ULB UNIVERSITÉ LIBRE DE BRUXELLES

# FACULTÉ DE LETTRES, TRADUCTION ET COMMUNICATION

**STIC-B545 – Traitement Automatique de Corpus - 202526**

## Rapport de travail final (TP4)

**KAMGA DZE Alban**

**Enseignant :** DE WILDE Max

**Assistant** : Lebailly Denis

Année académique 2025-2026

# Table des matières

**Introduction** ..... 3

**1. Constitution du sous-corpus** ..... 3

**2. Eploration et prétraitement du sous corpus** ..... 4

**3. Analyse lexicale et fréquentielle** ..... 5

**a) L'analyse des unigrammes** ..... 5

**b) Analyse des bigrammes** ..... 6

**c) Observations** ..... 6

**4. Extraction de mots-clés et visualisation du nuage de mots** ..... 6

**a) Extraction de mots-clés par la méthode YAKE** ..... 7

**b) Visualisation par nuage de mots** ..... 7

**5. Analyse des entités nommées (NER)** ..... 8

**a) Entités de type PERSON (Personnes)** ..... 9

**b) Entités de type LOC (Lieux)** ..... 9

**c) Entités de type ORG (Organisations)** ..... 10

**6. Clustering** ..... 10

**a) Choix du nombre de clusters (K)** ..... 11

**b) Interprétation des clusters** ..... 12

**7. Word embeddings** ..... 12

**8. Analyse de sentiment** ..... 14

**9. Conclusion** ..... 15

**10. Références bibliographiques** ..... 15

# Introduction

L'analyse de corpus de presse numérisée occupe aujourd'hui une place centrale dans les humanités numériques. Grâce aux progrès de la numérisation, de l'OCR et du traitement automatique des langues (TAL), il est désormais possible d'étudier de grandes quantités de documents historiques de manière rapide et systématique. Le projet **CAMille**, s'inscrit pleinement dans cette dynamique : il met à disposition un ensemble structuré de journaux belges francophones couvrant plus d'un siècle d'histoire médiatique, offrant ainsi un terrain d'enquête exceptionnel pour étudier l'évolution des représentations, des pratiques journalistiques et des discours médiatiques.

Dans le cadre de ce travail, nous nous intéressons à la manière dont la presse belge, en particulier le quotidien ***Le Soir*** a couvert un événement sportif majeur : **les Jeux Olympiques de Paris 1924**. Il s'agit d'un sujet daté, délimité, riche en enjeux politiques, sportifs et médiatiques, tout en demeurant suffisamment large pour permettre une analyse textuelle variée.

Les Jeux Olympiques de 1924 représentent un moment important de l'histoire du sport international, marqué par l'affirmation du mouvement olympique, la médiatisation croissante des compétitions, et l'utilisation de l'événement comme vecteur de prestige national. Comprendre comment un journal influent comme *Le Soir* a présenté cet événement permet d'éclairer La manière dont le sport était décrit et valorisé dans la presse belge de l'époque, les stratégies narratives mobilisées pour couvrir un événement international, L'évolution du lexique sportif au début du XXe siècle et la place des athlètes, des nations et des performances dans le discours médiatique.

## 1. Constitution du sous-corpus

Conformément aux consignes de ce travail, une requête ciblée a été effectuée sur la plateforme CAMille en combinant :

Requête utilisée :

- Le terme de recherche **« jeux olympiques »**,

Filtre appliqué :

- L'année **1924**,
- Le journal ***Le Soir***.

Ces filtres permettent d'obtenir un nombre de résultats (964 résultats) compris dans la plage recommandée (entre 500 et 999 documents). Ce sous-corpus forme notre base de l'analyse quantitative et qualitative réalisée dans ce travail.

Une fois les résultats exportés au format ZIP, les fichiers textuels sont structurés en un corpus qui sera traité dans un notebook Python. Les techniques mobilisées incluent : analyse de

fréquences, extraction de mots-clés, reconnaissance d'entités nommées (NER), clustering, construction d'embeddings, analyse de sentiment, etc.

L'objectif central de ce travail est double : un Objectif analytique qui nous amène à Explorer le sous-corpus afin de comprendre la manière dont le journal *Le Soir* a couvert les Jeux Olympiques de Paris 1924, à travers : l'évolution lexicale du discours, les thématiques dominantes, les entités les plus mentionnées. Le second objectif est un Objectif méthodologique, il sera question ici de pouvoir faire une analyse pertinente et efficace de notre corpus en s'appuyant sur les techniques vues durant le cours **(**nettoyage et préparation du corpus, extraction d'information, ...)

## 2. Eploration et prétraitement du sous corpus

Nous débutons L'analyse automatique de ce sous corpus textuel par une phase préalable d'exploration et de prétraitement, visant à normaliser les textes et à réduire le bruit linguistique susceptible de perturber les méthodes statistiques et algorithmiques appliquées par la suite. Dans le cadre de ce travail, le prétraitement a été conçu de manière progressive.

La première étape du prétraitement a consisté en un nettoyage approfondi des contenus textuels. Les textes extraits contiennent en effet de nombreux éléments non linguistiques ou parasites, tels que des signes typographiques spécifiques à la presse ancienne, des marqueurs techniques issus de l'export CAMille (par exemple <kw>Jeux</kw>), ainsi que des caractères de ponctuation ou symboles décoratifs.

Les opérations suivantes ont été appliquées :

- Conversion de l'ensemble du texte en minuscules ;
- Suppression des balises techniques et artefacts de balisage introduits lors de l'indexation ou de l'extraction des documents ;
- Élimination des caractères spéciaux, signes de ponctuation et chiffres, qui n'apportent pas d'information sémantique pertinente dans les analyses de fréquence, de clustering ou d'embeddings ;
- Suppression des mots vides (stopwords) du français, à l'aide des listes fournies par les bibliothèques NLP, afin de se concentrer sur les unités lexicales porteuses de sens.

Ce nettoyage permet de réduire significativement le bruit observé lors des premières analyses de fréquences, où dominaient initialement des symboles ou fragments non informatifs.

Une fois les textes nettoyés, une phase de segmentation linguistique a été appliquée. La **tokenisation** consiste à découper chaque texte en unités lexicales élémentaires (tokens), généralement assimilées aux mots. Cette opération est indispensable pour l'ensemble des traitements statistiques ultérieurs, tels que le calcul de fréquences, la vectorisation TF-IDF ou l'entraînement de modèles Word2Vec.

Enfin, le choix de travailler principalement à partir d'une colonne de texte nettoyé distincte du texte brut permet de conserver une traçabilité claire entre les données originales et les données transformées, facilitant ainsi l'interprétation des résultats et leur discussion dans les sections suivantes du rapport.

# 3. Analyse lexicale et fréquentielle

Cette étape permet d'identifier les termes dominants et les thématiques récurrentes propres aux articles relatifs aux Jeux Olympiques de 1924. Cette analyse repose sur le corpus prétraité, c'est-à-dire nettoyé et tokenisé, afin de garantir la pertinence linguistique des résultats obtenus. Comme le soulignent Manning, Raghavan et Schütze (2008), l'analyse fréquentielle constitue une méthode fondamentale en traitement automatique des langues pour dégager les régularités lexicales d'un corpus et mettre en évidence ses structures thématiques dominantes (Manning, 2008).

## a) L'analyse des unigrammes

L'analyse des unigrammes consiste à examiner la fréquence d'apparition des mots pris individuellement dans l'ensemble du corpus. Les résultats obtenus mettent en évidence une forte dominance de termes directement liés au champ sportif et géographique des Jeux Olympiques.

Parmi les mots les plus fréquents, on observe en tête **« jeux »** (4353 occurrences) et **« olympiques »** (649 occurrences), confirmant sans ambiguïté que le corpus est fortement centré sur l'événement olympique. D'autres termes comme **« lutte »**, **« victoire »**, **« finale »**, **« concours »**, **« bat »** ou **« mètres »** renvoient clairement aux compétitions sportives, aux résultats et aux disciplines athlétiques.

La présence marquée de noms de lieux et de pays tels que **« Bruxelles »**, **« Paris »**, **« Belgique »** et **« France »** reflète quant à elle la dimension géographique du discours journalistique, mettant en relation les sites de compétitions, les pays participants et le contexte international des Jeux.

Enfin, certains termes plus génériques comme **« première »**, **« deuxième »**, **« heures »**, **« après »** ou **« contre »** témoignent du style narratif et chronologique propre aux comptes rendus sportifs, où les classements, les temps et les confrontations jouent un rôle central.

Dans l'ensemble, l'analyse des unigrammes révèle un corpus fortement structuré autour de trois pôles principaux : l'événement olympique, la compétition sportive et l'ancrage géographique.

### b) Analyse des bigrammes

L'analyse des bigrammes permet d'aller au-delà des mots isolés en examinant les associations lexicales fréquentes entre deux termes consécutifs. Cette approche met en lumière des expressions figées ou des segments de discours récurrents, souvent porteurs d'un sens plus précis que les unigrammes seuls.

Les bigrammes les plus représentatifs observés dans le corpus incluent notamment **« seul marathon »**, **« marathon flandres »**, **« jeux olympiques »**, **« programme fête »**, **« fête athlétique »** ou encore **« critérium international »**. Ces associations confirment le rôle central des épreuves sportives spécifiques (comme le marathon des Flandres) et des événements organisés autour des Jeux.

Certains bigrammes suivent une logique narrative et descriptive, tels que **« disputé fin »**, **« fin mois »**, **« mois août »**, qui peuvent traduisent la volonté des journalistes de situer précisément les compétitions dans le temps. D'autres, comme **« programme fête »** ou **« épreuves suivantes »**, illustrent la structuration du discours journalistique autour de l'annonce et de la description des épreuves.

L'analyse des bigrammes apporte ainsi une vision plus contextualisée du corpus, révélant des formules récurrentes et des enchaînements lexicaux caractéristiques du langage de la presse sportive du début du XX$^{e}$ siècle.

### c) Observations

Croisée avec l'analyse des unigrammes, l'étude des bigrammes permet de confirmer la cohérence thématique du corpus. Les fréquences lexicales observées montrent que les articles ne se limitent pas à une simple mention des Jeux Olympiques, mais développent un discours riche autour des épreuves, des résultats, des lieux et de l'organisation des compétitions.

Cette analyse fréquentielle met également en évidence le rôle du journalisme sportif comme vecteur de narration : au-delà de l'information brute, les textes structurent les événements dans le temps, mettent en valeur les performances et inscrivent les compétitions dans un cadre festif et international.

Les résultats de cette section constituent une base essentielle pour les analyses ultérieures, notamment l'extraction de mots-clés, le clustering thématique et l'entraînement de modèles d'embeddings.

## 4. Extraction de mots-clés et visualisation du nuage de mots

Cette section vise à identifier les thématiques dominantes du sous-corpus consacré à la couverture des Jeux Olympiques de Paris 1924 dans le journal *Le Soir*, en combinant des méthodes statistiques classiques et des approches plus avancées d'extraction

automatique de mots-clés. Elle est complétée par une visualisation synthétique sous forme de nuage de mots.

## a) Extraction de mots-clés par la méthode YAKE

L'algorithme **YAKE (Yet Another Keyword Extractor)** a été utilisé afin d'extraire des mots-clés directement à partir du texte, sans recourir à un corpus de référence externe.

YAKE s'appuie sur plusieurs critères internes (position du mot, fréquence, cooccurrence, contexte) et permet d'identifier des expressions multi-mots particulièrement informatives, comme le décrivent Campos et al. dans leur présentation de l'algorithme, conçu pour fonctionner efficacement sur des corpus spécialisés (Campos, 2020).

Dans ce travail, nous avons préférez concentrée l'analyse sur les bigrammes (expressions de deux mots), jugés plus sémantiquement riches que les unigrammes isolés.

Les résultats obtenus font apparaître des expressions telles que :

- *Jeux olympiques*,
- *Marathon des Flandres*,
- *Records mondiaux*,
- *sporting chronique*,
- *Nation records*,
- *Organisation olympique*.

Ces expressions confirment la structuration du corpus autour des épreuves emblématiques, des records sportifs, des rubriques journalistiques spécialisées, Et de la dimension institutionnelle des Jeux.

A la différence des fréquences simples, les fréquences simples mettent surtout en évidence le vocabulaire dominant du corpus (*jeux, lutte, mètres, victoire, Bruxelles, Paris*), tandis que YAKE permettent de dégager des thématiques plus spécifiques, parfois moins visibles à première vue, comme les enjeux organisationnels, les rubriques sportives ou les expressions figées propres au discours journalistique.

## b) Visualisation par nuage de mots

Afin de proposer une synthèse visuelle du vocabulaire dominant, un nuage de mots global a été généré à partir des fréquences des termes nettoyés du corpus.

La visualisation fait apparaître de manière très lisible :

- La centralité des termes *jeux* et *olympiques*,

- L'importance des notions de performance sportive (*mètres, finale, victoire, concours*),
- La présence marquée de références **géographiques** (*Bruxelles, Paris, Belgique, France*),
- Ainsi que des termes liés à l'organisation et au calendrier (*heures, juillet, dimanche*).

Ce nuage de mots constitue un outil de lecture exploratoire rapide, permettant de confirmer visuellement les tendances observées dans les analyses quantitatives précédentes.


Figure1 : Nuage de mots sous corpus Jeux Olympique Paris 1924

## 5. Analyse des entités nommées (NER)

L'analyse des entités nommées (Named Entity Recognition – NER) vise à identifier automatiquement, au sein du corpus, les acteurs, lieux et organisations les plus fréquemment mentionnés dans les articles du journal *Le Soir* relatifs aux Jeux Olympiques de 1924.

Pour cette analyse, nous avons utilisé la bibliothèque **spaCy**, appliqué au texte préalablement nettoyé (clean_text). Les entités extraites ont ensuite été regroupées par type (PERSON, LOC, ORG) conformément aux principes généraux de la reconnaissance d'entités nommées décrits dans la littérature en traitement automatique des langues (Lample, 2016).

## a) Entités de type PERSON (Personnes)

L'extraction des entités de type PERSON permet d'identifier les individus athlètes, personnalités sportives ou figures publiques les plus présents dans le discours journalistique.

Les résultats montrent que les entités les plus fréquemment détectées sont :

Jumet daurel,Lucien samosate, rebecq, Etienne fou, paume lebrun, quaregnon david, houdeng, ransart, coubetin, lebout, valéry, watson, martin, henri, Tompson, françois...

Plusieurs observations peuvent être formulées :

- Une part significative des entités reconnues comme **PERSON** correspond en réalité à des **noms propres locaux**, ce qui illustre les limites du modèle NER lorsqu'il est appliqué à des textes journalistiques anciens.
- Les athlètes ne sont pas toujours désignés par leur nom complet, mais parfois par des surnoms, des noms partiels, ce qui complique leur identification automatique.
- Malgré ces limites, cette extraction met en évidence l'importance accordée aux figures individuelles dans la couverture médiatique des Jeux Olympiques, en particulier dans les comptes rendus de compétitions et de résultats.

## b) Entités de type LOC (Lieux)

Les entités de type **LOC** constituent un élément central de l'analyse, dans la mesure où les Jeux Olympiques sont un événement fortement ancré dans des espaces géographiques précis.

Les lieux les plus fréquemment mentionnés dans le corpus sont :

Bruxelles, Paris, Belgique, France, Gilly, Angleterre, Italie, Charleroi, Londres, États-Unis, Gand, Canada, Pologne, Liège...

Ces résultats confirment plusieurs éléments structurants du corpus :

- **Paris** et **Bruxelles** occupent une place centrale, reflétant à la fois le lieu de l'événement olympique et le point de vue du journal belge *Le Soir*.
- La forte présence de pays et de villes européennes (France, Belgique, Angleterre, Italie) souligne la dimension internationale mais majoritairement européenne des Jeux de 1924.

- Les références fréquentes à des villes belges (Charleroi, Gilly, Liège) montrent l'ancrage local du discours journalistique, notamment dans la valorisation des athlètes et clubs nationaux.

L'analyse des entités de lieu met ainsi en évidence une articulation constante entre dimension internationale de l'événement et perspective nationale du journal.

### c) Entités de type ORG (Organisations)

L'extraction des entités de type **ORG** vise à identifier les institutions, clubs, fédérations ou organisations impliquées dans les Jeux Olympiques.

Les résultats obtenus sont cependant plus hétérogènes :

ford, fiat, yacht club belgique, agence rossel, automobile club milan, lcr, ...

Plusieurs limites apparaissent clairement :

- Le modèle NER tend à classer comme organisations des marques commerciales ou des termes ambigus, parfois éloignés du champ sportif.
- Certaines entités correspondent à des regroupements lexicaux imparfaits (ex. *france belgique*), révélant les difficultés du modèle à segmenter correctement les entités dans des textes anciens.
- Les organisations sportives officielles sont moins systématiquement identifiées que les lieux ou les personnes, ce qui pourrait peut-être s'expliquer par des formulations journalistiques moins standardisées à l'époque.

## 6. Clustering

Afin d'identifier des regroupements thématiques d'articles au sein du sous-corpus consacré aux Jeux Olympiques de 1924, une approche de clustering non supervisé a été mise en œuvre.

Les textes ont d'abord été vectorisés à l'aide de la méthode **TF-IDF** (Term Frequency – Inverse Document Frequency), appliquée aux textes préalablement nettoyés (clean_text). Cette représentation permet de pondérer les termes en fonction de leur importance relative dans chaque document, tout en atténuant l'impact des mots trop fréquents dans l'ensemble du corpus.

Sur cette matrice TF-IDF, l'algorithme **K-Means** a ensuite été appliqué. **K-Means** initialement proposé par Lloyd (1982), vise à partitionner un ensemble de documents représentés dans un espace vectoriel en K groupes en minimisant la variance intra-cluster.

## a) Choix du nombre de clusters (K)

Plusieurs valeurs de *K* ont été testées (de 2 à 5 clusters), conformément aux pratiques vues en cours.

Le choix final s'est porté sur **K = 3**, pour les raisons suivantes :

- Chaque cluster présente une cohérence lexicale identifiable, facilitant l'interprétation thématique ;
- Un nombre plus élevé de clusters conduisait à une fragmentation excessive des thèmes, tandis qu'un nombre plus faible fusionnait des contenus distincts.

La figure ci-dessus illustre la projection des documents TF-IDF, colorés selon leur appartenance à l'un des trois clusters.

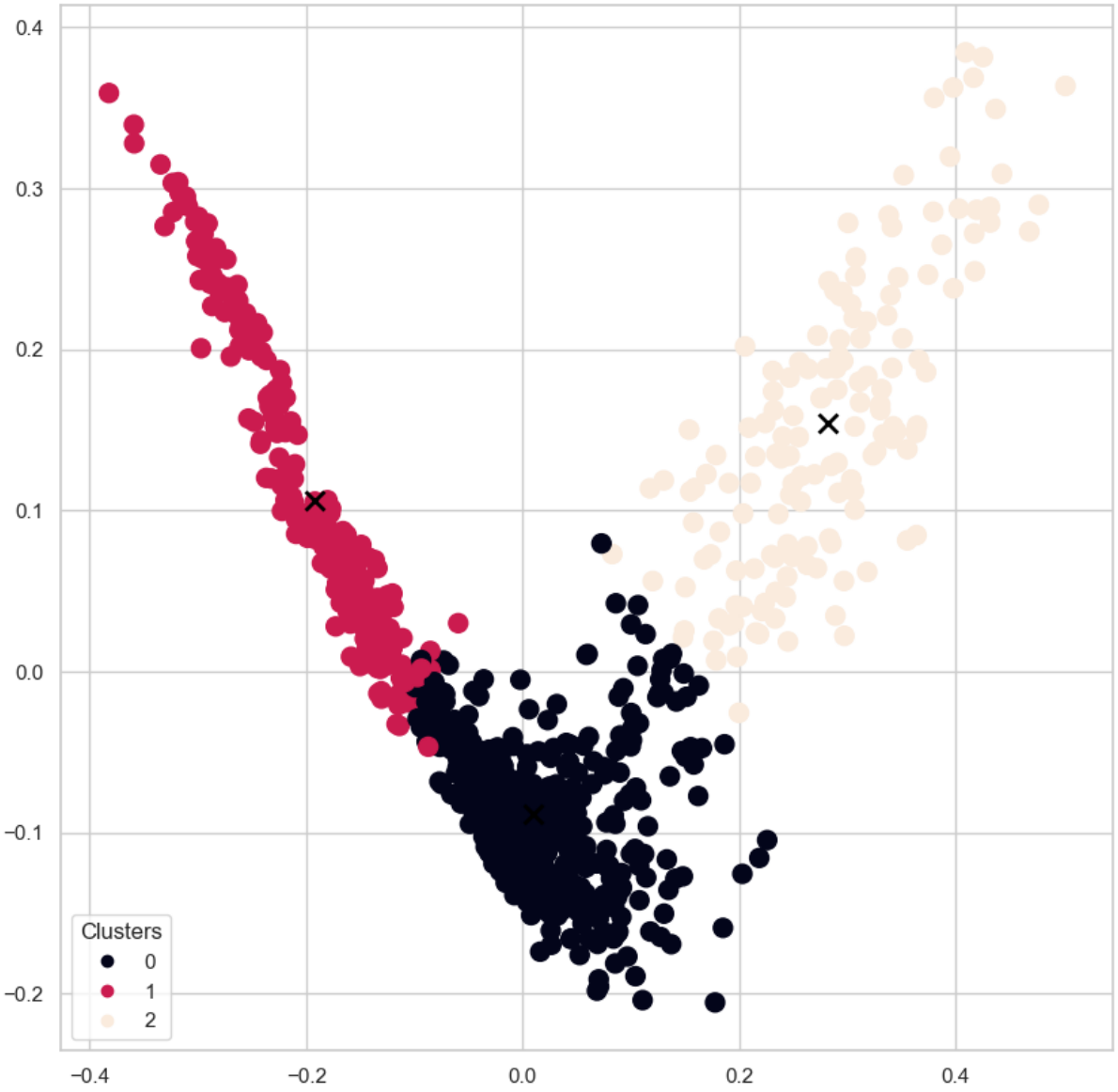


Figure 2 : représentation des clusters

### b) Interprétation des clusters

L'analyse des termes dominants associés à chaque cluster permet de proposer l'interprétation suivante :

❖ **Cluster 0 : Discours institutionnel et contexte international**

Ce cluster regroupe principalement des textes plus analytiques, traitant du cadre institutionnel, historique et international des Jeux Olympiques. On y retrouve notamment :

*france, belgique, paris, olympiques, international, heures, après*.

Ces articles mettent l'accent sur la portée symbolique et internationale des Jeux, les relations entre nations, ainsi que sur le rôle des Jeux Olympiques dans le contexte politique et culturel de l'époque.

❖ **Cluster 1 : Organisation, économie et logistique**

Ce cluster rassemble des textes liés à l'organisation matérielle et économique des Jeux : billetterie, finances, agences, ventes, logistique et aspects administratifs. Les termes dominants sont par exemple :

*prix, agence, vendre, louer, maison, rue, état, rossel*.

Ces articles s'éloignent du strict compte-rendu sportif pour aborder les enjeux pratiques et économiques entourant l'événement olympique.

❖ **Cluster 2 : Résultats sportifs et performances**

Le troisième cluster correspond principalement des articles centrés sur les résultats des compétitions, les classements et les performances sportives.

Les mots les plus représentatifs incluent notamment :

*victoire, finale, concours, bat, mètres, première, deuxième, lutte*.

Il s'agit d'articles à forte dominante sportive, décrivant les épreuves, les vainqueurs et les performances chiffrées, souvent dans un style factuel et descriptif.

## 7. Word embeddings

Dans cette section, une approche par word embeddings a été mobilisée afin d'explorer les relations sémantiques fines entre les termes du corpus. l'algorithme Word2Vec proposé par Mikolov et al. (2013) a été utilisé. Cet algorithme

apprend les représentations vectorielles des mots à partir de leur contexte d'apparition, selon l'hypothèse distributionnelle selon laquelle les mots apparaissant dans des contextes similaires ont des significations proches (Mikolov, 2013).

Nous avons pris le soin d'entrainer deux modèles Word2Vec distincts sur le sous-corpus consacré aux Jeux Olympiques de 1924, à partir des textes préalablement nettoyés et tokenisés :

- Modèle 1 : entraîné avec des paramètres favorisant une capture plus fine des cooccurrences locales ;
- Modèle 2 : entraîné avec des paramètres légèrement différents (fenêtre contextuelle et/ou seuils de fréquence), afin d'évaluer l'impact de ces choix sur les similarités sémantiques produites.

Une première exploitation des modèles consiste à calculer la similarité sémantique entre deux termes, mesurée ici par la similarité cosinus entre leurs vecteurs respectifs.

À titre d'exemple, la similarité entre les termes « *paris* » et « *londres* » a été calculée pour chacun des deux modèles :

- Modèle 1 : similarité ≈ 0,78
- Modèle 2 : similarité ≈ 0,35

Ce résultat met en évidence des comportements distincts entre les deux modèles. Dans le premier cas, la forte similarité suggère que *Paris* et *Londres* apparaissent dans des contextes proches au sein du corpus, ce qui est cohérent avec le discours journalistique de l'époque, où les grandes capitales européennes sont fréquemment évoquées conjointement dans des contextes sportifs, diplomatiques ou organisationnels.

Dans le second modèle, la similarité plus faible indique une représentation plus différenciée des deux villes, possiblement liée à une fenêtre contextuelle plus restreinte.

L'analyse a ensuite été approfondie en examinant, pour un terme donné, les mots les plus proches dans l'espace vectoriel, à l'aide de la fonction most_similar.

Pour le terme « *course* », les résultats obtenus sont particulièrement révélateurs.

**Modèle 1 :**

Les mots les plus proches incluent notamment :

***Vitesse*** **;** ***kilomètres*** **;** ***distance*** **;** ***relais*** **;** ***handicap*** **;** ***scolaires*** **;** ***mètres*** **;** ***gagnée*** **;**

Ces associations sont fortement cohérentes avec le champ lexical de l'athlétisme, et traduisent une représentation sémantique orientée vers les épreuves sportives, les performances et les distances.

**Modèle 2**

Pour le second modèle, les mots les plus proches de *« course »* sont par exemple :

***Ski ; épreuve* ; *brasse* ; *poursuite* ; *classements* ; *étapes* ; *parcours***

Ici, la représentation est plus hétérogène, intégrant des disciplines sportives variées et des notions plus générales liées à la compétition.

Ces différences confirment que les word embeddings ne constituent pas une représentation « objective » du sens, mais le résultat d'un apprentissage statistique dépendant du corpus et des paramètres choisis. Elles soulignent également l'intérêt de comparer plusieurs modèles afin d'éviter des interprétations trop univoques.

## 8. Analyse de sentiment

L'analyse de sentiment vise à évaluer le ton émotionnel des articles composant le sous-corpus consacré aux Jeux Olympiques de Paris 1924, afin de déterminer si les discours journalistiques sont majoritairement positifs, négatifs ou neutres. Afin de comparer les résultats et de tester une approche plus récente, un modèle de classification de sentiment pré-entraîné (basé sur des architectures de type Transformer) a été utilisé ponctuellement.

Cette méthode fournit une étiquette catégorielle (*POSITIVE* / *NEGATIVE*) accompagnée d'un score de confiance. L'analyse a été appliquée à des phrases extraits directement du sous corpus.

Les tests effectués montrent une forte dominance de tons neutres ou positifs, ce qui est cohérent avec la nature principalement informative et sportive du corpus.

**Exemples de textes à tonalité neutre ou positive**

- *« Les Jeux Olympiques de Paris ont débuté le 4 mai 1924. »* **Neutre et parfaitement objectif** (énoncé factuel, absence de jugement de valeur)
- *« Aujourd'hui, cependant, nous faisons trêve de notre incrédulité et nous enregistrons deux très belles performances accomplies par Avne Borg. »* **70 % positif, fortement subjectif** (lexique valorisant : *très belles performances*)
- *« Paris est en pleines Olympiades, ce qui signifie que les Jeux Olympiques ont lieu cette année en France. »*

  **Faiblement positif, modérément subjectif**

Ces résultats confirment que la presse met largement en avant : les **performances sportives**,les **records**, les **victoires**, et l'enthousiasme lié à l'événement olympique.

L'analyse de sentiment appliquée à un corpus journalistique de 1924 présente aussi quelques limites méthodologiques :

Le français des années 1920 utilise des tournures, expressions et connotations parfois différentes du français contemporain ; Pouvant mener à une interprétation totalement différente par les modèles.

Le style de la presse de l'époque est souvent : plus narratif, plus descriptif, moins explicitement émotionnel. Cela entraîne une surreprésentation de la catégorie *neutre*.

## 9. Conclusion

Cette analyse transversale du sous-corpus consacré aux Jeux Olympiques de Paris 1924 dans le journal *Le Soir* a permis de mettre en évidence la richesse lexicale et thématique de la couverture journalistique de l'événement. Les méthodes de traitement automatique du langage (fréquences, mots-clés, entités nommées, clustering, embeddings et analyse de sentiment) ont révélé une forte dominance du registre sportif, structurée autour des compétitions, des résultats et des lieux emblématiques, tout en laissant apparaître des dimensions économiques, organisationnelles et internationales.

Toutefois, certaines limites apparaissent clairement. Les outils contemporains de TAL (reconnaissance d'entités, analyse de sentiment, embeddings) sont parfois mis en difficulté par un corpus historique, dont les formes linguistiques et typographiques diffèrent du français actuel. Les résultats obtenus doivent donc être interprétés avec précaution et contextualisés. Néanmoins, cette étude met en évidence le potentiel des méthodes computationnelles pour l'exploration de vastes archives de presse.

## 10. Références bibliographiques

Campos, R. .. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*. Récupéré sur https://doi.org/10.1016/j.ins.2019.09.013

Lample. (2016). Neural Architectures for Named Entity Recognition. Récupéré sur https://aclanthology.org/N16-1030/

Manning, C. D. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. Récupéré sur https://nlp.stanford.edu/IR-book/

Mikolov, T. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*. Récupéré sur https://arxiv.org/abs/1301.3781