Natural Language Processing (NLP)

3 — Mots et tokens

Georges-André Silber

2023/2024

École des mines de Paris

Mots

Qu’est-ce qu’un mot ?

- Notion mal définie
- Problème de la séparation des mots d’une phrase
- Espaces, ponctuations ?
- Lemmatisation, racinisation pour commencer à ”classifier”

Lemmatisation (lemmatization)

- Réduction des mots à leur forme canonique (le lemme)
- « *avoir* » depuis « *eussions eu* »
- « *des avions* » vs « *nous avions* »

continu	continu
continua	continuer
continuait	continuer
continuant	continuer
continuation	continuation
continuations	continuation
continue	continu | continuer

Racinisation (stemming)

- Regroupement des mots par racine commune
- ”Lemmatisation” simplifiée

continu	continu
continua	continu
continuait	continu
continuant	continu
continuation	continu
continuations	continu
continue	continu

Délimitation d'une phrase

Séparation des phrases d'un texte. À l'écrit, la ponctuation ou la casse permet en général de séparer les phrases, mais des complications peuvent être causées par les abréviations utilisant un point, ou les citations comportant des ponctuations à l'intérieur d'une phrase, etc.

Séparation des mots

Dans la langue parlée, les phrases ne sont qu'une chaîne de phonèmes, où l'espace typographique n'est pas prononcé. Par exemple, « *un bon appartement chaud* » et « *un Bonaparte manchot* » sont identiques d'un point de vue phonétique.

Outil de base : les expressions régulières

Application : expressions régulières

- Expressions régulières par génération d'un automate fini (Ken Thompson).
- `grep`, `lex`, analyseur lexical
- `https://regexcrossword.com`
- Python : `import re`
- hyperscan

Exemple 1 : utilisation dans un IDE

```
#define MAX_URI_COUNTRY 3
#define MAX_URI_CORPUS 5
#define MAX_URI_NATURE 70
#define MAX_URI_YEAR 5
#define MAX_URI_MONTH 3
#define MAX_URI_DAY 3
#define MAX_URI_NUMBER 30
#define MAX_URI_VERSION 9
#define MAX_URI 256

MAX_(\w+)
$1_MAX
```

Exemple 2 : découpage d'un arrêt de cour d'appel d'Agen

```
intro_re = re.compile(
  r'^(?P<intro>.*?)(?=('
  r'<p>\s*A\s+rendu\s+l.arrêt\s+((réputé\s+)?
  r'contradictoire|par\s+défaut)'
  r'|<p>\s*EXPOS(É|E)\s*DU\s*LITIGE'
  r'|<p>A rendu réputé l.arrêt réputé contradictoire'
  r'))',
  re.UNICODE|re.DOTALL|re.MULTILINE|re.IGNORECASE)
decision_re = re.compile(
  r'(?P<decision>(<p>par\s*ces\s*motifs).*)$',
  re.U|re.DOTALL|re.MULTILINE|re.IGNORECASE)
```

Exemple 3 : numéros d'alinéas (Legistix)

```
alinea_number = (
  r"("
  r"\w\)(?=\s+)"
  r"|\d{1,2}°(\s+bis)?(?=\.?\s+)"
  r"|\d{1,2}(\s+bis)?(?=\.?\s+)"
  r"|[IVX]+(?=\.\s+)"
  r")"
)
```

Travaux pratiques 1

Extraction de texte structuré depuis un PDF

- Github, Gitlab
- Étape 0 : installer Docker, faire fonctionner le Dockerfile
- Étape 1 : extraire le texte du PDF (poppler), récupérer des données
- Étape 2 : compléter le script Python `text2md` (regexps)
- Étape 3 : compléter le script Python `md2xml` (regexps)
- À rendre avant dimanche 7 janvier à 23h59
- Livrable : pull request (Github), merge request (Gitlab) ou patch (`git diff`)
- Voir les instructions