Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE LA TRANSITION
ÉCOLOGIQUE ET SOLIDAIRE

MINISTÈRE
DE LA COHÉSION
DES TERRITOIRES

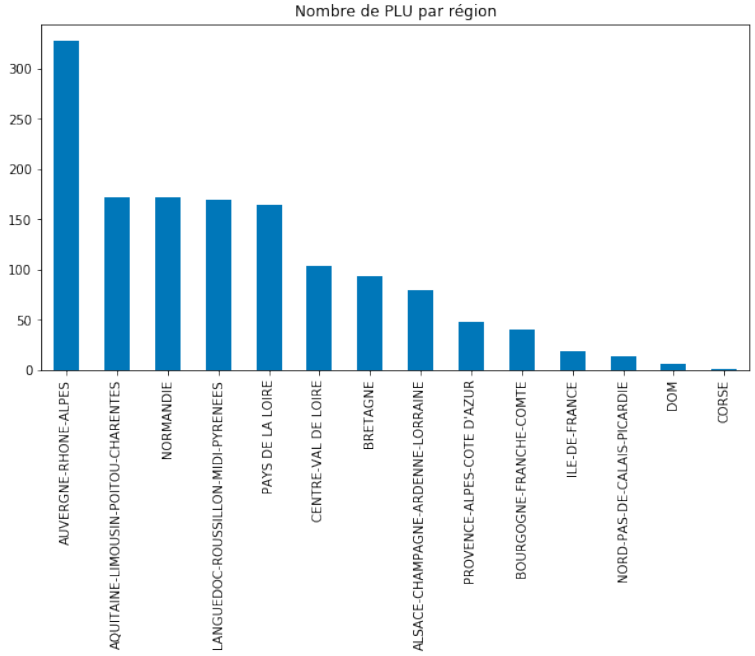
Projet SmartPLU

Rapport technique – Version du 25 juin 2018
Numen Digital

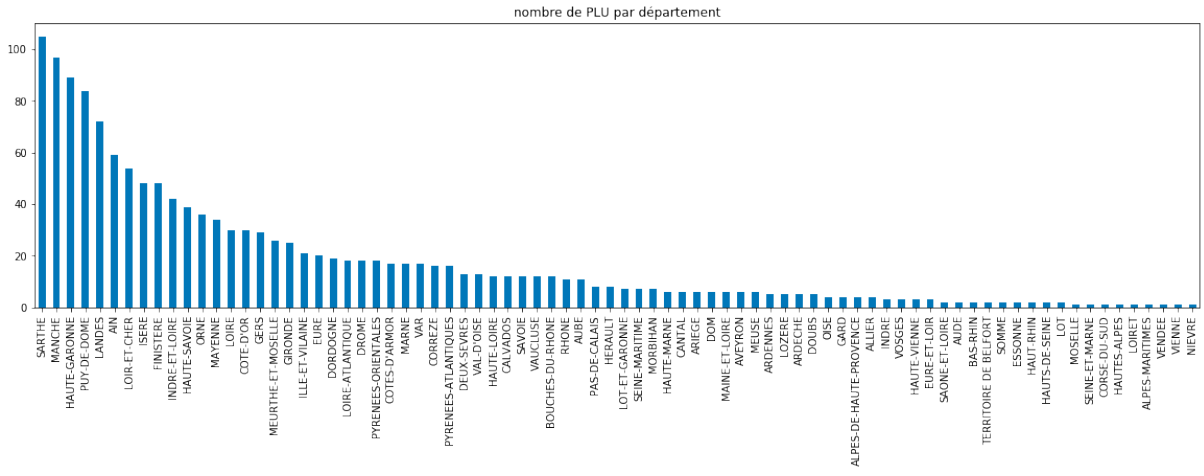
1. Composition du corpus PLU

Le corpus des PLU traités à ce jour et dans cette version du rapport technique est composé de 2493 PLU collectés sur GeoPortail ainsi que ceux ayant été utilisés comme source pour la création de la base CartoPLU+. Un robot collecte actuellement 70 PLU supplémentaires par semaine. Ces PLU seront intégrés au corpus pour la livraison finale.

La répartition des PLU par région est la suivante :



La répartition des PLU par département est la suivante :



1. Pré-traitement des PLU

a. Conversion en XML

Les PLU collectés sont au format PDF et doivent être transformés dans un format structuré pour être traités automatiquement. Nous avons utilisé ABBYY FineReader pour convertir les fichiers PDF en fichier de type XML contenant à la fois le contenu et la mise en page des documents.

Les PDF peuvent être de 2 types différents :

- Les PDF-Texte, pour lesquels le contenu textuel est disponible à l'intérieur du format PDF. Ces PDF sont généralement le résultat d'une conversion d'un document électronique de type Microsoft Word.
- Les PDF-Image, pour lesquels ne contenu textuel n'est pas disponible : seule une image du document est disponible dans le format PDF. Ces PDF sont généralement le résultat numérisation du document papier (scanner).

Pour le deuxième type de PDF, il est nécessaire de convertir l'image en texte à l'aide d'un logiciel d'OCR. ABBYY FineReader permet de faire cette conversion. Cependant, un taux d'erreur d'OCR assez important peut être observé lors de la conversion. Pour garantir la qualité des traitements linguistiques qui devaient être réalisés sur le texte des PLU, nous avons décidé d'écarter les PLU présentant un taux d'erreur OCR trop important. Pour ce faire, nous avons estimé le nombre de mots « suspects », c'est-à-dire les mots susceptibles de contenir des erreurs OCR, cette information étant produite par ABBYY FineReader. A partir de la distribution des taux de mots suspects par rapport au type de PDF (image ou texte), nous avons déterminé un seuil de rejet en dessous duquel les documents sont considérés comme trop bruités. Ce seuil a été fixé à 5% de mots suspects, ce qui permet de conserver 70% des documents du corpus initial pour les traitements linguistiques.

b. Structuration en article

L'extraction des données pour la modélisation des règles nécessite de connaitre la structure des PLU en zones et en articles. Cette structure peut se déduire de l'identification des titres et de leur contenu. Afin de structurer les PLU en articles nous avons identifié les titres à base de règles simples (expressions régulières). Ces règles ont permis d'identifier les titres correspondant à des articles et d'extraire la zone correspondant ainsi que le numéro d'article. La structuration en article a été ajoutée à la structure XML issue de la conversion en XML des PDF.

Livrable 1

L'ensemble de ces PLU convertis du format PDF en un format XML structuré dans lesquels les titres ont été identifiés constitue un livrable.

2. Analyse lexicale et sémantique

a. Pré-traitement du texte

La diversité des formats d'origine des documents (PDF-texte ou PDF-image) et le traitement par OCR pour la transformation en XML a pour conséquence la présence de bruit dans le

contenu textuel des PLU. Une phase de pré-traitement est donc nécessaire pour réduire ce niveau de bruit. Cette phase inclut aussi des traitements de normalisation permettant de réduire la diversité des formes de mots. Les pré-traitements réalisés sont les suivants :

- **Extraction du texte** des fichiers XML ;
- **Suppression des caractères spéciaux** et des symboles qui correspondent soit à des erreurs en sortie d'OCR soit à des caractères de mise en forme (puces) ;
- **Suppression des sauts de lignes**, tabulation, espaces multiples ;
- **Suppression de la ponctuation ;**
- **Conversion en minuscule ;**
- **Correction orthographique** : la phase d'OCR introduit 2 types de bruits majoritaires : la suppression des accents et l'insertion d'espaces à l'intérieur des mots. Pour corriger ces erreurs, nous avons appliqué une correction orthographique adaptée à ce type de bruit et basée sur un dictionnaire ;
- **Recomposition des mots composées** : la phase d'OCR peut aussi séparer les mots composés ou supprimer leur tiret. Une particularité du corpus PLU est la présence de nombreux toponymes composés qui ne peuvent être recomposés par la phase de correction orthographique car ils ne sont pas présents dans le dictionnaire. Une procédure spéciale de correction des mots composés a été mise en place ;
- **Etiquetage morphosyntaxique** (*POS tagging*) : cette étape permet de typer chaque mot avec sa catégorie grammaticale dans la phrase.
- **Lemmatisation** : cette étape consiste à remplacer les formes fléchies des mots par leur forme non fléchie : par exemple, remplacer les noms au pluriel par les noms au singulier, remplacer les formes conjuguées des verbes par leur forme à l'infinitif. Cette étape permet de réduire le nombre de formes de mots différentes présent dans le corpus. Deux outils ont été évalués : NLTK[1] et spacy[2]. Après une inspection visuelle, nous avons conclu que les performances de ces outils étaient similaires.

Un exemple de pré-traitement est illustré sur le texte suivant :

```
La hauteur des constructions ne pourra excéder 10 mètres au faîtage
et 7 mètres à l'égout ou à l'acrotère.
La règle précédente ne s'applique pas :
• pour les constructions, installations, ouvrages et équipements,
dits « techniques », liés ou nécessaires au fonctionnement des
services et équipements publics, collectifs ou d'intérêt général
(transformateurs, relais, stations de pompage, de refoulement ou de
traitement d'eaux usées, ...) ;
• pour les équipements collectifs publics ou d'intérêt collectif
nécessitant par leur fonction une hauteur plus importante.
```

Le texte pré-traité est :

```
hauteur construction pouvoir excéder mètre faîtage mètre égout
acrotère
règle précédent appliquer
construction installation ouvrage équipement dire technique lier
nécessaire fonctionnement service équipement public collectif
```

[1] https://www.nltk.org/

[2] https://spacy.io/

```
intérêt général transformateur relais station pompage refoulement
traitement eau user
équipement collectif public intérêt collectif nécessiter fonction
hauteur plus important
```

b. Apprentissage de représentations distribuées de mots (word embedding)

La première méthode d'analyse sémantique appliquée aux textes du corpus de PLU est un apprentissage de représentation distribuée de mots du corpus. Une représentation distribuée des mots est une représentation dans un espace vectoriel de grande dimension (quelques centaines). L'objectif de l'apprentissage est de construire une représentation qui traduit les proximités sémantiques des mots en une proximité dans le nouvel espace. Deux types d'algorithmes ont été testés :

- Word2vec[3] : ce modèle apprend une représentation vectorielle avec comme objectif de prédire la *probabilité* d'apparition d'un mot en fonction des mots du contexte.
- FastText[4] : ce modèle reprend les principes du modèle word2vec mais y ajoute une modélisation des séquences de caractères composant les mots (*n-gram* de caractères). Ce modèle permet de mieux estimer la représentation des mots rares, de modéliser les mots hors vocabulaires (mots non vus lors de la phase d'entraînement) et de prendre en compte la morphologie des mots (suffixe, préfixe, racine).

Le principe d'apprentissage de ces représentations est illustré sur la Figure 1.

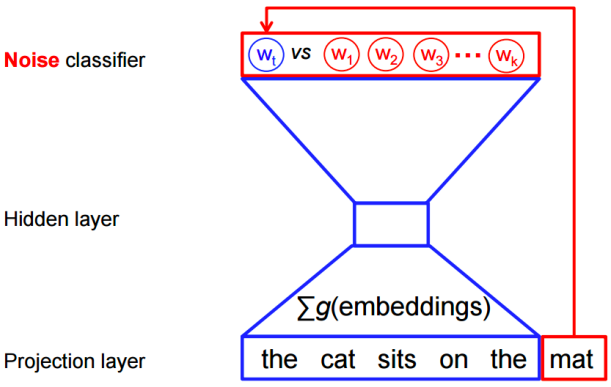


Figure 1: Apprentissage d'une représentation distribuée des mots (word embedding). La représentation de chaque mot est optimisée afin de pouvoir prédire avec une plus forte probabilité le mot correct dans la séquence des mots des phrases du corpus d'apprentissage. Sur le schéma, « mat » le mot correct noté w_t doit être prédit avec une plus forte probabilité que les mots incorrects (noise) notés $w_1 \ldots w_k$.

Ces modèles comportent un certain de nombre de paramètres dont les deux principaux sont :

- La taille de l'espace de représentation : elle est généralement comprise entre 100 et 600 et dépend à la fois de la taille du vocabulaire à modéliser et de la quantité de données disponible pour l'apprentissage. Etant donné la taille modérée du corpus PLU et étant donné que nous n'avons pas de métrique objective pour estimer la qualité des

[3] Efficient Estimation of Word Representations in Vector Space, Mikolov *et al.*, 2013
[4] Enriching Word Vectors with Subword Information, Bojanowski *et al.*, 2016 et https://fasttext.cc/

embedddings, nous avons choisi une taille de valeur 100 après avoir testé plusieurs valeurs et réalisé une inspection visuelle du résultat. Nous avons conclu que la taille de la représentation avait un impact faible sur le résultat en visualisation.

- Le choix du contexte : *cbow* ou *skipgram.* L'apprentissage de la représentation s'effectue de telle manière qu'il est possible de prédire la probabilité d'occurrence d'un mot en fonction de son contexte. Le type de contexte peut être défini de deux manières : le modèle se base soit sur tous les mots du contexte pour prédire un mot donné (modèle *cbow*) soit sur un seul mot pour prédire chaque mot du contexte (modèle *skipgram*). La différence est entre ces deux algorithmes est illustrée sur la
- Figure 2. Après une évaluation des deux méthodes (*cbow* et *skipgram*) nous avons retenu la méthodes cbow car elle est plus rapide à entrainer et les résultats nous semblaient similaires sachant que nous n'avions pas de mesure objective de qualité pour comparer les 2 types de contextes de manière systématique.
- Le nombre minimal d'occurrences d'un mot : ce paramètre permet de sélectionner les mots ayant un nombre minimal d'occurrences, garantissant ainsi une bonne estimation de leur représentation. Cette fréquence minimale est habituellement fixée à 5, mais ce nombre peut être augmenté afin de réduire le nombre de mots lors de l'affichage des représentations dans l'interface graphique.

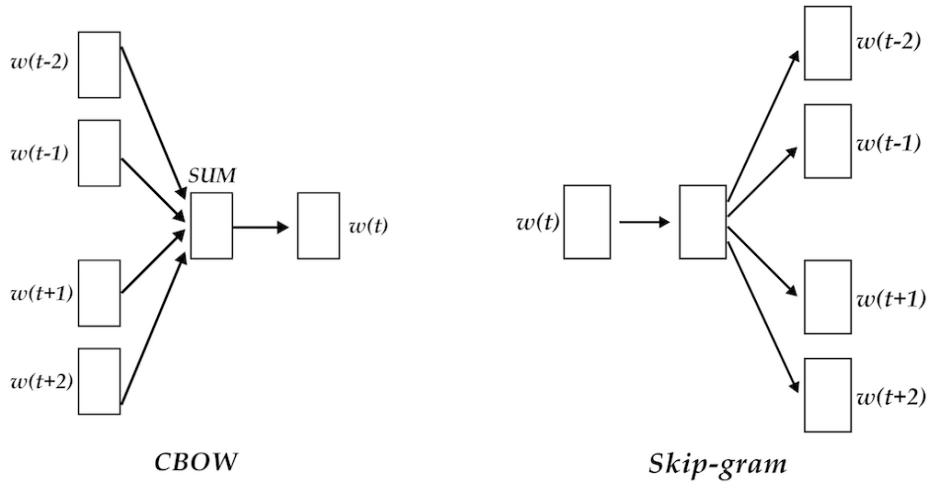


Figure 2 : Illustration des modèles CBOW et Skipgram.

Les représentations distribuées ont été entrainées sur un corpus de 1896 PLU. Les représentations apprises peuvent être visualisées grâce à l'interface *Embedding Projector*[5] de Tensorflow[6], la librairie de *deep learning* de Google. Les vecteurs de mots étant en grande dimension (100), il faut opérer une réduction de dimension à 3 pour pouvoir les visualiser. Plusieurs méthodes de réduction de dimension sont disponibles dans l'interface. Sur la Figure 3, on peut visualiser les mots les plus proches du mot *publicitaire.* On note la proximité avec les mots *enseigne, publicité, totem, néon* qui sont bien dans le champ lexical de la publicité. On

[5] https://projector.tensorflow.org/
[6] https://www.tensorflow.org/

note aussi le terme *enseigner* qui correspond à une lemmatisation sans doute erronée du nom commun enseigne.

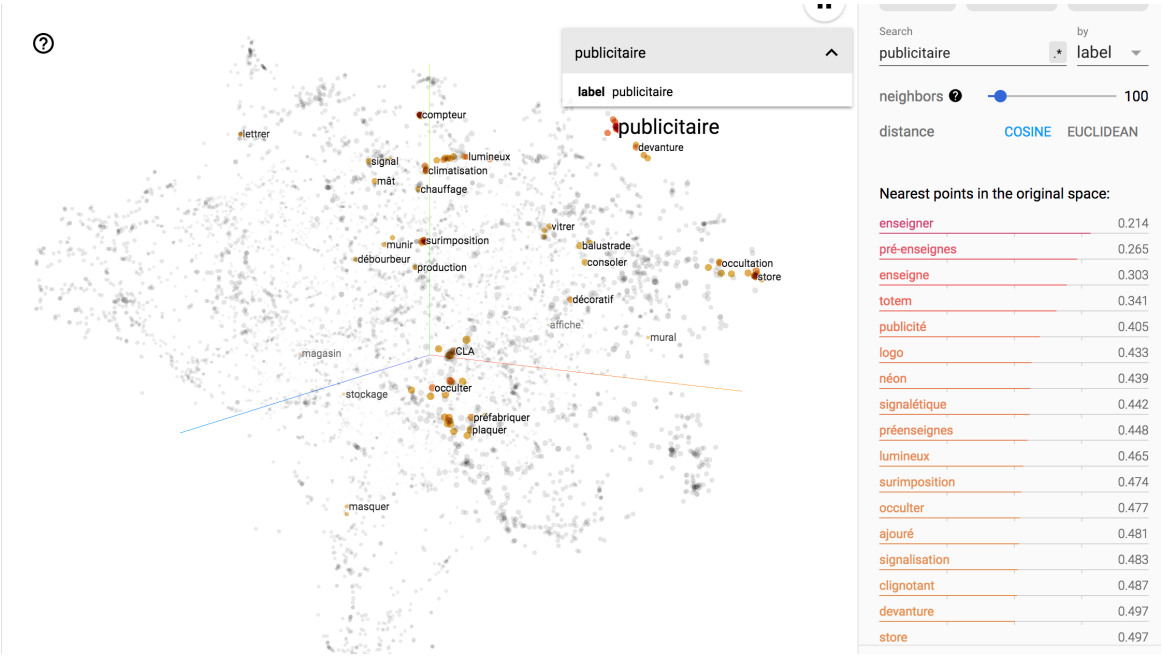


Figure 3 : interface de visualisation des représentations distribuées des mots entrainées sur le corpus PLU après réduction de dimension par tSNE (Emdedding Projector). La visualisation 3D permet d'explorer les clusters de termes, la liste des termes les plus proches dans l'espace d'origine est donnée après sélection d'un mot.

c. Extraction des termes et de leurs relations

Afin de proposer un outil d'aide à la construction d'un thésaurus spécialisé sur le vocabulaire des PLU, nous avons extrait une liste des termes les plus fréquents avec leurs relations lexicales. L'interface de présentation de cette liste est présentée sur la Figure 4 pour le premier terme de la liste.

- **voie** :: [59106 contexts, frequency rank : 1]

RELATED WORDS

- **JACCARD ::** accès, implantation, terrain, mètre, construction, condition, bâtiment, alignement, cas, hauteur, voirie, aménagement, stationnement, secteur, emprise, disposition, façade, limite, extension, installation, minimum, espace, règlement, sol, usage, desserte, habitation, parcelle, projet, opération.
- **WORD2VEC ::** publique, alignement, automobile, route, circulation, accès, minimum, retrait, mètre, priver.
- **FASTTEXT ::** publique, république, priver, circulation, voirie, dessert, automobile, oblique, arriver, accès.

EXPRESSIONS

- **NOUNS ::** alignement voie, accès voie, voie emprise, rapport voie.
- **ADJ ::** voie publique.
- **VERBS ::** voie ouvertes.

Figure 4: Interface de présentation de la liste des termes les plus fréquents avec leurs relations lexicales. Related words : les mots partageant les mêmes contextes lexicaux (jaccard) ou les plus proches dans les espaces de représentation distribuées (words embedding selon les méthodes word2vec ou fasttext). Expressions : contextes les plus fréquents dans lesquels le terme apparait, pour différentes catégories grammaticales de contexte (nouns, adj, verbs).

Les termes sélectionnés sont les noms communs les plus fréquents du corpus. Les 500 termes les plus fréquents sont présentés dans l'interface. Le nombre de contextes correspond au nombre d'occurrences du terme dans le corpus.
Les termes liés sont les termes qui partagent le plus de contexte avec le terme de l'entrée. Cette proximité de contexte est basée sur une métrique Jaccard[7] qui calcule le rapport du nombre de contextes communs entre les termes et le nombre de contextes dans l'union des contextes des 2 termes. Si x et y sont deux termes et X et Y leurs contextes respectifs (un mot à gauche ou à droite), la mesure de Jaccard est donnée par $\frac{|X \cap Y|}{|X \cup Y|}$. L'interface présente les termes ayant la mesure de Jaccard la plus forte avec le terme de l'entrée (*JACCARD ::*). D'autres méthodes d'extraction de collocations comme celles basées sur des tests statistiques[8] pourraient aussi être explorées.

Les termes les plus proches du terme de l'entrée dans l'espace de représentation distribuée apprise sur le corpus des PLU sont listés pour les deux méthodes *word2vec* et *fasttext*. Les deux méthodes cherchent à modéliser la proximité sémantique des termes mais la méthodes *fasttext* prend en compte les séquences de lettres des mots (*ngram* de caractères) ce que ne fait pas la méthode word2vec. Il en résulte des listes de mots différentes pour les deux méthodes, comme illustré pour le mot *aménagement* :

- **WORD2VEC ::** amener, infrastructure, réaménagement, amélioration, propre, fouille, création, exploitant, réhabilitation, cadrer.
- **FASTTEXT ::** réaménagement, aménageur, natation, aménager, réaménager, récemment, nécessairement, amendement, fixation, événement.

On voit que la méthode *fasttext* rapproche des termes qui partagent des séquences de lettres avec le terme *aménagement*, sans toutefois être très proches sémantiquement : récemment, nécessairement, amendement. Les termes *natation* et *fixation* sont aussi rapprochés alors que leur relation au terme *aménagement* n'est pas évidente : il peut s'agir d'erreurs dues à la faible taille du corpus d'apprentissage.

Les expressions les plus fréquentes contenant le terme de l'entrée sont listés par catégorie grammaticale : nom, adjectif et verbe. Ces listes sont simplement obtenues en filtrant les 10 contextes les plus fréquents par leur catégorie grammaticale donnée par l'étiqueteur morphosyntaxique[9].

3. Rapprochement des PLU

Dans cette section, nous décrivons les travaux réalisés pour visualiser la proximité des PLU en termes de rédaction. L'objectif est de rapprocher les articles de PLU similaires quant à leur

[7] Méthode proposée par Grefenstette, *Exploration in Automatic Thesaurus Discovery*, 1994.
[8] Voir Manning et Schütze, *Foundations of Statistical Language Processing*, 2003.
[9] Méthode proposée par Justeson et Katz, *Natural Language Engineering*, 1995.

contenu et à la forme de leur rédaction. Pour ce faire, nous avons préparé les PLU de la manière suivante :

- Extraction du texte des zones urbaines des PLU par sous-zones ;
- Pré-traitement et nettoyage du texte (identique à celui réalisé pour l'analyse syntaxique) ;
- Rapprochement des codes régions/communes issus des PLU avec les listes nominatives des régions/communes ;
- Extraction du numéro d'article par expression régulière ;
- Analyse des fréquences de zones, sélections des zones les plus fréquentes pour la poursuite de l'étude (voir Figure 5 et Figure 6)
- Construction d'une représentation vectorielle de type TF-IDF pour chaque article
- Export et visualisation des articles dans *Embedding Projector*

La représentation TF-IDF est une méthode simple de représentation vectorielle de document. Elle consiste à sélectionner N mots différents dans l'ensemble de tous les documents, puis à représenter chaque document par un vecteur fixe de taille N dont les valeurs sont proportionnelles à la fréquence du mot correspondant dans le document. La fréquence d'un mot dans un document est appelée *term frequency, TF*. Les N mots sélectionnés sont généralement les N mots les plus fréquents dans le corpus, après suppression des mots vides (articles, pronom, prépositions). Afin de défavoriser les mots qui apparaissent souvent dans le corpus mais qui ne sont pas spécifique à un document, chaque mot dans le vecteur est pondéré par l'inverse du nombre de documents dans lequel il apparait. C'est le facteur *inverse document frequency, IDF*. Le produit de ces deux facteurs donne la représentation vectorielle TF-IDF.

Analyse des articles

L'extraction des types de zone nous a permis de faire une analyse des fréquences d'articles par type de zone :

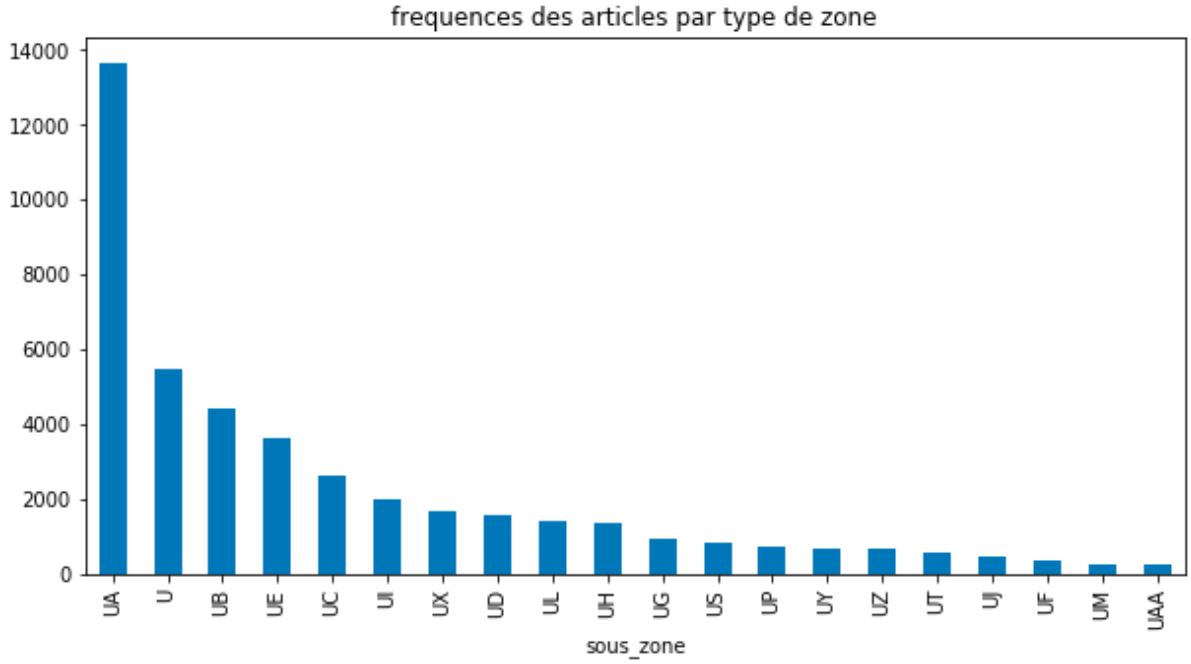


Figure 5: Fréquence des articles (tout numéro) par type de zone.

La zone UA est la plus représentée en termes de nombre d'articles et on note la présence des zones U non typées en seconde position.

L'extraction des numéros d'article nous a permis de mener une analyse des différents types d'articles. Pour rappel, la liste des articles est

U.1 : Occupations et utilisations du sol interdites ;
U.2 : Occupations et utilisations du sol soumises à des conditions particulières.
U.3 : Accès et voirie ;
U.4 : Desserte par les réseaux ;
U.5 : Caractéristiques des terrains ;
U.6 : Implantation des constructions par rapport aux voies et emprises publiques
U.7 : Implantation des constructions par rapport aux limites séparatives
U.8 : Implantation des constructions les unes par rapport aux autres sur une même propriété ;
U.9 : Emprise au sol ;
U.10 : Hauteur maximum des constructions ;
U.11 : Aspect extérieur ;
U.12 : Stationnement ;
U.13 : Espaces libres et plantations, espaces boisés classés.
U.14 : Coefficient d'occupation du sol.
U.15 et U.16 : supprimés.

- **Fréquence des articles par numéro d'article** : on note des variations par numéro d'articles, les articles 13, 14 étant les moins représentés, 15 et 16 étant encore présents dans certains PLU.

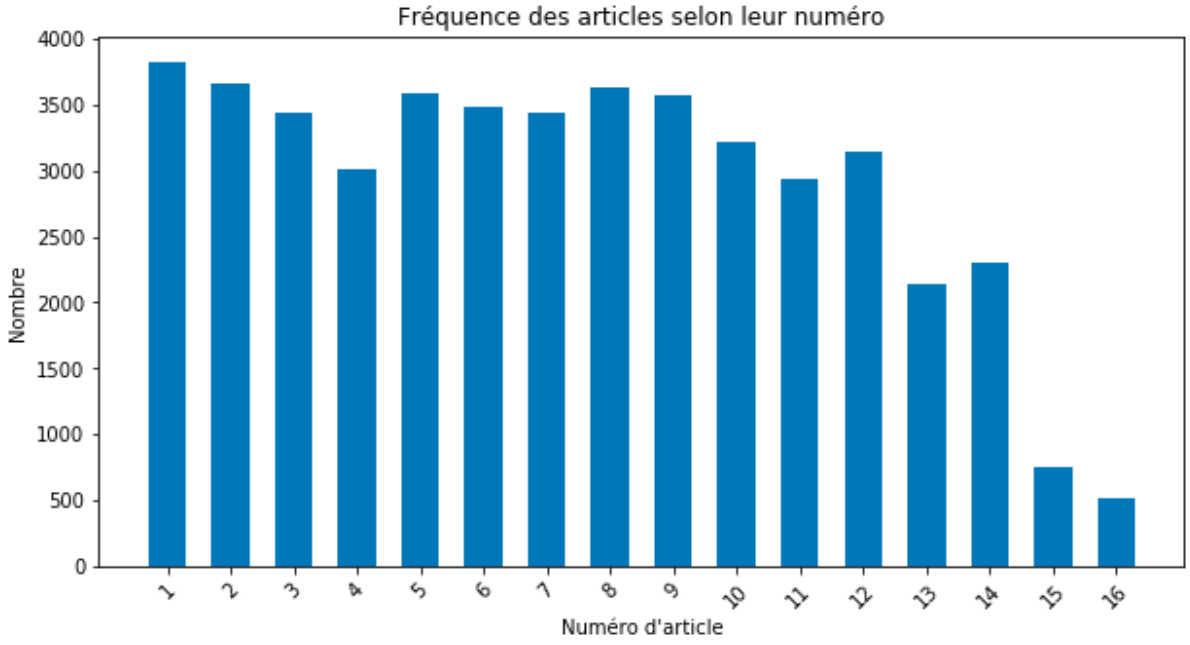
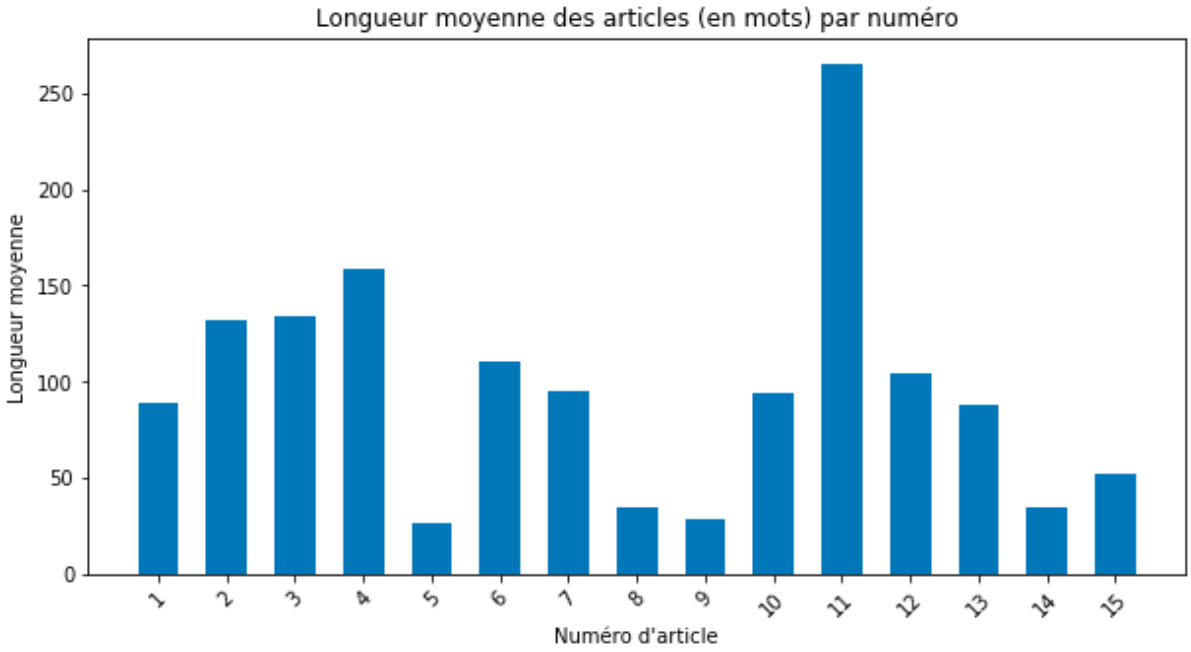


Figure 6: Fréquence des articles selon leur numéro.

- **Longueur moyenne des articles par numéro** : on note de fortes différences entre les longueurs moyennes articles par numéro, les articles 11 étant les plus longs, 5, 8 et 9 étant les plus courts.



Représentation des proximités des articles

La représentation vectorielle des articles nous permet de les représenter graphiquement afin de faire apparaitre des groupes d'articles proches et de visualiser les régions associées. On note que pour de nombreux articles, les articles les plus proches viennent de la même région.

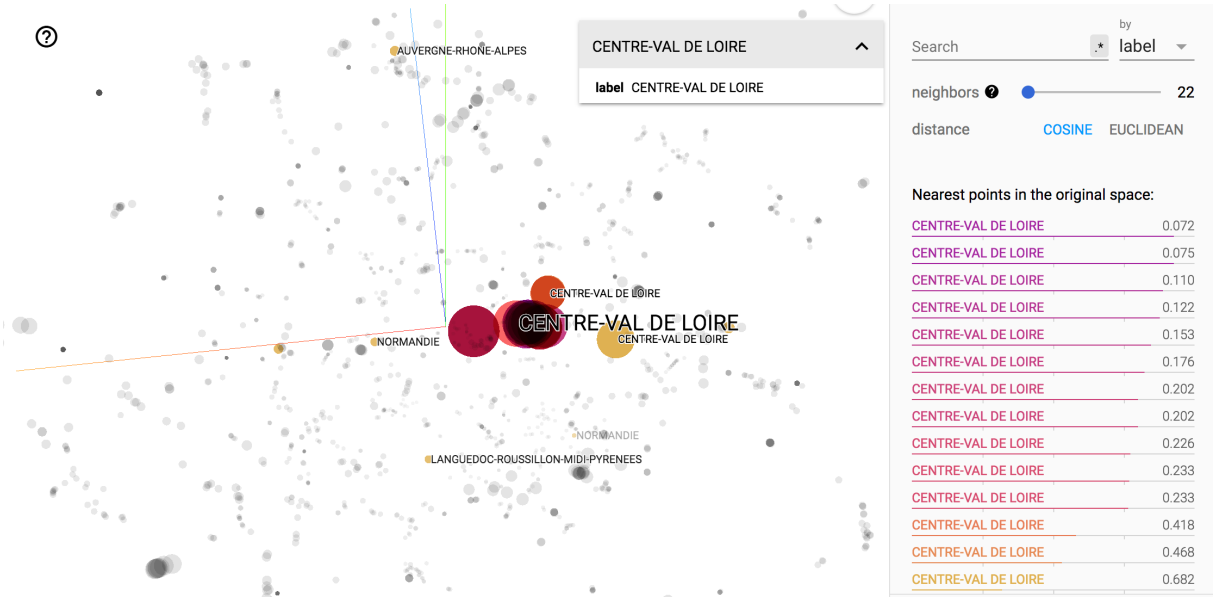


Figure 7: Visualisation des proximités des articles: on note que les articles les plus proches de l'article sélectionné proviennent de PLU de la même région (Centre-Val de Loire).

Annexes

Description des librairies testés

StanfordNLP

StanfordNLP (https://nlp.stanford.edu/software/) est un ensemble d'outils développés par l'équipe de Traitement automatique du langage naturel (TALN) de l'université de Stanford aux Etats-Unis et distribués gratuitement sous licence GNU GPLv3.
Ces outils couvrent une grande partie du champ d'application du TALN : analyse syntaxique, étiquetage morpho-syntaxique, extraction d'entités nommées, résolution de coréférences, segmentation en phrase et mots, extraction d'entités temporelles. Ces outils sont développés en java. Les algorithmes implémentés datent du début des années 2000 pour les plus anciens et sont pour la plupart basés sur des modèles statistiques. Des modèles pré-entrainés pour différentes langues et des interface dans différents langage de programmation sont aussi disponibles.

Spacy

Spacy (https://spacy.io/) est un outil de traitement automatique de la langue développé pour être à la fois facile à utiliser et très performant. Il est écrit en python/cython. Il permet d'effectuer des traitements de segmentation en mots et en phrases, d'étiquetage morpho-syntaxique, d'extraction d'entités, d'analyse de dépendances. Il est disponible dans 28 langues et propose aussi des vecteurs de mots (*words embedding*) déjà entrainés pour certaines langues et des fonctionalités d'affichage. Il peut être combiné avec l'outil *prodigy* pour l'annotation de données et l'apprentissage incrémental de modèles.