Analyse de résumés de données

Travail à rendre

Ce projet consiste à implémenter une approche d'exploration de données et d'extraction de connaissances. Vous allez réaliser ce projet en binôme et vous disposez de trois séances de deux heures encadrées. Vous rendrez un rapport de projet ainsi que le code de votre implémentation.

Principe de l'approche

L'objectif de votre travail est de développer des méthodes d'extraction de connaissances à partir de données brutes (des objets décrits dans un fichier *csv* e.g.). La particularité de l'approche est de reposer sur une première étape de réécriture des données à l'aide d'un vocabulaire utilisateur. Ce vocabulaire, comme l'illustre la figure 1, est composé de partitions floues. Une partition floue est une discrétisation d'un domaine de définition à l'aide d'ensembles flous.

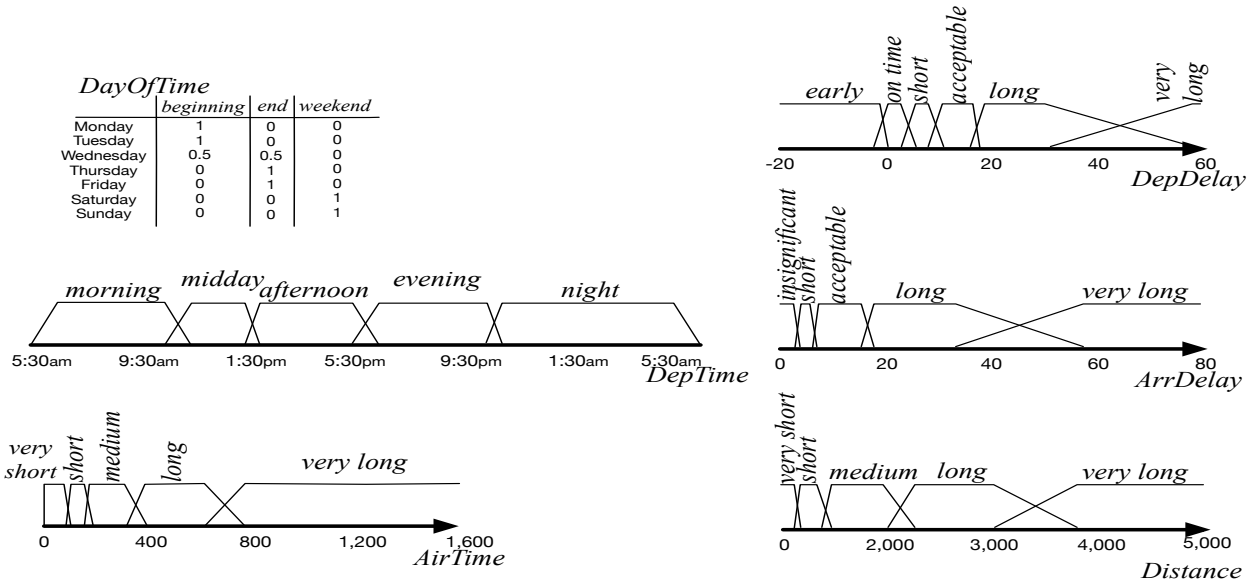

Figure 1 Extrait d'un vocabulaire flou décrivant des vols d'avion

Préparation

Récupérez l'archive *baseProject_BDA..tgz* sur l'ENT.

Cette archive contient un répertoire *Data/* dans lequel vous trouverez deux jeux de données de tests (extrait_2008.csv et test.csv). Une fois votre programme finalisé, vous pourrez le tester sur un jeu complet *2008.csv* (mis à disposition dans un répertoire partagé) contenant la description de plusieurs millions de vols domestiques aux USA en. Vous trouverez également un fichier nommé *FlightsVoc.txt* contenant la description textuelle d'un vocabulaire flou sur les vols d'avion. Enfin, dans le répertoire Src/, vous trouverez une base de code pour effectuer la réécriture des données (vols) selon un vocabulaire. Vous pouvez tester ce code de la façon suivante :

```
python rewriterFromCSV.py ../Data/FlightsVoc.txt ../Data/test.csv
```

Etape 1 : Réécriture des données

Une fois le vocabulaire chargé en mémoire, définissez une méthode de réécriture des t-uples selon le vocabulaire. L'objectif est d'aboutir à un vecteur de réécriture qui décrit pour chaque terme à quel degré il couvre l'ensemble des objets.

Cette couverture correspond à la somme des degrés de satisfaction de chaque objet vis-à-vis de l'élément de vocabulaire, somme ensuite normalisée par le nombre de tuples du jeu de données (en gros une moyenne).

Etape 2 : Exploration de données

Définissez une nouvelle méthode de réécriture des données permettant de ne résumer que les données satisfaisant un ensemble de termes (degré de satisfaction strictement supérieur à un seuil de satisfaction fixé, 0 par défaut). La méthode que vous devez écrire prend donc en paramètre une liste d'identifiants de termes à considérer de manière conjonctive et un seuil de satisfaction).

Cette méthode vous permettra par exemple de résumer uniquement les vols partant de l'ouest et couvrant une longue distance, etc.

Etape 3 : Extraction de connaissances

A partir du vecteur de réécriture du jeu de données complet (désigné par R) et d'un vecteur de réécriture d'un sous-ensemble d'objets (notons-le *R'*) satisfaisant une certaine condition (disons *v*), il est possible d'extraire des connaissances permettant de mieux comprendre les propriétés des objets qui satisfont *v*.

Termes corrélés : Vous allez tout d'abord identifier les termes corrélés à un terme *v*, c'est-à-dire les propriétés impliquées par *v*. La corrélation entre le terme v utilisé pour sélectionner des données et un autre terme v' présent dans le vecteur de réécriture (R') est quantifié par la formule suivante :

$$assoc(v,v') = \begin{cases} 0 & \text{if } dep(v,v') \leq 1 \\ 1 - \frac{1}{dep(v,v')} & \text{otherwise} \end{cases}$$

où

$$dep(v,v') = \frac{cover(v', R_v)}{cover(v', R)}.$$

$cover(v', R_v)$ correspond à la couverture du terme v' dans le vecteur de réécriture Rv (somme des degrés de satisfaction de v' divisée par le nombre de termes résumés dans Rv).

Termes atypiques : Vous allez désormais identifier les termes atypiques dans Rv, c'est-à-dire les termes surprenants. Un terme v' est considéré comme surprenant s'il couvre une minorité des données d'un ensemble (*1-cover(v',R)*) et est éloigné des autres termes (*d(v,v')*) de la même partition qui couvrent de manière majoritaire (*cover(v,R)*) l'ensemble de données concerné.

$$D(v', \mathcal{R}) = \max_{v \in \mathcal{R}} min(d(v,v'), cover(v,R), 1 - cover(v', R))$$

où d(v,v') est une mesure de distance entre les termes v et v'. Si l'attribut concerné par v et v' est numérique alors d(v,v') est égal au nombre d'éléments de partition qui les sépare divisé par le nombre d'éléments de partition moins un. Si l'attribut est catégoriel, la distance est 1 dès que les termes sont différents, 0 sinon. Évidemment, $cover(v', \mathcal{R})$ doit être strictement positif.

Etape 4 (Bonus) : Visualisation de données et de connaissances

Vous disposez désormais de fonctionnalités de résumé (Etape 1), d'exploration des données (Etape 2) et d'extraction de connaissances (Etape 3). Proposez une stratégie de visualisation des termes linguistiques qui apparaissent dans un résumé R et transformez cette vue en interface d'exploration.

Implémentez des stratégies pour visualiser les termes corrélés et atypiques d'un résumé représentant un sous-ensemble des données. Vous pouvez vous inspirer des travaux présents sur le site *d3.js* ou bien rester en Python en proposant des représentations plus simples à base d'histogrammes (module *matplotlib*) ou de nuages de mots (module *wordcloud*).