UNIVERSITÉ LIBRE DE BRUXELLES

Faculté de Lettres, Traduction et Communication

STICB-545 | Traitement automatique de corpus

TP2

Extraction d'information

PITISCI Antonio
MA-STIC2
Matricule ULB 000592990
antonio.pitisci@ulb.be

Année académique 2024-2025

1 Contexte

Le présent rapport fait état de la démarche d'application des techniques d'enrichissement présentées lors du cours théorique sur le sous-corpus *CAMille (Le Soir)* – ce dernier étant mis à disposition sur l'*UV*.

L'objectif principal de ce second travail pratique était d'analyser un ensemble d'articles d'une année au choix entre 1887 et 1970 – ici l'année 1969, comptant 100 articles – en vue d'extraire des informations telles que les mots-clés, les entités nommées, les termes les plus fréquents et, finalement, de conduire une analyse de sentiment sur des échantillons de texte sélectionnés.

Le rapport se propose, en premier lieu, de présenter la structure du *notebook* et, ensuite, de dresser un bilan succinct des analyses effectuées ; à chaque étape, je vous ferai part de mes constatations et découvertes, en essayant d'interpréter les résultats de chaque évaluation sous un angle aussi critique que possible.

1.1 Structure du *notebook*

Comme le préconise l'énoncé, mon *notebook* rassemble toutes les étapes[1] / modifications de chaque *script* abordé en cours. Celui-ci est structuré en 5 sections principales, comme suit :

1. **Importations**
   Cette cellule initiale permet d'importer toutes les bibliothèques nécessaires à la bonne exécution du *script* ;
2. **Étape 3 | Extraction des mots-clés**
   Modification du *notebook* `s1_keywords.ipynb` pour extraire uniquement les mots clés relatifs à l'année 1969 ;
3. **Étape 4 | *Stopwords* et nuage de mots**
   Modification du *notebook* `s2_wordcloud.ipynb` pour enrichir de manière itérative la liste de *stopwords* et générer le nuage de mots pour l'année choisie ;
4. **Étape 5 | Entités nommées**
   Modification du *notebook* `s2_ner.ipynb` pour trouver les entités nommées principales (personnes, organisations et lieux) pertinentes pour cette année ;
5. **Étape 6 | Analyse de sentiment**
   Analyse de sentiment effectuée sur 10 phrases sélectionnées arbitrairement dans les articles de l'année choisie à l'aide de blocs de code du *notebook* `s4_sentiment.ipynb` ; résumé et présentation des résultats dans un tableau.

2 Bilan des analysés effectuées

2.1 Imports

Ne m'y connaissant pas vraiment en *Python* et ne sachant pas si c'est une pratique recommandée, j'ai préféré réunir toutes les importations dans une seule cellule au début du *notebook*, afin de centraliser toutes les bibliothèques nécessaires à l'exécution du code et de les rendre plus visibles d'un seul coup d'œil.

[1] Par souci de clarté, chaque "étape" correspond à son point dans l'énoncé.

J'ai considéré qu'il s'agissait d'une bonne solution car, étant donné que certaines bibliothèques sont utilisées à plusieurs étapes du projet, cela permet d'éviter les doublons, de réduire le temps d'exécution et de garder le *notebook* plus propre.

Les téléchargements de ressources externes – et notamment `nltk.download('stopwords')` et `!python -m spacy download fr_core_news_md` – n'ont pas été inclus dans cette section, mais isolés dans des cellules distinctes, placées là où elles sont respectivement nécessaires.

En outre, les commentaires que j'ai ajoutés permettent de répartir les différentes bibliothèques en fonction des étapes où elles sont utilisées.

2.2 Étape 3 | Extraction des mots-clés

La troisième étape porte sur l'extraction de mots-clés liés à l'année 1969 à l'aide de *Yake*, une approche non supervisée pour l'extraction automatique de mots-clés à l'aide de caractéristiques textuelles.

Le *script* :

1. Instancie d'abord l'extracteur de mots-clés ;
2. Identifie et liste tous les fichiers contenant "1969" dans leur nom et les copie dans un dossier dédié, nommé "Articles 1969" ;
3. Applique l'extracteur de mots-clés.

2.2.1 Interprétation des résultats

En regardant de près les mots-clés extraits, j'ai immédiatement remarqué leur diversité thématique, couvrant des sujets variés : sport, culture, économie, emploi, ... On y trouve des références à des événements culturels, ainsi qu'à des institutions ; toutefois, ce qui ressort surtout, c'est un intérêt marqué pour les conditions de travail et les tendances socio-économiques de l'époque.

2.3 Étape 4 | *Stopwords* et nuage de mots

La quatrième étape a pour but de générer un nuage de mots uniquement après avoir filtré les mots vides (*stopwords*).

Pour ce faire, le *script* :

1. Initialise la liste `sw` avec les *stopwords* français prédéfinis à partir de la bibliothèque `NLTK` et y ajoute une liste de *stopwords* supplémentaires ;
2. Définit une nouvelle liste de *stopwords*[2] à ajouter à l'ensemble `sw` ;
3. Parcourt – grâce à une boucle – chaque mot dans `new_stopwords` et l'ajoute à `sw` (uniquement s'il n'est pas déjà présent !) ; `sw` est ensuite convertie en un ensemble (`set`) pour éliminer tout doublon éventuel.

[2] Ces *stopwords* ont été sélectionnés aléatoirement à partir de la liste présente sur la page https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.txt.

4. Crée un fichier `.txt` unique contenant le texte de tous les journaux de 1969, le stocke dans le dossier "Article 1969" et le nettoie à l'aide d'une fonction de nettoyage ;
5. Génère un nuage de mots à partir du fichier nettoyé.

2.3.1 Interprétation du nuage de mots

Le nuage de mots ci-dessous montre les mots restants.

Le nuage de mots :

- mentionne fréquemment des lieux et leurs événements respectifs ;
- suggère une forte présence d'annonces immobilières et d'offres de services et de moyens de contact ;
- présente des références temporelles et des personnes.

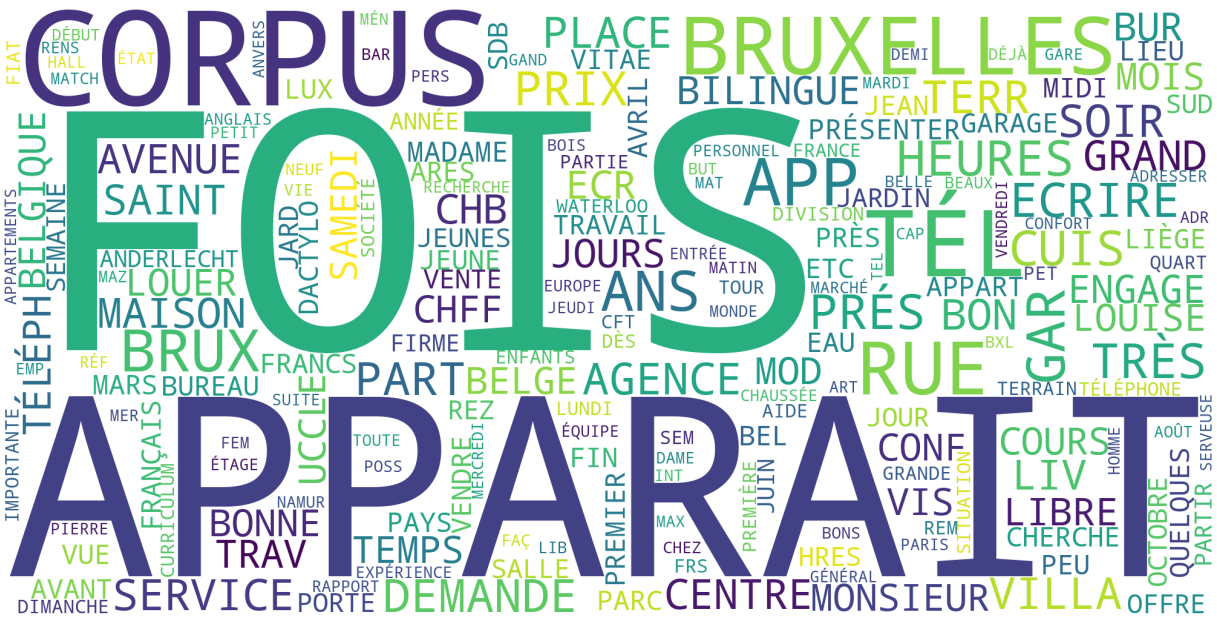

FIG. 1 : Nuage de mots pour l'année 1969.

2.4 Étape 5 | Entités nommées

La cinquième étape vise a identifier les entités nommées principales (personnes, organisations et lieux), présentes dans les articles de 1969.

Le *script* :

1. Télécharge d'abord le modèle *SpaCy* `fr_core_news_md` pour le français et le charge dans la variable `nlp` ;
2. Charge et traite les articles de 1969 (désormais contenus dans un seul fichier `.txt`) ;
3. Procède à l'extraction des entités nommées, en comptant les entités de type `PER`, `ORG` et `LOC` ;
4. Trie les différentes entités et enregistre les résultats dans 3 fichiers `.txt` distincts dans le dossier "Articles 1969".

2.4.1 Interprétation des résultats

`PER`

- Noms historiques corrects ;
- Nombreuses erreurs de reconnaissance (notamment à cause de la qualité du texte et de l'OCR) ;
- Mauvaise classification de certains termes en tant qu'entité de type "personne" ;
- Occurrences répétitives.

`ORG`

- Mauvaise classification de certains termes en tant qu'entité de type "organisation" ;
- Problèmes d'OCR ;
- Difficulté de déterminer, sans contexte supplémentaire, si certaines entités sont réellement des organisations.

`LOC`

- Noms de pays, noms de ville et noms de lieux spécifiques ;
- Occurrences répétitives ;
- Combinaisons de mots ou des phrases parfois dépourvues de sens ;
- Mauvaise classification de certains termes en tant qu'entité de type "lieu".

2.5 Étape 6 | Analyse de sentiment

L'analyse de sentiment a été réalisée sur 10 phrases choisies arbitrairement dans les articles de 1969, à l'aide de *TextBlob-FR*[3].

Le *script* :

- Initialise, tout d'abord, un objet `Blobber` ;
- Définit la fonction `get_sentiment` ;
- Effectue l'analyse de sentiment sur la liste contenant les 10 phrases choisies ;
- Stocke les résultats dans une liste de dictionnaires ;
- Crée un tableau (*dataframe*) `pandas` qui affiche ces derniers.

2.5.1 Interprétation des résultats

Je suis globalement d'accord avec les résultats, bien que je trouve que certaines phrases reconnues comme positives devaient être reconnues comme négatives et vice versa.

[3] *TextBlob* est une librairie *Python* destinée à effectuer des tâches usuelles de TAL.

Les phrases soumises à l'analyse varient entre sentiments positifs, sentiments négatifs et sentiments neutres (ces derniers se traduisant peut-être comme les plus difficiles à évaluer pour le modèle, en ce qu'ils cachent souvent une critique ou une inquiétude voilée) et révèlent donc une diversité d'émotions.

	Polarité	Subjectivité
0	11% positive	22% subjective
1	neutral	perfectly objective
2	15% negative	44% subjective
3	10% positive	20% subjective
4	50% positive	75% subjective
5	25% positive	10% subjective
6	9% positive	28% subjective
7	17% negative	5% subjective
8	45% positive	10% subjective
9	19% positive	20% subjective

FIG. 2 : Tableau exposant les résultats de l'analyse de sentiment.