Surname…… Name……..ID number……

Text Mining and Sentiment Analysis

24th July 2023

TO BE READ BEFORE STARTING

Perform the exercises using the white space on the exam paper, anything else will be evaluated.

---

1. With reference to sentiment analysis, answer the following questions.

1.a What is sentiment analysis (SA)? What are the two main approaches to implement SA?

1.b Describe the dictionary-based approach and list 3 drawbacks.

1.c Why the Udpipe approach can improve dictionary-based SA?

1.d Describe Lemmatization and Stemming. Why are they useful in machine learning algorithm to implement SA?

2. With reference to string manipulation, answer the following questions.

2.a What is a regular expression?

2.b List at least three common uses of regular expressions

2.c What do we mean by metacharacters?

The following exercises have to be solved using the R software. Write an R script that starts with a comment line with your name and surname and that performs the requested tasks. Answer the questions using the space provided below. Once you have completed the exercises, you should upoload the script on the e-learning webpage. Save the script file with name corresponding to your surname and ID number.

3. The tibble **movies.rds** contains a list of Marvel movies and some corresponding characteristics:

- **Title**: Movie title
- **Date**: Release date
- **Phase**: Marvel Cinematic Universe phase

Write proper code to perform the following steps and answer the questions:
3a. Filter the tibble to select only films about "Avengers". Report the list of selected titles.

3b. Using a proper function, count the number of movies with a subtitle (i.e., titles with ":"). Report the result below.

3c. What is the average length of titles for each Phase? Report the result below.

3d. Using regular expressions extract and store in two columns the month and the year from the Date.
3e. What is the most popular month to release a movie? Report the result below.

4. The tibble **moviesTopic.rds** contains a collection of movies' plots:

- **doc_id:** Document id
- **text:** Plot

Your objective is to implement a topic modelling analysis in order to classify the movie saga.
Write proper code to perform the following steps and answer the questions:

4a. Read the data in R and create a tibble.
4b. In order to analyse the topic of the variable *text*, transform it in tidy format and remove the stop words. How many words have been removed in this process?

4c. Produce a frequency distribution for the words. Which are the three most common words used in the *text*? How many times do they appear?

4d. Produce a frequency distribution for the words in each document. Then, convert the tidy data into the dtm data format.

4e. Implement the LDA model with 4 topics. Use as seed 654.

4f. Extract the beta matrix and plot the 7 most probable words by topic.

4g. What are the topics about?

4h. Extract the gamma matrix. Compare and comment the values for the document number 1 and 18.