Linguistique de corpus

Outils et méthodes de traitement de corpus

Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris XI, 91403 Orsay Cedex
pap@limsi.fr

mercredi 01 mars 2017 / L3 - Cours S2-4

Pratique Unitex
Tokenization

L'importation d'un document dans Unitex passe toujours par une étape de

1. normalisation
2. tokenization

Les séparateurs sont : $\backslash s$ $\backslash t$ $\backslash n$

Pratique Unitex
Tokenization

Les règles de normalisation des séparateurs :

1. $[\backslash s\ \backslash t\ \backslash n]* \backslash n\ [\backslash s\ \backslash t\ \backslash n]* \rightarrow \backslash n$
2. $[\backslash s\ \backslash t]* \rightarrow \backslash s$

Le résultat de la normalisation d'un texte `foo.txt` test stoké dans un repertoire `foo_snt` .
situé dans le même répertoire que le fichier d'origine.
Attention, avec le répertoire `foo_snt` il y a aussi un fichier `foo.snt` qui est créé, au même niveau de l'arborescence de fichiers.

Exemple, soit le texte suivant dans le repertoire `/tmp/`

```
  pap@joffo:/tmp$ cat foo.txt
Il fait beau.
 Je vais me promener dans la campagne.
```

Son importation dans unitex provoque la création de 2 fichier `_snt` et `.snt`

```
pap@joffo:/tmp$ ls -l | grep foo
drwxr-xr-x 2 pap     grpiles 4096 Mar 12 23:30 foo_snt
-rw-r--r-- 1 pap     grpiles  116 Mar 12 23:30 foo.snt
-rw-r--r-- 1 pap     grpiles   53 Mar 12 23:29 foo.txt
```

```
pap@joffo:/tmp$ ls -l foo_snt
total 68
-rw-r--r-- 1 pap grpiles   2 Mar 12 23:30 dlc
-rw-r--r-- 1 pap grpiles   8 Mar 12 23:30 dlc.n
-rw-r--r-- 1 pap grpiles 720 Mar 12 23:30 dlf
-rw-r--r-- 1 pap grpiles  10 Mar 12 23:30 dlf.n
-rw-r--r-- 1 pap grpiles   8 Mar 12 23:30 enter.pos
-rw-r--r-- 1 pap grpiles   2 Mar 12 23:30 err
-rw-r--r-- 1 pap grpiles   8 Mar 12 23:30 err.n
-rw-r--r-- 1 pap grpiles  24 Mar 12 23:30 snt_offsets.pos
-rw-r--r-- 1 pap grpiles  22 Mar 12 23:30 stat_dic.n
-rw-r--r-- 1 pap grpiles 158 Mar 12 23:30 stats.n
-rw-r--r-- 1 pap grpiles   2 Mar 12 23:30 tags_err
-rw-r--r-- 1 pap grpiles   8 Mar 12 23:30 tags_err.n
-rw-r--r-- 1 pap grpiles  10 Mar 12 23:30 tags.ind
-rw-r--r-- 1 pap grpiles  92 Mar 12 23:30 text.cod
-rw-r--r-- 1 pap grpiles 198 Mar 12 23:30 tok_by_alph.txt
-rw-r--r-- 1 pap grpiles 198 Mar 12 23:30 tok_by_freq.txt
-rw-r--r-- 1 pap grpiles 168 Mar 12 23:30 tokens.txt
pap@joffo:/tmp$
```

Attention l'affichage de `foo.snt` dans le terminal avec `cat` ne donne rien. Utiliser un éditeur comme vi, emacs, gedit, kate...

```
pap@joffo:/tmp$ vi /tmp/foo.snt


^L"/tmp/foo.snt" [converted][dos] 2L, 59CIl fait beau.
{S}Je vais me promener dans la campagne.
~
~
~
~
1,1All2^G
```

Pratique Unitex
Segmentation en phrases

La segmentation en énoncés est expliqué aux pages 34 du manuel Unitex v3.1 français).
Par défaut les espaces sont optionnels entre le textes de 2 boîtes d'un graphe, cependant l'utilisation du caractère # permet de les interdire, et la séquence " " les rend obligatoires.

Pratique Unitex
L'alphabet

Unitex utilise une représentation Unicode Little-Endian (cf p. 28 et 270 du manuel unitex français).
Les charactères d'un langue sont définis dans 2 fichiers situés sous la racine de l'arborescence spécifique à cette langue, par ex. English/alphabet et English/sorted_alphabet
Dans le premier fichier, on trouve les caractères (1 par ligne), avec 3 formats possibles :

1. 1 multi-caractère, la ligne commence alors par #,
2. majuscule, et minuscule pour 1 caractères (e.g. Aa),
3. ou 1 seul caractère sur la ligne

Le second fichier contient l'ordre des caractères, plusieurs caractères situés sur une même ligne sont considéres de même ordre (e.g. Cc si l'on ne veut pas distinguer majuscule de minuscule).
Voir pages 260, 261 et 262 du manuel Unitex.

Pratique Unitex

Normalisation des formes non ambiguës

La normalisation des formes non amiguës se fait avec des automates qui acceptent les mêmes symboles que pour la segmentation en énoncés, en mode “remplacement” (cf p 39).
Elle se trouve dans le répertoire :
/(home directory)/(active language)/Graphs/Preprocessing/Replace
et se nomme Replace.fst2.

Pratique Unitex
Filtrage dictionnairique

Le filtrage dictionairique est décrit dans le manuel Unitex. Il produit 3 fichiers :

1. liste des mots simples
2. liste des mots composés
3. liste des mots inconnus. Pour les langues agglutinatives (e.g. Allemand, les mots formés de plusieurs mots connnus sont retirés de cette liste automatiquement).

Pratique Unitex
Normalisation

En regardant la normalisation du roman “Le tour du Monde en 80 jours” faite lorsque le document est importé dans Unitex, on s’apperçoit que les énoncés de dialogue ne sont pas segmentés commme des énoncés. Nous allons donc ajouter un filtre de normalisation qui identifie les dialogues dans et ajoute un frontière de phrase {S} a la fin de chaque énoncé en créant un automate transducteur.

1. dans un premier temps sans ses soucier des balises d’énoncé existantes
2. en prenant compte les balises d’énoncé existantes, lorsqu’elles sont présente au milieu d’un énoncé