University of Ioannina,
Computer Science & Engineering Department

Data Mining – CS 059
Fall 2013

Final Project

In this class you learned a variety of techniques for analyzing data such as Frequent Itemset Mining and Association Rules, Clustering, Classification, Coverage, Ranking. The goal of the project is to apply the tools that you learned in a real-life setting.

For the project you will use the data from the Yelp Dataset Challenge. Yelp is a site that has reviews on different venues (mostly restaurants) from real users. Users write a review, and give a rating in stars from 1 (bad) to 5 (excellent). They also rate other people's reviews, and they can check-in at a venue. The dataset contains information about businesses in Phoenix Arizona. Spend some time on the Yelp site to get a feel how the site works, and the type of content it contains. Also read through the instructions on the data challenge page to understand what information is contained in the data.

For the class project you will propose some analysis on the data to extract some interesting information. The type of analysis that you will do is entirely up to you: it can be some form of association rule mining, some kind of clustering that you believe will yield something interesting, some classification task, or some type of ranking. The goal is to find something useful from the data at hand, or conclude that a specific type of analysis cannot find anything interesting. You should clearly state your goal and how you will evaluate it. As part of your project you will hand in a report with an analysis of what you discovered.

The project will have the following steps:

Step 1: Download the data from Yelp Dataset Challenge page, and decompress it locally (use the `tar xvf` command). You should get a collection of JSON files. Spend some time to understand how Yelp works, and what information is contained in the files.

Step 2: The data is in JSON format, so you will need to build a parser. Build a simple program that reads one of the files and extracts some of the fields. There are several existing libraries in different languages for parsing JSON data in different languages (e.g., look at the json page). You should use one of those in your code. The goal of this step is to familiarize yourselves with these libraries.

Step 3: Create a project proposal for what kind of analysis you want to do. The project proposal should have the following parts:

1. What you want to test or find in the data and why.
2. How you plan to do what you want.
3. How you will evaluate what you did.

For the purpose of the proposal you should address these three points briefly and at a high level.

Step 4: Implement your proposal, and write a report where you describe in detail what you did with respect to the three points above, and how you dealt with different problems that came up. You should also describe what your experiments show.

You will hand in your code and the report. There will also be an examination on the final project.