Quelques suggestions de travaux

Proposé par Luc Lamontagne
Automne 2018

Instructions :

- Travail en équipe de 1 à 3 personnes.
- Objectifs :
    - Mener une expérimentation pour approfondir vous connaissance dans un sujet relié à la matière étudiée dans le cours.
    - Se familiariser avec un nouveau domaine du traitement automatique de la langue naturelle.
    - Entreprendre ou poursuivre des travaux de recherche reliés à un projet de maîtrise ou de doctorat.
- Pour les étudiants souhaitant aborder un sujet libre, me faire parvenir une description par courriel d'ici le **27 novembre prochain** pour obtenir mon approbation ou des suggestions.
- Pour ceux qui n'ont pas choisi de sujet, vous trouverez dans les prochains paragraphes quelques propositions de travaux.
- Rapport et logiciel (le cas échéant) à remettre le **17 décembre**.
- Ce travail est noté sur 100 et vaut 20% de la note de session.

1. Système question-réponse (QA)

Le chapitre 8 du livre *Taming Text* (voir document sur le site du cours) décrit une implémentation de système QA en Java en s'appuyant principalement sur *OpenNLP* et *Solr*.

Proposition : Refaire en Python les différents modules décrits dans ce chapitre de livre. Un exemple de configuration logicielle qui pourrait être utilisé pour accomplir cette tâche est :

- Moteur de recherche : *Solr* ou *ElasticSearch*
- Collection: un sous-ensemble de Wikipédia est disponible avec le code de *Taming Text*.
- Classification de texte : Scikit-learn
- Prétraitement : NTLK, *Spacy* ou *Stanford Core NLP* (avec interface Python)

Dans votre rapport, je vous demande d'aborder les points suivants :

- Expliquez la démarche que vous avez suivie.
- Indiquez les difficultés que vous avez rencontrées et les limitations de votre approche.
- Construisez une banque de questions et évaluez la performance du système. Présentez les résultats de votre évaluation.

À remettre : votre code, votre rapport, la collection de documents et le jeu de test que vous avez utilisés.

2. Classification de texte et analyse de sentiments dans les conversations

EmoContext est une compétition actuellement ouverte qui consiste à faire l'analyse de conversations afin de détecter les émotions qui y sont présentes. Voir le lien suivant pour plus d'informations : https://www.humanizing-ai.com/emocontext.html

Proposition : Utilisez le corpus rendu disponible pour cette compétition afin de mener une expérimentation. Le corpus est disponible sur le site du cours. Vous pourriez réutiliser ce que vous avez fait pour le Travail pratique 2 (TP2) et l'adapter afin de traiter ces séquences d'échanges. Sinon vous pourriez en profiter pour explorer de nouveaux outils et comparer les résultats avec ceux développés au TP2.

Dans votre rapport, vous devriez aborder les points suivants :

- Expliquez la démarche que vous avez suivie.
- Indiquez les difficultés que vous avez rencontrées et les limitations de votre approche.
- Présentez les résultats que vous avez obtenus avec les jeux de données de la compétition.

À remettre : votre code et votre rapport.

3. Systèmes de dialogue : comparaison de *ParlAI* vs. *PyDial* et expérimentations.

De nouveaux environnements pour mener de travaux de recherche dans le domaine des systèmes de dialogues ont été rendus disponibles au cours de la dernière année. Je vous propose une étude exploration afin de vous familiariser avec deux de ces systèmes :

- La plateforme *ParlAI* de Facebook est un environnement qui offre différentes approches (par ex. recherche d'information, *memory networks*, réseau profond LSTM) pour étudier différentes approches de dialogue automatisé. De plus, plusieurs jeux de données sont disponibles sur le site. Voir https://github.com/facebookresearch/ParlAI

- Le système *PyDial*, développé par une équipe de recherche de l'Université Cambridge, est une librairie qui fournit de nombreux outils pour étudier les approches récentes dans les systèmes de gestion dialogue. Voir http://www.camdial.org/pydial/Docs/

Proposition : Je vous propose de mener une étude en 2 étapes :

- Prenez connaissance des 2 environnements et décrivez-les sommairement dans votre rapport (les grandes lignes de chacun). De plus, expliquez les différences et les recoupements que vous retrouvez dans ces plateformes.
- À l'aide d'un de ces 2 environnements (par ex. une tâche spécifique de dialogue fourni par *ParlAI*), menez une expérimentation et expliquez les résultats que vous avez obtenus.

À remettre : votre code, votre rapport et le corpus utilisé pour mener vos expérimentations.

Attention : Le niveau de difficulté de ce sujet est assez élevé et exige un bon niveau de familiarité avec les domaines de l'intelligence artificielle et l'apprentissage automatique.

4. Extraction d'informations

Dans le cadre de la compétition *SemEval* 2017, un corpus a été rendu disponible pour mener des expérimentations en extraction d'information et de classification de relations. Le corpus contient plusieurs résumés d'articles scientifiques et la tâche consiste à repérer des mots clés de différentes catégories (par ex. algorithmes ou procédé). Un corpus similaire est également disponible pour la compétition. Voir les liens suivants :

https://scienceie.github.io
https://competitions.codalab.org/competitions/17422#learn_the_details-overview

Proposition : Utiliser ce corpus afin de mener une expérimentation en extraction d'information. Je vous laisse entière liberté sur le choix des outils à utiliser. Si vous réutilisez un logiciel conçu spécialement pour cette compétition, je vous demande d'apporter quelques modifications et de mesurer l'impact de ces modifications aux résultats obtenus.

À remettre : votre code et votre rapport.

***Attention* :** Le niveau de difficulté de cette tâche peut être également élevé.