. `task_1_search_extract_transform_tweets.R`
  2. `task_2_word_count_in_tweets.R`

Submit all files via the corresponding exercise in ILIAS.