université de Caen
Basse-Normandie

TP « NLP & KM »

Master 2 IDC

Frédérik Bilhaut (frederik.bilhaut@unicaen.fr)

La finalité du TP est de réaliser un moteur de recherche sémantique en tirant parti d'un modèle de *word embeddings*. et/ou d'une extraction des entités nommées. Fonctionnellement, l'objectif et de permettre à l'utilisateur d'un moteur de recherche :

- De trouver des résultats par similarité de sens. Par exemple en cherchant le mot « vélo » on s'attend à trouver des documents qui contiennent des mots comme « cycle », « VTT », etc. Pour rappel, les word embeddings permettent de représenter des mots par des vecteurs qui peuvent être comparés pour évaluer leur proximité sémantique.

- De trouver des résultats par entité nommée. Par exemple je recherche les documents qui parlent de « Apple » en tant qu'entreprise et non pas du fruit. Pour rappel, l'extraction d'entités nommées (ou NER) permet de repérer les segments qui désignent ces entités et de les désambiguiser.

Pour ce faire nous n'allons bien sûr pas tout ré-écrire en partant de zéro : nous allons utiliser des briques existantes :

- Pour la partie NLP : Spacy

- Pour la partie moteur de recherche, au choix : soit une API publique de recherche, soit Whoosh.

D'autres composants pourront éventuellement être utilisées pour parfaire le système si vous souhaitez aller plus loin.

Dans un premier temps les modèles « par défaut » de Spacy seront utilisés, dans un second temps on cherchera autant que possible à personnaliser les modèles sur votre propre corpus, en particulier pour les *embeddings*.

Etapes du travail

- Se familiariser avec Spacy : installer le SDK, un ou plusieurs modèles, consulter la documentation (type « Getting started »), faire quelques expériences « à la main » sur des petits textes. Essayer notamment la récupération des entités nommées, et la proximité vectorielle (cf. « Vectors.most_similar »).

- Faire son choix en termes de moteur de recherche. Une possibilité est d'exploiter une API publique du type Emploi Store, par exemple. Une autre possibilité est d'implémenter votre propre moteur avec Whoosh. Dans ce cas il vous faudra également récupérer un corpus pour tester, par exemple celui du Grand débat. Rq : la solution de faire votre propre moteur est un peu plus complexe, mais offre beaucoup plus de possibilités (pour l'entraînement de votre propre modèle, utiliser les entités comme facettes, etc.).

- Implémenter l'expansion de requêtes de la manière qui vous paraît la plus appropriée selon la solution que vous avez retenue. Dans tous les cas, l'idée est « d'augmenter » automatiquement la requête de l'utilisateur avec des mots proches en termes d'embeddings.

- Implémenter la recherche par entités. L'idée est que l'utilisateur spécifique le nom et le type d'une entité recherchée, et de filtrer les résultats de recherche en conésquences.

- Apprendre vos propres modèles.