**Data Mining and Word Clouds**

**Exercise 1.**

Select a tweeter user of your interest. Briefly (1-2 sentences) reason your choice in comment of the code. E.g.: # “For this Assignment I have chosen [Basti Schweinsteiger ‏account @BSchweinsteiger](https://twitter.com/BSchweinsteiger) because he is one of my favorite football players”. You can pick up a user from here: <https://trendingdeutschland.com/tweeps>

* 1. Scrape 1000 tweets of this user. Present them as a dataframe in R and show the first 6 rows of this dataset.
  2. How many “likes“ and “retweets” average post from this user becomes? Print out top 5 most “liked” posts.
  3. Return 10 most often referenced accounts within your sample of tweets. Referenced accounts can be recognized by @useraccount. Example from Obama tweets: *Thank you for your leadership* ***@RepHalRogers****. This epidemic doesn't discriminate between red or blue, so it's up to all of us to do our part.* Plot the frequency of 10 most often referenced accounts.
  4. When were the extracted tweets reported? Plot a histogram of tweet counts with time on x-axis. Hint: **ymd\_hms()** function may be useful. It transforms dates stored as character vectors in year, month, day, hour, minute, second format to POSIXct objects
  5. What devices/services (e.g., Web, Android, Iphone) were used to post tweets? Plot histogram by source (=device/service).
  6. Inspect the content of your sample of tweets. Do necessary text transformations and clean the text as if you want to present the 1000 tweets as a word cloud. Explore the standard set of English/German stopwords, e.g. here <https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords> and add at least 2 more stopwords that have in your opinion little value for your sample. Consider the language of the tweets when choosing between English/German stopwords set.
  7. Generate a term-document matrix and print out 10 most frequently used words as a table.
  8. Create a bar plot for the 10 most frequently used words.
  9. Generate the word cloud so that maximum number of words to be plotted is 20 and minimum frequency of the plotted word is 2. Print out 1 word cloud in the color palette of your choice and 1 word cloud in black-and-white colors.

**Exercise 2.**

1. Select a tweeter user, somehow related to a twitter user of your interest in Task 1. It may be a friend, business partner, rival / competitor, advisor or any other person from a similar field. E.g. [Cristiano Ronaldo and Lionel Messi, Kim Kardashian](https://www.skybet.de/de/news/fussball/ronaldo-vs-messi-der-ultimative-vergleich-zahlen-statistik.shtml)[and](https://www.skybet.de/de/news/fussball/ronaldo-vs-messi-der-ultimative-vergleich-zahlen-statistik.shtml)[Beyoncé](https://www.skybet.de/de/news/fussball/ronaldo-vs-messi-der-ultimative-vergleich-zahlen-statistik.shtml). Shortly comment on your choice. Scrape 1000 tweets of this second user.
2. Inspect the content, do necessary text transformations and clean the text using the standard set of English (German) stopwords thus preparing to present word clouds for both speeches. Print out the word cloud for the tweets of the 1st user in red colors and the word cloud for the 2nd user in blue colors, set maximum number of words to be plotted and minimum frequency by yourself. Shortly #comment on your choices.
3. Generate a term-document matrix for the tweets of each user and print out 6 most frequently used words for the tweets of each user.
4. Go to <https://rpubs.com/brandonkopp/creating-word-clouds-in-r> or use any other source to get acquainted with Comparison Cloud. Generate Comparison Cloud using **comparison.cloud()** function for the tweets of 2 users. Set the argument **max.words** on your own. Shortly #comment on the output, e.g.: “#*From the comparison cloud, we can see that issues like Iraq were more front-and-center in 2008 than in 2016. We also see ISIL, which didn’t exist (at least by that name) in 2008, pop up in President Obama’s speech. “Change” was used more by President Obama and, interestingly, “hope” was used more often in President Bush’s 2008 speech*“
5. Get acquainted with Commonality Cloud. Generate Commonality Cloud for the tweets of 2 users. Shortly comment on the output.

**Exercise 3.**

Create a word cloud of a web page: <https://www.uni-potsdam.de/de/social-media-krasnova.html>   
Do text cleaning, if necessary. Set the color palette, maximum number of words to be plotted and minimum frequency by yourself. Shortly #comment on the output.   
Hint: explore, e.g. <http://www.sthda.com/english/wiki/print.php?id=159> how to do it.

General tips to successfully complete the assignment:

* For text mining, you can use both **tm** or **tidytext** package.
* Please shortly comment on what you are doing, to what task the code is written, etc.
* Do not forget to call a necessary package using the **library ()** function.
* Submit a report as a .pdf document.