TP – Information Retrieval from the Web

1) Choose 10 short articles on the Web about the schizophrenia disease. They can be found on Wikipedia, blogs, specialized websites, etc. Extract these texts and save them in .txt files.

2) Write a program to compute the term-document incidence matrix and the inverted index representation for the document collection you created.

   Note that for the tokenization part, you can use the Stanford Tokenizer (http://nlp.stanford.edu/software/tokenizer.shtml). We suggest you to eliminate the stop words before creating the matrix.

   No programming language is imposed, you are free to use the language you prefer.

3) For the document collection, execute and report the results for these queries:

   a. schizophrenia AND drug
   b. for AND NOT(drug OR approach)
   c. (treatment OR medication) AND NOT (risk OR concern)

4) Write a report (min. 2 pages – max. 5 pages) with all the details about the implementation and the obtained results.

SEND BOTH THE REPORT AND ALL THE MATERIAL
(I.E., SOURCE CODE, DOCUMENTS, ETC) IN A .zip FILE TO villata@i3s.unice.fr

DEADLINE: MAY 27TH, 2016