Text Mining and Sentiment Analysis

Prof. Annamaria Bianchi
A.Y. 2024/2025

Lecture 5

3 March 2025

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Scienze Economiche

Exercises

Exercise 1. Generate a string with the following elements:

```
"Text Mining","Big-Data","Data Science","Math101","Stat01"
```

a. Which is the length of each string? Use a proper function.
b. Convert the strings to lower characters using a proper function and save it.
c. Combine the elements into a string of length 1. Separate the elements by '; '.
d. Combine the elements into a vector of dimension 2: the first element should contain the first three strings, the second element the remaining ones. Separate the combined elements by '/'.
e. Create a tibble containing the strings and the first and last letter of each of them.
f. Create a regular expression to find the words containing numbers.
g. Create a regular expression to find the words containing three numbers.
h. Create a regular expression to find the strings made by two words.
i. Count the number of strings with numbers using a proper function.
l. Substitute the word "text" with the word "data".
m. Remove all numbers from the strings using a proper function.

Exercises

Exercise 2 [Exam July 23]. 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:

2a. Filter the tibble to select only films about “Avengers”. Report the list of selected titles.

2b. Using a proper function, count the number of movies with a subtitle (i.e., titles with “:”). Report the result below.

2c. What is the average length of titles for each Phase? Report the result below.

2d. Using regular expressions extract and store in two columns the month and the year from the Date.

2e. What is the most popular month to release a movie? Report the result below.

Exercises

Exercise 3. The tibble **df.rds** contains a list of countries and the corresponding continent. Write proper code to perform the following steps and answer the questions:

3a. What is the average length of the country names as they appear in the dataset? Use a stringr function to compute it and report it below.

3b. Extract the first letter of each country's name and produce a frequency plot. What can you say about the frequency distribution?

3c. What countries have the word "and" as part of their name?

3d. Identify and delete all instances of "," from the country names.

3e. Only one country has "x" in its name. Which is it?

3f. Create a variable counting the number of "a"s in the country names. What is the country that has the most "a"s in its name?