Big data computing - 2020/2021

Homework 2, due within December 23st, 11.59pm

You must hand in your homework by the due date and time by an email to the instructor (becchetti@diag.uniroma1.it) that will contain as attachment a zip containing: 1) your code, ii) a README.txt that explains how to run it, iii) One /two pages describing your design choices (and their motivations) and a brief description of your findings.

Important: the subject of your email should be: [BD] [Last_name First_name] HW2

Assignment

Download the dataset available here

The dataset contains two files. File `corpus.txt` contains a collection of short texts from 2 different topics, one per line. The file `labels.txt` contains the corresponding (numeric) labels. You should use the labels only to assess the quality of your clustering of the documents (see below). Your tasks are the following:

1. Use an unsupervised clustering algorithm to cluster the short texts (each short text corresponds to a point of course). You should at least use a variant of K-means (see below) and try to use SVD/PCA, possibly on a sample of the dataset (again, see below).
2. Once you have a clustering of reasonable accuracy (note that we have to classes, so a random classifier would achieve expected accuracy 50%), identify the most important terms for each of the two topics you identified, rendering them in a way that reflects their relative importance (e.g., use word clouds).

Tips and tricks

Try to do your best. The data might not help you: they are relatively large, applying certain techniques might turn less straightforward than you may suppose. For example, memory may not suffice to store the entire dataset in main memory. You are in open waters, do your best to achieve what you can. In the end, we want to understand what the documents talk about. Below a few tips and things you may try:

1. Feel free to reuse/modify/improve any of the notebooks that were made available in the past weeks.
2. You might consider applying scalable variants of k-means(++). `sklearn` provides some tools that scale to large datasets. One of these is MiniBatchKMeans. The idea of the algorithm is explained here and in reference [1] below.
3. Given the task at hand, another trick you might attempt is sampling a subset of the original data, so that you still achieve comparable results (in terms of identifying the topics), but your dataset becomes smaller. How large should the sample be? How should you sample? Well, that is up to you, but you do not have to think of anything exotic, only be careful that your sample should not be too small. Sampling might be a good idea in the case of SVD/PCA, unless you find a way to keep the entire dataset without killing your machine (or you have a lot of memory).

What you should submit

1. Your code, with a README.txt that explains how to run it
2. One /two pages describing i) your design choices and their motivations, ii) a brief description of your findings

References

[1] Web Scale K-Means clustering D. Sculley, *Proceedings of the 19th international conference on World wide web* (2010)