Università
della
Svizzera
italiana

**Facoltà
di scienze
informatiche**

Data Analytics

Academic Year 2018-19

Course Project N.01: Reuters

Prof. Fabio Crestani

TAs: Esteban A. Ríssola & Manajit Chakraborty

For this assignment you will work individually to carry out simple tasks of data analysis given a specific dataset. The goal of this assignment is to use Python and complementary libraries on a given dataset in order to *explore* and *analyze* the given data and *draw conclusions*.

Description

The dataset contains Reuters news. Documents should be turned into word feature vectors.

This dataset contains a set of documents with their word occurrence patterns. Your goal is to cluster the documents based on occurrence patterns of words. There are class labels for 5 different entities such as 'people', 'places', and topics. Your tasks are to:

- Explore and describe the data (*i.e.*, standard descriptive statistics, visualize the variables with different graphs, draw distributions and histograms of variables, are there outliers? Any interesting observation? Any correlations? Etc.)
- Pre-process the data (*i.e.*, handle and fill unknowns if there are any, etc.)
- Use at least two different clustering algorithm and analyze the results. What is the most optimal number of clusters?
- Evaluate and compare the performance of the models.

Submission procedure and evaluation

You should produce a report of your work and its evaluation along with the source code. It will be a concise explanation of how you tackled the different tasks, the reasons of your choices, successive conclusions, plots you produced, results of the decisions and their accuracy, *etc*.

Use Jupyter Notebook to produce results of the commands in a single .ipynb file. For more information check: https://jupyter.org/documentation

The report (max 5 pages) and the code of the project need to be submitted via iCorsi.

Please, upload all the required items in a single file and name it following the structure: **noProject_FirstnameLastname.[zip|tar.gz|7z]**. For instance, 05_EstebanRissola.tar.gz

The dataset regarding this project can be downloaded from: http://ir.inf.usi.ch/da-datasets/