

Introduction

Afin d'orienter les étudiants arrivants en première année de licence vers un parcours qui leur semble le plus adapté, il serait intéressant de pouvoir leur fournir un aperçu des chemins suivis par leur prédécesseur.

À cet effet, notre solution devra pouvoir répondre aux problématiques suivantes :

- Comment catégoriser les différentes filières universitaires ?
- Quel secteur l'étudiant sera-t-il susceptible de pouvoir intégrer ?
- Fournir une modélisation de la trajectoire représentée par les différents parcours.

Pour se faire, la solution proposée devra être en mesure de traiter une base de données constituée de documents. Afin de tenter d'apporter une réponse aux deux premières problématiques. La première partie de ce mémoire se concentrera sur l'étude des différentes méthodes de clustering existantes afin de potentiellement apporter un premier élément de réponse concernant cette problématique de catégorisation des données. De plus, pour cette première étape de catégorisation nous nous baserons sur les informations suivantes :

- Un mot clé représentant une filière
- L'ordre dans lequel ceux-ci sont rencontrés

Dans la suite de ce document, seront étudiés les différents types de clusters existants ainsi que les différents algorithmes. Dans un second temps, nous étudierons les types de clusters existants ainsi que les différentes méthodes de clustering et algorithmes. En outre, le choix de l'algorithme se fera en fonctions des critères suivants :

- Capacité à traiter un jeu de données numérique.
- La difficulté d'implémentation.
- La sensibilité au bruit.
- La scalabilité.

Enfin, nous terminerons ce document par un chapitre de conclusion.