Session 3: Problem Solving

(*continues from Session 2*)

Objectives

To experiment with text vectorization and implementation in Natural Language Processing.

Tasks

1. Load the CSV file with the companies' data into a data frame.

2. Do you have a text feature of a company? If you don't, add as a feature the description of companies' activities.
3. Apply one of the popular measures of a distance between vectors to find out which two companies are closest in activities.

4. Extend it with a column containing the web addresses of the companies. Scrap information from the web pages, read the text paragraphs only.
5. Try to search for some specific information you find relevant applying cosine similarity measure.

6. Explore the results.