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

Enoncés du TP 04

Analyse Lexicale, Syntaxique, Sémantique, Extraction d'Entités et Analyse des sentiments avec spaCy

Enoncés de l'exercice

Objectif

Utiliser la bibliothèque **spaCy** en Python pour effectuer une analyse NLP complète sur le paragraphe suivant, en appliquant les concepts et les techniques du cours présenté dans le chapitre 3.

Paragraphe

"Le ministre de l'Économie a annoncé à Alger des réformes importantes pour stimuler l'investissement dans le secteur des technologies de l'information et de la communication, ce qui a suscité un vif enthousiasme parmi les entrepreneurs algériens. Cependant, certains experts restent sceptiques quant à l'impact de ces mesures sur le marché du travail. Par ailleurs, la banque centrale a révisé ses taux d'intérêt pour encourager l'épargne, une décision qui a été accueillie avec prudence par les acteurs financiers. En outre, la ville d'Oran se prépare activement à accueillir le prochain sommet international sur les énergies renouvelables, un événement qui promet de mettre en lumière les avancées de l'Algérie dans ce domaine. Durant la conférence, le ministre a levé son verre en l'honneur de l'innovation, tandis qu'un assistant ajustait le poignet de sa chemise avant de lui remettre un livre contenant la lettre officielle d'invitation."

الفقرة باللغة العربية

"أعلن وزير الاقتصاد بالجزائر العاصمة عن إصلاحات مهمة لتحفيز الاستثمار في قطاع تكنولوجيا المعلومات والاتصال، الأمر الذي أثار حماسا كبيرا لدى رواد الأعمال الجزائريين. إلا أن بعض الخبراء ما زالوا يشككون بشأن تأثير هذه الإجراءات على سوق العمل. علاوة على ذلك، فإن البنك المركزي قام بمراجعة أسعار الفائدة لتشجيع الادخار، وهو القرار الذي استقبل بحذر من قبل الفاعلين في مجال المال و الأعمال، وبالإضافة إلى ذلك، تستعد مدينة وهران بنشاط لاستضافة القمة الدولية المقبلة حول الطاقات المتجددة، وهو الحدث الذي يعد بتسليط الضوء على التقدم الذي أحرزته الجزائر في مجال الطاقة المتجددة. وخلال المؤتمر رفع الوزير كأسه تكريما للابتكار فيما قام أحد المساعدين بتعديل كف قميصه قبل أن يسلمه كتابا يحتوي على رسالة الدعوة الرسمية. "

Questions

1. **Tokenisation et Analyse Morphologique :**

- Tokenisez le paragraphe et affichez chaque token avec sa catégorie morphologique.
- Identifiez les racines, préfixes et suffixes des mots "réformes", "enthousiasme" et "épargne".

2. **Analyse Syntaxique :**

- Analysez la structure syntaxique du paragraphe et affichez les relations de dépendance entre les mots.
- Visualisez l'arbre de dépendance syntaxique du paragraphe.

3. **Exploration des Mots Polysémiques ou Homonymes et Désambiguïsation Sémantique :**

- Identifiez les mots polysémiques ou homonymes dans le paragraphe.
- Discutez de la manière dont le contexte pourrait aider à déterminer leur sens spécifique. (Note : Cette tâche sera plus théorique, car spaCy ne fournit pas de fonctionnalité directe pour la désambiguïsation sémantique.)

4. **Extraction d'Entités Nommées :**

- Extrayez les entités nommées du paragraphe et identifiez leur type (par exemple, personne, lieu, organisation, etc.).
- Affichez les entités extraites et leurs types.

5. **Exploration et Analyse du Sentiment :**

- Identifiez les phrases ou expressions qui semblent exprimer un sentiment positif ou négatif.
- Explorez la manière dont vous pourriez utiliser spaCy, en combinaison avec une approche externe ou un modèle supplémentaire, pour évaluer le sentiment global du paragraphe.
  - Pour analyser le sentiment en français, vous pouvez utiliser le modèle pré-entraîné **CamemBERT**, basé sur l'architecture **BERT** et adapté à la langue française. Utilisez la bibliothèque **`transformers`** pour charger le modèle **CamemBERT** et son **tokenizer**.
  - Assurez-vous d'avoir installé les bibliothèques `transformers` et `torch` pour travailler avec CamemBERT.
  - Préparez le texte que vous souhaitez analyser en le tokenisant à l'aide du tokenizer de **CamemBERT**, puis passez-le au modèle pour obtenir les scores de sentiment.
  - Appliquez la fonction **softmax** aux scores obtenus pour convertir les **logits** en **probabilités**, qui représentent la probabilité que le texte exprime un sentiment positif ou négatif. Interprétez ces scores pour déterminer le sentiment global du texte. Les scores vous donneront une indication sur la tendance du sentiment exprimé par le texte.

Rappels

1. **Question 3 :** La différence principale entre les mots polysémiques et les homonymes en français est la suivante :

- Mots polysémiques : Un mot ayant plusieurs sens liés ou dérivés d'une même origine. Exemple : "vol" (action de voler ou acte de dérober).
- Homonymes : Des mots qui se prononcent ou s'écrivent de la même manière mais ont des origines et des sens différents. Exemple : "verre" (récipient) et "vert" (couleur).

2. **Question 5 :** Les **logits** sont des termes utilisés en apprentissage automatique, en particulier dans le contexte des réseaux de neurones et de la classification, ce sont les scores bruts non normalisés obtenus à partir de la dernière couche d'un modèle de réseau de neurones avant d'appliquer une fonction d'activation comme la fonction **softmax**

Consignes :

- Assurez-vous d'avoir installé spaCy et téléchargé le modèle de langue française.
- Importez spaCy et chargez le modèle de langue au début de votre script Python.
- Utilisez l'objet **`nlp`** pour traiter le paragraphe et créer un objet **`Doc`.**
- Parcourez les tokens dans l'objet **`Doc`** pour accéder à leurs attributs et réaliser les analyses demandées.
- Utilisez les outils de visualisation de spaCy pour afficher l'arbre de dépendance syntaxique.