Project Proposal
Data Mining 2019

Instructions

The project should be made in groups of 2 persons (or 1 person only) and should be delivered **by the date of the second written exam**. For the delivery, you should create a google drive directory that has the following subdirectories:

- Data, that keeps your datasets
- Docs, which will contain your final report
- Src: which contains your source code

Once the folder has been created it should be shared with the instructor (velgias@unitn.it)

The report should follow the ACM template that can be found here:

https://www.acm.org/publications/proceedings-template

It is strictly forbidden to modify the template. Modified templates will be marked as 0. On the first page of the report where is the place of the affiliation of the authors, you should include under your name the following: Your matricola, Your email, Your year of studies, your program of studies (e.g., EIT Embedded Systems), as well as whether you intent to continue in industry or you would be interested for a PhD (write simply "PhD intention", or "Industry intention" or "Do not know yet". The length of the report is hard to specify but as a general guideline it should be around 8 pages.

Introduction

In a typical supermarket, customers buy different kinds of ingredients (food) for the recipes they are planning to prepare at home. We would like to identify types of customers based on the food they eat. This is challenging for many reasons. First because the same ingredient may be used in different recipes or for the same recipe different ingredients may be used. Furthermore, a customer may buy only some of the elements he/she needs (because she may have the rest at home) which means that it is not clear what recipe a person wants to make when buying a specific item.

For each "kind" of customer a basic textual description should be provided that gives a human understanding of what this kind of customers are about.

Datasets:

To produce the kinds of customers that exist we assume the existence of a market basket dataset, as well as a dataset that provides for each recipe its ingredients. One example of the former may be found here:

https://www.kaggle.com/irfanasrullah/groceries

while the second here:

https://www.kaggle.com/kaggle/recipe-ingredients-dataset

These are real datasets, but it is equally important to have tests with some synthetic data.

Software

You are free to use any software that fits you best

Report:

Your report should contain the specific sections:

1) Introduction
   a) A general description of the problem
2) Related work
   a) Articles and algorithms that are related to the problem at hand
3) Problem statement
   a) Formal definition of the problem as we saw in the case of the algorithms presented in class. Input, Output, Variables, etc.
4) Solution
   a) This is the detailed description of your solution. Provide the algorithm also in pseudocode.
5) Experimental Evaluation
   a) Experiments with Real and Synthetic Datasets
   b) Comparison with a base line algorithm or solution of your choice
6) Conclusion
   a) Wrapping up the main contributions and results of your work.

Material for the Written Exam

The written exam will be a series of questions on the theory that we have learned throughout the semester. The topics from the book:

Chapter 1	Data Mining
Chapter 2	Map-Reduce and the New Software Stack
Chapter 3	Finding Similar Items
Chapter 6	Frequent Itemsets
Chapter 7	Clustering
Chapter 8	Advertising on the Web
Chapter 9	Recommendation Systems
Chapter 11	Dimensionality Reduction

From the extra Material available on google classroom (that are not part of the book)

The Statistics chapter on how to compute statistical significance
The SPARK as described on the slides
The noSQL Slides
The Entity Linkage Slides
The Data Integration Slides