Université de Montréal

IFT6285 (TALN) — Devoir10
Extraction d'information

Contact :
Philippe Langlais
RALI/DIRO
Université de Montréal
+1 514 343 61 11 ext: 47494
felipe@iro.umontreal.ca
http://www.iro.umontreal.ca/~felipe/

dernière compilation : 28 novembre 2020 (15:26)

Contexte

L'extraction d'information est un sujet d'intérêt qui revêt différentes technologies (depuis les expressions régulières jusqu'à des étiqueteurs neuronaux) et est utile à différentes applications (recherche d'information, text2Ontology, etc.). Pour ce dernier sujet, vous allez réaliser un extracteur très simple dans la lignée de ceux demandés dans le projet 2. Les données sur lesquelles vous allez l'appliquer sont les **phrases définitoires** de Wikipedia, c'est à dire la première phrase d'un article : comme *Jarvis Cocker, également surnommé Darren Spooner, est un chanteur, musicien, animateur de radio britannique, né le 19 septembre 1963, à Sheffield en Angleterre, et leader du groupe britpop Pulp.* pour l'article Jarvis_Cocker.

À faire

1. Utilisez l'API wikipedia afin de consulter des articles de Wikipedia, soit répondant à une requête particulière (`search`) soit pris au hasard (`random`). Attention, cette API est simple d'usage, mais accède à l'information d'une page via requêtes HTTP (lent). Pour chaque article, récupérez la phrase définitoire que vous analyserez en dépendances avec Spacy. Vous pouvez utilisez displacy pour visualiser les arbres correspondants.
2. Parcourez (programmatiquement) les analyses en dépendances ainsi produites afin d'y rechercher la racine r de l'arbre qui domine (via une relation `cop`) un mot ayant pour lemme *être* ainsi qu'un mot (ou groupe de mots) s via la relation `nsubj`. L'extraction s IS_A r sera alors proposée. Il s'agit d'un extracteur très simple qui ne fonctionnera pas à tout coup. Par exemple pour la phrase définitoire donnée en exemple, on peut extraire *(Jarvis Cocker, IS_A, chanteur)* de l'analyse en dépendances suivante où *chanteur* est la racine qui gouverne la copule *est* ainsi que l'entité nommée *Jarvis Cocker*. (les spans d'entités nommées ont ici été regroupés en un seul token).
3. Produisez un fichier au format texte extractions qui contient dans un format texte, [1] le titre de l'article, la phrase définitoire et les relations

1. format non normatif.

extraites de cette phrase. Si vous n'améliorez pas l'extracteur proposé, alors au plus une extraction sera proposée par phrase définitoire. Ce fichier doit contenir au moins 100 extractions.

4. Analysez les extractions produites. Améliorez éventuellement votre extracteur.

Vous devez produire un **rapport** d'au plus 3 pages (format pdf, en français ou en anglais) résumant vos découvertes ainsi qu'un fichier contenant les extractions.

Remise

La remise est à faire sur Studium sous le libellé `devoir10`. Vous devez remettre votre rapport (format pdf, texte en anglais ou en français) ainsi qu'un fichier d'extractions dans une archive (gzip, tar, tar.gz) dont le nom est préfixé de `devoir10-name1` ou `devoir10-name1-name2` selon que vous remettez seul ou a deux, où `name1` et `name2` sont à remplacer par l'identité des personnes faisant la remise (`prénom_nom`). Assurez vous que le nom des personnes impliquées dans le devoir soit indiqué sur tous les documents remis (code et ressources). Le devoir est à remettre en groupe d'au plus deux personnes au plus tard dimanche 13 décembre à 23h59.