RAPPORT DESCRIPTIF : MESURE DE SIMILARITÉ JACCARD

ENSEIGNANT

Dr Serge Sonfack Sounchio

Etudiants

CAOMPAORE Omar Bassy

KAFANDO Aminata

SAWADOGO Jacques Hamidou

Introduction

I. **Objectifs**

II. **Méthodologie et Outils utilisés**

III. **Tests réalisés**

IV. **Résultats**

V. **Perspectives**

Conclusion

Introduction

La mesure de similarité entre deux phrases constitue un enjeu majeur en Traitement Automatique du Langage. Ce projet vise à développer une application capable d'évaluer le degré de similarité entre deux phrases en combinant plusieurs techniques : prétraitement linguistique, enrichissement lexical via WordNet et utilisation d'un indice de similarité de Jaccard amélioré.

I. Objectifs

Pour mener a bien notre projet des objectifs ont été fixé :

- Mettre en place un système capable de comparer deux phrases en termes de sens.

- Appliquer des techniques de prétraitement : suppression des stopwords, lemmatisation, normalisation.

- Exploiter WordNet pour identifier des synonymes et des formes dérivées.

- Améliorer la mesure de similarité en enrichissant l'indice de Jaccard.

- Réaliser des tests afin de valider la fiabilité du système.

II. Méthodologie

1. Prétraitement des phrases

Chaque phrase est convertie en minuscules, débarrassée de la ponctuation, tokenisée, nettoyée des stopwords et lemmatisée.

2. Recherche des synonymes et formes dérivées

WordNet permet d'identifier des synonymes ainsi que des variantes morphologiques des mots, afin d'établir une comparaison plus riche et plus juste.

3. Calcul de la similarité

La similarité est calculée via un Jaccard enrichi prenant en compte les mots équivalents (synonymes, dérivés) et les fréquences (Counter).

4. Outils et bibliothèques utilisés

- Python
    - NLTK : stopwords, lemmatisation, WordNet
    - Collections (Counter)

III. Tests réalisés

Plusieurs tests ont été réalisés :

- phrases identiques ;
- phrases contenant des synonymes ;
- phrases avec variations morphologiques ;
- phrases inversant les rôles sujets/objets ;
- phrases sans aucun lien ;
- tests de robustesse à la casse et à la ponctuation.

IV. Résultats

Les tests montrent que :

- L'ajout de WordNet améliore nettement la prise en compte du sens réel des mots.
- La lemmatisation réduit les variations morphologiques,
- L'indice de Jaccard enrichi offre une comparaison plus fidèle qu'un Jaccard classique.

V. Perspectives

- Intégration de modèles avancés (BERT, SBERT, E5).
- Création d'une interface web ou mobile.
- Enrichissement des ressources lexicales.
- Prise en compte de la structure syntaxique.

- Amélioration des performances pour un traitement à grande échelle.

Conclusion

Ce projet a permis de mettre en place un outil fonctionnel de comparaison sémantique basé sur des techniques linguistiques fondamentales et sur WordNet. Les résultats obtenus sont cohérents et démontrent la pertinence de cette approche pour des tâches telles que la détection de paraphrases ou l'analyse de similarité textuelle.