Group: Gonzalo Espinosa, Marc Valentí

Web scrapping:

The purpose of our project is to create a database about the Harry Potter characters. The link used is http://harrypotter.wikia.com and is in HTML format.

It is expected to be about 3000 Harry Potter characters.

As we get deeper in our scraping exercise, we are going to decide which variables are we interested in collecting.

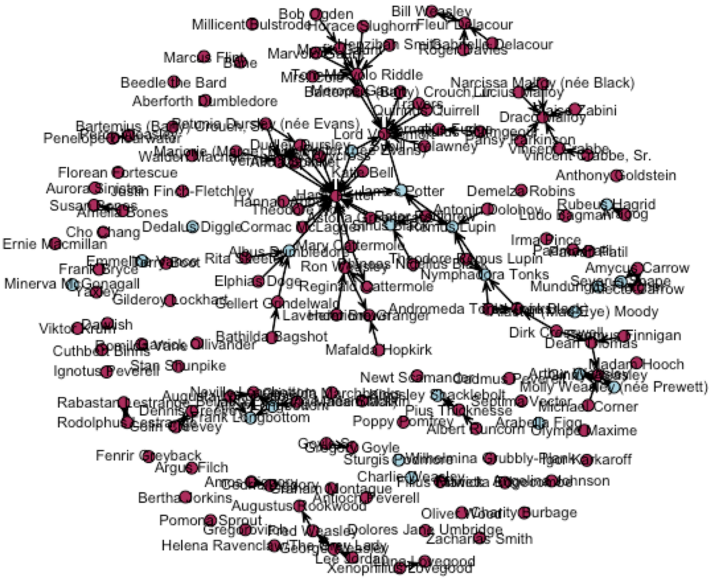
If all the characters are considered, the number of variables will be less. If this is the case, for a given character, variables such as a small description, house, gender, occupation and loyalty will be considered.

If on the other hand, we focus on just creating a dataset with the main characters – the ones that have many information-, variables can be more informative. Apart from the mentioned before, we could consider the family members, and the description given for every book.

Posterior analysis:

If possible, our idea is to create a Shiny App (related to last year course). This Shiny app will be basically doing Network Analysis and simple descriptive analysis. The user will be able to:

- Choose a character and see its characteristics and descriptions and how it is related to the others.
- Visualize a network graph of the period he/she is interested in. This network graph will be created from inspecting the descriptions of every character and see if the other characters are mentioned. This network can also be a dynamic one, having a different network for every book. See the example we have created from scrapping this Wikipedia webpage: https://en.wikipedia.org/wiki/List_of_Harry_Potter_characters



At the end, we'll upload the dataset on the internet so that other users can explore it.