

Description du projet

Dans le cadre du projet professionnel carto-stages IDéaL on a besoin de constituer une base de données à partir du site <https://dumas.ccsd.cnrs.fr/>. La base de données vise à donner une idée sur les stages effectués par les étudiants dans le domaine du TAL. Pour atteindre cet objectif, nous avons besoin de collecter différents types d'informations sur ce site: titre de la thèse, nom de l'auteur, organisme de rattachement et année de stage.

Méthodes utilisées

Notre projet est divisé en deux parties: collecte d'informations et tri.

1. Collecte d'informations

Afin d'obtenir les données du site, nous utilisons la librairie Scrapy. Le site DUMAS permet d'affiner la recherche de la thèse par domaine. Malheureusement, le domaine du TAL n'a pas de page dans la classification des disciplines sur le site, c'est pourquoi on va utiliser trois liens différents pour collecter un maximum de données. Tout d'abord on utilise deux liens du domaine informatique qui correspondent à notre recherche (script `informatics.py`):

Informatique et langage

https://dumas.ccsd.cnrs.fr/search/index/?q=%2A&domain_t=info.info-cl

Traitement du texte et du document

https://dumas.ccsd.cnrs.fr/search/index/?q=%2A&domain_t=info.info-tt

On utilise le troisième lien avec le domaine linguistique entièrement pour trier les résultats de collecte plus tard (scripts `ling_re.py` et `ling_spacy.py`)

https://dumas.ccsd.cnrs.fr/search/index/?q=%2A&domain_t=shs.langue

☐  dumas-00841377v1 Mémoires
Allison Sanders. **Développer une méthode de recueil de données sur la perception sociolinguistique : comment les jeunes perçoivent-ils l'occitan dans le Sud-Ouest de la France ?**
Sciences de l'Homme et Société. 2013

☐  dumas-00714038v1 Mémoires
Anouk Darne. **L'enseignement / apprentissage des déterminants. Analyse d'erreurs chez un public polonais et élaboration d'activités de remédiation**
Linguistique. 2011

☐

1

2

3

4

5

6

7

8

9

10

⏮

⏭

↕ Tri ▼

☰ Nombre ▼

🔧 Outils ▼

Notre recherche s'effectue en largeur, c'est-à-dire sur un principe de pagination et ce jusqu'à la dernière page et en profondeur, avec l'idée d'ouvrir chaque thèse sur la page initiale. Pour obtenir les informations complètes on a besoin de parcourir toutes les pages du même lien.

Dans cette optique de travail, nous extrayons les liens vers la page suivante des balises de pagination et répétons le parsing récursivement (`def parse(self, res)`). Cette opération

est effectuée seulement sur la page du domaine linguistique, les pages "Informatique et langage" et "Traitement du texte et du document" ont tous les résultats sur une page.

```
#extract link to the next page and parse
k = res.css('ul.pagination.pagination-sm')[0]
next_page = k.css('li')[-2].css('a::attr(href)').get()
if next_page is not None:
    yield res.follow(next_page, callback=self.parse)
```

En même temps on a besoin de collecter les informations sur l'auteur, le sujet de la thèse, l'année et l'organisme de rattachement qui se trouvent sur la page de la thèse. On passe à la page de thèse et y collectons ces informations encore une fois grâce à la fonction récursive (`def parse_link(self, res)`)

Projet TourInFlux. Annotation des expressions temporelles

Lucie Drat ¹ Détails

1 UGA UFR LLASIC SLFLE - Université Grenoble Alpes - UFR Langage, lettres et arts du spectacle, information et communication - Dpt Sciences du langage et français langue étrangère

en

fr

Résumé : Ce mémoire professionnel présente le travail effectué au sein du laboratoire L3i de l'université de La Rochelle. Ce stage a pris part au projet TourInFlux dont le but est la création d'un tableau de bord réunissant les informations disponibles sur les différents territoires de France pour les acteurs du tourisme. La contribution au projet était le développement d'une grammaire hors-contexte pour la détection et l'annotation des expressions temporelles dans les documents touristiques. Le travail effectué a permis également de souligner l'importance de l'étude linguistique dans le choix du schéma d'annotation, de l'écriture du guide d'annotation ainsi que du développement de l'annotateur automatique, peu importe la méthode choisie pour cela.

en

fr

Mots-clés : Schéma et guide d'annotation Corpus touristiques Expressions temporelles Grammaire hors-contexte Extraction automatique d'information

Type de document : Mémoires

Domaine : Sciences de l'Homme et Société / Linguistique

2. Tri

Les informations collectées sur les pages informatiques n'ont pas besoin d'être triées grâce à la classification préalable faite sur le site. Par contre, nous obtenons 307 résultats de la page linguistique qui ne correspondent pas tous à notre recherche. Pour trier ces résultats, on utilise 2 librairies différentes: re et Spacy et on compare les résultats.

⇒ re — Opérations à base d'expressions rationnelles
Fichier ling_re.py

Afin d'obtenir les résultats qui correspondent au domaine TAL, nous utilisons des mots-clés dans la recherche d'expressions régulières:

`p = re.compile(r'\b(automatique|NLP|TAL|TALN|ingénierie)').` Nous limitons notre recherche à deux champs sur la page de la thèse: "résumé" et "mots-clés". Si le champ contient les mot-clés, on attribue vrai à ce résultat et faux dans l'autre cas. Finalement, on obtient un tableau de booléens où chaque élément correspond à l'information retenue (sujet, auteur, année et organisme). Le tableau de booléens a la même longueur que les tableaux contenant les informations. On enregistre ces informations dans un fichier csv seulement dans le cas où l'élément correspondant de boollist a vrai comme valeur. Le résultat est enregistré dans un fichier ling_re.csv.

⇒ **Spacy — rule-based matcher**
Fichier ling_spacy.py

Nous utilisons la même méthode de recherche avec la librairie Spacy. Au lieu d'utiliser l'expression régulière, on utilise le pattern (ex. `pattern1 = [{"LOWER": "automatique"}]`). Encore une fois on cherche les mots-clés dans les champs "résumé" et "mots-clés" sur la page et on leur attribue la valeur de tableau booléen. Le résultat est enregistré dans un fichier `ling_spacy.csv`.

3. Comparaison

Pour comparer les résultats de deux scripts (`ling_re.py` et `ling_spacy.py`) on a besoin de créer le corpus de référence. Pour faire ça, nous avons collecté des informations directement du lien contenant le mot-clé dans la requête (`script standard.py`)

Rechercher?

[+ Recherche avancée...](#)


☐

Tri

Nombre

Outils

☐



dumas-00561995v1

Mémoires

Gaëlle Chabert. **SMS et TAL : kL 1Trè* ? (*SMS et TAL : Quel intérêt ?)**
Linguistique. 2010

Par exemple, dans ce cas on récupère tous les résultats de recherche de l'acronyme "TAL". Comme ça on peut créer un tableau avec tous les résultats qui donnent le moteur de recherche du site sans avoir collecté et trié les données inutiles. Le résultat est enregistré dans le fichier `standard.csv`. Il contient 28 lignes, tandis que `ling_re.csv` a 24 lignes et `ling_spacy.csv` a 23 lignes. Il est possible que le corpus de référence ait plus de résultats parce que le moteur de recherche du site prend en compte toute la page et pas seulement les champs "résumé" et "mots-clés". En ce qui concerne le résultat obtenu, il ne varie pas beaucoup malgré le fait d'utiliser deux méthodes différentes. Cependant, nous devons admettre que "matcher" de la librairie Spacy prend au moins 10 fois plus de temps pour trier les données. Nous pouvons conclure que l'utilisation du pattern "matcher" peut être utile dans les recherches plus fines qui incluent des catégories grammaticales précises. Pour une recherche simple de mots isolés, regex semble être plus pratique en termes de temps.

4. Difficultés rencontrées

```
#ignore redirection error
handle_httpstatus_list = [302], [200]

def start_requests(self):
    yield scrapy.Request(url = self.url, dont_filter=True, callback=self.parse)

ROBOTSTXT_OBEY = False
```

1) Le serveur CNRS est protégé du scraping, il redirige la requête vers la page qui n'existe pas. Nous avons donc dû passer un temps considérable à tester différentes méthodes pour gérer la redirection.

```
FEED_EXPORT_ENCODING = 'utf-8'
```

2) Scrapy ne reconnaît pas les caractères accentués, on a besoin de changer les settings
3) Le balisage sur la page de la thèse se change si le titre est écrit en deux langues, il a fallu réécrire la fonction et collecter les informations de la page initiale où les balises sont les mêmes dans tous les cas.