Extracting Social Networks from Literary Text

Moses Boudourides[*1]

[1]University of Patras – Greece

Abstract

In this workshop, we intend to use certain natural language processing, text analysis and and machine learning techniques in order to extract social networks from texts of literary fiction (or other texts such as archives, biographies, transcripts of movies or theatrical plays etc.). In the workshop, we are going to elaborate on a mixed automatic and manual approach, which is driven by the extraction of three entities: (1) character names, (2) a direct and indirect speech attribution and (3) a categorization of attributes to characters and quoted speech (and other referrals).

(1) Text is processed with the Stanford Named Entity Recognition (NER) tagger that seeks to locate and classify textual elements into pre-defined categories such as the names of persons, organizations, locations, expressions of times, various quantities, etc. The outcome of this task is to identify the actors of the social network to be extracted from the text.

(2) Although there are many methods for direct and indirect speech attribution (so that one might be able to identify occurrences when one character refers to or interacts with another one inside quoted speeches), due to existing tilme limitations in the workshop, we are going to follow a "primitive" approach in which a first automatic retrieve of all speech quotes will be subsequently annotated manually so that one might be able to get a first (rough) approximation of the extracted social network ties among the previously identified actors. A particular (and rather easily implemented) case of this task is when quoted speech consists of (well defined) conversational chunks, as it is the case in many theatrical plays (or transcripts of movies etc.).

(3) Similarly, due to the same limitations of presenting the general methodology of textual social networks extraction in the workshop, the distribution of attributes and attitudes to characters (actors) and the assignment of a certain categorical tags (labels) to speech interactions or other referrals (ties) will be done by annotators who process manually the set of sampled quotations (and possibly resorting to the context of the plot of the text). Nevertheless, in certain cases, one is able to apply the machine learning technique of sentiment analysis in order to obtain a signed social network, by evaluating the affective state of characters (actors) and the emotional effect of their speech or observational interactions or judgements (ties).

An example of such a social network extraction on Shakespeare's Macbeth is given here:

[*]Speaker