# PROGRAMMING FOR COGNITIVE SCIENCES

## Laboratory 4: **Text-mining**

## Tasks:

1. Load the following description of the Cognitive technologies field into RStudio. You can download in from the Platform.

   The proposed field of study Cognitive Technologies allows you to gain practical skills in the use of modern technologies used in Industry 4.0, modern business services and Smart City management. The field of study is prepared under the KATAMARAN program from the National Agency for Academic Exchange (NAWA), in partnership with the Kiev National University of Building and Architecture - KNUBA (Ukraine). Students during 4-semester Master's studies will have the option of carrying out stationary classes and distance learning. The assumptions of the project are in line with the mission of the Silesian University of Technology - educating highly qualified staff for a knowledge-based society and economy - and the university's vision - preparing the elite of society and supporting the dynamic development of the economy in the spirit of ethical values. By becoming a graduate of the field, you can, among others, become: a specialist in interaction with artificial intelligence systems; a business process analyst; a specialist in cognitive and decision-making processes; a specialist in designing solutions for Industry 4.0; a Smart City specialist; data scientist or a research or research and teaching employee.

   a) Find the positions of the word “specialist” in the text above. How many times does it appear?
   b) Extract the first 100 signs and print them.
   c) Find where the first sentence ends and extract it.
   d) Check how many digits there are in the text.
   e) Substitute "4-semester" with "four-semester".
   f) Check how many sentences there are in the text.
   Hint: All sentences end with a dot. After each sentence (apart from the last one) there is a space.
   g) Check how many words there are in the text.
   h) Check how many signs there are in the text.
   i) Check how many unique signs there are in the text.
   j) Print 10 most and least frequent signs.
   k) Translate the text to lowercase.
   l) Export the resulting text to a TXT file.

2. Load the tweets.RData file into RStudio. It is a data frame containing the tweets published by the RDataMining profile in a certain time range.
   a) What is the number of tweets?
   b) Change "-" to “/” in the publication dates.
   c) Create a new column containing only the year of publication.

d) Create a violin plot of retweet counts per year. Mark years with different colours (parameter `fill`).

e) Check how many words there are in each tweet. Hint: if you have a list, in which each element is a vector of words used in a tweet, you may use the `lengths()` function.

f) Count how many times each word appears in all tweets considered together.

g) Make a wordcloud plot with words that appeared at least 10 times.