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?

Sentiment Analysis is the process of extracting an author's emotional intent from text.

The two main approaches are:

- The lexicon-based approach: It relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It involves calculating orientation for a document from the semantic orientation of words or phrases in the document. dictionary or corpus-based.
- The machine learning approach: supervised or unsupervised. Applies ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features. Supervised (labelled dataset is available) or unsupervised (labelled dataset is not available, we learn the latent structure of data).

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

The dictionary-based approach considers the text as a combination of its individual words, and the sentiment content of the whole text as the sum of the sentiment content of the individual words.
This approach performs well with short and concise text but it has some drawbacks
3 of these:

- Requires powerful linguistic resources that are not always available
- inability to find opinion words with domain and context specific orientations
- Urban slang and abbreviation
- Sarcasm
- It is based on unigrams
- Does not consider qualifiers before a word (e.g. «not good»)

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

Because we can take into account negation words and amplifiers. Negation word is a word that would lead to the opposite polarity. Amplifiers introduce a stronger tone in the sentence

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

Are two alternative methodologies for word normalization;

• Lemmatization: is the process of determining the lemma of a word based on the context and identifying the part of seech of each word. This is done by implementing Machine Learning algorithms and can be computationally expensive;

• Stemming: is the process of reducing the word to its «stem» (or root) eliminating the suffix. It is less computational expensive than lemmatization;

Both methods are used in order to reduce the dimensionality and speed up the estimation processes when implementing ML algorithms.

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

2.a What is a regular expression?

A regular expression is a pattern that describes a set of strings

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

Common uses:

- test if a phone number has the correct number of digits
- if a date follows a specific format (e.g. mm/dd/yy),
- if an email address is in a valid format, or if a password has numbers and special characters.
- search a document for gray spelt either as “gray” or “grey”.
- search a document and replace all occurrences of “Will”, “Bill”, or “W.” with William.
- count the number of times in a document that the word “analysis” is immediately preceded by the words “data”, “computer” or “statistical” only in those cases.
- convert a comma-delimited file into a tab-delimited file or find duplicate words in a text.

2.c What do we mean by metacharacters?

A metacharacter is a character that has a special meaning when used inside a regular expression.

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.

```
# A tibble: 4 x 3
  Title                  Date           Phase
  <chr>                  <chr>          <dbl>
1 Marvel's The Avengers  May 4, 2012       1
2 Avengers: Age of Ultron May 1, 2015      2
3 Avengers: Infinity War  April 27, 2018   3
4 Avengers: Endgame       April 26, 2019   3
```

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

11

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

```
# A tibble: 4 x 2
  Phase   avg
  <dbl> <dbl>
1     1  16
2     2  19.7
3     3  19.7
4     4  20.8
> |
```

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.

May with 8 movies.

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?

   1288 -616

   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?

   word n

   <chr> <int>

   1 rocky 18

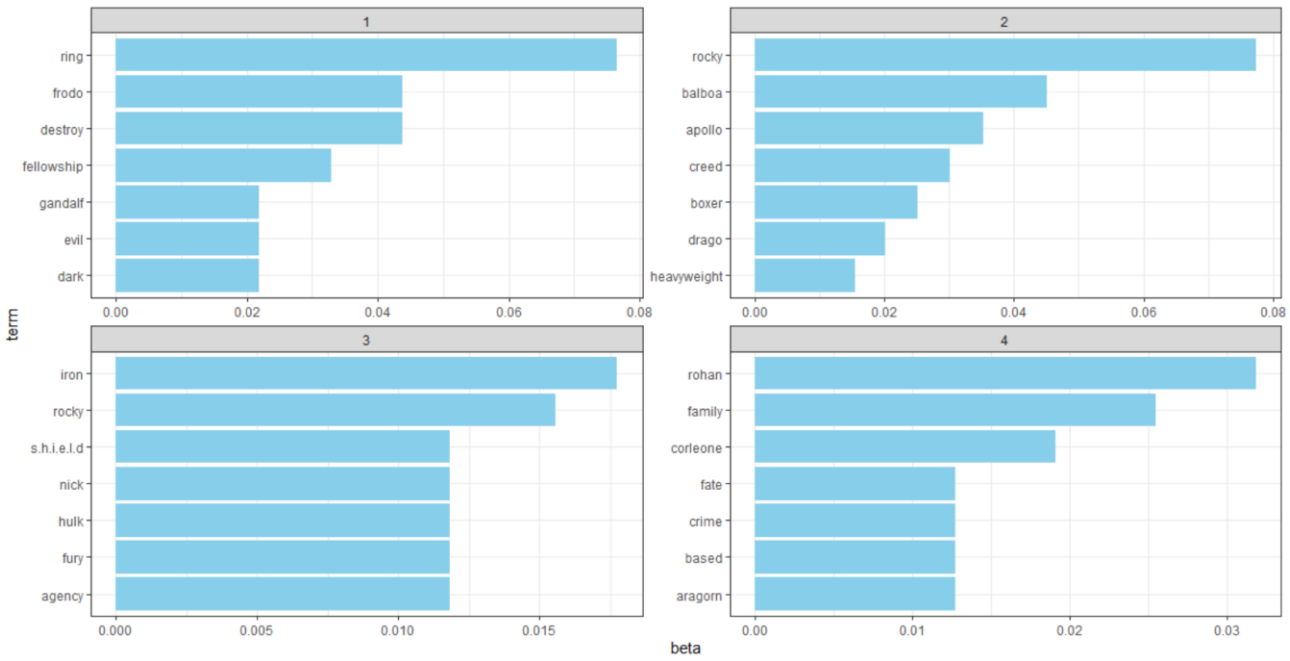
   2 ring 11

   3 balboa 9

   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?

The plots are about the Avengers of Marvel's saga, The Lord of Rings saga and The Godfather saga.

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

Document 1 his prevalently about the 4 topic (Godfather - it is correct). While document 18 is prevalently about topic 3 (0.7 Avengers) and topic 1 (0.3 The Lord of Rings). However, the document is from the Lord of Rings saga.