Guillaume Lemaitre

Dataset Repositories

The aim of this first coursework is to find three dataset repositories as follows:

- One that specialises in financial data
- One that specialises in time series data
- One that specialises in anything else

1. Financial dataset repository:

The first sort of dataset is financial data. The main financial data known is stock market data. With this data, we can make some forecasts. A financial dataset repository concerning stock market can be Yahoo Finance! The URL of this repository is the following: http://finance.yahoo.com/. For each company, Yahoo suggests an historical quote. This data are available on several years, so there are not only financial but also time series dataset. Each quote is divided on seven fields. The fields are:

- Date which is the date of the quote
- Open which is the value of the stock at the beginning of the day
- High which is the highest value of the stock at the day
- Low which is the lowest value of the stock at the day
- Close which is the value of the stock at the end of the day
- Volume which is the exchange of stock during the day
- Adjusted value which is the adjusted value of the stock at the end of the day.

An example of this data is:

Date,Open,High,Low,Close,Volume,Adj Close
2009-09-25,9706.68,9781.73,9605.19,9665.19,4507090000,9665.19
2009-09-24,9749.99,9836.82,9637.53,9707.44,5505610000,9707.44
2009-09-23,9830.63,9937.72,9724.90,9748.55,5531930000,9748.55
2009-09-22,9779.61,9890.71,9742.96,9829.87,5246600000,9829.87
2009-09-21,9818.61,9846.12,9688.40,9778.86,4615280000,9778.86
2009-09-18,9784.75,9898.57,9751.27,9820.20,5607970000,9820.20

The data are in CVS format.

2. Time series dataset repository:

The second sort of dataset is time series dataset. Time series datasets are used to make some estimations and predictions. These estimations and predictions are used to miscellaneous fields. These fields can be sports, financial or concerning organisations of countries. A time series dataset repository regarding countries is the data provided by United Nation. This dataset repository is named UNdata. We can find it at this URL: http://data.un.org/. The different data concern domains like:

- Education
- Energy
- Environment
- Food and Agriculture
- Health
- Human Development
- Indicator databases
- Industry
- Information and Communication Technology
- Labour
- National Accounts
- Population
- Refugees
- Trade
- Tourism

Each time series are collected year-by-year. Each field of record is specific what you want to estimate or compute. By example, to study population of city the different fields will be:

- Country
- Year
- Area
- Sex
- City
- City type
- Record type
- Reliability
- Source Year
- Value

We can see that the field “Year” is the most important because it justifies appellation “Time series”. The data are in CVS, XML, ASCII formats.

3. Other dataset repository:

The last dataset repository chosen is specific at spread service which is internet. On internet, lot of people use the tool “Wiki” and the best known is “Wikipedia”. “Wikimedia” gives a lot of dataset on

the “Wiki”. You can download the data at the follow URL: http://download.wikimedia.org/. “Wikimedia” provide information like:

- Page content
- Page-to-page link list
- Image metadata
- Misc bits
- Log data
- Dump metadata
- Multi-language dumps

The data are in CVS format.