Université Lumière Lyon II
Master 2 Datamining

Projet fouille de texte

Analyse d'un réseau social de personnes romanesques

Elaboré par :
Amine TAHIRI
Maram ROMDHANE

Sous la direction de :
M. Julien Velcin

Année universitaire :
2016-2017

Table des matières

1 Résumé

Ce travail s'inscrit dans le cadre d'un projet académique lié au module 'Text mining'. Ce projet vise principalement à analyser un ou plusieurs romans de la série "les rois maudits" : Il s'agit d'analyser un réseau social de personnes romanesques, en examinant les différentes relations entre les différents personnages de la saga, pour pouvoir ensuite les visualiser sous forme d'un graphe explicitant ce réseau social.

2 Introduction et contexte du projet

2.1 Introduction

Actuellement, les différents outils informatiques et modèle mathématique se sont avérées bien efficaces dans le traitement automatique des contenus des œuvres littéraires, dont généralement l'analyse manuelle s'avère quasiment impossible. Pour cette raison, des outils d'analyse à base de graphes ont été mises en œuvre, notamment pour les cas d'analyses des réseaux de personnages dans les romans, là ou est amené à déterminer et visualiser les différents relations existant entre les personnages.
Notre projet vise donc de procéder de manière analogue, pour analyser un réseau social de personnages d'un ou plusieurs romans de la série "Les rois maudits", en se basant sur les différentes techniques de Text mining.

2.2 Contexte du projet

L'objectif principal du projet consiste principalement à étudier le réseau social des personnages de la saga "Les rois maudits". Nous sommes donc amenés à :

1. Déterminer l'ensemble des personnages constituant le contexte du roman qui ont un intérêt pour l'étude.

2. Définir la méthode pour calculer les différentes relations qui existent entre les personnages de la saga.

3. Visualiser le réseau social des personnages résultant du calcul des relations entre les différents personnages.

Notre étude a été menée sur le premier roman de la saga, appelé " [Rois Maudits-1] Le Roi de fer - Druon,Maurice ".

3 Conception de la solution

Dans ce chapitre, nous allons expliciter la stratégie de résolution auxquelles nous avons opté pour arriver à notre solution finale d'analyse du réseau social de personnages du roman.

3.1 Préparation du corpus

La structure du texte brut se compose au premier lieu des métadonnées sur le roman, après un prologue, ensuite la succession des chapitres et finalement un répertoire. Un prétraitement sur la structure du texte brut s'avère nécessaire, dans ce sens on a vu le besoin d'améliorer la structure afin

d'aboutir a une structure de corpus plus explicative. Ce prétraitement de la structure consistait simplement à éliminer les métadonnées et le répertoire puisqu'ils citent les personnages dans un contexte générale loin de celui de l'histoire et qui vas biaisé notre texte. Cette transformation est basé sur la définition d'une expression régulière.

Le deuxième prétraitement qu'on effectué sur la structure du texte, c'est de définir les lignes de notre texte du livre comme étant des chapitres, et supprimer les ligne vides ce qui vas nous aider à avoir des corpus plus significatif.

Le dernier prétraitement tiens le point sur la forme textuelle des personnages qui constituent l'axe principale de notre étude, après l'analyse par la lecture de la forme textuelles des personnages on a conclue que cette forme est non unique, parfois on trouve des différents noms pour un seul personnage. On a unifier les noms des personnages d'une manière manuelle en remplacent les les différents noms par un seul nom.
Tous ces prétraitements sont nommés "manuelSagaProcessing".

Finalement on à transformé les noms composés des personnages au noms atomique afin d'éviter la perte de ces noms pendant la phase de nettoyage du corpus, vue leur particularités.

3.2 Extraction des personnages

L'extraction des personnages va être d'une manière manuelle en se basent sur un dictionnaire de l'ensemble des personnages de la saga.

3.3 Extraction des réseaux des personnages

Création du corpus

Après avoir préparer la bonne structure du texte on crée notre corpus qui définis les chapitres du roman.

Prétraitement du corpus

Le prétraitement du corpus consistait en, le nettoyage du corpus tout en suppriment les nombres, les ponctuations et les "stops-words".

Matrice terme-document des personnages

Cette partie consiste à calculer la matrice terme-document de pondération tf-idf, et la fusionner avec le dictionnaire des personnages de la série afin d'extraire la matrice terme-document juste pour les personnages du roman.

Relations entre les personnages

Après avoir définie la matrice terme-document des personnages du roman, la relation entre les personnages est définie par l'intensité de co-ccurence, en remplace les pondérations tf-idf supérieur à zéro par des 1, pour finalement calculer la matrice d'adjacence décrivant cette intensité en multiplient la matrice terme-document des personnages par sa transposé.

3.4 Visualisation du réseau des relations entre les personnages

Après avoir déterminé comment les différents personnages se positionnent entre eux nous avons construit notre graphe de relations basé sur l'intensité de co-occurence. La figure 1 illustre le réseau social des personnages. Pour visualiser le graphe, nous avons utilisé la librairie igraph de R.

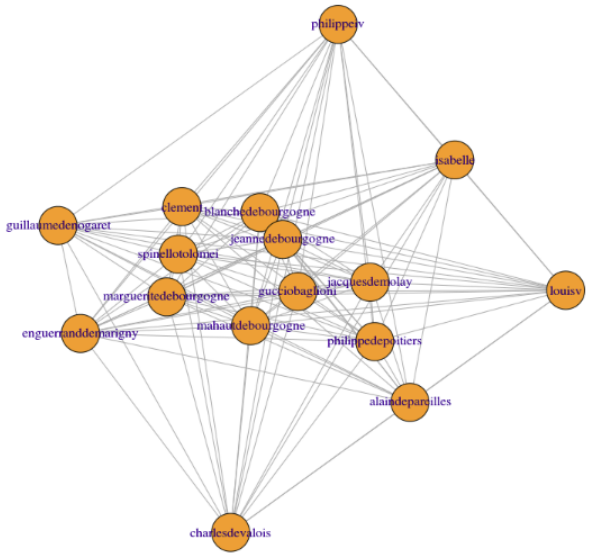

FIGURE 1 – Réseau social des personnes romanesques

4 Conclusion et perspectives

En guise de conclusion, Ce projet nous a permis de bien assimiler les concepts de base du "text mining" et qu' on peut se poser devant des exemples de texte très particulier, ce qui est le cas avec le roman "les rois maudits". Ainsi afin d'aboutir à des résultats correctes une compréhension préalable de la forme de texte s'avère un point important, notamment la structure du texte et les noms des personnages qui construisait notre axe principale d'étude.

Nétographie

[N3] https ://cran.r-project.org/web/packages/igraph/igraph.pdf