M2 NLP

UE901 EC2 : Data Mining

Rapport de Projet

Autrices :
Asmaa Demny
Cécile Macaire
Ludivine Robert

1 Février 2021

Table des matières

1 Introduction

Le but de ce projet est d'analyser et comprendre le contenu d'un ensemble de données en utilisant des techniques de *pattern mining* et de *règles d'association*. L'objectif du projet est de pouvoir observer certaines cooccurrences fréquentes entre les attributs.

L'étude des données est basée sur l'algorithme Apriori [1] ; pour la présente analyse nous utiliserons uniquement l'algorithme FPGrowth (*_itemsets* et *_association_rules*).

Pour manipuler l'algorithme, nous avons utilisé SPMF et les ensembles de données décrits ci-dessous.

SPMF [2] est un logiciel open-source et une bibliothèque de Data Mining écrite en JAVA, spécialisée dans le pattern mining.

L'ensemble de données étudié est une enquête sur les personnes vivant en France, plus précisément dans le Grand-Est. Il provient de l'INSEE (Institut National de la Statistique et des Études Économiques. Chaque ligne décrit une personne/famille. La collection nationale de l'ensemble de données contient 3,3 millions d'individus. La base de données de la région du Grand-Est contient 1 474 560 enregistrements. Chaque individu est décrit par 57 attributs multivalués qui sont, soit des valeurs symboliques, soit des valeurs numériques. Certains attributs ont plus de 120 valeurs différentes. L'ensemble de données doit donc être simplifié au maximum. Nous avons sélectionné et simplifié les attributs qui semblent le plus intéressants d'un point de vue sociologique (âge, catégories socioprofessionnelles, mode de vie, etc.).

AGER20	Âge en années révolues (âge au dernier anniversaire) en 13 classes d'âge (détaillées autour de 20 ans)
ARRIVR	Période d'arrivée en France
BATI	Aspect du bâti (DOM)
CS1	Catégorie socioprofessionnelle en 8 postes
DIPL_15	Diplôme le plus élevé
EMPL	Condition d'emploi
INAT	Indicateur de nationalité
MODV	Mode de vie
NATN12	Nationalité à la naissance des Français en 12 postes
NBPI	Nombres de pièces du logement
NPERR	Nombre de personnes du ménage (regroupées)
SEXE	Sexe
STOCD	Statut d'occupation détaillé du logement
SURF	Superficie du logement
TACT	Type d'activité
TYPL	Type de logement

Table 1 – Liste des attributs sélectionnés.

2 Expérimentations

L'ensemble des scripts et résultats sont disponibles dans le répertoire GITHUB à l'adresse `https://github.com/macairececile/data-mining-project`.

2.1 Question 1 : Préparation des données pour SPMF

La première étape a constitué à écrire un programme python pour préparer les données SPMF.

Le script *encode_SPMF.py* prend en entrée le fichier de données GrandEst. Il sélectionne les données de chaque ligne correspondant aux attributs que nous avons sélectionné (cf. table 1). Ces données filtrées seront enregistrées dans un nouveau fichier sous la forme *GrandEst_filter*. A partir de ce dernier, les attributs multivalués sont convertis en attributs à identifiant unique. Par exemple, l'attribut *AGER20_64*, qui correspond un individu ayant en âge révolu entre 55 à 64 ans a, comme identifiant unique, la valeur 2. L'encodage a été enregistré dans un fichier binaire qui servira au décodage après le lancement des différents algorithmes de SPMF. Enfin, l'ensemble des données encodées ont été triées dans l'ordre croissant pour chaque ligne (un individu/une famille). La figure 1 présente le début des données encodées.

```
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
1 3 5 6 7 8 9 10 11 13 14 16 18 19 20 21
1 3 5 6 7 9 14 17 18 19 20 22 23 24 25 26
1 3 7 9 11 12 16 27 28 29 30 31 32 33 34 35
1 3 7 9 11 16 19 26 29 30 31 32 33 34 36 37
1 3 7 9 12 16 23 26 29 31 32 34 35 36 38 39
1 3 4 6 7 9 19 24 25 40 41 42 43 44 45 46
1 3 7 9 19 24 25 26 29 34 35 36 42 43 44 47
1 3 7 9 12 24 25 29 34 35 36 42 43 44 46 48
1 3 7 8 9 11 12 16 24 29 34 35 38 49 50 51
1 3 7 9 11 12 16 24 26 27 29 30 34 38 49 50
```

FIGURE 1 – Début du fichier de données encodées.

2.2 Question 2 : Itemsets avec SPMF et décodage

Pour extraire les itemsets en utilisant l'algorithme FPGrowth_itemsets, nous avons utilisé la ligne de commande suivante :

```
java -jar spmf.jar run FPGrowth_itemsets GrandEst_encode.txt
output_spmf_item_fpg_15.txt 0.15
```

où *GrandEst_encode.txt* est le fichier en entrée comprenant les données encodées, *output_spmf_item_fpg_15.txt* est le fichier comprenant les résultats générés par l'algorithme, et enfin *0.15* correspondant à la valeur du support. Le début du fichier généré par FPGrowth_itemsets est visible dans la figure 2.

```
45 #SUP: 242023
9 45 #SUP: 222620
7 9 45 #SUP: 222620
7 45 #SUP: 222620
2 45 #SUP: 230388
2 3 45 #SUP: 228765
3 45 #SUP: 240119
40 45 #SUP: 242023
9 40 45 #SUP: 222620
7 9 40 45 #SUP: 222620
7 40 45 #SUP: 222620
2 40 45 #SUP: 230388
2 3 40 45 #SUP: 228765
3 40 45 #SUP: 240119
6 40 45 #SUP: 242023
```

FIGURE 2 – Extrait des itemset générés par l'algorithme FPGrowth_itemsets avec un support de 15%.

Nous avons testé par la suite plusieurs valeurs de support : 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80% et 90%. La table 2 présente les statistiques observées par chaque support avec FPGrowth_itemsets.

Support (en %)	15	20	25	30	40	50	60	70	80	90
Nombre d'itemsets fréquents	1330	542	305	157	113	34	16	15	15	1

TABLE 2 – Statistiques observées par les différents supports avec FPGrowth_itemsets.

Le script *decode_SPMF.py* permet de décoder le fichier généré. Il utilise en effet le dictionnaire pour retrouver l'attribut associé à l'identifiant unique. La figure 3 montre le contenu décodé des itemsets trouvés par FPGrowth_itemsets lorsque le support est paramétré à 80%. 15 itemsets fréquents sont présents.

```
NATN12_01 # SUP: 1293417
ARRIVR_Z NATN12_01 # SUP: 1265095
ARRIVR_Z BATI_Z NATN12_01 # SUP: 1239863
BATI_Z NATN12_01 # SUP: 1267347
INAT_11 NATN12_01 # SUP: 1293417
ARRIVR_Z INAT_11 NATN12_01 # SUP: 1265095
ARRIVR_Z BATI_Z INAT_11 NATN12_01 # SUP: 1239863
BATI_Z INAT_11 NATN12_01 # SUP: 1267347
INAT_11 # SUP: 1293417
ARRIVR_Z INAT_11 # SUP: 1265095
ARRIVR_Z BATI_Z INAT_11 # SUP: 1239863
BATI_Z INAT_11 # SUP: 1267347
ARRIVR_Z # SUP: 1293933
ARRIVR_Z BATI_Z # SUP: 1268355
BATI_Z # SUP: 1444294
```

FIGURE 3 – Itemsets générés par l'algorithme FPGrowth_itemsets avec un support de 80% et décodés.

On constate, dans un premier temps, que l'itemset le plus observé concerne la nationalité, ici, *NATN12_01* est associé aux français•e•s de naissance. Le même phénomène est visible avec l'item *ARRIVR_Z* correspondant aux individus nés en France (métropole, DOM, TOM ou COM). Ces deux items sont liés à *BATI_Z* signifiant "logement ordinaire en France métropolitaine", lui même associé à l'item *INAT_11* pour "français•e•s de naissance".

Nous avons réduit le support pour obtenir plus d'itemsets différents, et ainsi pouvoir les analyser.

2.3 Question 3 : Choix des itemsets

En analysant les données décodées lorsque le support est paramétré à 15%, nous avons pu extraire une dizaine d'entre eux qui nous semblaient pertinent dans le domaine étudié, grâce au script *analyse_resultats.py*.

- **BATI_Z TYPL_2 SEXE_2 # SUP : 338114**
  BATI_Z SEXE_1 TYPL_2 # SUP : 302794
  Ces deux premiers ensembles nous indiquent que les femmes vivent plus souvent dans un appartement par rapport aux hommes.
- **BATI_Z SURF_7 STOCD_10 # SUP : 277670**
  L'ensemble ici présenté nous indique que la personne est propriétaire du logement, et que ce dernier est ordinaire avec une surface de plus de 120 m^2.
- **SURF_7 STOCD_10 TYPL_1 # SUP : 263794**
  Cet ensemble d'items nous informe qu'une partie des logements ordinaires dans le Grand-Est en 2016 matérialisés par l'item *BATI_Z* ont une surface de plus de 120 m^2 (*SURF_7*) et sont principalement des logements habités par son acquéreur (propriétaire).
- **ARRIVR_Z BATI_Z INAT_11 NATN12_01 SURF_5 # SUP : 312823**
  Par ces items, nous apprenons qu'une partie des individus nés en France métropolitaine habitent dans un logement ordinaire en France d'une surface de 80 à 100 m^2 (*SURF_5*).
- **BATI_Z INAT_11 NATN12_01 NBPI_05 # SUP : 314526**
  Par cet ensemble, nous constatons que les français•e•s de naissance ont un logement comprenant 5 pièces. Cette observation est corrélée avec la surface du logement précédemment expliquée.
- **NBPI_05 STOCD_10 # SUP : 249826**
  Enfin, nous apprenons qu'une majorité d'individus sont propriétaires d'un bien immobilier de 5 pièces.
- **ARRIVR_Z SEXE_2 TYPL_1 # SUP : 401668**
  ARRIVR_Z SEXE_1 TYPL_1 # SUP : 391318
  Ces itemsets nous permettent de voir, premièrement, qu'en général les personnes du Grand-Est sont nées en France et habitent une maison, et deuxièmement, que les femmes sont un peu plus nombreuses que les hommes dans cette situation.
- **ARRIVR_Z TYPL_1 EMPL_16 TACT_11 # SUP : 236366**
  Par cet itemset, nous pouvons dire qu'en générale les personnes ayant un emploi sans limite de durée (CDI, etc.) habitent dans une maison.

— **CS1_8 DIPL_15_Z TACT_23 # SUP : 242023**
Cet itemset nous permet de dire que le Grand-Est regroupe un grand nombre de personnes qui ne sont pas encore classées dans une catégorie socioprofessionnelle ni diplômées, ce sont celles de moins de 14 ans, c'est-à-dire les jeunes.
— **DIPL_15_A SEXE_2 # SUP : 226831**
Aussi, dans le Grand-Est, se sont principalement des femmes qui n'ont aucun diplôme ou au mieux BEPC, brevet des collèges ou DNB.
— **ARRIVR_Z EMPL_16 TACT_11 # SUP : 398978**
Ces items nous montrent que les individus nés en France sont majoritairement actifs, donc ayant un emploi, et que ce type d'emploi est sans limite de durée, CDI ou titulaire de la fonction publique.
— **MODV_40 NPERR_2 # SUP : 270568**
Ce dernier petit ensemble nous informe que les foyers dont les membres sont un couple sans enfants, de 40 ans ou plus vivent à deux ; et cette situation est fréquente dans le Grand-Est en 2016.

2.4 Question 4 : Extraction des règles d'association et analyse

De la même manière que précédemment pour les itemsets, nous avons utilisé la ligne de commande ci-dessous pour extraire les règles d'association :

```
java -jar spmf.jar run FPGrowth_association_rules GrandEst_encode.txt
output_asr_sup20_conf20.txt 20% 20%
```

Le début du fichier sortie *output_asr_sup20_conf20.txt* est montré dans la figure 4.

```
3 ==> 2 #SUP: 1268355 #CONF: 0.8781833892545423
2 ==> 3 #SUP: 1268355 #CONF: 0.980232361335556
4 ==> 2 #SUP: 425621 #CONF: 0.8821672553018621
2 ==> 4 #SUP: 425621 #CONF: 0.3289358877159791
5 ==> 2 #SUP: 319866 #CONF: 0.8093017607157225
2 ==> 5 #SUP: 319866 #CONF: 0.24720445339905545
6 ==> 2 #SUP: 774910 #CONF: 0.8723428446632873
2 ==> 6 #SUP: 774910 #CONF: 0.5988795401307486
7 ==> 2 #SUP: 1265095 #CONF: 0.9781029629268828
2 ==> 7 #SUP: 1265095 #CONF: 0.9777129109467028
9 ==> 2 #SUP: 1265095 #CONF: 0.9781029629268828
2 ==> 9 #SUP: 1265095 #CONF: 0.9777129109467028
10 ==> 2 #SUP: 314735 #CONF: 0.9000323140583312
2 ==> 10 #SUP: 314735 #CONF: 0.2432390239680107
```

FIGURE 4 – Extrait des règles d'association générées par l'algorithme FP-Growth_association_rules avec un support et une confidence de 20%.

À nouveau, le script *decode_SPMF.py* permet de décoder le fichier généré. La figure 5 affiche le début du contenu décodé des règles d'association trouvées par FP-Growth_association_rules avec un support et une confidence fixées à 20%.

```
BATI_Z ==> ARRIVR_Z # SUP: 1268355 # CONF: 0.8781833892545423
ARRIVR_Z ==> BATI_Z # SUP: 1268355 # CONF: 0.980232361335556
CS1_8 ==> ARRIVR_Z # SUP: 425621 # CONF: 0.8821672553018621
ARRIVR_Z ==> CS1_8 # SUP: 425621 # CONF: 0.3289358877159791
DIPL_15_A ==> ARRIVR_Z # SUP: 319866 # CONF: 0.8093017607157225
ARRIVR_Z ==> DIPL_15_A # SUP: 319866 # CONF: 0.24720445339905545
EMPL_ZZ ==> ARRIVR_Z # SUP: 774910 # CONF: 0.8723428446632873
ARRIVR_Z ==> EMPL_ZZ # SUP: 774910 # CONF: 0.5988795401307486
INAT_11 ==> ARRIVR_Z # SUP: 1265095 # CONF: 0.9781029629268828
ARRIVR_Z ==> INAT_11 # SUP: 1265095 # CONF: 0.9777129109467028
NATN12_01 ==> ARRIVR_Z # SUP: 1265095 # CONF: 0.9781029629268828
ARRIVR_Z ==> NATN12_01 # SUP: 1265095 # CONF: 0.9777129109467028
NBPI_05 ==> ARRIVR_Z # SUP: 314735 # CONF: 0.9000323140583312
ARRIVR_Z ==> NBPI_05 # SUP: 314735 # CONF: 0.2432390239680107
```

FIGURE 5 – Règles d'association générées par l'algorithme FPGrowth_association_rules avec un support et une confidence de 20% décodées.

Nous avions testé plusieurs valeurs de support et confidence : 80 & 90%, 80 & 80%, 80 & 60%, 40 & 30% et 20 & 20%. La table 3 présente les statistiques observées par chaque support et confidence avec FPGrowth_association_rules.

Support (en %)	80	80	80	40	20
Confidence (en %)	90	80	60	30	20
Nombre de règles fréquentes	43	50	50	994	1892

TABLE 3 – Statistiques observées par les différents supports et confidences avec FP-Growth_association_rules.

Enfin, nous avons également utilisé le script *analyse_resultats.py* pour sélectionner les règles qui nous semblaient intéressantes en fonction de l'attribut que nous souhaitions analyser.

On constate que les règles dont la valeur de confidence est la plus élevée concerne la nationalité et le fait d'être né en France ; par exemple, la règle d'association

INAT_11 ==> NATN12_01 # SUP : 1293417 # CONF : 1.0

Nous avons donc fait le choix de générer les règles d'association en prenant un indice de confidence plus bas, de même pour le support, pour obtenir des informations plus variées.

— **TYPL_1 ==> STOCD_10 # SUP : 3983940 # CONF : 0.837**
Cette première règle nous amène à dire que, dans 90% des cas, les habitants de maison en sont propriétaires.

— **TACT_11 ==> SEXE_1 # SUP : 306457 # CONF : 0.523**
Ici on voit que les personnes ayant un emploi de la catégorie des actifs sont souvent des hommes.
TACT_11 ==> TYPL_1 # SUP : 326877 # CONF : 0.557
Cette règle nous indique qu'une personne active au niveau de l'emploi vit dans une maison (*TYPL_1*).
TACT_11 ==> STOCD_10 # SUP : 337372 # CONF : 0.575
En regardant les règles relatives à l'attribut *TACT*, on peut voir qu'en moyenne les actifs ayant un emploi, y compris sous apprentissage ou en stage rémunéré du Grand-Est sont des hommes, qu'ils habitent une maison et enfin qu'ils en sont propriétaires.

— **NATN12_01 ==> SURF_5 # SUP : 319649 # CONF : 0.247**
L'attribut *SURF* n'est pas très présent dans notre ensemble, mais nous constatons tout de même que, dans 25 % des cas, les français•e•s de naissances ont une superficie de logement de 80 à 100 m^2 (*SURF_5*).

— **STOCD_10 ==> SEXE_1 # SUP : 390023 # CONF : 0.489**
Cette règle nous montre la répartition des hommes propriétaires d'un logement.
STOCD_10 ==> SEXE_2 # SUP : 408019 # CONF : 0.511
Par l'attribut *STOCD*, nous observons simplement que les propriétaires sont plus souvent de sexe féminin, dans le Grand-Est en 2016, mais l'écart est minime.

— **SEXE_1 ==> ARRIVR_Z # SUP : 627505 # CONF : 0.878**
La règle indique qu'une personne est née en France, et que cette dernière est un

homme.
SEXE_2 ==> ARRIVR_Z # SUP : 666428 # CONF : 0.878
Également, *SEXE* nous apprend que les hommes et femmes du Grand-Est sont nés en France (métropole, DOM, TOM ou COM), dans environ 90% des cas.

— **ARRIVR_Z ==> NBPI_05 # SUP : 314735 # CONF : 0.243**
ARRIVR_Z ==> NBPI_04 # SUP : 320449 # CONF : 0.247
Ici, les personnes nées en France, dans 25 % des situations, habitent un logement de 4 ou 5 pièces.

— **INAT_11 ==> SEXE_1 # SUP : 626612 # CONF : 0.484**
INAT_11 ==> SEXE_2 # SUP : 666805 # CONF : 0.515
Ces deux règles montrent la répartition des sexes chez les français•e•s de naissance. Dans la région Grand-Est, plus de femmes sont présentes.

— **CS1_7 ==> EMPL_ZZ # SUP : 312145 # CONF : 1.0**
CS1_8 ==> EMPL_ZZ # SUP : 482472 # CONF : 1.0
Par cette présente règle, on constate que l'ensemble des retraités n'ont pas d'emploi, de même pour les autres personnes sans activité professionnelle. Effectivement, étant une vérité générale, la valeur de confidence ici est logique et cohérente dans les deux cas.

— **EMPL_ZZ ==> TYPL_2 # SUP : 389257 # CONF : 0.438**
EMPL_ZZ ==> TYPL_1 # SUP : 466109 # CONF : 0.524
Être sans activité professionnelle implique de vivre dans un appartement dans plus de 43% des cas et de vivre dans une maison dans plus de 52% des cas.

— **DIPL_15_A ==> EMPL_ZZ # SUP : 306748 # CONF : 0.776**
Avec une confidence de 77%, une personne ne possédant aucun diplôme ou au mieux BEPC, brevet des collèges ou DNB a une condition d'emploi sans objet.

3 Conclusion

Pour résumer, la réalisation de ce projet s'est divisée en trois parties :

1. Encodage des données.
2. Décodage des données.
3. Analyse des résultats.

L'encodage des données nous a permis de préparer les données pour être utilisés avec SPMF. Le choix des attributs s'est focalisé sur les aspects sociaux ou encore économiques des personnes vivant dans la région Grand-Est. En effet, nous avons pris en considération entre autre leur sexe, nationalité, condition de travail ou encore type de logement. Nous avons constaté que les différentes pièces qui peuvent se trouver dans un logement n'étaient pas forcément intéressantes car elles ne donnent pas d'informations décrivant la situation d'une personne (par exemple l'attribut WC a été écarté car, en général, tous les logements en possèdent au moins un).

En ce qui concerne le décodage des données, cette étape nous a permis d'analyser les itemsets et les règles d'associations extraits par FPGrowth et ainsi obtenir les situations les plus récurrentes dans le Grand-Est.

D'après le décodage des nos données, nous pouvons affirmer que :
— Dans la région Grand-Est, plus de femmes sont présentes que d'hommes.
— Dans 90% des cas, les habitants des maisons en sont propriétaires.
— Les propriétaires des maisons sont souvent des retraitées.
— Si le propriétaire d'un logement n'est pas retraité, il est majoritairement actif et a un emploi sans limite de durée.

Du fait de la puissance computationelle limitée de nos ordinateurs, il n'a pas été possible de diminuer les valeurs de support et de confidence. Il aurait été intéressant de conserver plus d'attributs pour en apprendre davantage sur la composition d'un logement (contient une baignoire ou douche, type de chauffage, etc.).

Références

[1] Rakesh Agrawal, Tomasz Imieliński et Arun Swami. « Mining association rules between sets of items in large databases ». In : *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. 1993, p. 207-216.

[2] Philippe Fournier-Viger et al. « The SPMF open-source data mining library version 2 ». In : *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, p. 36-40.