Warm-Up 07: Regular Expressions

Stat 133, Fall 2018, Prof. Sanchez

Due date: Nov-13 (before midnight)

The main purpose of this assignment is to work with strings. More especifically, you will practice with some *basic/intermediate* manipulations of strings and regular expressions.

General Instructions

- Write your narrative and code in an `Rmd` (R markdown) file.
- Name this file as `warmup07-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `warmup07-gaston-sanchez.Rmd`).
- Submit your Rmd and html files to bCourses.

Data "Emotion in Text"

You'll be working with the data file `text-emotion.csv` available in the course github repository. The original source is the data set "Emotion in Text" from the website Crowd Flower Data for Everyone https://www.crowdflower.com/data-for-everyone/

The file contains four columns:

- `tweet_id`: tweet identifier
- `sentiment`: class or sentiment label
- `author`: username author of the tweet
- `content`: content of the tweet

In your `Rmd` file write R code to do computations in order to answer each of the following questions

1) Number of characters per tweet

- Count the number of characters in the tweet contents; create a vector for this purpose. It is possible that you find tweets containing more than 140 characters. This has to do with the so-called *predefined XML entities* such as
  - `&amp;` which represents an ampersand `&`
  - `&quot;` which represent quotes `"`
  - `&lt;` which represents less-than symbol `<`
  - `&gt;` which represents greater-than symbol `>`

- Display the `summary()` of the vector obtained above.
- Likewise, graph a histogram of these counts. To plot the histogram, use a bin width of 5 units: 1-5, 6-10, 11-15, 16-20, etc. In other words: the first bin involves tweets between 1 and 5 characters (inclusive), the second bin involves tweets containing between 6 and 10 characters (inclusive), and so on.
- Are there any tweets with 0 characters? (*write a command that answers this question*).
- Are there any tweets with 1 character? If yes (*write commands that answer these questions*):
  - how many?
  - what is their content?
  - what is their location (i.e. index or position)?
- What is the tweet with the most characters (i.e. max length)? (*write a command that answers these questions*).
  - the number of characters
  - display its content
  - what is its location (i.e. index or position)?

2) Sentiment

- What are the different types of sentiments (i.e. categories)? (*write a command that answers this question*)
- Compute the frequencies (i.e. counts) of each sentiment (and display these frequencies).
- Graph the relative frequencies (i.e. proportions) with a horizontal barplot (bars horizontally oriented) in decreasing order, including names of sentiment types.
- Sentiment and length of tweets: compute a table with the average length of characters per sentiment (i.e. average number of characters for `neutral` tweets, for `happy` tweets, etc.). Display this table.

3) Author (usernames)

According to Twitter, usernames:

- cannot be longer than 15 characters
- can only contain alphanumeric characters (letters A-Z, numbers 0-9) with the exception of underscores (i.e. cannot contain any symbols, dashes or spaces, except underscores)
- *If you want to know more about twitter usernames, visit:*

https://help.twitter.com/en/managing-your-account/twitter-username-rules

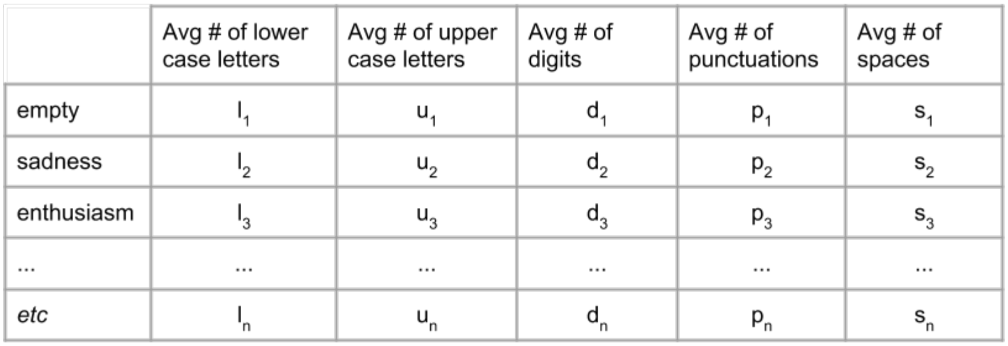
Confirm that the values in column `author` follow each of the rules for valid usernames:

- No longer than 15 characters *(if you find usernames longer than 15 characters, display them)*
- Contain alphanumeric characters and underscores *(if you find usernames containing other symbols, display them)*
- What is the number of characters of the shortest usernames? And what are the names of these authors? *(write commands to answer these questions)*

4) Various Symbols and Strings

- How many tweets contain at least one caret symbol `"^"` *(write a command to answer this question).*
- How many tweets contain three or more consecutive dollar symbols `"$"` *(write a command to answer this question).*
- How many tweets do NOT contain the characters `"a"` or `"A"` *(write a command to answer this question).*
- Display the first 10 elements of the tweets that do NOT contain the characters `"a"` or `"A"` *(write a command to answer this question).*
- Number of exclamation symbols `"!"`: compute a vector with the number of exclamation symbols in each tweet, and display its `summary()`.
- What's the tweet (content) with the largest number of exclamation symbols `!`? Display its content. *(write a command to answer this question)*
- How many tweets contain the *individual* strings `"omg"` or `"OMG"` *(write a command to answer this question).* For example:
  - `omg I just saw them again` (this would be a match)
  - `OMG I just saw them again` (this would be a match)
  - `I just saw them again omg` (this would be a match)
  - `I just saw them again OMG` (this would be a match)
  - `I just saw them omg can't believe it` (this would be a match)
  - `I just saw them OMG can't believe it` (this would be a match)
  - `omg: I just saw them again` (this would NOT be a match)
  - `OMG,I just saw them again` (this would NOT be a match)
  - `I just saw them again omg!!!` (this would NOT be a match)
  - `I just saw them again omgomgomg` (this would NOT be a match)
  - `I just saw them again lol-omg!!!` (this would NOT be a match)

5) Table of Average Number of Patterns by Sentiment



	Avg # of lower case letters	Avg # of upper case letters	Avg # of digits	Avg # of punctuations	Avg # of spaces
empty	l_1	u_1	d_1	p_1	s_1
sadness	l_2	u_2	d_2	p_2	s_2
enthusiasm	l_3	u_3	d_3	p_3	s_3
...	...	...	...	...	...
etc	l_n	u_n	d_n	p_n	s_n

Write code to create (and display) a table (e.g. data frame, tibble, matrix) in which the rows correspond to the unique types of sentiments, and the columns correspond to:

1. average number of lower case letters
2. average number of upper case letters
3. average number of digits
4. average number of punctuation symbols
5. average number of spaces

Hint: POSIX character classes are your friends (e.g. `"[[:xdigit:]]"`).