Propositions de sujets pour le 3e travail pratique

Automne 2020
Luc Lamontagne

Sujet 1 – Choix libre de sujets

L'une des options suivantes:

- Mener une expérimentation en lien avec votre sujet de recherche de maîtrise ou de doctorat. Cela doit être pertinent au TALN (NLP).
- Mener une expérimentation reliée à votre emploi. Cela doit être pertinent au TALN.
- Reproduire les résultats d'un article du domaine du traitement automatique de la langue naturelle et faire une analyse de ces résultats.
- Faire une revue de littérature d'un sujet qui vous intéresse. Il est surprenant de constater que cette option s'avère souvent plus difficile qu'on le croit.

Date limite pour me soumettre votre choix de projet: au plus tard le 1er décembre.
Format de la proposition : Document PDF décrivant en quelques phrases ce que vous voulez faire.
Remise de la proposition : Déposez le document PDF dans la boîte de remise du 3e travail au plus tard le 1er décembre.
Mise en garde : Éviter de choisir un sujet qui nécessiterait un temps considérable à accomplir (par ex. 100 heures de travail) ou un sujet trop complexe. Je pourrai vous donner une rétroaction sur votre proposition.

Sujet 2 – Reconnaissance d'entités nommées (NER)

Il est possible d'entraîner un classificateur de séquences afin de repérer les entités nommées d'un domaine d'application particulier (par ex. des concepts scientifiques ou de la médecine). On vous propose ici de mener ici une expérimentation afin d'entraîner un extracteur d'information à partir de textes d'entraînement.

2 corpus de textes seront rendus disponibles pour ce thème. Un premier s'inspire de la compétition *SemEval* 2017 et vise à repérer dans des résumés d'articles scientifiques des mots clés désignant des méthodes, des tâches, des matériaux et des termes scientifiques. Un 2e corpus contient des noms de maladies que vous devrez apprendre à extraire.

Vous avez totale liberté sur le choix d'approche à adopter. Une approche simple serait d'utiliser Spacy qui permet l'entraînement d'un module NER. Une autre serait de récupérer du code sur le Web et de l'adapter afin de mener votre expérimentation. Vous pouvez réutiliser du code, mais il est important de décrire clairement votre démarche dans votre rapport. L'objectif ici est de se familiariser avec ce domaine en d'en savoir plus sur les classificateurs de séquences, pas de battre le *state-of-the-art*.

Sujet 3 – Système question-réponse

Le chapitre 8 du livre *Taming Text* (voir document PDF sur le site du cours) décrit une implémentation d'un système de question-réponse en Java en s'appuyant principalement sur *OpenNLP* et *Solr*. La proposition consiste à refaire en Python les différents modules décrits dans ce chapitre de livre. Un exemple de configuration logicielle qui pourrait être utilisée pour accomplir cette tâche est :

- Classificateur de questions : Scikit-learn ou PyTorch
- Moteur de recherche : *Whoosh* ou *ElasticSearch*
- Collection de documents: un sous-ensemble de Wikipédia
- Reconnaissance d'entités nommées et prétraitement de textes*: Spacy*

Contre-exemples de sujet valide – Analyse de sentiment, classification de textes avec Scikit-learn et classification de noms

Comme nous avons épuisé le thème de l'analyse de sentiment dans le 2e travail pratique, ce sujet n'est pas autorisé pour le 3e travail pratique.

Dans le même ordre d'idée, si vous souhaitez mener des travaux en classification de textes (au niveau des mots ou des caractères), il vous faudra obtenir mon approbation auparavant pour m'assurer de l'originalité du sujet et d'éviter de répéter ce qui a été fait dans le cours.

Livrables à remettre pour ce travail pratique (valide pour tous les sujets)

a) Un rapport entre 5-10 pages indiquant:
   - La démarche que vous avez préconisée pour chacune des étapes de votre travail.
   - Une description de vos choix logiciels et du code que vous avez développé.
   - Une évaluation qualitative ou quantitative des résultats que vous obtenez.
   - Des exemples de résultats positifs et négatifs que vous avez obtenus.

b) Le code que vous avez développé pour mener ce projet. Il n'y a aucune contrainte particulière pour ce travail pratique. Vous avez la liberté de choisir les librairies que vous jugerez utiles pour réaliser vos travaux. Le but est d'apprendre, de découvrir, d'élargir vos horizons. Merci de me consulter en cas de doute.

Date de remise du 3e travail : Le vendredi 18 décembre 2020