Software Engineering Project 1 (Comp 10050)

Assignment 4 – Social network analysis of novels

Aim of the assignment: *to create an application to extract social networks from text.*

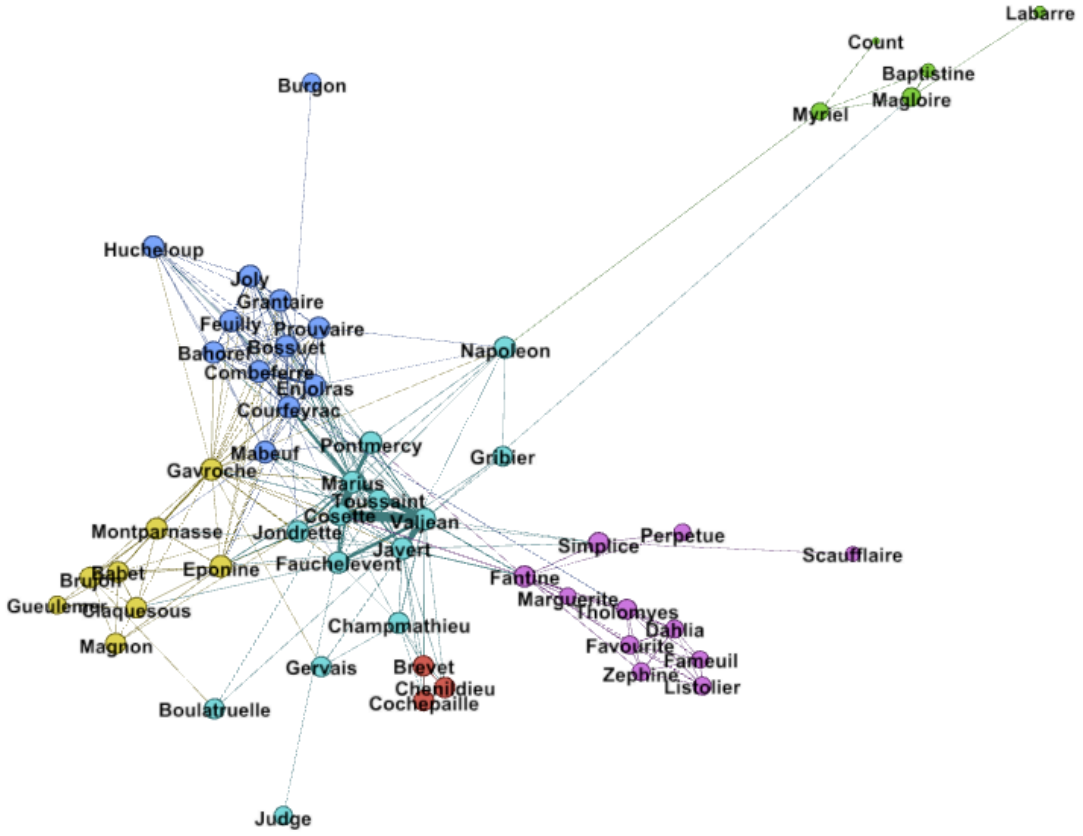

Figure 1. A sample network extracted from Les Miserables.

A. Detailed Specification

The graph in the figure above shows a social network extracted from the text of Les Miserables - the epic novel by Victor Hugo. The graph is constructed by identifying character names that co-occur in the text of the novel and then constructing a network where the weight of an edge between two individuals is proportional to their co-occurrence count. The network is produced using the Gephi software (gephi.org) so the main output of the software to be written is a list of co-occurrences.

Your submission should include your source code and output based on an analysis of the english language text of Les Miserables that is available on the Moodle page. The Moodle page also contains shorter extracts from the full text that can be used for testing. There are also two files containing lists of character names, Les-Mis-Names.txt contains a fairly complete list and Les-Mis-Names-20.txt contains a list of the 20 most prominent characters.

B. Outline for a Possible Solution

Here is an outline for a possible implementation.

1. Load the names into an array.
2. Read the text file a line at a time and, for each name, record the lines numbers on which it occurs.
    2.1. Use `char * strstr (const char *lineOfText, const char *name)` to search for occurrences of name in lineOfText.
    2.2. Use a linked list to record the line numbers for each name.
3. Scan through these lists of line numbers to find co-occurrences.
    3.1. You might decide that two names occurring within five lines of each other counts as a co-occurrence.
    3.2. For every co-occurrence found write the pair of names to a file.
4. The main output is this long file with a pair of names on each line. This file can be read by Gephi.

For example, if the text contains the following 4 lines:

"Madame Fabantou seems to me to be better," went on M. Leblanc, casting his eyes on the eccentric costume of the Jondrette woman, as she stood between him and the door, as though already guarding the exit, and gazed at him in an attitude of menace and almost of combat.

And Fabantou, Leblanc and Jondrette are in the list of names then the output should contain the following co-occurrences:

Fabantou, Leblanc
Fabantou, Jondrette
Leblanc, Jondrette

The Gephi software can build a network from a co-occurrence list.

C. Code Design Requirements

You should use the code for managing an array of lists to record the line numbers on which names have been found.

D. Your Submission

Document your code

You must comment your solution as follows:

1. You should include a short comment at the start of your main c source file which describes generally how the code works (e.g. describe inputs for the game etc).
2. For each function, you should describe (in a few sentences) the purpose of the function, any parameters of the function and possible return types the function may have.

Submitting your solution

You should submit your solution through the COMP10050 Moodle page. The submission should comprise:

1. a single source file,
2. a data file showing sample output from the program (the co-occurrence list)
3. an image from Gephi showing the network extracted from the text.