Big data computing - 2019/2020

Homework 2, due date: Wednesday, December 11th, 11.59pm

Handing in your homework

You must hand your homework by the due date and time by an email to the instructor that will contain as attachment a zip with i) your code and ii) a pdf with a short (1-2 pages) report on the implementation AND your answers to the theory questions.

Please remember that subject of your email should be: [BD] [Last_name First_name] HW 2

Assignment 1

Consider the lab on Finding the number of common friends using Spark, that is now published in the section Resources -> Lecture Notes of the course's Web site. You should modify the solution provided there to solve the problem described below.

In this application, we are given a file (available in the "Resources/Other stuff" section of the course's Web site), representing a sample of a the LiveJournal social network that you find at https://snap.stanford.edu/data/soc-LiveJournal1.html>. The network is undirected and is described by a `tab`-separated text file with the following format:

```
7    0,31993,40218,40433,1357,21843
```

The first number is the id of a node of the network. It is follow by a comma-separated list of its neighbours. The original dataset was used and is described in the following papers:

- L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. KDD, 2006.
- J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1) 29--123, 2009.

Your assignment

Your are asked to generate a tab-separated output text file, that for each node pair contains the *list* of common friends. For example, if nodes 2 and 25 have nodes {5, 7, 9, 13} as common neighbours, the output file has to contain the following line: `2,15 [5, 7, 9, 13]`, where the blank is a `tab`

Assignment 2

In this assignment, you should again address the topic distillation/keyword identification addressed in the following lab that we saw together in class. In more detail:

1. Test all three approaches presented in the notebook on the entire dataset. If you look at the dataset description, you will probably notice that there are 6 main topics and 20 topics in total. Retrieve the 10 most important keywords for the main i topics, for $i = 6$ and $i = 20$. For each value of i, try all 3 approaches we considered. Note that i will define the number of cluster/principal components you need to consider.

   Tip: note that, while you need to run k-means many times, you need to compute SVD only once

2. Repeat the same experiments, this time using PCA. Note two things:
   - PCA just requires centering the data, so you really need not change your code, but only the input matrix
   - The rows of the input matrix will now contain positive, negative and possibly 0 entries.

Assignment 3

1. Assume that A is a *square invertible*, n-dimensional matrix, with SVD $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Show that the inverse of A is $B = \sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^T$.

   Hint: recall the properties of the matrices U and V.

2. Suppose again that A is square and has SVD $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, but this time A is *not necessarily invertible*. Let again $B = \sum_{i=1}^{r} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^T$. Show that $BA\mathbf{x} = \mathbf{x}$, for every vector $\mathbf{x}$ that can be expressed as a linear combination of the right singular vectors of A. I.e., we consider vectors of the form $\mathbf{x} = \sum_{i=1}^{r} \alpha_i \mathbf{v}_i$ (we say that $\mathbf{x}$ is in the *span* of the right singular vectors of A). B is called the *pseudo-inverse* of A and can play the role of A^{-1} in many applications.