Conservatoire National des Arts et Métiers
Certificat de spécialisation analyste de données massives

21 février 2017
Rapport de projet RCP216

TOPIC MODELLING
ASSOCIATIONS RECONNUES D'UTILITÉ PUBLIQUE

Sébastien Gardoll

Ingénierie de la fouille et de la visualisation de données massives
Responsable : Michel Crucianu

Table des matières

1 Associations reconnues d'utilité publique **3**
1.1 Analyse de l'énoncé . . . 3

2 Traitements des données textuelles **3**
2.1 Traitements données textuelles . . . 4
2.2 Mise en forme des données . . . 4
2.3 Segmentation et tris . . . 5

3 Vectorisation **6**
3.1 TF-IDF . . . 6
3.2 ESA . . . 6
3.3 Word2Vec . . . 6
3.4 Mise en œuvre . . . 6
3.5 ACP . . . 7

4 Classification automatique **7**
4.1 Nombre de groupes . . . 8
4.2 Distribution des associations par groupes . . . 8
4.3 Distribution des catégories par groupes . . . 8
4.4 Classification peu pertinente . . . 10

5 Réduction de dimensions **10**
5.1 Paramètres . . . 11
5.2 Mise en œuvre . . . 11

6 Termes pertinents **11**
6.1 Exclusions de termes . . . 11
6.2 Résultats . . . 12

7 Conclusion **13**

A Liste des termes exclus **15**

B Déploiement **15**
B.1 Version des logiciels . . . 15
B.2 Dépendances logiciel . . . 15
B.3 Structure . . . 16

C Bibliographie **17**
C.1 Word2Vec . . . 17

D Résultats **17**

1 Associations reconnues d'utilité publique

Dans le cadre du cours d'ingénierie de la fouille et de la visualisation de données massives, j'ai choisi le sujet numéro 16 portant sur l'analyse des associations reconnues d'utilité publique (lien). Ce rapport présente une étude de ces associations qui sont caractérisées, en autre, par une description et une catégorie. Le premier objectif de l'étude porte sur la classification automatique de ces associations selon leur description. Le deuxième objectif a pour but de trouver les termes, issus des descriptions, pertinents vis à vis de la catégorie des associations. Les détails techniques d'implémentation sont décrits dans l'annexe B.

1.1 Analyse de l'énoncé

L'énoncé du projet :

> « Associations Reconnues d'Utilité Publique : à partir des données ouvertes de https ://www.data.gouv.fr/fr/datasets/associations-reconnues-d-utilite-publique/, appliquez une classification automatique à partir du contenu de l'attribut textuel (colonne Objet). Cherchez les termes issus de ces descriptions textuelles (colonne Objet) qui sont les plus pertinents pour chaque catégorie déclarée (colonne Catégorie). »

L'énoncé demande d'appliquer une classification automatique sur le contenu de la description des associations. Cette description étant textuelle, la première partie de ce rapport va détailler les étapes menant à la vectorisation de cette description. La deuxième partie porte sur la classification automatique par kmeans et présente la distribution des groupes trouvés selon les catégories des associations qui en font partie et la distribution des catégories selon ces groupes. La troisième partie de ce document répond à la question de la pertinence des termes utilisés dans la description des associations vis à vis des catégories d'association. Elle détaille le test du χ^2 appliqué aux données afin de sélectionner uniquement les termes pertinents pour chaque catégorie. La dernière partie de ce rapport propose une conclusion de cette étude.

2 Traitements des données textuelles

Les données du projet se présentent sous la forme d'un fichier au format Excel (daté de juin 2016. Depuis janvier 2017 le format est l'Open Document Spreadsheet). Chaque ligne du tableau représente l'enregistrement d'une association. Les métadonnées intéressantes des enregistrements sont l'identifiant unique de l'association, la description textuelle et la catégorie (une seule) caractérisant l'activité de l'association.

Les données sont donc des chaînes de caractères comportant des caractères alphanumériques, divers symboles (ponctuation, opérateur plus, esperluette, abréviations) ainsi que des caractères de mise en forme (sauts de ligne, tabulations et espaces).

Avant d'appliquer une méthode de classification sur des données textuelles, il convient de les traiter par une série d'opérations. La sous section suivante présente brièvement quelque unes de ces opérations. Les sous sections suivantes détaillent les opérations pertinentes pour le projet.

2.1 Traitements données textuelles

Ces traitements sont regroupés en quatre ensembles d'opérations :

- La collecte des données : les données, proposées par l'état français, sont en libre accès à cette adresse.
- La mise en forme des données : opération développée dans la section éponyme (2.2).
- L'extraction d'entités primaires, l'étiquetage grammatical et la lemmatisation sont abordés dans la section 2.3.
- L'extraction d'entités nommées, la résolution référentielle, l'analyse syntaxique ne présentent pas d'intérêt particulier pour les objectifs demandés. En effet, la classification automatique des associations ou la caractérisation des catégories d'association à partir de leur description ne nécessitent pas la compréhension du texte. Quant à l'extraction d'informations, l'étude du projet n'a pas pour objet la mise en relation avec des modèles de compréhension.

A noter que les descriptions des associations présentent une bonne conformité grammaticale et une syntaxe correcte.

2.2 Mise en forme des données

Afin de simplifier la lecture des enregistrements d'association sous Spark, la stratégie adoptée, a été de transformer le fichier au format Excel, en fichier texte CSV (fonction intégrée à Excel) avec comme séparateur le point-virgule. Cependant, certaines descriptions comportent des points-virgules. Au préalable, ces points-virgules ont donc été transformés en virgule (fonction remplacer d'Excel). A l'aide d'un éditeur de texte, le fichier CSV obtenu a été réencodé en UTF-8 et le caractère de saut de ligne choisi selon la convention d'Unix, afin d'uniformiser le texte. Le reste de la mise en forme des données est effectué sous Spark.

A cause de la présence de saut de ligne dans la description de certaines associations, il n'est pas possible de lire ligne par ligne le fichier CSV. Cependant, la transformation en CSV garantie, pour un enregistrement, un nombre constant de valeurs de métadonnée (même vide), 13, séparées par un point-virgule. Le contenu du fichier est donc lu dans son intégralité et placé en mémoire puis découpé selon le caractère séparateur. Chaque enregistrement d'association est reconstitué en regroupant un ensemble de 13 valeurs de métadonnée. Des 13 valeurs, seules celles de l'identifiant, de la description et de la catégorie sont gardés. Les caractères alphabétique des valeurs de catégorie et d'identifiant sont systématiquement passés en minuscule, tandis que les valeurs de description font l'objet des traitements suivants :

- les caractères alphabétiques passés en minuscule.
- les symboles et abréviations (en très petit nombre) sont explicités dans leur équivalent en langue française ('et' pour '&', 'plus' pour '+' et 'sur' pour 's/').
- les caractères de saut de ligne sont remplacés par un caractère d'espacement.

Les enregistrements ne possédant pas d'identifiant, de description ou de catégorie ainsi que les enregistrements en double, sont éliminés du corpus. Si la description et l'identifiant de l'asso-

ciation reste sous forme de chaînes de caractères, la catégorie est condensée (fonction de hash SHA1) sous forme d'un entier codé sur quatre octets. Cette transformation est nécessaire pour le filtrage des termes par le test du χ^2 qui est développé dans la section 5. Le nombre de catégorie étant très petit (29), le risque de collision est quasiment nul avec un codage sur quatre octets.

A l'issue de cette opération, les enregistrements sont organisés en mémoire sous forme de tuples dans un RDD. Les tuples sont composés de la description et de l'identifiant d'association, deux chaînes de caractères distinctes, et du condensat de la catégorie de l'association.

On compte :

- 1895 enregistrements d'association au total
- 29 catégories
- 16 enregistrements dont la catégorie est manquante (dont un en double)
- 5 enregistrements dont le description est manquante
- 1874 enregistrements montés en mémoire

Cette mise en forme des données assure en premier lieu la montée en mémoire d'informations cohérentes puis la préparation des informations aux traitements présentés dans la section suivante.

2.3 Segmentation et tris

Ces derniers traitements ont pour but de segmenter chaque description d'association en un ensemble de termes susceptibles d'être pertinents. Ces traitements s'appuient en grande partie sur l'application de la même solution utilisée dans le TP d'étiquetage morpho-syntaxique et lemmatisation pour le français. Cette solution est basée sur le Stanford Core NLP, les travaux de Monsieur Ahmet Aker et le projet de Monsieur Fabrice Lebel. Elle effectue la segmentation des descriptions en terme, l'étiquetage morpho-syntaxique des termes et leur lemmatisation. L'étiquetage permet un premier tri : seuls les noms, verbes, adverbes et adjectifs sont retenus. Les autres unités morphologiques comme la ponctuation ou les articles, non pertinents, ne sont pas retenus. La lemmatisation en transformant les termes dans leur forme canonique, supprime les éventuels déclinaisons des termes et permet une comparaison des termes appartenant à des descriptions différentes. Cette dernière propriété est primordiale pour la classification automatique et le test du χ^2 .

Enfin trois derniers tris sont effectués sur les termes :

- Les termes formés d'un seul caractère sont éliminés de l'étude. En effet, les articles apostrophés (exemple : l') passent le tri sur l'étiquette morpho-syntaxique.
- Les enregistrements d'association dont la description ne comporte qu'un seul terme, sont éliminés de l'étude afin d'éviter des points isolés dans l'espace des descriptions vectorisées.
- Certains termes que j'ai jugés non pertinents (développement dans la section 6.1) sont exclus du vocabulaire de termes de l'étude.

A l'issue de la segmentation et des tris, le vocabulaire de termes de l'étude compte 3812 termes (3907 sans l'exclusion de termes). Deux enregistrements ont été exclus (total porté à 1872).

3 Vectorisation

La vectorisation des termes des descriptions d'association offre la possibilité d'utiliser les méthodes de classification automatique implémentées dans Spark. Il existe différentes techniques de vectorisation plus ou moins en adéquation avec l'énoncé du projet. En effet, le deuxième objectif de l'énoncé précise bien que les termes pertinents vis à vis des catégories, sont issus de la description. Autrement dit, la technique de vectorisation doit impérativement être réversible afin de retrouver les termes.

Voici une brève revue des techniques et leur adéquation avec ce projet :

3.1 TF-IDF

TF-IDF est une technique classique de vectorisation basée sur la fréquence d'utilisation des termes. Chaque description est un vecteur dont les composantes sont les termes du vocabulaire et les valeurs de ses composantes sont leur poids TF-IDF. Cette technique, en gardant en mémoire une table de correspondance entre indices de vecteur et termes, permet l'opération inverse nécessaire au deuxième objectif.

3.2 ESA

L'analyse sémantique explicite (ESA) représente chaque terme par leur pondération (par exemple TF-IDF) calculée à partir de leur présence dans les concepts issus d'un très grand corpus (par exemple Wikipedia). Cette méthode n'est pas réversible car les composantes du vecteur obtenu ne sont pas les termes mais les concepts du corpus choisi.

3.3 Word2Vec

Cette méthode est basée sur la représentation des termes par leur contexte au sein du texte. Comme pour ESA, l'opération inverse n'est pas réalisable car les composantes du vecteur obtenu par cette méthode ne sont pas les termes mais les contextes des termes.

3.4 Mise en œuvre

La technique TF-IDF apparaît comme la plus pertinente face aux objectifs de ce projet. Elle est mise en œuvre grâce à la réutilisation de la pile du projet Cloud9 (la méthode termDocumentMatrix) vue dans le TP de fouille de données textuelles. Cette pile a l'avantage de renvoyer la table de correspondante indices - termes. A noter que la bibliothèque de Spark propose également une pile de calcul de TF-IDF (HashTF et IDF) mais son désavantage est d'obliger un espace vectoriel de dimension bien plus grand que le nombre de termes de cette étude. En effet, elle calcule l'indice des termes en appliquant une fonction de hash sur ces derniers. Or le vecteur obtenu dépasse les limites techniques de Spark (maximum 65535 dimensions) pour l'Analyse en Composantes Principales (ACP) qui est abordée en section 3.5.

La réduction de dimensions par analyse sémantique latente (LSA), n'a pas été expérimentée par manque de temps. En effet, cette réduction rend l'opération inverse plus complexe à cause de la gestion de la correspondance entre concepts latents et termes. Cette technique aurait présentée l'avantage de réduire le « bruit » à l'aide des concepts latents.

3.5 ACP

La projection des descriptions vectorisées sur les deux premiers axes factoriels donne de très bons renseignements sur la distribution des descriptions vectorisées.

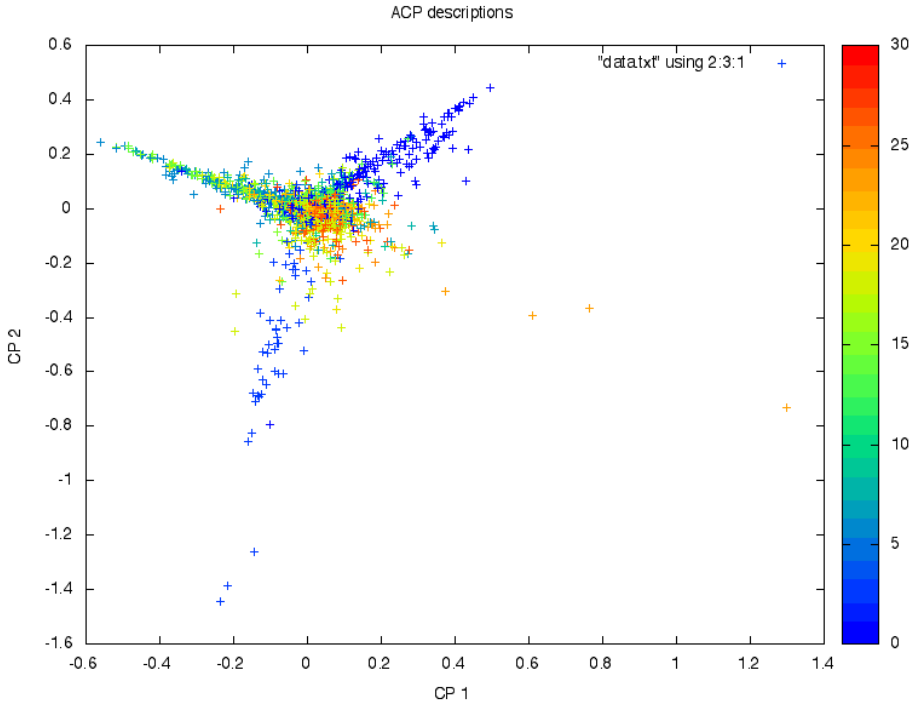


Figure 1 : projection sur deux axes

La figure 1 présente un nuage de points dont la couleur indique la catégorie de l'association. Les points sont très ramassés autour de l'origine. En effet, les valeurs TF-IDF sont faibles. La projection indique, comme il sera démontré dans la section 4, que la classification automatique ne peut pas donner de groupes bien délimités.

4 Classification automatique

Le premier objectif de cette étude consiste à classifier les associations selon leur description.

A l'issue de l'étape précédente, chaque enregistrement d'association est sous la forme d'un tuple composé du condensat de la catégorie de l'association et du vecteur modélisant les termes de la description de l'association (vecteur description).

La mise œuvre de la classification est simple car Spark implémente l'algorithme kmeans. Celui-ci prend en paramètre, k le nombre de groupes à trouver, le RDD de vecteurs descriptions et le nombre maximum d'itérations sans convergence.

Si le nombre maximum d'itérations sans atteindre la convergence est arbitrairement choisi à 100 et ne semble pas être atteint, le choix du nombre de groupes est généralement sensible.

4.1 Nombre de groupes

En première approche, on peut attendre qu'il y ait autant de groupes que de catégories. Cependant, la plupart des catégories ont des thèmes en communs. Par exemple le thème (et le mot) de la science est commun à plusieurs catégories comme la « recherche », la « culture », la « culture et science », « éducation et formation », etc. La figure 2 montre l'évolution du kmeans cost (somme des distances entre les points et le centre de groupe le plus proche).

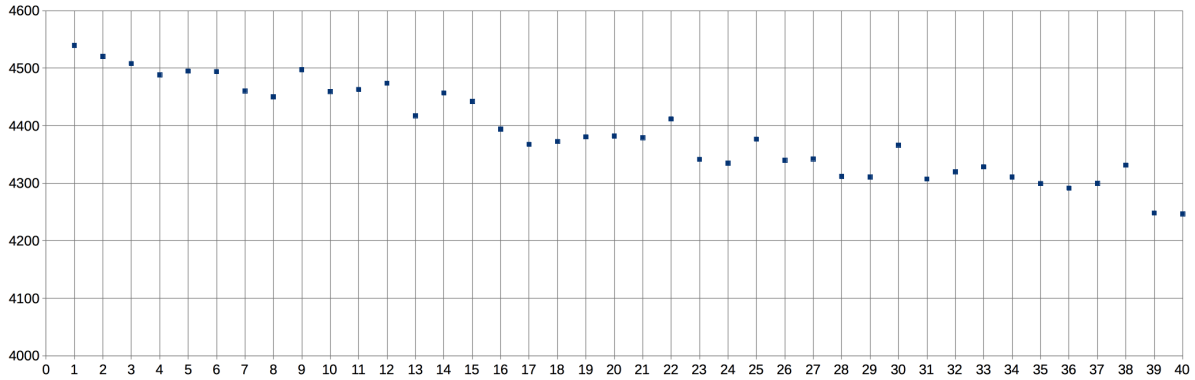


Figure 2 : évolution du kmeans cost selon le nombre de groupes

Cette figure montre une tendance décroissante à mesure que le nombre de groupes augmente. Lorsque le nombre de groupe à rechercher est égal au nombre de points, le coût est nul. Il n'y a donc pas optimum avec un nombre de groupes moins important que le nombre de catégories.

Par la suite, le nombre de groupes est fixé au même nombre de catégories (29).

4.2 Distribution des associations par groupes

L'algorithme du kmeans attribue à chaque vecteur description un unique groupe. La figure 3 donne le nombre d'associations par groupe. Sur les 29 groupes, un seul réunit la quasi totalité des descriptions. Les déductions de la projection ACP (figure 1) sont confirmées, la classification automatique n'est pas du tout satisfaisante.

4.3 Distribution des catégories par groupes

Comme les vecteurs descriptions sont rattachés à une catégorie, il est possible de montrer le nombre de groupes dans lesquels les descriptions de même catégorie se répartissent.

Pour la majorité des catégories, le nombre de groupes ne dépassent pas deux. Pour les catégories pour lesquelles les descriptions sont affectées à un unique groupe, il s'agit du plus grand groupe trouvé sur la figure 3.

La répartition sur plusieurs groupes suggère des associations de même catégorie aux activités relativement différentes les unes des autres. La catégorie « culture et sciences » illustrent bien

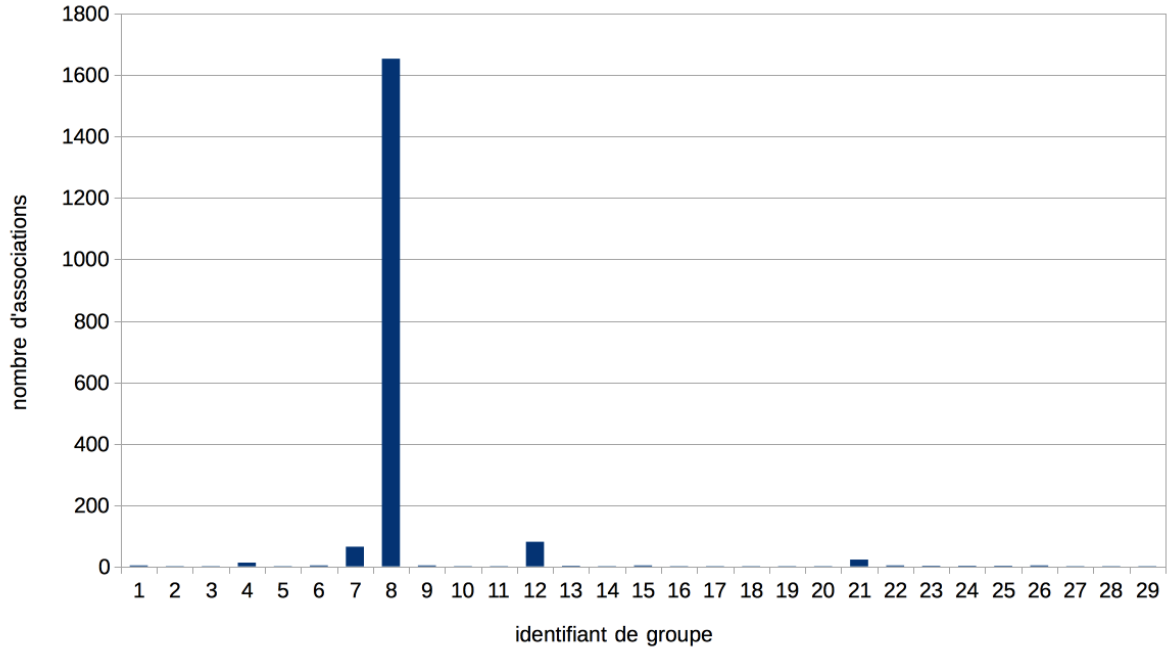


Figure 3 : nombre d'enregistrements d'association par groupe

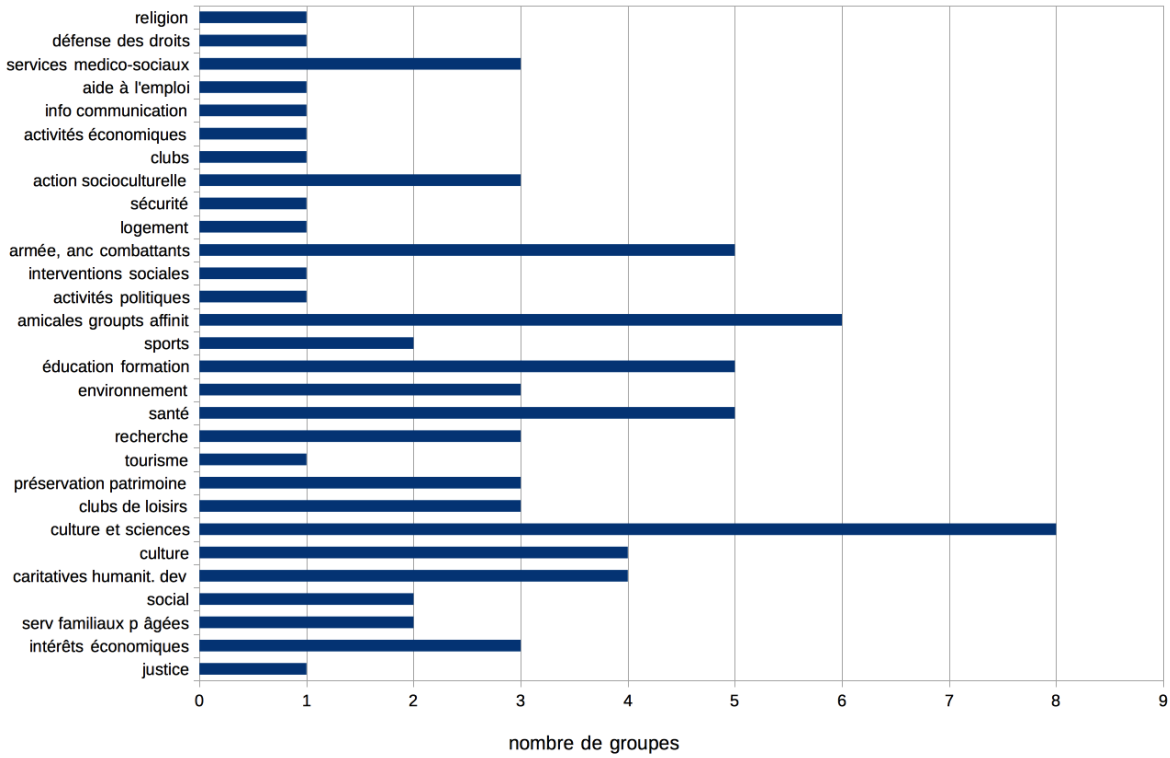


Figure 4 : répartition des catégories par groupe

cette hypothèse. En effet, le champ lexical de cette catégorie est très grand.

4.4 Classification peu pertinente

Comme le montre les précédents résultats, la classification automatique des associations selon leur description ne donne pas de résultats pertinents. En effet, même si les descriptions d'association contiennent des termes différents, voir spécifiques à elles seules, les descriptions vectorisées forment pratiquement qu'un seul groupe car leurs composantes ont des valeurs TF-IDF faibles et surtout proches. Les descriptions vectorisées s'amassent autour de zéro, comme le montre la projection ACP (1).

Les poids TF-IDF faibles s'expliquent généralement par l'emploi de termes communs à tous les documents mais pour cette étude, les termes propres aux descriptions (nom de localités, termes techniques, etc.) sont pénalisées par la taille courte des descriptions. Si l'on reprend la formule TF-IDF pour un terme donné, on a :

$$TF \cdot IDF = \frac{n}{N} \cdot \log\left(\frac{M}{m}\right)$$

avec n la fréquence du terme dans la description considérée, N le nombre de termes dans la description, M le nombre de descriptions, m le nombre de descriptions où apparaissent le terme considéré.

Si l'on s'intéresse qu'aux termes propres aux descriptions, on a m égal à 1. En prenant connaissance des descriptions, on observe qu'en généralement, n est égal à 1 car les descriptions étant très courtes, les termes propres ne sont pas répétés. On peut déjà en déduire que TF est faible et qu'IDF est le même pour tous les termes propres.

En prenant N égale à 15 en moyenne et M égal à 1872, on trouve un TF-IDF égal à 0,218. Cette valeur est proche de celles des termes propres (roche et posay) trouvées pour la catégorie social dont il n'existe qu'une seule description (voir 9).

Cette méthode de vectorisation n'est certainement pas adaptée pour de courts documents. Quelques publications (voir annexe C.1) suggèrent que la technique Word2Vec est mieux indiquée dans ce cas, cependant je n'ai pas eu le temps de l'expérimenter.

5 Réduction de dimensions

Pour rappel, le deuxième objectif est de trouver les termes, issus des descriptions, pertinents vis à vis des catégories d'association.

Nous avons vu en sous section 4.3 que les catégories partagent des termes en communs.

Il s'agit donc d'éliminer ces termes en communs qui apportent peu d'information spécifique sur la catégorie de l'association. Spark implémente justement une technique de réduction de dimension supervisée basée sur le test du χ^2 (ChiSelector).

5.1 Paramètres

ChiSelector admet deux paramètres, les vecteurs labellisés (par une catégorie) à réduire et un paramètre dont la nature dépend de la modalité du choix des dimensions à conserver :

- Le nombre de dimensions : le paramètre précise le nombre de dimensions à garder
- Le pourcentage de dimensions à garder : il s'agit d'une variante du premier
- Le risque α : le risque de rejeter à tort l'hypothèse nulle du test χ^2

La paramètre risque α est uniquement disponible dans la version 2.1.0 de Spark.

Si les deux premières modalités obligent l'utilisateur à fixer un nombre arbitraire de dimensions, le risque α présente l'avantage de sélectionner de façon plus objective les dimensions à conserver. Traditionnellement, le risque α est pris à 1 ou 5%. Cette étude utilise la modalité de sélection au risque α pris à 5%.

5.2 Mise en œuvre

Bien que la version 2.1.0 de ChiSelector prend le risque α en paramètre, j'ai décidé d'implémenter un algorithme inspiré de celui-ci mais compatible avec les versions antérieurs, considérant que l'usage de cette seule fonctionnalité ne justifiait pas la mise à jour de Spark.

La sélection de dimension par le test du χ^2 réduit le nombre de composantes des vecteurs descriptions et modifie donc les valeurs d'indice de ses composantes. La table correspondance entre les indices des vecteurs et des termes est alors mise à jour en recalculant les indices impactés par les composantes éliminées.

Les données en mémoire ne posent pas problème particulier. Sous forme de tuples liant la catégorie au vecteur description, une simple opération map permet de les transformer en vecteurs labellisés.

6 Termes pertinents

L'ensemble des termes pertinents pour chaque catégorie est simplement calculé à partir du centre de gravité de l'ensemble de vecteurs descriptions réduits et regroupés selon leur catégorie. Compte tenu de la relative différence d'activités (et donc de description) de certaines associations de même catégorie (par exemple « culture et science »), il m'a semblé que le centre de gravité est un bon représentant de l'ensemble des descriptions pour une catégorie donnée.

6.1 Exclusions de termes

Les premiers ensembles de termes contenaient quelques mots (par exemple : tous) dont la signification n'est pas discriminante vis à vis des catégories. En effet, la réduction de dimensions par le test du χ^2 reste un filtrage statistique et non sémantique. En constituant une liste de termes à exclure (voir annexe A) après quelques itérations à analyser les résultats, complétée par la

concaténation de plusieurs listes de stop words (lien ici et là), les ensembles sont débarrassés d'un maximum de termes non pertinents.

6.2 Résultats

Pour la plupart des ensembles, les dix premiers termes ayant le plus grand score TF-IDF font partie du champ lexical de la catégorie qu'ils caractérisent. Cependant, quelques ensembles présentent des termes très spécifiques à une seule description. Les figures suivantes présentent quelques graphes composés des dix premiers termes ayant le plus grand score TF-IDF. Les autres graphes sont disponibles en annexe D

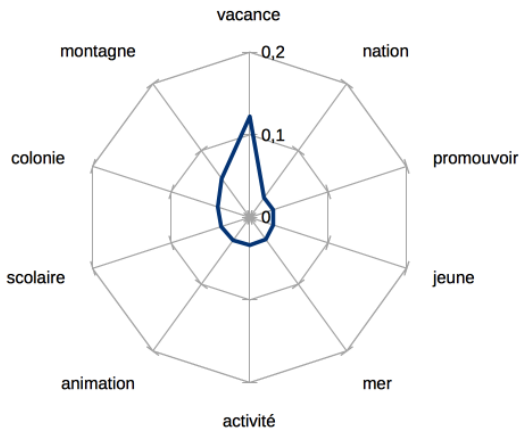


Figure 5 : termes pour la catégorie action socioculturelle

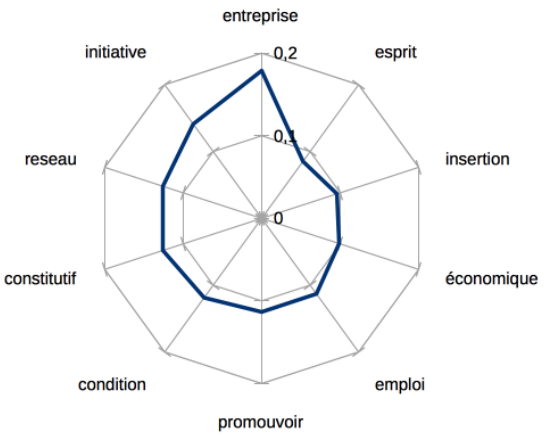


Figure 6 : termes pour la catégorie aide à l'emploi

Les figures 5, 6 et 7 montrent des ensembles de termes qui caractérisent plutôt bien leur catégorie.

La figure 8 présente les termes pour la catégorie activités économiques. Elle illustre le cas où les descriptions d'une catégorie emploient en majorité des termes relativement bien distribués dans les descriptions des autres catégories. L'ensemble de termes de cette catégorie se retrouve dépouillé de ses termes sémantiquement significatifs mais statistiquement non pertinents. Il ne reste donc plus que des termes très spécifiques, présents généralement que dans une seule description, des termes peu représentatifs de la catégorie (olm et onm). Ces termes arrivent à se retrouver parmi les dix premiers car les descriptions sont courtes (voir explications à la sous section 4.4).

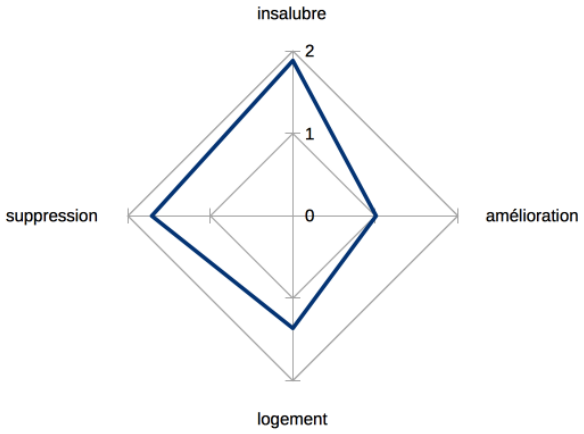


Figure 7 : termes pour la catégorie logement

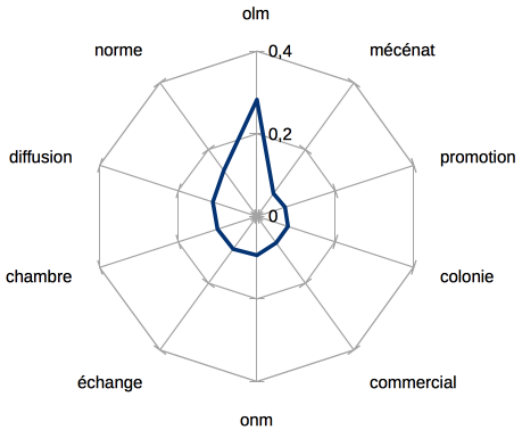


Figure 8 : termes pour la catégorie activités économiques

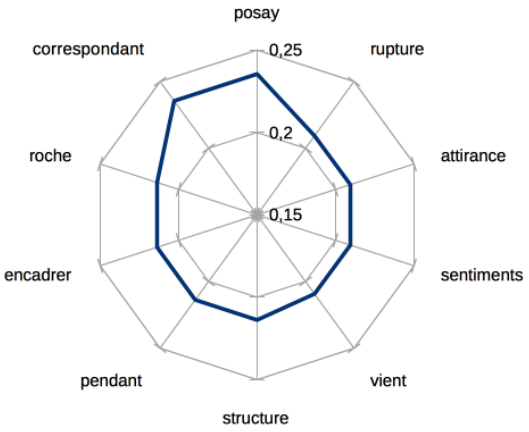


Figure 9 : termes pour la catégorie social

Comme il n'y a qu'une seule association rangée sous la catégorie social, le graphe de cette catégorie (figure 9) montre également le même effet décrit pour la catégorie activité économiques, en plus exacerbé.

7 Conclusion

Cette étude comporte deux objectifs : la classification automatique sur les descriptions d'association et la recherche des termes pertinents par catégorie. La classification automatique n'a pas donné de résultats satisfaisants. En effet, la méthode de vectorisation TF-IDF, utilisée pour les deux objectifs, n'est pas adaptée aux courts documents comme ces descriptions. La méthode de vectorisation Word2Vec plus indiquée dans ce cas de figure, aurait dû être expérimentée mais séparément du deuxième objectif. Le deuxième objectif, implémenté à l'aide du test du χ^2 , a donné de bon résultats mis à part quelques singularités. Ces singularités, des termes propres uniquement à une description, s'expliquent également par le choix de la méthode TF-IDF et la taille très courte des descriptions.

ANNEXES

A Liste des termes exclus

Ces termes, trouvés dans les premiers résultats de l'étude ont été exclus car leur signification n'est pas particulièrement en rapport avec les catégories d'association. Par soucis d'exhaustivité, cette liste est complétée par les termes de deux listes de stop words téléchargeables aux adresses suivantes : http ://www.ranks.nl/stopwords/french et

http ://snowball.tartarus.org/algorithms/french/stop.txt

Termes exclus :

- plus
- ont
- outre
- ayant
- tous
- tout
- toutes
- toute
- grâce

Au final, le nombre de termes exclus s'élève à 543.

B Déploiement

B.1 Version des logiciels

Ce projet a été exécuté avec :

- un MacBook pro sous Mac OSX Mavricks, processeur 2,3 GHz Intel Core i7 et 8 Go RAM
- Spark 2.0.1 mais normalement le code spark devrait être portable sur la version 1.6.x (les data frames ne sont pas utilisées).
- Scala 2.11.8
- Java 8
- Bash 3.2.53 pour les scripts shell

B.2 Dépendances logiciel

Ce projet dépend des bibliothèques suivantes :

- package LSA de Cloud9 empaqueté pour la version 2.0.1 de Spark, source disponibles sur la page du TP de fouille de données textuelle :

  http ://cedric.cnam.fr/vertigo/Cours/RCP216/tpFouilleTexte.html

- le package myopennlp.jar disponible sur le site du TP d'étiquetage morpho-syntaxique et lemmatisation pour le français :

  http ://cedric.cnam.fr/vertigo/Cours/RCP216/tpLemmatisationFr.html

B.3 Structure

- DataAnalyzer.scala est le fichier de code principal. Les arguments en ligne de commande sont listés en exécutant avec l'argument suivant : –help
- arup1606_pretraite.csv est le fichier de données après les traitements manuels décrits à la section 2.2. Le fichier original arup1606_original.xlsx est donné à titre indicatif.
- cluster_optim.sh est le script de recherche d'optimisation de nombre de groupes pour kmeans.
- data_analyzer.sbt est le fichier de compilation de DataAnalyzer.scala
- empty_excluded_lemmas_list.txt est la liste de termes exclus, utilisée par défaut. Elle ne contient pas de terme.
- excluded_lemmas_list.txt contient seulement les termes exclus trouvés dans les premiers résultats.
- more_excluded_lemmas_list.txt contient les termes de excluded_lemmas_list.txt plus les termes des listes de stop words (voir annexe A)
- generate_pca_plots.sh est un fichier de script pour générer dans un répertoire nommé plots les graphiques d'ACP
- pca_gnuplot_cmds.gp est le fichier de commande gnuplot pour dessiner les graphiques d'ACP
- settings.sh est un fichier de configuration pour les scripts shell.
- les fichiers ods sont les fichiers au format Libre Office contenant les résultats numériques et graphiques.

A noter que par défaut, DataAnalyzer ne fait que lire le fichier arup1606_pretraite.csv, localisé dans le même répertoire que lui, vectoriser et enregistrer les données vectorisées dans un répertoire nommé data. Par défaut, aucun terme n'est exclus (liste empty_excluded_lemmas_list.txt).

exemple de commande d'exécution :

```
spark-submit --driver-memory 8000m --class "DataAnalyzer" \
--jars "lib/myopennlp.jar,lib/lsa-1.0.0-jar-with-dependencies.jar" \
--master local[4] target/scala-2.11/data-analyzer_2.11-1.0.jar \
--resume-opt 1 --significance-lvl 0.05
```

C Bibliographie

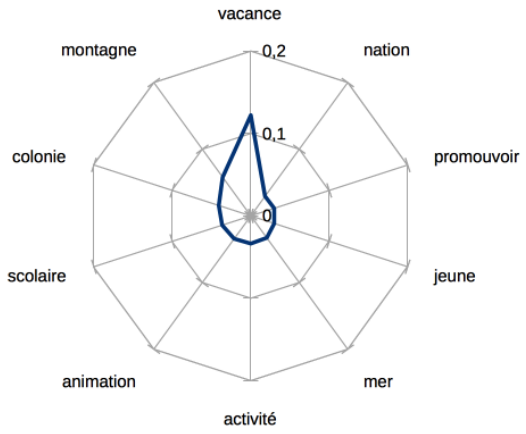
C.1 Word2Vec

”Short Text Similarity with Word Embeddings”, Kenter et de Rijke. https ://staff.fnwi.uva.nl/m.derijke/wp-content/papercite-data/pdf/kenter-short-2015.pdf

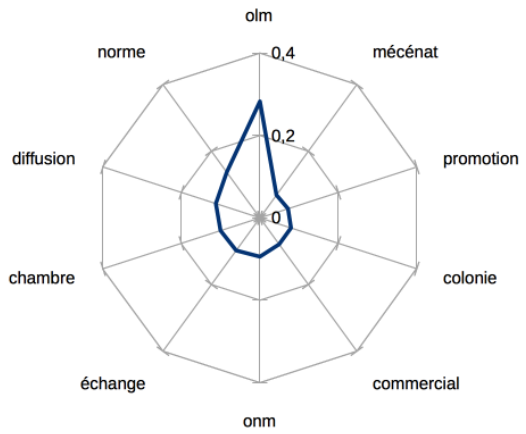
”Short text categorization by smoothing word distribution”, Fukumoto et Suzuki. http ://ltc.amu.edu.pl/book/pap 2.pdf

D Résultats

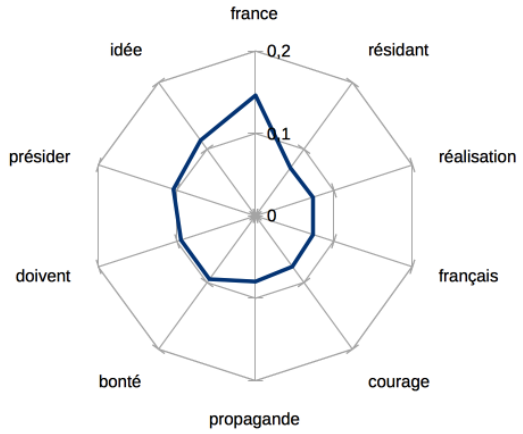
Graphiques des dix termes les plus pertinents pour chaque catégorie. Les valeurs numériques données sont les poids TF-IDF des termes.



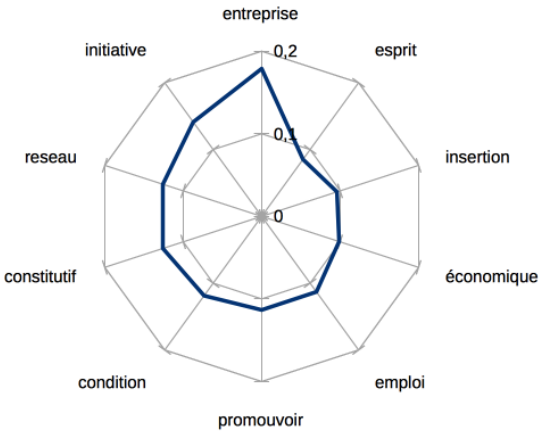
termes pour la catégorie action socio-culturelle



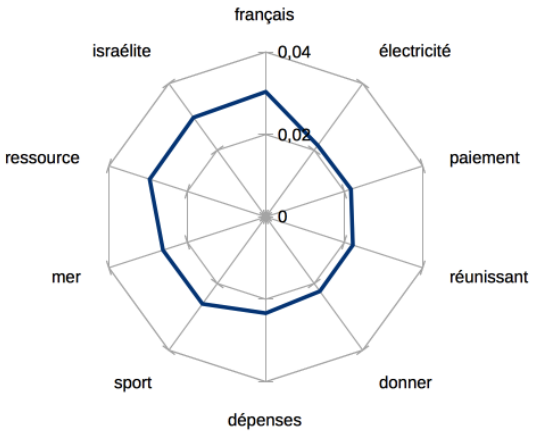
termes pour la catégorie activités économiques



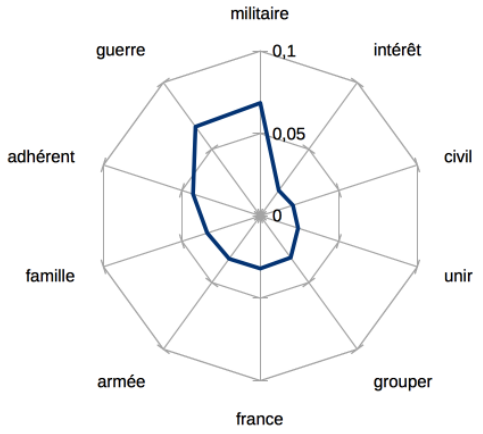
termes pour la catégorie activités politiques



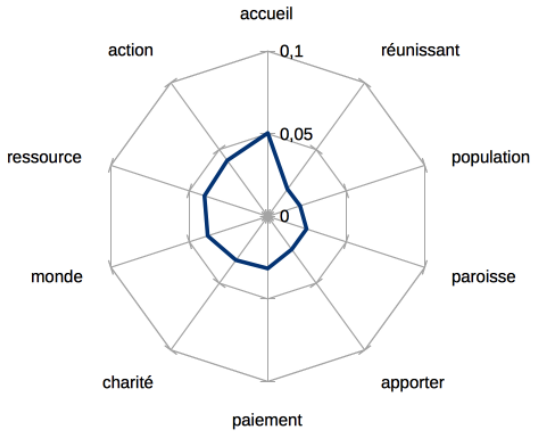
termes pour la catégorie aide à l'emploi



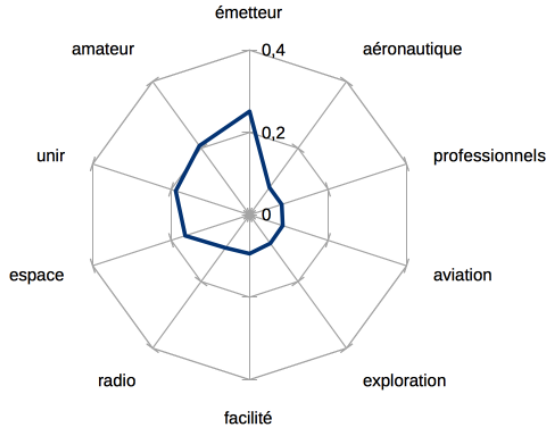
termes pour la catégorie amicales groupts affinit



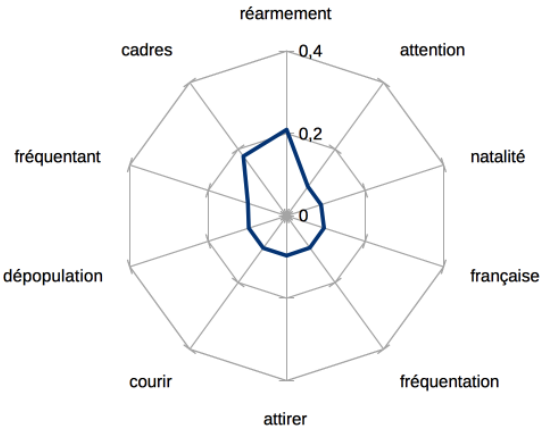
termes pour la catégorie armée anc combattants



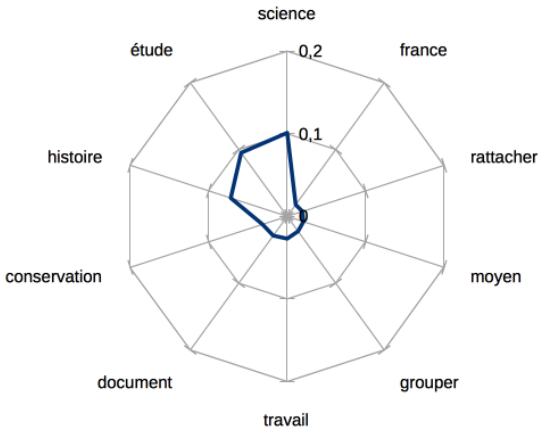
termes pour la catégorie caritatives humanit dev



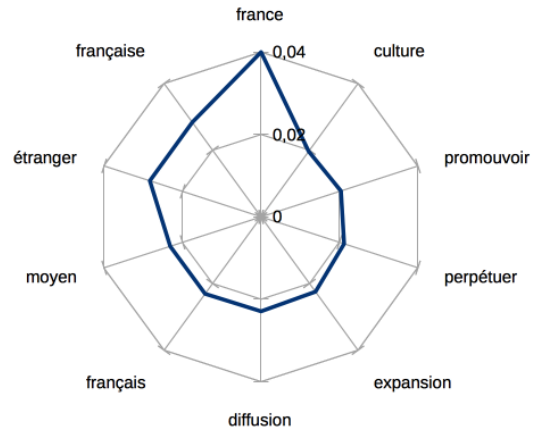
termes pour la catégorie clubs et loisir



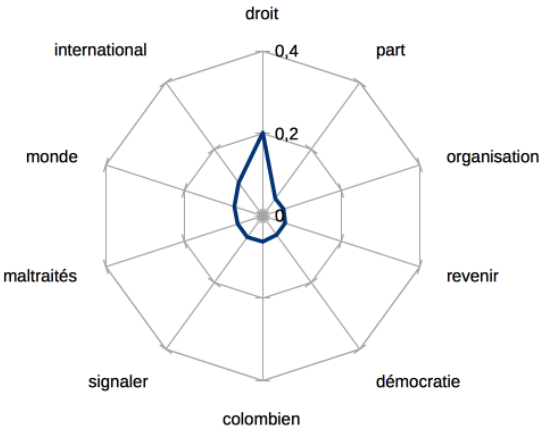
termes pour la catégorie clubs



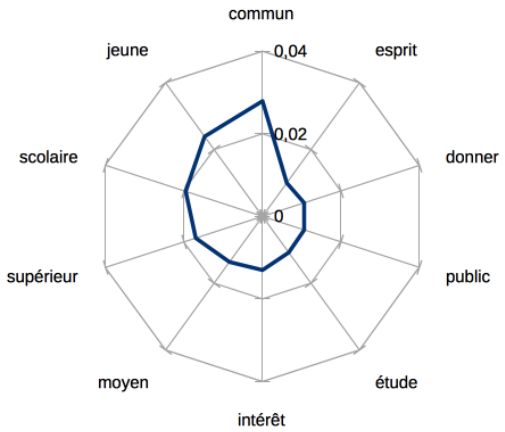
termes pour la catégorie culture et sciences



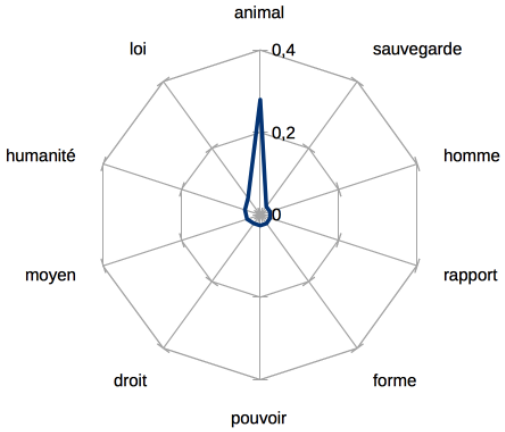
termes pour la catégorie culture



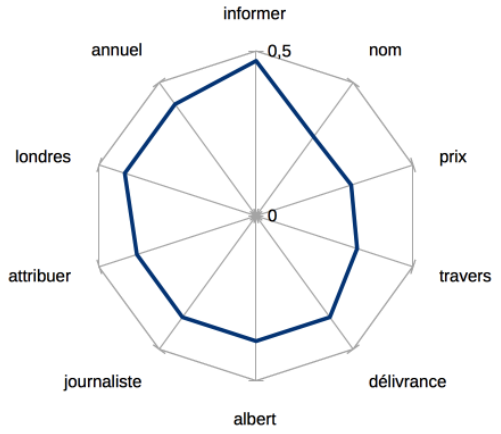
termes pour la catégorie défanse des droits



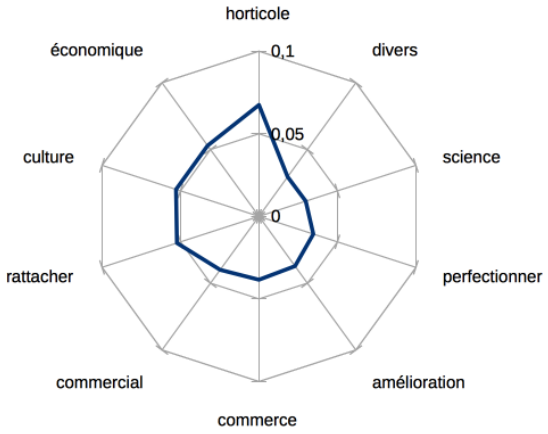
termes pour la catégorie éducation formation



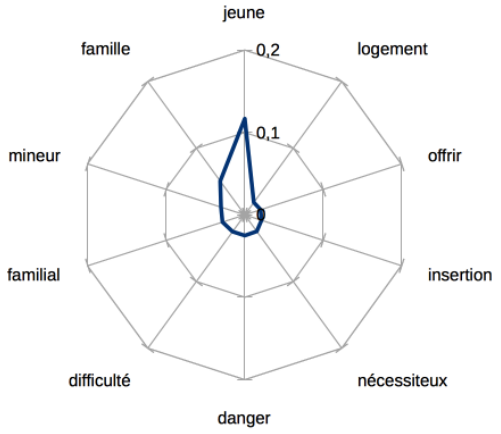
termes pour la catégorie environnement



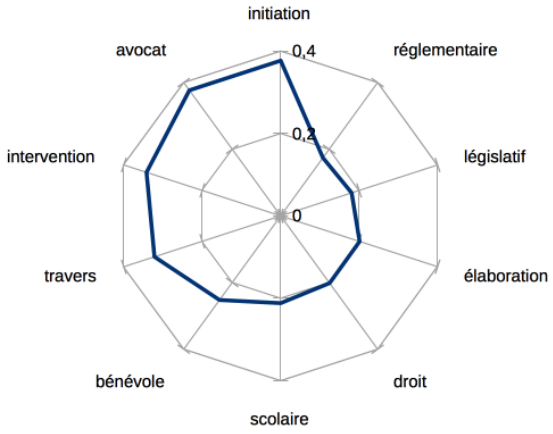
termes pour la catégorie info communication



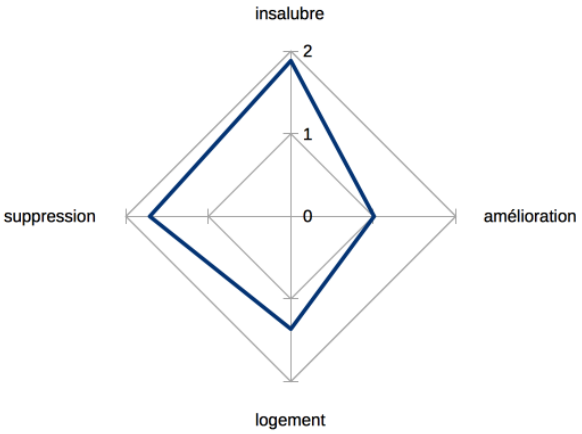
termes pour la catégorie intérêts économiques



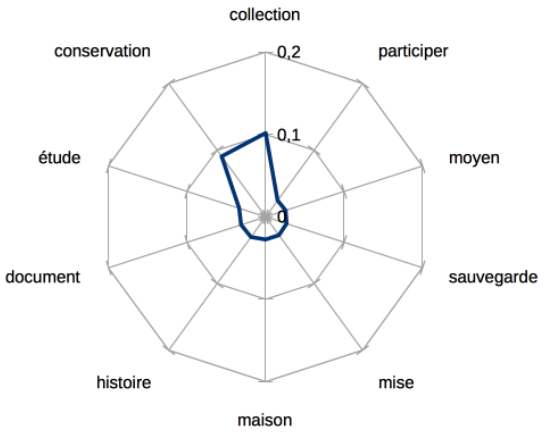
termes pour la catégorie interventions sociales



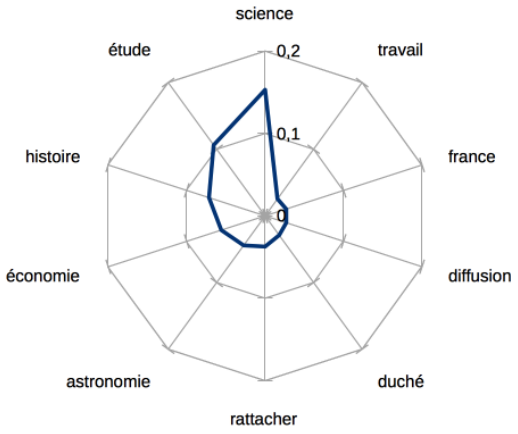
termes pour la catégorie justice



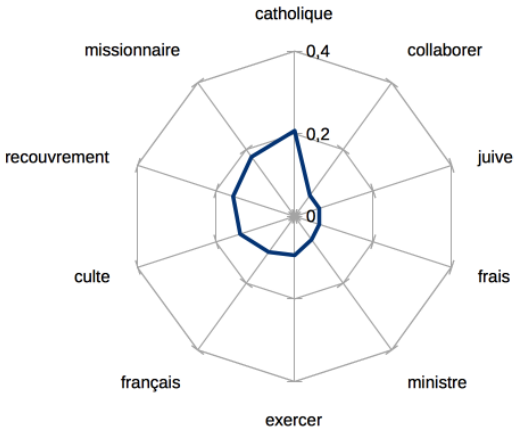
termes pour la catégorie logement



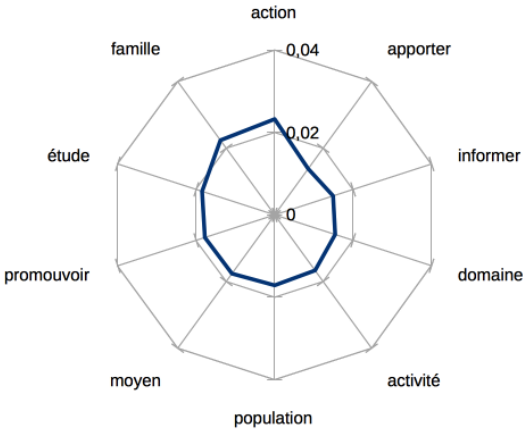
termes pour la catégorie préservation patrimoine



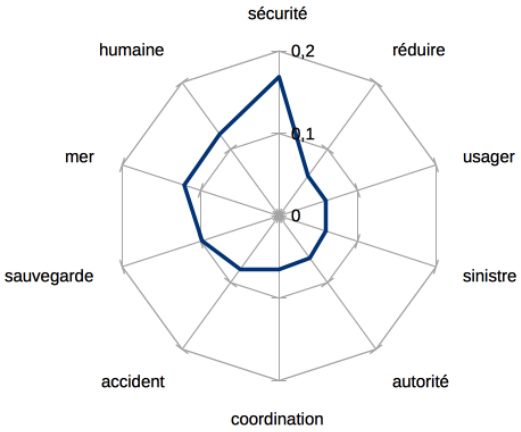
termes pour la catégorie recherche



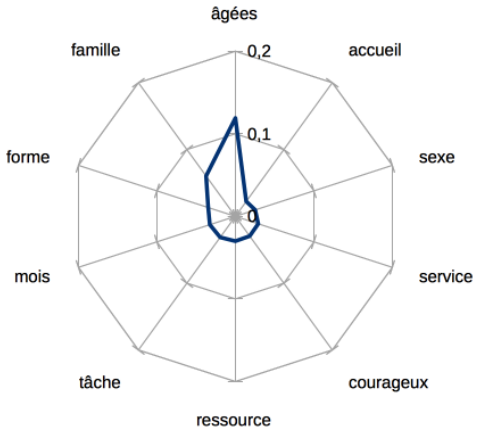
termes pour la catégorie religion



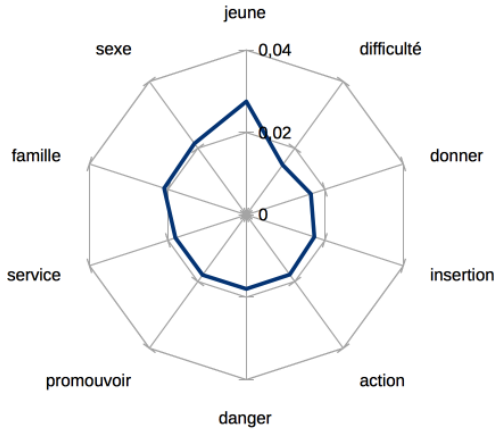
termes pour la catégorie santé



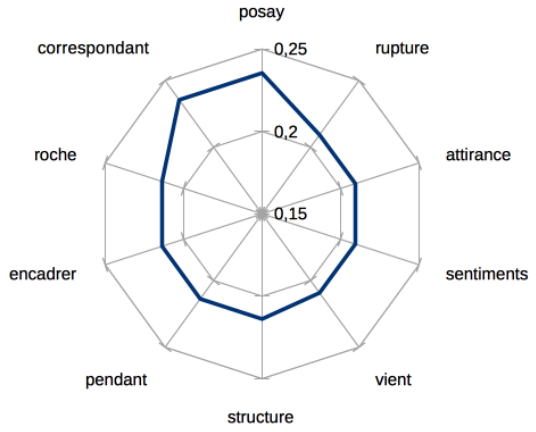
termes pour la catégorie sécurité



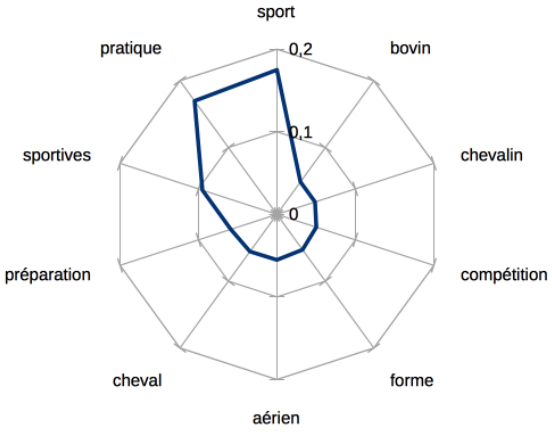
termes pour la catégorie services familiaux personnes âgées



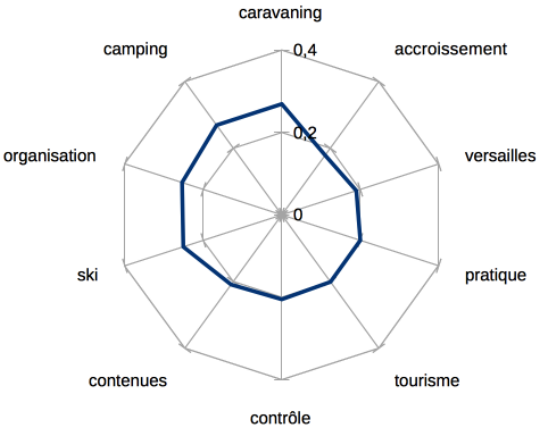
termes pour la catégorie services médico-scociaux



termes pour la catégorie social



termes pour la catégorie sports



termes pour la catégorie tourisme

Table des figures

1 projection sur deux axes 7
2 évolution du kmeans cost selon le nombre de groupes 8
3 nombre d'enregistrements d'association par groupe 9
4 répartition des catégories par groupe 9
5 termes pour la catégorie action socioculturelle 12
6 termes pour la catégorie aide à l'emploi 12
7 termes pour la catégorie logement 13
8 termes pour la catégorie activités économiques 13
9 termes pour la catégorie social 14