Université Constantine2 – Abdelhamid Mehri

Faculté des Nouvelles Technologies de l'Information et de la Communication
Département Informatique Fondamentale et ses Applications

Module : WANLP | M1-SDIA

Projets du module WANLP

Liste et description des projets proposés

Projet 01 : Analyse de sentiment sur les réseaux sociaux en langue arabe

Objectif :
Développer une plateforme web capable d'analyser les sentiments exprimés dans les publications sur les réseaux sociaux en arabe. La plateforme devra extraire les tweets ou les publications Facebook liés à un sujet spécifique et déterminer si les sentiments exprimés sont positifs, négatifs ou neutres.

Fonctionnalités clés :
- Extraction automatique des publications en langue arabe en fonction de mots-clés.
- Prétraitement des textes pour normaliser la langue et gérer les spécificités.
- Classification des sentiments à l'aide de modèles d'apprentissage automatique.
- Visualisation des résultats sous forme de graphiques et de statistiques sur l'interface web.

Sources et ressources à utiliser :
- Corpus : Collecte de publications sur les réseaux sociaux en arabe.
- Lexique : Utilisation de lexiques de sentiments en arabe pour l'analyse de sentiments.
- Réseaux Sociaux : Extraction de données à partir de Twitter, Facebook ou d'autres plateformes populaires.

Projet 02 : Classification automatique d'articles de presse en arabes

Objectif :
Créer une plateforme web qui classifie automatiquement les articles de presse en langue arabes selon des catégories thématiques prédéfinies (politique, sport, culture, etc.). La plateforme devra identifier et classer les articles après reconnaissance du sujet ou de la thématique.

Fonctionnalités clés :
- Prétraitement des articles pour extraire les caractéristiques linguistiques.
- Utilisation de techniques de classification pour attribuer un article à une catégorie spécifique.
- Interface web permettant aux utilisateurs de soumettre des articles et de visualiser les résultats de classification.

Sources et ressources à utiliser :
- Corpus : Collecte d'articles de presse en langue arabes.
- Datasets Structurés : Utilisation de datasets annotés avec les catégories thématiques et les dialectes correspondants.
- Bases de Données Lexicales : Utilisation de bases de données lexicales spécifiques à la langue arabes pour la classification.

Projet 03 : Résumé automatique de textes en arabe

Objectif :
Développer une plateforme web qui génère des résumés concis et pertinents à partir de longs articles ou documents en arabe littéraire. Le système devra extraire les points clés du texte et les présenter sous forme de résumé.

Fonctionnalités clés :

- Prétraitement du texte pour identifier les phrases et les mots clés importants.
- Utilisation d'algorithmes de réduction de la redondance pour générer un résumé cohérent.
- Interface web permettant aux utilisateurs de soumettre des textes et de recevoir des résumés.

Sources et ressources à utiliser :

- Corpus : Collecte de longs documents ou articles en arabe littéraire et en dialectes.
- Datasets Structurés : Utilisation de datasets contenant des paires de textes et de résumés pour l'entraînement des modèles.
- Bases de Données Lexicales : Utilisation de bases de données lexicales en arabe pour l'identification des mots-clés.

Projet 04 : Reconnaissance et désambiguïsation d'entités nommées en arabe

Objectif :
Construire une plateforme web qui identifie et désambiguïse les entités nommées (personnes, lieux, organisations) dans des textes en arabe, en exploitant des bases de données lexicales adaptées à la langue arabe et à ses dialectes.

Fonctionnalités clés :

- Prétraitement du texte pour extraire les entités nommées potentielles.
- Utilisation de bases de données lexicales pour désambiguïser les entités en fonction du contexte.
- Interface web permettant aux utilisateurs de soumettre des textes et de visualiser les entités nommées identifiées.

Sources et ressources à utiliser :

- Corpus : Collecte de textes en arabe contenant des entités nommées.
- Bases de Données Lexicales : Utilisation de bases de données lexicales en arabe pour la désambiguïsation des entités.
- Datasets Structurés : Utilisation de datasets annotés pour l'entraînement des modèles de reconnaissance d'entités.

Projet 05 : Traduction automatique entre l'arabe littéraire et les dialectes Algériens

Objectif :
Développer sur une plateforme web pour la traduction automatique entre l'arabe littéraire et différents dialectes algériens. La plateforme devra prendre en compte la compréhension sémantique et la préservation du contexte dans le processus de traduction.

Fonctionnalités clés :

- Prétraitement des textes pour gérer les spécificités linguistiques de l'arabe littéraire et des dialectes.
- Utilisation de modèles de traduction automatique adaptés à la variété linguistique de la langue arabe.
- Interface web permettant aux utilisateurs de saisir des textes en arabe littéraire ou en dialecte et de recevoir la traduction correspondante.

Sources et ressources à utiliser :

- Corpus : Collecte de paires de textes en arabe littéraire et en dialectes pour l'entraînement des modèles de traduction.
- Bases de Données Lexicales : Utilisation de bases de données lexicales bilingues pour la traduction entre l'arabe littéraire et les dialectes.

Projet 06 : Chatbot intelligent pour le service client en arabe

Objectif :

Créer un chatbot intelligent intégré à une plateforme web qui répond aux questions fréquemment posées par les clients en arabe, en utilisant des techniques de traitement du langage naturel pour comprendre les requêtes et fournir des réponses pertinentes.

Fonctionnalités clés :

- Analyse des questions et requêtes des utilisateurs en langue arabe pour en extraire l'intention et les mots-clés.
- Génération automatique de réponses pertinentes basées sur une base de connaissances préétablie.
- Intégration du chatbot dans une plateforme web pour une interaction en temps réel avec les utilisateurs.

Sources et ressources à utiliser :

- Corpus : Collecte de dialogues et de questions-réponses en arabe pour l'entraînement du chatbot.
- Bases de Données Lexicales : Utilisation de bases de données lexicales en arabe pour la compréhension des requêtes des utilisateurs.
- Réseaux Sociaux : Intégration avec des plateformes de réseaux sociaux pour une interaction en temps réel.

Projet 07 : Analyse de la cohérence textuelle en arabe

Objectif :

Développer une plateforme web qui évalue la cohérence et la logique d'un texte en arabe, en identifiant les passages incohérents ou contradictoires et en suggérant des améliorations pour renforcer la clarté du discours.

Fonctionnalités clés :

- Prétraitement du texte pour analyser la structure syntaxique et sémantique des phrases.
- Utilisation d'algorithmes pour détecter les incohérences et les contradictions dans le texte.
- Interface web permettant aux utilisateurs de soumettre des textes et de recevoir des suggestions d'amélioration.

Sources et ressources à utiliser :

- Corpus : Collecte de textes en arabe pour l'analyse de la cohérence.
- Datasets Structurés : Utilisation de datasets annotés indiquant les incohérences dans les textes.
- Bases de Données Lexicales : Utilisation de bases de données lexicales en arabe pour l'analyse sémantique des textes.

Projet 08 : Génération automatique de questions à partir de textes en arabe

Objectif :

Développer une plateforme web capable de générer automatiquement des questions pertinentes à partir d'un passage de texte en arabe, utile pour la création de quiz ou d'exercices éducatifs.

Fonctionnalités clés :

- Prétraitement du texte pour identifier les informations clés et les concepts importants.
- Développement d'algorithmes pour générer différents types de questions en fonction du contenu du texte.
- Interface web permettant aux utilisateurs de soumettre des textes et de recevoir les questions générées.

Sources et ressources à utiliser :

- Corpus : Collecte de textes en arabe accompagnés de questions pertinentes.
- Datasets Structurés : Utilisation de datasets contenant des paires de textes et de questions pour l'entraînement des modèles.
- Bases de Données Lexicales : Utilisation de bases de données lexicales en arabe pour l'identification des concepts clés dans les textes.

Projet 09 : Détection de plagiat dans les documents en arabe

Objectif :

Développer une plateforme web de détection de plagiat capable d'analyser des documents textuels en arabe et d'identifier les similitudes significatives qui pourraient indiquer une possible tentative de plagiat.

Fonctionnalités clés :

- Prétraitement des documents pour normaliser le texte et extraire les caractéristiques pertinentes.
- Utilisation d'algorithmes pour comparer les documents et calculer les scores de similarité.
- Génération de rapports de plagiat indiquant les passages suspects et les sources potentielles de plagiat.
- Interface web permettant aux utilisateurs de soumettre des documents et de visualiser les résultats de l'analyse de plagiat.

Sources et ressources à utiliser :

- Corpus : Collecte de documents en arabe pour l'analyse de similarité.
- Datasets Structurés : Utilisation de datasets contenant des paires de documents originaux et plagiés pour l'entraînement des modèles.
- Bases de Données Lexicales : Utilisation de bases de données lexicales en arabe pour la normalisation des textes.

Remarques importantes pour tous les projets

1. **Bonus pour l'intégration de l'Arabizi :** Chaque projet doit se concentrer sur une tâche spécifique de traitement du langage naturel en langue arabe. L'exploration et l'utilisation de l'arabizi sont encouragées lorsque cela est pertinent pour le projet. Un bonus sera attribué aux équipes qui intègrent de manière innovante ce type de texte dans leur travail.

2. **Bonus pour la collecte de données :** Les données utilisées comme entrée pour chaque projet peuvent être obtenues à partir de diverses sources et ressources en langue arabe, telles que les lexiques, les datasets structurés, les bases de données lexicales et les réseaux sociaux. Si ces ressources ne sont pas directement disponibles, les équipes sont encouragées à collecter leurs propres données en utilisant des méthodes reconnues en NLP. Un bonus sera accordé aux équipes qui fournissent des efforts supplémentaires pour cette tâche de collecte de données.

Livrable, Calendrier, Consignes et Evaluation des projets

Livrables

1. **Code Source :** Un dépôt Git contenant tout le code source du projet, bien commenté et organisé, avec des instructions pour la configuration et l'exécution.
2. **Présentation PowerPoint :** Une présentation détaillant l'approche utilisée, les méthodes de prétraitement du texte, les algorithmes spécifiques au projet, et les résultats obtenus, mettant en évidence la contribution de chaque membre de l'équipe.
3. **Plateforme Web :** Une interface web conviviale permettant aux utilisateurs d'interagir avec le système développé, de soumettre des données et de visualiser les résultats.

Évaluation

1. **Présentation du Projet :** 30 minutes de présentation suivies de 10 minutes de questions-réponses et de débat. Évaluation basée sur la clarté, l'organisation, la qualité technique et la capacité à répondre aux questions.
2. **Qualité du Code :** Évaluation de la clarté, de la structure, de l'efficacité et de la documentation du code source.
3. **Fonctionnalité du Système :** Évaluation de la précision, de la pertinence et de la diversité des fonctionnalités développées par le système.
4. **Contribution Individuelle :** Évaluation de la participation et de la contribution de chaque membre de l'équipe.
5. **Pondération sur la Note Finale :** Le projet sera comptabilisé dans l'évaluation continue du module WANLP, L'évaluation continue comptera pour **40% de la note** finale dans la note finale du module.

Calendrier sur 10 semaines

- Semaines 1-2 : Compréhension, prise en main, planification et répartition des taches du projet.
- Semaines 3-4 : Recherche et sélection des sources, outils et librairies, début du développement.
- Semaines 5-6 : Développement des fonctionnalités principales du système.
- Semaines 7-8 : Intégration des fonctionnalités et développement de l'interface web.
- Semaines 9-10 : Améliorations et finalisation du projet. Préparation de la présentation.

Outils et Librairies à Utiliser

- Python : Langage de programmation principal.
- NLTK (Natural Language Toolkit) : Bibliothèque Python pour le traitement du langage naturel.
- SpaCy : Bibliothèque Python pour des tâches avancées de NLP.
- Scikit-learn : Bibliothèque Python pour l'apprentissage automatique.
- Pandas : Bibliothèque Python pour la manipulation et l'analyse de données.
- Flask ou Django : Frameworks Python pour le développement de l'interface web.
- Matplotlib et Seaborn : Bibliothèques Python pour la visualisation de données.

Ressources Supplémentaires

- Documentation officielle des librairies Python utilisées.
- Support de Cours, TD et TP du module WANL
- Support de cours et autres ressources et sur l'Apprentissage Automatique.
- Accès à des corpus et des datasets spécifiques à la langue arabe et à ses dialectes.
- Forums et communautés en ligne pour le partage de connaissances et la résolution de problèmes liée aux projets sur le NLP.