TP4: Mining Lab

Mauro Sozio
Oana Balalau (TA), Maximilien Danisch (TA)
J.B. Griesner (TA), Raphael Charbey (TA)
firstname.lastname@telecom-paristech.fr

This is our last lab session, which will not be evaluated.

Question 1

Implement a k-means algorithm to cluster the collection of documents stored in text.txt (one document per line):

- use TF-IDF to turn the collection of documents into vectors;
- consider all values in $[1, 10]$ for k. Plot the SSE as a function of k and choose the optimal value for k.
- if your code is too slow, you can select only the most frequent words (say less than 1000 words) in order to reduce the number of dimensions.

Question 2

Check "manually" a few texts in the different clusters you obtained and try to label each cluster according to the topic (e.g. 'sport', 'entertainment'). Design an automatic method to label your clusters and implement it.