Université de Montréal

IFT6285 (TALN) — Projet 2
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 : 23 novembre 2020 (21:22)

Contexte

Dans le projet 1, vous avez étudié un aspect important du TALN qu'est la classification. Dans ce projet vous allez être aux prises avec un autre aspect du TALN qui est moins documenté académiquement : celui d'apprivoiser un *dataset* et d'en extraire des informations pertinentes. On parle le plus souvent d'**extraction d'information** pour désigner un ensemble de techniques visant à extraire du texte toute sorte d'information en vue d'aider à structurer ce texte ou à enrichir une base de connaissance.

Je vous encourage vivement dans ce projet à utiliser la plateforme Spacy qui vous offre des briques de base du traitement des langues (découpage de phrases, de mots, reconnaissance d'entités nommées, analyse syntaxique, etc.) pour le français. Une API de recherche de patrons vous permet de plus d'écrire des patrons portant sur des séquences de mots et leurs caractéristiques (étiquette morphosyntaxique, forme orthographique, relations de dépendance, etc.). Plus à ce sujet ici.

Données

Vous avez accès à un corpus d'environ 10k nouvelles publiées par le devoir entre septembre 2018 et août 2019 (un fichier json par nouvelle). Je vous demanderai pour des raisons que j'exposerai en cours de ne pas utiliser ce *dataset* à d'autres fins que le présent projet. Chaque nouvelle porte sur un sujet particulier (ex : sport) et vous trouverez dans les métadonnées quelques informations pertinentes. Les données sont téléchargeables ici (18Mo) et sont également installées sur les machines du DIRO dans le répertoire (certains éléments de ce path ne sont pas en lecture, mais les ressources le sont) :
`/u/felipe/HTML/IFT6285-Automne2020/__243cfdjsdh/ift6285-le-devoir-articles/`

qui contient des fichiers avec des noms amicaux comme :

`ffb79bcb6aafe8ba5f72672a12a8e79a064090e600e3a397b7a484fe5db4b445.json`
`ffba9c9b61593b611ae07cda2e312bd26c4c29e610751d6b793a3969af634917.json`

Toutes les nouvelles sont au format json, encodées en UTF-8.

Voici un exemple de nouvelle déformatée (texte) où les métadonnées sont indiquées en gras :

> **date** : 2019-06-21T00 :00 :00-04 :00
> **labels** : Moyen-Orient | Académie française | Égypte (Pays) | pays arabe
> **breadcrumb** : /lire
> **url** : https://www.ledevoir.com/lire/557204/le-bateau-ivre-d-amin-maalouf
> Le 9 juin 1967, à 18 h 50, heure du Caire, la radio d'État égyptienne diffuse le discours de démission du président Gamal Abdel Nasser, au sortir de ce qu'on baptisera la « guerre des Six Jours ». À 1200 kilomètres de la « Voix des Arabes », à l'Université Saint-Joseph de Beyrouth, un jeune homme attend ses résultats d'examens. . . .

Et voici les 5 *breadcrumb* les plus fréquents dans le *dataset* :
`/societe` (1095), `/economie` (1038), `/sports` (956), `/politique/quebec` (701), `/politique/canada` (648).

Les méta-données peuvent vous servir pour évaluer (partiellement) vos extracteurs, mais dans le projet, vous allez vous concentrer sur les données non structurées (le contenu textuel des nouvelles).

À faire

Votre but est d'extraire le plus d'informations fiables possible des textes distribués. On se concentrera ici sur de l'information extraite à même les textes des nouvelles (à opposer par exemple à l'extraction d'informations d'une base de données existante que l'on rechercherait dans les textes). Cela signifie que vous devez regarder des nouvelles, repérer les types d'information qui s'y trouvent et mettre au point ou utiliser des **extracteurs**. Aucune contrainte quant à la nature des informations extraites n'est imposée. Vous tenterez cependant d'évaluer la qualité de cette information extraite par des moyens que vous décrirez et ajusterez en conséquence vos extracteurs pour qu'ils produisent le moins de bruit possible.

Suit une **liste non exhaustive** d'informations que vous pouvez extraire.

1. Spacy vous donne accès à de multiples informations ; faites-en usage ! Vous pouvez notamment extraire les **entités nommées** comme les personnes, les lieux, les organisations ou autres (ex : données chiffrées, etc). Vous remarquerez que de lister les entités identifiées par Spacy (si vous utilisez Spacy) amène du bruit. La même entité (ex : `Olivier Royant`) peut par exemple se décliner en plusieurs variantes (ex : `O. Royan`, `M. Royan`, etc.), et certaines entités peuvent être mal découpées (ex : `Match` pour `Paris Match`). Votre but est bien sûr de réduire le bruit.
2. Les acronymes et leur(s) forme(s) longue(s) comme :
   — PIB : **p**roduit **i**ntérieur **b**rut
   — AQLPA : **A**ssociation **q**uébécoise de **l**utte contre la **p**ollution **a**tmosphérique

   Une approche simple consiste à repérer des séquences courtes de lettres majuscules à **aligner** (un alignment est indiqué en gras dans les exemples ci-dessus) avec des séquences de mots dans les textes (souvent dans le voisinage où l'acronyme est rencontré).
3. En consultant les textes, vous identifierez rapidement de nombreux **patrons d'extraction** potentiels. Implémentez certains d'entre eux et tenter de les affiner (filtrer le bruit, trier, etc.). Bien sûr, il existe des ressources déjà compilées qui listent une partie de cette information, mais votre but est de récupérer cette information directement des textes disponibles. Voici à titre d'exemple des informations relativement faciles à extraire :
   — la **fonction** occupée par des personnes, comme par exemple :
      — `Heiko Haas` qui en 2019 était ministre allemand des affaires étrangères, ce que vous pouvez capturer d'une phrase comme celle-ci : `...notamment le ministre allemand des Affaires étrangères, Heiko Haas, ...`,
      — `Marie-Claire Blais` est une écrivaine québécoise, ce que l'on peut extraire d'une phrase comme `L'écrivaine québécoise Marie-Claire Blais a décroché le Grand Prix du livre de Montréal.`
   — les fléaux sociétaires comme : les changements climatiques, le profilage racial du SPVM, la pollution atmosphérique que l'on peut

obtenir à l'aide d'un patron comme `lutte contre X`, où `X` est un fléau candidat,
- les noms de rues, de boulevards, d'avenues,
- des **relations taxonomiques** comme l'hyponymie (`est_un`) ou la méronymie (`est_une_partie_de`). Voici des exemples d'extractions possibles :

```
... une superstar comme Gad Elmaleh...
   =>  Gad Elmaleh IS_A superstar
... un pays comme le Burkina Faso ...
   => Burkina Faso IS_A pays
... avec des joueurs comme le quart Anthony Calvillo,
le receveur Ben Cahoon, le joueur de ligne offensive
Scott Flory et le botteur Damon Duval ...
   => Anthony Calvillo IS_A joueur,
   => le quart IS_A_TYPE_OF joueur,
   etc.
```

La liste des informations que l'on peut extraire est grande, mais il faut prendre conscience que chaque extracteur peut s'avérer en soit un objet d'étude comme par exemple, extraire des relations taxonomiques (**roller-etal-2018-hearst**), des acronymes (**ACRO**) ou encore des adresses dans des sites Web commerciaux (**ADRESSES**). Votre but est donc de vous limiter à certains types d'information et d'explorer ce qu'il est possible d'obtenir et avec quelle fiabilité.

4. Utilisez les liens de dépendance entre les mots ou groupes de mots de façon à extraire certains triplets (ou tuples) comme ceux-ci qui pourraient être extraits de l'extrait de nouvelle donné en exemple :
   - (`la radio d'Etat Égyptienne`, `diffuse`, `le discours de démission du président Gamal Abdel Nasser`)
   - (`Université Saint-Joseph`, `est_dans`, `Beyrouth`)
   - (`Gamal Abdel Nasser`, `démissionne`, `le 9 juin 1967`)

   Extraire de tels triplets est l'objet de l'extraction d'**information ouverte** et fait l'objet de nombreux travaux (**niklaus-etal-2018-survey**). Il n'existe pas à ma connaissance d'extracteur ouvert libre pour le français. En développer un dans le cadre de ce projet est un travail trop important. Vous pouvez néanmoins lister par exemple les relations verbales impliquant les personnes que vous avez identifiées, comme : `L'écrivaine québécoise Marie-Claire Blais`, `a_décroché`, `le Grand`

`Prix du livre de Montréal`. Vous trouverez peut-être une source d'inspiration dans cet article.

5. Bien sûr, vous pourriez entraîner un classificateur à produire certaines métadonnées, de manière à pouvoir les prédire lorsqu'elles sont absentes. L'**objet du projet 2 n'est** cependant **pas de développer un classificateur** de plus : vous avez déjà donné ! Aussi, il ne serait pas forcément pertinent de comparer entre eux pleins de classificateurs. En revanche, utiliser l'information extraite pour alimenter un classificateur que l'on cherchera à être le plus léger (# de paramètres) possible pourrait être un moyen parmi d'autres d'évaluer l'information extraite.

Vous devez produire un **rapport d'au plus 6 pages** (format pdf, en français ou en anglais) qui contient au moins les informations suivantes :

— une description des données et leur particularité,
— une description des extracteurs implémentés (les plus intéressants) et leur évaluation,
— une mesure de la densité d'information extraite.

N'hésitez surtout pas à positionner votre approche dans la littérature consacrée.

Une semaine environ avant la fin du projet, quelques nouvelles seront disponibles sur lesquelles vous appliquerez vos extracteurs. Pour chaque fichier de nouvelle, vous produirez un fichier texte pourtant le même nom (avec l'extension `.out`). Le format (que vous documenterez dans un fichier `readme`) doit être suffisamment simple pour qu'une tierce personne (moi par exemple) puisse comprendre la nature des informations extraites et puisse rapidement en apprécier la qualité. Il conviendra donc de classer l'information en ordre décroissant de pertinence (un concept laissé ici volontairement vague). Vous limiterez le cas échéant l'information produite de façon à ne pas dépasser 5ko par fichier annoté.

Notation

La notation sera corrélée à votre curiosité, à l'analyse des techniques déployées. La clarté et l'informativité de vos rapports sont comme toujours deux critères importants.

Remise

La remise est à faire sur Studium sous le libellé `projet2`. Vous devez remettre votre code, votre rapport (format pdf, texte en anglais ou en français), ainsi que les fichiers de sortie (`*.out`) et `readme` dans une archive (gzip, tar, tar.gz) dont le nom est préfixé de `projet2-noms`, où `noms` sont à remplacer par l'identité des personnes (`prénom_nom`) impliquées dans le projet. Donc si j'avais à remettre seul mon solutionnaire au projet2, je le ferais sous le nom `projet2-philippe_langlais.tar.gz`. Assurez vous que le nom des personnes impliquées dans le projet soit indiqué sur tous les documents remis. Le projet est à remettre en groupe d'au plus trois personnes au plus tard dimanche 20 décembre à 23h59.
Note : Aucun modèle n'est demandé : juste le rapport, le code, les fichiers d'information extraites et le fichier `readme`.