PROJECT PROPOSAL: WEB SCRAPING

Andreu Schoenenberger López

Goal

We will extract information about the Database of Radiological Incidents and Related Events collected from 1896 to 2013 approx. (last major event being Fukushima incident) by Wm. Robert Johnston and stored in the following url: http://www.johnstonsarchive.net/nuclear/radevents/

So there can be several goals for this project, but I think the most important ones are:

- To see the overall (or per country) evolution of deaths due to radiological events (direct deaths in the incident, not posterior deaths) per year.
- To see the overall (or per country) evolution of leaked radiation that can lead to posterior deaths or health issues related to exposure of radiation per year.
- Also, less obvious, if number of incidents worldwide follow a tendency of decrease or increase per year. Knowing that there is more use of nuclear energy but also more safety protocols and knowledge of consequences of nuclear incidents.

To achieve this goal we will use the list of deadliest incidents, list of criticality accidents, list of naval reactor accidents, list of criminal incidents and list of nuclear test accidents. All lists are stored in the previous url, as well as the table listing that describe the codes used to classify incidents in the previous tables.

Data source and web technology

Data source will be a list of tables (containing numeric and text variables) and text information used to classify the type of incident. The web is based on html.

Technology used for downloading

I will use R with the following packages (could change depending on needs): *rvest, XML, RCurl*. Then, functions of those packages will allow us to download the information stored in the web.

Cleaning the Data

To clean the data we will make use of R tools that, for instance, allow us to select columns that we want, merge tables, etc. But the values of the columns of interest will be cleaned using regular expressions to extract the most relevant information.

For example, for the column year of all tables we will have a string like *11 Feb 1945*, we are not really interested in the day and month so we will extract the last 4 digits to obtain the year of the incident, which is the value we are looking for. Similar process will be used in other columns. Also, when constructing the table of codes (code of incident + description of the code) that will be extracted as text by downloading the html page.

Analyzing extracted information

At the end we will have a single table containing year, location, code of incident (if available, also with description), deaths (if available), injuries (if available) and radiation dose emitted (if available). That will be the table used to obtain the summary goals stated at the beginning. We can make use of exploratory analysis (figures), statistical tests and probably modelling, but that will depend on the data extracted which we still don’t have.