Chapitre 1

Implémentation

1.0.1 Objectifs

Notre but étant de pouvoir proposer aux étudiants de première année les éventuelles voix qu'ils sont susceptibles de poursuivre tout au long de leurs études. Le premier objectif de cette étude est dans un premier temps de pouvoir catégoriser les curriculums vitæ des étudiants précédent en fonction des filières que ceux-ci ont suivi. Cette première phase de catégorisation nous permettra par la suite de conditionner l'algorithme utilisé. Cependant la réelle problématique est que pour un CV donné, il sera nécessaire de prendre en considération chaque ligne du parcours d'un ancien étudiant comme une occurrence unique et trouver une méthode permettant de représenter graphiquement l'ensemble de celles-ci afin de pouvoir représenter un parcours professionnel.

1.0.2 Méthode proposée

Le jeu de données

Pour cette étude, le jeu de données utilisé contiendra environ 2000 CV d'un projet pouvant être trouver à l'adresse suivante :

— `https://github.com/JAIJANYANI/Automated-Resume-Screening-System`

Un premier tri a été effectué afin de filtrer les documents en fonction du format utiliser afin de ne retenir que ceux au format **.doc** ou **.docx** pour faciliter les traitements visant ensuite à les catégoriser. En outre ce premier tri permettra par la suite de minimiser la charge de calcul afin que les étapes de traitement et d'analyse ne soient pas impacter par un temps de traitement trop important. A noter que bien que cette première étape de tri aurait pu être réaliser de façon manuelle, l'objectif final étant de proposer cette solution sous la forme d'un programme, il est plus intéressant que cette phase de tri puisse s'effectuer de façon automatique une fois le jeu de données fourni.

Choix de l'algorithme

Afin de pouvoir choisir l'algorithme le plus adapté à notre cas, celui-ci doit pouvoir répondre aux critères suivant :

— Capacité à traiter des documents

— Temps d'exécution
— Capacité à traiter d'important jeux de données

Comme vu précédemment, l'algorithme K-means bien que celui-ci puisse traiter des documents texte, il impose toutefois de connaitre le nombre de clusters souhaités afin de pouvoir effectuer son traitement.

1.0.3 Problèmatiques restantes