Une image contenant texte

Description générée automatiquement

**Faculté de lettres, Traduction et communication**

**STIC-B545 : Traitement Automatique de Corpus**

**TP2 : Extraction d’information**

**ATCHOFFO GUEMTSA Sylvestre**

**MATRICULE : 000556104**

**IDENTIFIANT GIT : Sylvestre4**

**Projet réaliser sous la direction de**

**Monsieur Max De Wilde et de Monsieur Denis Lebailly**

**Année académique 2024-2025**

**TABLE DES MATIÈRES**

[I. Introduction 2](#_Toc181714755)

[II. Téléchargement et décompression du corpus 2](#_Toc181714756)

[III. Sélection de l'année et Extraction de mots-clés 2](#_Toc181714757)

[IV. Nuage de mots 3](#_Toc181714758)

[V. Reconnaissance d'entités nommées 4](#_Toc181714759)

[VI. Analyse de sentiment 6](#_Toc181714760)

[VII. Conclusion 7](#_Toc181714761)

# Introduction

Le traitement automatique du texte (TAT) est un champ de recherche en pleine évolution, qui vise à extraire des informations utiles et pertinentes à partir de grandes quantités de données textuelles. Dans ce TP, nous avons mis en œuvre plusieurs techniques de TAT, comme l'extraction de mots-clés, la reconnaissance d'entités nommées (NER) et l'analyse de sentiment, pour analyser un corpus d'articles issus du journal *Le Soir*. Ce corpus, appelé CAMille, regroupe des articles publiés entre 1887 et 1970. Nous nous sommes concentrés sur une étude approfondie de l’année 1955.

L’année 1955 est marquée par des transformations importantes, tant au niveau national qu'international. Dans un contexte de Guerre Froide, les tensions internationales s’intensifient, tandis qu’aux États-Unis, les mouvements pour les droits civiques prennent de l’ampleur. En Belgique, ces dynamiques mondiales influencent les débats politiques et économiques. En analysant les articles de cette époque, nous cherchons à révéler les préoccupations et les thèmes prédominants dans le discours médiatique de 1955.

# Téléchargement et décompression du corpus

Le corpus CAMille a été téléchargé et extrait dans le dossier tac/data/txt, qui regroupe un total de 9893 fichiers texte.

# Sélection de l'année et Extraction de mots-clés

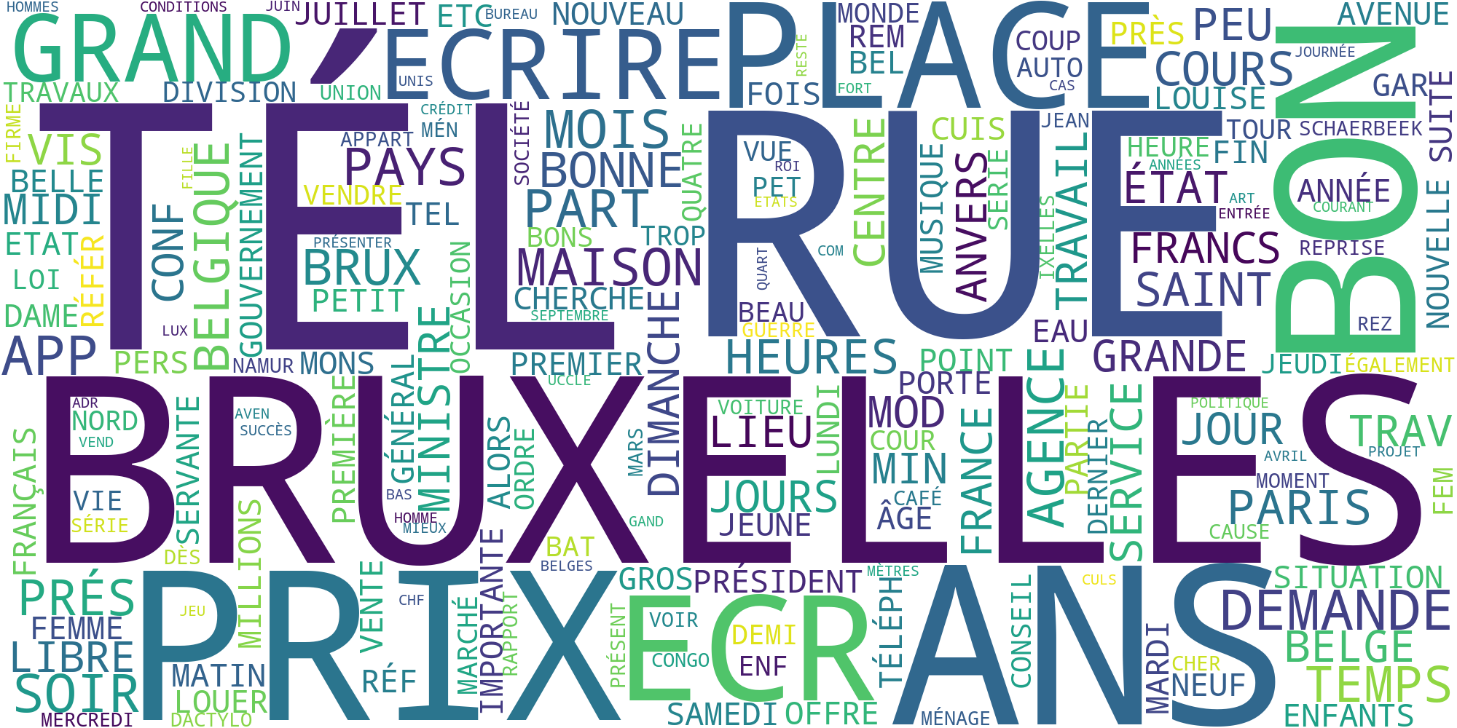
Les mots-clés extraits pour l’année 1955 révèlent des thèmes dominants qui reflètent les préoccupations et les intérêts médiatiques de l’époque. Les termes les plus significatifs incluent **"rue,"** **"Bruxelles,"** **"notaire,"** et **"chaussée,"** qui soulignent un intérêt marqué pour les aspects immobiliers et commerciaux, particulièrement dans la région de Bruxelles. Des expressions comme **"notaire Verhaegen,"** **"centiares,"** **"contenance,"** et **"adjugera"** suggèrent que de nombreux articles étaient centrés sur des ventes de propriétés et des transactions notariales, indiquant une forte présence d’annonces immobilières dans le corpus.

En outre, les mots **"kilo,"** **"avoine,"** **"seigle,"** **"colza,"** et **"orge"** montrent une concentration sur les denrées agricoles. Cela laisse penser que des articles consacrés au marché agricole étaient publiés régulièrement, possiblement sous forme de rapports de récoltes ou de prix pour des produits agricoles importants à cette époque.

D’autres termes, comme **"paix,"** **"indigène,"** et **"Van,"** suggèrent également la présence d’articles portant sur des sujets de société et de géopolitique. Par exemple, le terme "indigène" pourrait faire allusion aux discussions coloniales encore en cours pour certaines régions, un sujet d'importance dans le contexte belge de l'époque.

En résumé, les mots-clés extraits mettent en évidence trois grands axes thématiques pour l'année 1955 : les transactions immobilières, les produits agricoles et leur commerce, ainsi que certains sujets de société et de géopolitique. Ces thèmes offrent un aperçu des centres d’intérêt médiatiques et des informations recherchées par les lecteurs durant cette période.

# Nuage de mots



Le nuage de mots généré pour l’année 1955 met en lumière les termes les plus fréquents dans le corpus d’articles de cette période. Les mots les plus visibles, tels que **"TÉL,"** **"RUE,"** **"BRUXELLES,"** et **"PRIX,"** révèlent les thèmes récurrents et les centres d'intérêt médiatiques de cette époque.

* **"TÉL"** : Ce terme, probablement une abréviation pour "téléphone," suggère une forte présence de contacts et d’annonces impliquant des numéros de téléphone, ce qui souligne l'importance des communications dans les articles de 1955.
* **"RUE"** et **"BRUXELLES"** : Ces mots indiquent un intérêt marqué pour les adresses et les localisations, particulièrement dans la région de Bruxelles, ce qui témoigne de l'importance des informations locales et des transactions immobilières pour le lectorat.
* **"PRIX"** et **"FRANCS"** : La fréquence de ces termes suggère que de nombreuses annonces ou informations sur les prix étaient publiées, couvrant probablement des biens immobiliers, des produits agricoles, ou autres articles de consommation.
* **"MAISON" et "GRAND"** : Ces mots semblent indiquer des descriptions de biens immobiliers ou d'objets, souvent dans un contexte d'achat ou de vente, soulignant le rôle des annonces dans la vie quotidienne des lecteurs.
* **"MINISTRE" et "GOUVERNEMENT"** : La présence de ces mots dans le nuage indique également un intérêt pour les sujets politiques, possiblement liés à des débats publics ou des décisions politiques influentes de l’époque.

En résumé, ce nuage de mots illustre les préoccupations principales des lecteurs de 1955 : les annonces immobilières, les prix et transactions, ainsi que des sujets de société et de politique. Ces résultats mettent en avant les sujets d'intérêt dans la société belge de cette année, en lien avec le contexte social et économique de l’époque.

# Reconnaissance d'entités nommées

Les entités nommées principales pour l'année 1955 révèlent des thématiques géographiques, institutionnelles et sociales significatives dans les articles de cette période.

Lieux (LOC)

Les lieux les plus mentionnés incluent "Bruxelles" (81 occurrences), "Belgique" (47 occurrences), "Anvers" (41 occurrences), et "Paris" (40 occurrences). Cette fréquence élevée pour des villes belges comme Bruxelles et Anvers, ainsi que pour des capitales étrangères comme Paris, montre un intérêt pour les informations locales et internationales, en particulier pour les villes avec des activités politiques, économiques, ou culturelles marquantes.

Des pays comme "Allemagne" (36 occurrences), "Angleterre" (29 occurrences), "France" (28 occurrences), et "Russie" (22 occurrences) apparaissent également de façon importante. Ces mentions suggèrent que les événements internationaux, notamment ceux liés à la politique et aux relations internationales, étaient des sujets récurrents dans les articles de l’époque.

Le "Congo" apparaît 22 fois, ce qui peut indiquer une couverture médiatique liée aux colonies belges et aux discussions autour de l’indépendance des pays africains.

D'autres lieux comme "Londres," "Genève," et "Rome" sont aussi mentionnés, ce qui pourrait refléter les liens diplomatiques et les influences culturelles ou économiques de ces villes en 1955.

Personnes (PER)

Parmi les noms propres, les résultats montrent une présence de termes généraux comme "Messieurs" (28 occurrences) ou encore des fragments de mots comme "Créd" et "Verger" en raison d'erreurs de segmentation. Toutefois, quelques noms spécifiques émergent, tels que "Van Zeeland" (7 occurrences) et "Pape" (10 occurrences).

Le nom "Van Zeeland" pourrait se référer à un personnage politique ou public belge, tandis que "Pape" fait probablement référence au chef de l'Église catholique, un sujet qui aurait attiré l’attention dans les articles de cette année.

Organisations (ORG)

Les organisations les plus mentionnées incluent des institutions politiques comme "Chambre" (30 occurrences) et "Sénat" (7 occurrences), ainsi que des entités économiques telles que la "Banque Nationale" (5 occurrences). Cela témoigne d'une couverture médiatique axée sur les affaires politiques et économiques de la Belgique.

La mention de "Conseil" (6 occurrences) et "Cour" (5 occurrences) indique également un intérêt pour les activités juridiques et administratives, suggérant que les articles de l’époque contenaient des informations sur les décisions de justice ou les politiques publiques.

Des organisations internationales et des agences de presse comme "Reuter" (5 occurrences) et "Guardian" (5 occurrences) sont également présentes, illustrant une ouverture vers des sources d'information internationales pour traiter des événements mondiaux.

En résumé, l'analyse des entités nommées principales pour 1955 met en évidence une focalisation sur les villes belges et étrangères importantes, des personnalités politiques et religieuses, et des institutions économiques et politiques. Cette répartition reflète les préoccupations de l'époque, qui se concentrent à la fois sur les affaires locales en Belgique et sur les événements internationaux influençant le contexte sociopolitique de l’époque.

# Analyse de sentiment

|  |  |  |
| --- | --- | --- |
| Phrases sélectionnées pour l'analyse | Polarité | Subjectivité |
| Apple cherche à acheter une start-up anglaise pour 1 milliard de dollars | **7% positive** | **0% objective** |
| Les voitures autonomes déplacent la responsabilité de l'assurance vers les constructeurs | **23% positive** | **45% subjective** |
| San Francisco envisage d'interdire les robots coursiers sur les trottoirs | **Non disponible** | **Non disponible** |
| Londres est une grande ville du Royaume-Uni | **Non disponible** | **Non disponible** |
| L’Italie choisit ArcelorMittal pour reprendre la plus grande aciérie d’Europe | **39% positive** | **26% subjective** |
| Apple lance HomePod parce qu'il se sent menacé par l'Echo d'Amazon | **Non disponible** | **Non disponible** |
| La France ne devrait pas manquer d'électricité cet été, même en cas de canicule | **Neutre** | **50% subjective** |
| Nouvelles attaques de Trump contre le maire de Londres | **10% positive** | **25% subjective** |
| Qui est le président de la France ? | **Non disponible** | **Non disponible** |
| Où est la capitale des États-Unis | **Non disponible** | **Non disponible** |

**Résumé des Tendances**

Dans l’ensemble, les phrases sélectionnées pour cette analyse montrent une majorité de ton neutre ou légèrement positif, avec une subjectivité qui varie en fonction du sujet abordé. Les phrases relatives aux annonces économiques et aux prévisions (comme les acquisitions d'entreprises ou les prévisions énergétiques) tendent à être factuelles et objectives. Celles qui touchent à des questions de responsabilité technologique ou d’interactions politiques sont plus subjectives, reflétant un intérêt pour les implications sociétales de ces sujets.

Ces résultats illustrent bien la diversité des contenus médiatiques de l’année 1955, allant des nouvelles économiques et technologiques aux sujets de société, chacun avec des niveaux variables de sentiment et d’objectivité.

# Conclusion

L’analyse des articles de l’année 1955 a révélé les thèmes dominants de cette période grâce aux techniques d’extraction de mots-clés, de reconnaissance d’entités nommées et d’analyse de sentiment. Les résultats montrent un intérêt marqué pour les transactions immobilières, les produits agricoles, ainsi que les actualités locales et internationales. L’analyse de sentiment souligne une majorité de ton neutre, reflétant un style informatif, avec quelques sujets plus subjectifs, notamment liés aux évolutions sociales et technologiques. Ce TP a démontré l’utilité des outils de traitement automatique de texte pour comprendre les préoccupations d’une époque. En somme, ces techniques offrent une perspective riche pour analyser les discours médiatiques et explorer les thèmes marquants du passé.