جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

Extraction d'information

Cours 8

Nassim ZELLAL

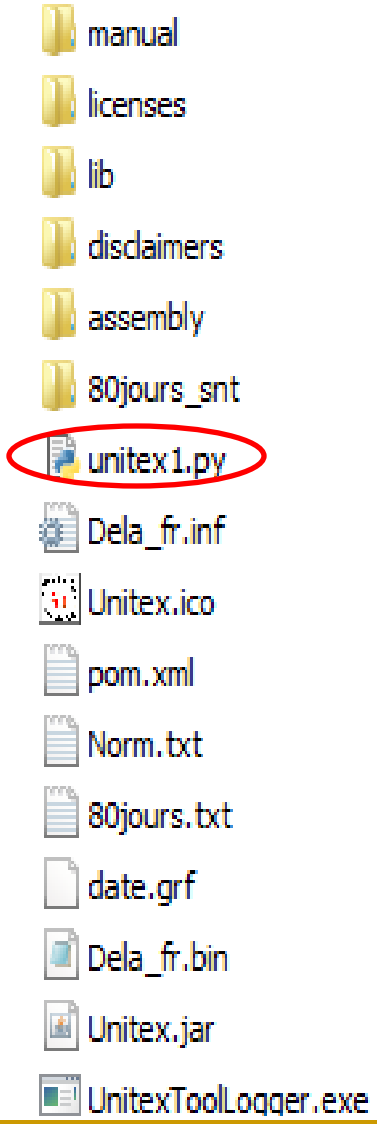
Lancer Unitex à partir du script « unitex1.py »

```
# ----------------------unitex1.py-----------------------------------------------#
import os
os.system("rd /s 80jours_snt")
os.mkdir("80jours_snt")
os.system("UnitexToolLogger Normalize 80jours.txt -r Norm.txt")
os.system("UnitexToolLogger Tokenize 80jours.snt")
os.system("UnitexToolLogger Dico -t 80jours.snt Dela_fr.bin")
os.system("UnitexToolLogger Grf2Fst2 date.grf")
os.system("UnitexToolLogger Locate -t 80jours.snt date.fst2 -L -I --all")
os.system("UnitexToolLogger Concord 80jours_snt/concord.ind -f \"Courrier new\" -s 12 -l 40 -r 55")
# ----------------------unitex1.py-----------------------------------------------#
```

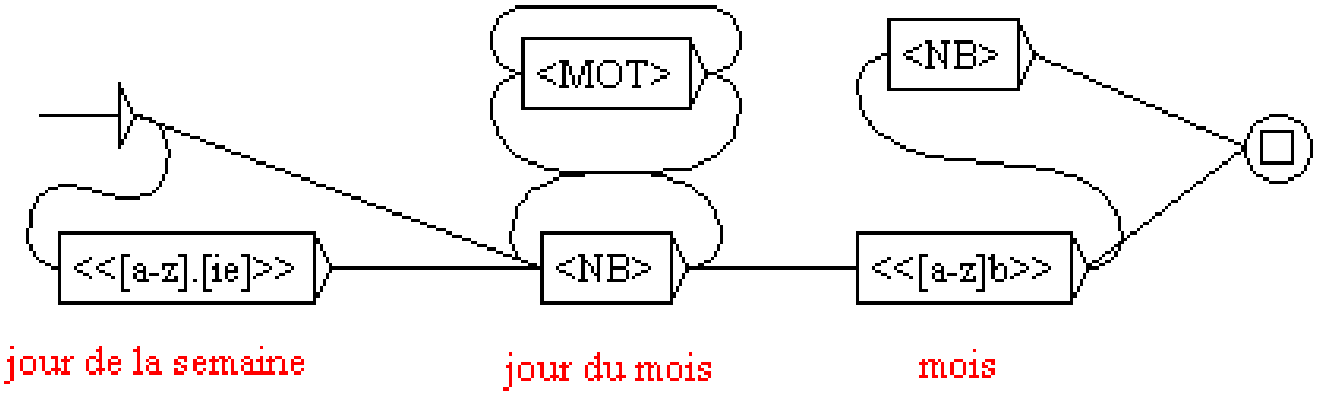
Lancer Unitex à partir du script « unitex1.py »

```
#----------------------unitex1.py-----------------------------------------------#
import os
from os import path
if path.exists("80jours_snt"):
    os.system("rd /s 80jours_snt")
os.mkdir("80jours_snt")
os.system("UnitexToolLogger Normalize 80jours.txt -r Norm.txt")
os.system("UnitexToolLogger Tokenize 80jours.snt")
os.system("UnitexToolLogger Dico -t 80jours.snt Dela_fr.bin")
os.system("UnitexToolLogger Grf2Fst2 date.grf")
os.system("UnitexToolLogger Locate -t 80jours.snt date.fst2 -L -I --all")
os.system("UnitexToolLogger Concord 80jours_snt/concord.ind -f \"Courrier new\" -s 12 -l 40 -r 55")
#----------------------unitex.py------------------------------------------------#
```

Lancer Unitex à partir du script « unitex1.py »



Appliquer le graphe « date.grf » d'Unitex à partir du script « unitex1.grf »

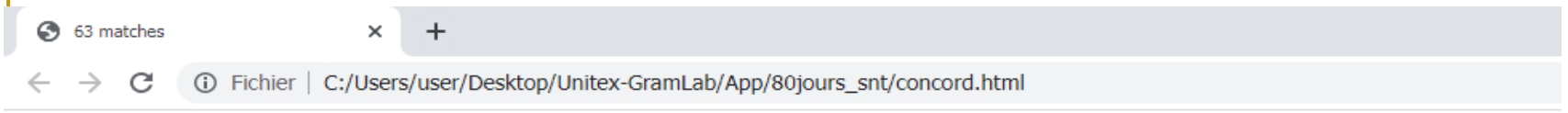


Lancer Unitex à partir du script « unitex1.py »

```
Done.
Initializing...
Counting tokens...
Applying dico  Dela_fr.bin...
Looking for simple words...
Looking for compound words...
First block...
Sorting and saving tag sequences...
Saving unknown words...
Done.
Compiling graph date.grf
Compilation has succeeded
Loading fst2...
Loading token list...
Loading morphological dictionaries...
Computing fst2 tags...
Loading dlf...
dlf: 10000 lines loaded...
Loading dlc...
Optimizing fst2 pattern tags...
Optimizing compound word dictionary...
Optimizing fst2...
Working...
100% done

63 matches
285 recognized units
(0.176% of the text is covered)
65550 exploration step
Done.
Loading concordance index...
Constructing concordance...
Done.
```

Résultat de l'extraction - « concord.html »

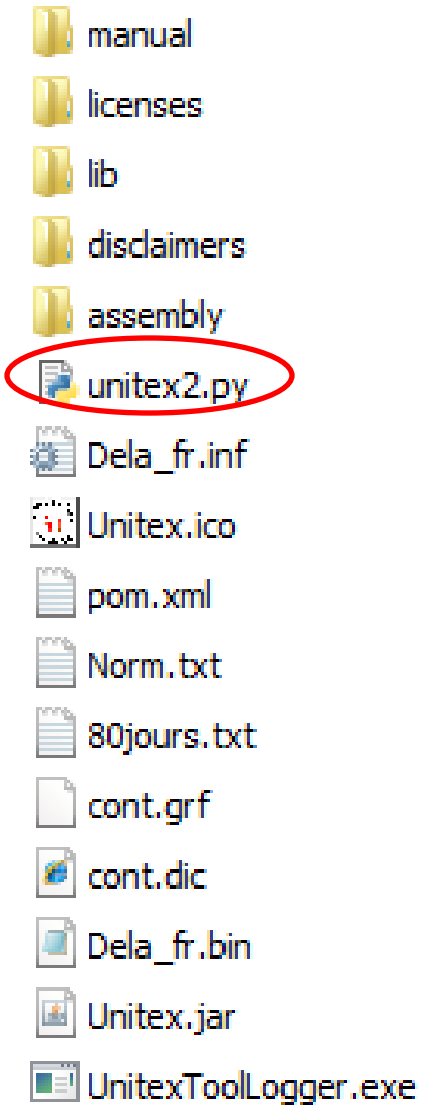


arité extraordinaires. Ce jour-là même, 2 octobre, Phileas Fogg avait donné son congé à James F
nt, onze heures vingt-neuf du matin, ce mercredi 2 octobre 1872, vous êtes à mon service. » Cel
ait accompli trois jours auparavant, le 29 septembre. Une liasse de bank-notes, formant l'énorm
er eût seulement levé la tête. Mais, le 29 septembre, les choses ne se passèrent pas tout à fai
d'Angleterre. Pendant cette journée du 29 septembre, un gentleman bien mis, de bonnes manières
ier de poche, puisque c'est aujourd'hui mercredi 2 octobre, je devrai être de retour à Londres,
, dans ce salon même du Reform-Club, le samedi 21 décembre, à huit heures quarante-cinq du soir
tre-vingts jours, répondit Mr. Fogg, le samedi 21 décembre 1872, à huit heures quarante-cinq mi
lir. En effet, un long article parut le 7 octobre dans le Bulletin de la Société royale de géog
he concernant le sieur Phileas Fogg. Le mercredi 9 octobre, on attendait pour onze heures du ma
les notes suivantes : " Quitté Londres, mercredi 2 octobre, 8 heures 45 soir. " Arrivé à Paris,
re, 8 heures 45 soir. " Arrivé à Paris, jeudi 3 octobre, 7 heures 20 matin. " Quitté Paris, jeu
in. " Arrivé par le Mont-Cenis à Turin, vendredi 4 octobre, 6 heures 35 matin. " Quitté Turin,
7 heures 20 matin. " Arrivé à Brindisi, samedi 5 octobre, 4 heures soir. " Embarqué sur le Mong
samedi, 5 heures soir. " Arrivé à Suez, mercredi 9 octobre, 11 heures matin. " Total des heures
par colonnes, qui indiquait - depuis le 2 octobre jusqu'au 21 décembre - le mois, le quantième,
ndiquait - depuis le 2 octobre jusqu'au 21 décembre - le mois, le quantième, le jour, les arriv
retard. Il inscrivit donc, ce jour-là, mercredi 9 octobre, son arrivée à Suez, qui, concordant
bay. Le lendemain du départ de Suez, le 10 octobre, ce ne fut pas sans un certain plaisir qu'il
e Mongolia, au lieu d'arriver à Aden le 15 octobre seulement au matin, y entrait le 14 au soir.

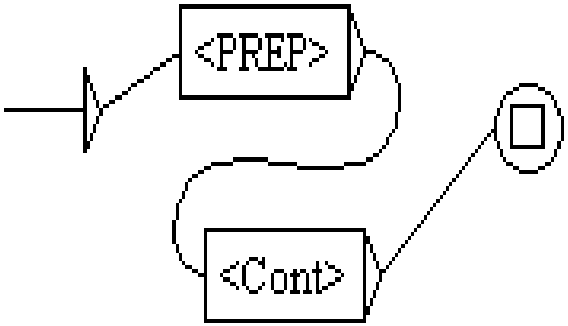
Lancer Unitex à partir du script « unitex2.py »

```
#----------------------unitex2.py-----------------------------------------------#
import os
os.system("rd /s 80jours_snt")
os.mkdir("80jours_snt")
os.system("UnitexToolLogger Normalize 80jours.txt -r Norm.txt")
os.system("UnitexToolLogger Tokenize 80jours.snt")
os.system("UnitexToolLogger Compress cont.dic")
os.system("UnitexToolLogger Dico -t 80jours.snt Dela_fr.bin cont.bin")
os.system("UnitexToolLogger Grf2Fst2 cont.grf")
os.system("UnitexToolLogger Locate -t 80jours.snt cont.fst2 -L -I --all")
os.system("UnitexToolLogger Concord 80jours_snt/concord.ind -f \"Courrier new\" -s 12 -l 40 -r 55")
#----------------------unitex2.py-----------------------------------------------#
```

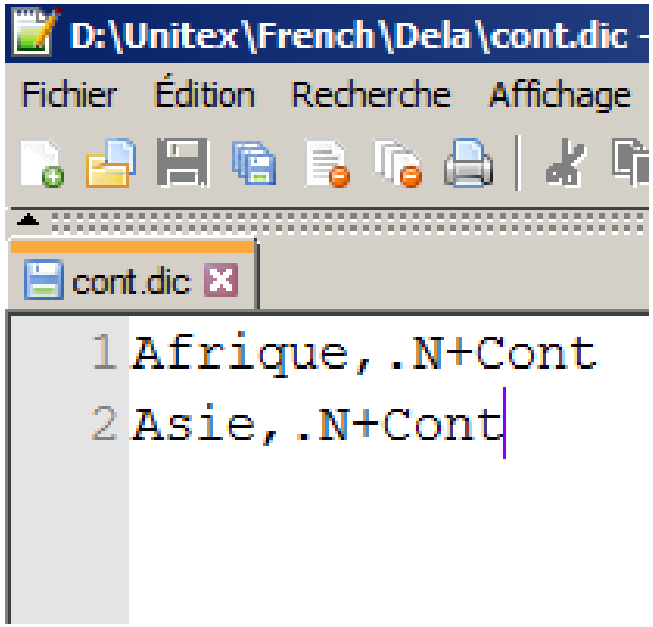
Lancer Unitex à partir du script « unitex2.py »



Application du graphe « cont.grf » d'Unitex à partir du script « unitex2.py »



Application du DELAF « cont.dic » d'Unitex à partir du script « unitex2.py »



Lancer Unitex à partir du script « unitex2.py »

```
Applying dico  Dela_fr.bin...
Looking for simple words...
Looking for compound words...
First block...
Applying dico  cont.bin...
Looking for simple words...
Looking for compound words...
First block...
Sorting and saving tag sequences...
Saving unknown words...
Done.
Compiling graph cont.grf
Compilation has succeeded
Loading fst2...
Loading token list...
Loading morphological dictionaries...
Computing fst2 tags...
Loading dlf...
dlf: 10000 lines loaded...
Loading dlc...
Optimizing fst2 pattern tags...
Optimizing compound word dictionary...
Optimizing fst2...
Working...
100% done

5 matches
15 recognized units
(0.009% of the text is covered)
```

Résultat de l'extraction - « concord.html »

gypte ? - En Égypte, parfaitement. - Et en Afrique ? - En Afrique. - En Afrique ! répéta Passep
ypte, parfaitement. - Et en Afrique ? - En Afrique. - En Afrique ! répéta Passepartout. Je ne p
ment. - Et en Afrique ? - En Afrique. - En Afrique ! répéta Passepartout. Je ne peux y croire.
ù prenez-vous Bombay ? - Dans l'Inde. - En Asie ? - Naturellement. - Diable ! C'est que je vais
ise. On lança des dépêches en Amérique, en Asie, pour avoir des nouvelles de Phileas Fogg ! On

Exercice

- **Écrire un script Python permettant d'extraire les entités médicales de type noms de médicaments par substance active ou par nom commercial à partir du fichier :**

  « corpus-medical.txt » encodé en UTF-8 sans BOM.
- **Mettre le résultat en minuscule et le rediriger vers le fichier « subst.txt ».**
- **Exemples d'entités médicales de type noms de médicaments (substance active ou nom commercial), se trouvant dans ce corpus :**
- **Simvastatine (substance active)**
- **Cytarabine (substance active)**
- **Zolpidem (substance active)**
- **Inexium (nom commercial)**
- **Plavix (nom commercial)**
- **Crestor (nom commercial)**

Exercice

- **Exemples extraits du corpus médical :**
- **simvastatine**
- **plavix**
- **buflomedil**
- **crestor**
- **voluven**
- **topalgic**
- **lasilix**
- **aspegic**
- **contramal**
- **inexium**
- **cytarabine**
- **idarubicine**
- **zolpidem**
- **lovenox**