Université Caen Normandie

Rapport Technologie du langage (L3 informatique)

Soumis par :
Emilien Huron
Florian Pépin
Tom David

20 octobre 2024
Campus 2 CAEN
1^{er} Semestre

Table des matières

1 Introduction **2**

2 Initialisation **3**
2.1 Corpus d'entrée . . . 3
2.2 Gold standard . . . 4
2.3 Métriques d'évaluation . . . 7
2.3.1 Méthodologie du calcul de similarité . . . 7
2.3.2 Calcul avec un seuil . . . 7
2.3.3 Utilité de la métrique . . . 7
2.3.4 Conclusion . . . 7

3 Approche naïve **8**
3.1 Conception de l'approche naïve . . . 8
3.1.1 Analyse comparative . . . 8
3.1.2 Traitement des résultats . . . 9
3.1.3 Application du seuil . . . 9
3.1.4 Extraction des résultats . . . 9
3.2 Evaluer notre approche naïve . . . 10
3.2.1 Méthodologie de l'évaluation . . . 10
3.3 Résultats obtenus . . . 11
3.3.1 Résultats du gold standard . . . 11
3.3.2 Résultats de l'approche naïve . . . 11
3.3.3 Comparaison de l'approche naïve avec le gold standard . . . 11
3.4 Analyse des résultats . . . 12
3.4.1 Métriques globales . . . 12
3.4.2 Métriques macro . . . 12
3.4.3 Conclusion . . . 12

4 Approche intelligente **13**
4.1 Conception de l'approche intelligente . . . 13
4.1.1 Préparation des données . . . 13
4.1.2 Vectorisation des textes avec TF-IDF . . . 13
4.1.3 Mesure de similarité cosinus . . . 13
4.2 Interprétation des résultats . . . 13
4.3 Optimisation et ajustements . . . 14
4.4 Conclusion . . . 14

5 Conclusion **15**

1 Introduction

Dans le cadre de notre projet en **technologie du langage**, nous avons développé une approche visant à **extraire des informations** à partir de sources textuelles.
Notre projet s'inscrit dans une démarche visant à répondre à la problématique suivante : **"Comment déterminer les similarités entre deux personnalités en s'appuyant sur leur page Wikipédia ?"**

Wikidata est une base de connaissances collaborative (ontologie) et libre, maintenue par la Wikimedia Foundation. Elle sert de dépôt central pour des données structurées, permettant de soutenir les projets de Wikimedia tels que Wikipédia, Wikivoyage ou encore Wikimedia Commons. Chaque élément de Wikidata est identifié par un code unique, par exemple Q42 pour Douglas Adams. Ce code contient des propriétés et des valeurs. Cela facilite ainsi l'intégration, la réutilisation et la connexion des informations entre différentes langues.

Afin de répondre à cette problématique, nous avons développé une application permettant de déterminer la ressemblance entre diverses célébrités. Pour cela, nous avons exploité les données disponibles de ces dernières sur leurs pages **Wikipédia** et **Wikidata**.
Les concepts étudiés pendant le cours, tels que la vectorisation de textes et les métriques de similarité, sont utilisés dans ce projet afin d'obtenir des résultats plus pertinents.

Notre étude a un double objectif. D'une part, montrer comment des méthodes de traitement du langage naturel peuvent être utilisées pour **effectuer des analyses détaillées** sur des personnalités publiques.
D'autre part, examiner **les liens et les ressemblances**, entre un certain nombre de personnes, qui ne sont pas visibles à travers une simple lecture des textes.

Nous avons **deux approches** pour montrer cela. D'une part, l'approche « **naïve** », qui évalue la similarité par le partage de liens.
D'autre part, l'approche « **intelligente** », qui transforme les textes en vecteurs pour une analyse comparative plus approfondie.

Ces techniques nous permettent d'explorer une grande diversité de données textuelles et de synthétiser des informations complexes en résultats plus facilement compréhensibles par l'humain.

2 Initialisation

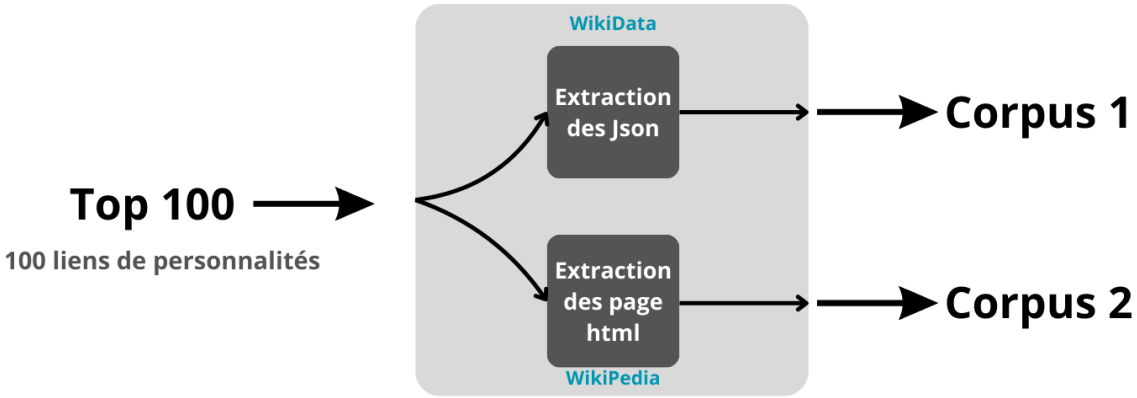
2.1 Corpus d'entrée

Pour créer notre corpus d'entrée, nous avons décidé de nous appuyer sur le contenu de certains articles Wikipédia. Nous avons analysé les célébrités les plus consultées sur la plateforme Wikipédia (version anglophone).

Ce corpus a été conçu dans le but d'établir une base de données solide permettant d'évaluer les points communs et les différences entre les différentes célébrités.
Pour notre projet, la sélection des données consiste à extraire un ensemble de 100 liens vers des pages de personnalités.
Chaque lien renvoie à une page **Wikipédia** , qui sera utilisée comme point de départ pour l'extraction des informations. La création de ce corpus est une étape primordiale dans notre projet, il se base sur deux sources en ligne : **WikiData** et **Wikipédia**.

Pour récupérer les informations des différentes personnes, nous avons utilisé **Wikidata**. C'est un outil qui permet d'extraire, à partir de pages Wikipédia, des données rangées dans différentes catégories. Toutes les données, de chaque personnalité, seront mises dans un fichier JSON. Chaque fichier est donc une ressource d'informations sur les différentes catégories des personnalités. Ces informations sont structurées et seront utilisées pour évaluer la similarité. Ces données JSON servent de base pour le gold standard.

Le deuxième corpus est formé à partir des textes extraits des pages **Wikipédia**. Ce corpus comprend des descriptions détaillées sur les différentes personnalités, ainsi que des liens liées à la personnalité en question. Ces données textuelles servent de base pour l'approche naïve.



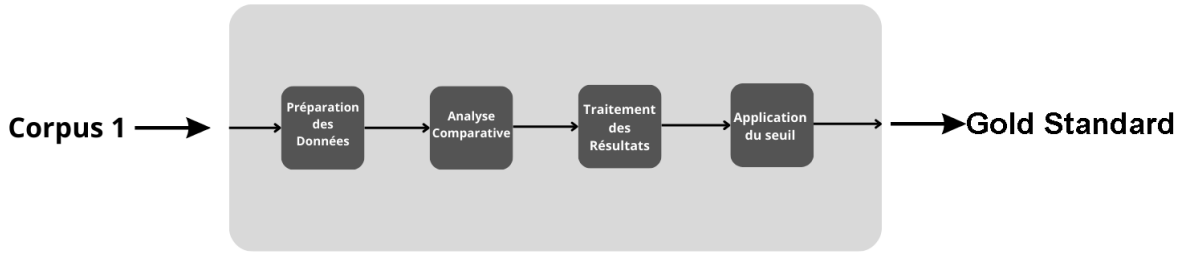
2.2 Gold standard

Le projet que nous avons développé utilise un gold standard construit à partir de données récupérées sur l'ontologie **WikiData**. Il permet de calculer un score de similarité entre plusieurs personnalités en fonction des catégories qu'elles ont en commun.

L'entrée du système est une liste des 100 personnalités les plus visitées sur **Wikipédia**. À partir de cette liste, le programme compare chaque couple de personnalités pour déterminer le nombre de catégories en commun. Ces catégories peuvent être constituées de la profession, du pays d'origine, ou de tout autre attributs.

Pour chaque couple de personnalités, un score de similarité est calculé en fonction du nombre de catégories communes.
De plus, nous avons caculé un seuil à partir duquel nous déterminons si un couple est similaire ou non. Pour faire cela nous nous basons sur le score de similarité vu plus haut. Dans le cas où la valeur se situe en dessous du seuil, le gold standard de ce couple sera "non similaire". Dans le cas contraire il sera "similaire".

Pour conclure, le gold standard que nous avons créé a pour but de définir un point de comparaison solide, fiable et également reproductible par des humains. Le but est de mesurer la similarité entre des personnalités à partir de leurs informations **WikiData**.



1. **Chargement des données** : Le programme prend un dossier contenant des fichiers JSON. Chaque fichier représente une personne avec des informations catégorisées. Il charge les fichiers en mémoire.

2. **Comparaison des couples** :

— Pour chaque paire de fichiers JSON, le programme identifie des catégories communes et les comparent entre elles.

— Deux calculs principaux sont effectués :

— *Proportion de catégories communes* :

$$\text{proportion\_communes} = \frac{\text{nombre de catégories communes}}{\text{nombre total de catégories}}$$

— *Score de similarité des catégories* :

— Pour chaque catégorie commune, si elles sont multivaluées, le score est calculé comme suit :

$$\text{score\_catégorie} = \frac{\text{nombre d'éléments communs}}{\text{nombre total d'éléments uniques}}$$

— Pour chaque catégorie commune, si elles sont monovaluées, un score de 1 est attribué, sinon un score de 0 est attribué.

— *Score moyen par catégorie* :

$$\text{score\_moyen} = \frac{\text{somme des scores des catégories}}{\text{nombre de catégories communes}}$$

— *Score final de similarité pour chaque couple* :

$$\text{score\_similarité} = \frac{\text{score moyen des catégories} + \text{proportion de catégories communes}}{2}$$

3. **Calcul du seuil** :

— Après avoir calculé les scores de similarité pour tous les couples, le programme calcule la moyenne des scores :

$$\mu = \frac{\sum \text{scores de similarité}}{\text{nombre total de couples}}$$

— Il calcule ensuite l'écart-type pour mesurer la dispersion des scores (cela permet d'éviter les disparités entre les personnes avec beaucoup et peu de catégories renseignées) :

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{\text{nombre total de couples}}}$$

— Enfin, le seuil est calculé pour en déduire la similarité :

$$\text{seuil} = \mu + \sigma$$

— Le seuil est limité par une valeur comprise entre 0 et 1 pour garantir qu'il reste dans des limites réalistes.

4. **Résultats finaux** :

— Le programme compare chaque score de similarité au seuil :

$$\text{similaire} = \text{vrai si score\_similarité} \geq \text{seuil sinon faux}$$

— Il produit ensuite un résultat indiquant si chaque couple est similaire ou non, en fonction du seuil calculé.

Ainsi, le seuil est calculé pour juger si les couples de personnes sont considérées comme similaires ou non.

2.3 Métriques d'évaluation

Dans notre projet, nous avons utilisé une métrique d'évaluation simple mais efficace pour comparer les personnalités à partir de leurs différentes catégories **Wikidata**.

2.3.1 Méthodologie du calcul de similarité

Pour chaque couple de personnalités, nous comparons les catégories **Wikidata** qu'ils ont en commun. Si deux personnalités partagent la même catégorie, nous ajoutons un point **(+1)** (ou on ajoute la proportion si la catégorie est multivalué) à leur score de similarité. Nous répétons ce procédé pour toutes les catégories de tous les couples. À la fin de la comparaison, chaque couple possède un score de similarité qui correspond au nombre total de catégories qu'ils ont en commun.

2.3.2 Calcul avec un seuil

Une fois que tous les scores de similarité, de chaque couple, sont calculés, nous appliquons, avec un seuil, une formule pour évaluer leur degré de similarité. Le seuil est une valeur définie qui détermine si un couple de personnalités peut être considéré comme similaire ou non.

Nous calculons ensuite le pourcentage de couples dont le score de similarité est supérieur ou égal à ce seuil. Ce pourcentage représente la proportion de couples de personnalités considérés comme similaires parmi l'ensemble des couples possibles.

2.3.3 Utilité de la métrique

Cette métrique simple permet d'identifier rapidement les relations entre les personnalités en s'appuyant sur les catégories **Wikidata**. L'utilisation d'un seuil permet de garder les couples les plus pertinents, ce qui facilite la modularité du gold standard et des différentes approches.

2.3.4 Conclusion

Pour résumer, les métriques "similaire" et "non similaire" se sont révélées les plus intuitives et les plus faciles à comprendre pour l'ensemble des gens. De plus, leur calcul est simple et le seuil automatisé offre une grande flexibilité pour l'appliqué sur n'importe quel corpus.

3 Approche naïve

3.1 Conception de l'approche naïve

Dans notre projet, nous utilisons une méthode naïve pour étudier la similarité entre différentes personnalités en analysant les liens partagés entre les pages **Wikipédia** liées à chaque personnalité. Le Corpus 2, regroupant tous les lien de **Wikipédia**, permet d'évaluer la similarité de deux textes en fonction des liens hypertextes qu'ils partagent.

3.1.1 Analyse comparative

Chaque entité est associée à un ensemble de liens extraits de sa page **Wikipédia** . Ces liens renvoient vers d'autres articles de **Wikipédia**. La similarité est identifiée en comparant les ensembles de liens entre différentes entités. L'idée est que si deux entités renvoient vers un nombre important d'articles **Wikipédia** communs, alors elles partagent probablement des caractéristiques similaires.

Les comparaisons entre les couples de fichiers sont effectuées selon les étapes suivantes :

— Les liens extraits de chaque fichier sont stockés dans des ensembles pour faciliter le calcul de l'intersection et de l'union.

— Le nombre de liens communs est calculé par l'intersection des ensembles :

$$\text{liens\_communs} = \text{ensemble de l'intersection des liens entre fichier1 et fichier2}$$

— Le nombre de liens uniques est calculé par l'union des ensembles :

$$\text{total\_liens\_uniques} = \text{ensemble de l'union des liens entre fichier1 et fichier2}$$

— La similarité entre les entités est ensuite mesurée :

$$\text{similarité} = \frac{\text{nombre de liens communs}}{\text{nombre de liens uniques}}$$

— Si l'ensemble des liens uniques est vide, la similarité est définie à 0.

3.1.2 Traitement des résultats

Ces comparaisons sont évaluées en fonction du nombre de liens communs. Cela permet de mesurer la similarité de manière brute, en se basant uniquement sur des références communes, sans tenir compte du contenu des textes eux-mêmes.

3.1.3 Application du seuil

Nous établissons un seuil afin de différencier les couples d'entités similaires des couples d'entités non similaires. On peut considérer deux entités comme étant similaires si elles partagent un certain pourcentage de liens en commun.

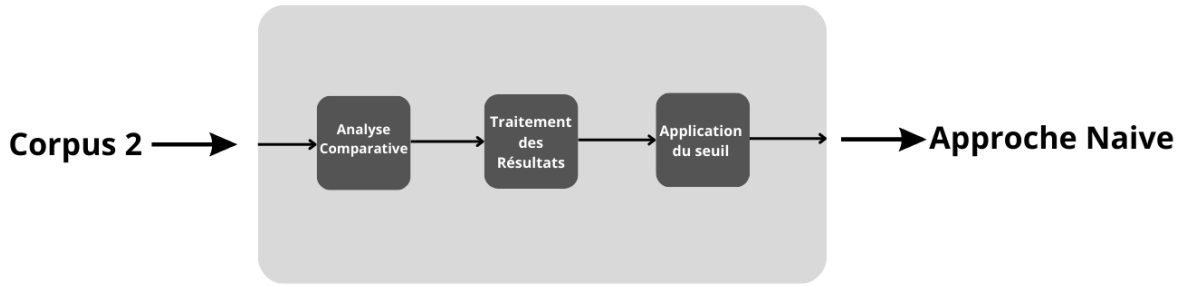
Le nombre de couples à garder est déterminé par le `gold standard`, qui correspond au pourcentage des couples les plus similaires à retenir. Par exemple, si le `gold standard` est à 10%, seules les 10% des couples ayant la plus grande similarité dans l'approche naïve seront marquées comme similaires.

3.1.4 Extraction des résultats

Les couples d'entités dépassant le seuil de similarité sont considérées comme similaires. Ces résultats peuvent ensuite être comparés à un *gold standard* pour une évaluation plus approfondie. Son utilisation permet d'évaluer la similarité en se basant sur le lien entre les entités dans un contexte donné, sans avoir besoin d'analyser le contenu textuel des pages elles-mêmes. Toutefois, cette approche présente certaines limites, puisqu'elle dépend uniquement des liens partagés, plutôt que du contenu réel des textes.

Enfin, pour évaluer la proportion de couples marqués comme similaires, le pourcentage est calculé comme suit :

$$\text{pourcentage\_vrai} = \frac{\text{nombre de couples similaires}}{\text{nombre total de couples}} \times 100$$



3.2 Evaluer notre approche naïve

Afin d'évaluer la performance de l'approche naïve par rapport à la méthode du gold standard, nous avons calculé un score de similarité pour chacune d'entre elles.
Ce pourcentage représente la proportion de couples de personnalités considérés comme similaires par chaque approche.

3.2.1 Méthodologie de l'évaluation

L'évaluation consiste à comparer les résultats de l'approche naïve avec ceux du **gold standard**. Pour chaque couple de personnalités, nous vérifions s'ils sont considérés comme similaires dans les deux approches.
Si un couple est reconnu comme similaire dans les deux cas, nous augmentons le pourcentage.

À la fin du processus, nous obtenons :

— **Le score du gold standard :** il s'agit du pourcentage de couples similaires selon le gold standard.

— **L'efficacité de l'approche naïve :** cela correspond au pourcentage de couples similaires trouvés par l'approche naïve, comparés à ceux identifiés par le gold standard.

3.3 Résultats obtenus

Les résultats de notre évaluation montrent des différences notables entre l'approche naïve et le gold standard, nous allons expliquer pourquoi il existe de telles différences.

3.3.1 Résultats du gold standard

Le gold standard, conçu à partir des catégories**Wikidata**, a identifié un pourcentage plus élevé de couples similaires, atteignant **13,49%**. Cela montre que le gold standard, basé sur des données plus précises, est beaucoup plus efficace que l'approche naïve que nous allons voir ci-dessous.

3.3.2 Résultats de l'approche naïve

L'approche naïve, basée sur les liens similaires entre les couples, a identifié un pourcentage de **13,49%** de couples de personnalités similaires ce qui est similaire au gold standard. Mais cela est normal, en fait le seuil de l'approche naïve est basé sur le résultat du **gold standard**. Néanmoins cela ne veut pas dire que les résultat sont identiques. Pour avoir une vision plus claire, nous allons calculer la comparaison des valeurs entre les deux méthodes.

3.3.3 Comparaison de l'approche naïve avec le gold standard

Enfin, lorsque nous comparons les résultats des deux méthodes, le pourcentage de similarité entre les couples identifiés comme similaires dans les deux approches est de **58.45 %**. Cela montre que la moitié des couples identifiés par l'approche naïve correspondent aux couples identifiés par le gold standard. Ce score montre bien que l'approche naïve est moins performante que le gold standard.

3.4 Analyse des résultats

Les résultats obtenus lors de la comparaison entre l'approche naïve et le gold standard montre plusieurs points intéressants :

3.4.1 Métriques globales

— Exactitude : 0,8892 (88,92%)- Cette valeur forte montre que l'approche naïve est efficace sur une grande proportion des couples, qu'elles soient similaires ou non.

— Précision, Rappel et F-mesure : 0,5845 (58,45%) - Ces valeurs montrent que l'approche naïve a des difficultés à identifier correctement les couples similaires.

3.4.2 Métriques macro

Les métriques macro : Précision, Rappel et F-mesure : 0,7923 (79,23%) sont plus élevées que les métriques globales.
Cela s'explique par la faible proportion de paires similaires dans le Gold Standard, qui ne représente que 13,34% dans notre cas.

3.4.3 Conclusion

L'approche naïve montre de bonnes performances globales, mais elles sont limitées en termes de précision et de rappel pour identifier les couples similaires. Ces résultats montrent qu'il y a une marge d'amélioration possible pour mieux identifier les couples similaires tout en maintenant une bonne performance sur les couples non similaires.

4 Approche intelligente

4.1 Conception de l'approche intelligente

L'approche intelligente sur laquelle nous allons partir pour notre projet repose sur la vectorisation des textes et l'analyse des vecteurs pour évaluer la similarité entre les personnalités sur Wikipédia. Cette approche comporte quatre phases : pour commencer la préparation des données, ensuite la vectorisation des textes, puis le calcul de la similarité entre les vecteurs, et pour finir une réflexion approfondie sur le choix des vecteurs les plus appropriés et les améliorations possibles.

4.1.1 Préparation des données

Avant de pouvoir appliquer TF-IDF, il est important de préparer correctement nos textes :

— Tokenisation : Diviser les textes en mots ou termes individuels. Cela implique de séparer les phrases en mots.

— Suppression des mots vides : Éliminer les mots vides qui n'apportent pas de valeur significative au texte, comme "et", "le", "à", etc.

— Lemmatisation : Réduire les mots à leur forme racine ou base pour améliorer la cohérence des vecteurs.

4.1.2 Vectorisation des textes avec TF-IDF

Une fois nos données préparées, on pourra appliquer TF-IDF pour vectoriser les textes, une méthode utile pour évaluer l'importance des mots dans les documents. Chaque document sera ainsi transformé en un vecteur où chaque dimension correspond à un terme spécifique du corpus, et la valeur représente le score TF-IDF du terme dans ce document.

4.1.3 Mesure de similarité cosinus

Le cosinus de similarité est choisi pour son efficacité à comparer des vecteurs, comme ceux générés par TF-IDF :

Cette mesure évalue l'angle entre deux vecteurs, fournissant un score qui montre leur proximité sémantique. Un score proche de 1 indique une forte similarité, tandis qu'un score proche de 0 indique une dissimilarité.

4.2 Interprétation des résultats

Pour vérifier nos résultat, nous allons utiliser la meme méthode pour l'approche naive, cela nous permettra de comparer le score de notre approche intelligente avec le gold Standard.

4.3 Optimisation et ajustements

Des ajustements seront faits sur les paramètres de TF-IDF basés sur les résultats pour affiner la précision.

— on pourra Analyser l'importance des différents termes inclus dans votre modèle TF-IDF et retirer ceux qui contribuent peu à la variation entre les documents

— Introduction de n-grams : Plutôt que de se limiter aux mots uniques, intégrer des bi-grams ou tri-grams peut capter des phrases ou des expressions spécifiques qui sont plus informatives pour l'analyse.

4.4 Conclusion

Notre approche intelligente, basée sur la vectorisation des textes via TF-IDF et l'analyse de similarité avec le cosinus de similarité, nous permet de mesurer efficacement les relations entre les personnalités sur **Wikipédia** . Cette méthode pourrait nous aide à explorer et à comprendre les nuances sémantiques entre différents textes, offrant une méthode pour des analyses textuelles plus poussées.

5 Conclusion

Cette approche a permis d'explorer la possibilité de déterminer les similarités entre des célébrités. Pour cela nous avons utilisé des notions vues pendant le cours, permettant de dégager des informations non triviales à partir de textes.
En construisant un corpus d'articles Wikipédia et en développant un gold standard, nous avons établi une base solide pour l'évaluation.
Pour le moment nos résultats montrent que notre approche naïve, basée sur les liens Wikipédia, identifie passablement les similarités. Néanmoins, nous remarquons que l'approche naïve ne semble pas les identifier de manière efficace. En réponse à ces limitations nous avons envisagé une approche dite intelligente qui se baserait sur des techniques de vectorisations de textes et de calculs de similarité, pour fournir une analyse plus précise. Mais il est important de souligner que cette approche est encore théorique et que sa mise en place et son efficacité nécessitent une évaluation et que rien nous dit que cela fonctionnera.