NLTK

Natural Language Toolkit

Informations générales

- Créé en 2001 par Steven Bird et Edward Loper, librairie open source
- Dernière mise à jour : Janvier 2023
- Version stable : 3.8.2
- Programmé en Python
- Compatible avec Python 3.7, 3.8, 3.9, 3.10, 3.11.
- Licenses : Apache License 2.0

Ressources

Documentation : https://www.nltk.org/index.html

Article : NLTK: The Natural Language Toolkit Edward Loper and Steven Bird, *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pp 62-69, Philadelphia, Association for Computational Linguistics. July 2002.

NLTK book : Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Quelles tâches ?

Language processing task	NLTK modules	Functionality
Accessing corpora	corpus	standardized interfaces to corpora and lexicons
String processing	tokenize, stem	tokenizers, sentence tokenizers, stemmers
Collocation discovery	collocations	t-test, chi-squared, point-wise mutual information
Part-of-speech tagging	tag	n-gram, backoff, Brill, HMM, TnT
Machine learning	classify, cluster, tbl	decision tree, maximum entropy, naive Bayes, EM, k-means
Chunking	chunk	regular expression, n-gram, named-entity
Parsing	parse, ccg	chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	sem, inference	lambda calculus, first-order logic, model checking
Evaluation metrics	metrics	precision, recall, agreement coefficients
Probability and estimation	probability	frequency distributions, smoothed probability distributions
Applications	app, chat	graphical concordancer, parsers, WordNet browser, chatbots
Linguistic fieldwork	toolbox	manipulate data in SIL Toolbox format

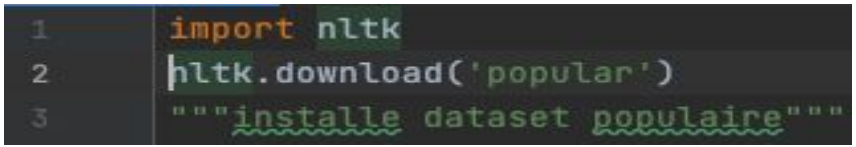
Source : NLTK book

Installation

En ligne de commande :

pip install - -user -U nltk

Nécessité d’installer des datasets :



Autres librairies recommandées par les auteurs :

numPy

matplotlib

Quelques exemples :

Tokenisation :

```
sentence = """At eight o'clock on Thursday morning Arthur didn't feel very good."""
tokens = nltk.word_tokenize(sentence)
print(tokens)
```

```
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning', 'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
```

```
phrase = """En grammaire, une phrase peut être considérée comme un ensemble autonome,
réunissant des unités syntaxiques organisées selon différents réseaux de relations plus ou moins complexes appelés
subordination, coordination ou juxtaposition.
L'autonomie peut être définie comme un prédicat
(en français, le plus souvent un verbe conjugué) associé à une modalité d'énonciation
(assertion, interrogation, injonction, exclamation dans un sens restreint, amorcé par un marqueur exclamatif)."""
french_tokens = nltk.word_tokenize(phrase,"french")
print(french_tokens)
```

```
['En', 'grammaire', ',', 'une', 'phrase', 'peut', 'être', 'considérée', 'comme', 'un', 'ensemble', 'autonome', ',', 'réunissant',
'des', 'unités', 'syntaxiques', 'organisées', 'selon', 'différents', 'réseaux', 'de', 'relations', 'plus', 'ou', 'moins', 'complexes',
'appelés', 'subordination', ',', 'coordination', 'ou', 'juxtaposition', '.', "L'autonomie", 'peut', 'être', 'définie', 'comme', 'un',
'prédicat', '(', 'en', 'français', ',', 'le', 'plus', 'souvent', 'un', 'verbe', 'conjugué', ')', 'associé', 'à', 'une', 'modalité',
"d'énonciation", '(', 'assertion', ',', 'interrogation', ',', 'injonction', ',', 'exclamation', 'dans', 'un', 'sens', 'restreint', ',',
'amorcé', 'par', 'un', 'marqueur', 'exclamatif', ')', '.']
```

Lemmatisation

```
from nltk.stem.wordnet import WordNetLemmatizer
wnl = WordNetLemmatizer()
```

```
for w in tokens:
    print(wnl.lemmatize(w))
for mt in french_tokens:
    print(wnl.lemmatize(mt))
```

```
At
eight
o'clock
on
Thursday
morning
Arthur
did
n't
feel
very
good
.
```

```
En
grammaire
,
une
phrase
peut
être
considérée
comme
un
ensemble
autonome
,
réunissant
de
unités
syntaxiques
organisées
selon
différents
réseaux
de
relation
plus
ou
moins
complex
appelés
subordination
,
coordination
ou
juxtaposition
.
```

note : Lemmatize retourne l'input inchangé s'il n'est pas retrouvé dans WordNet

POS

Attention : POS tagging uniquement disponible en anglais et russe

```
tagged = nltk.pos_tag(tokens)
print(tagged)
```

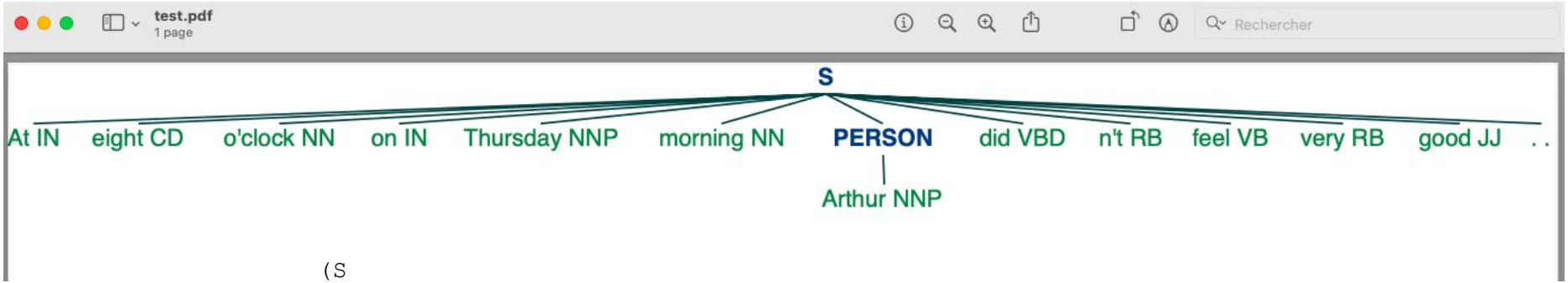
```
[('At', 'IN'), ('eight',
'CD'), ("o'clock", 'NN'),
('on', 'IN'), ('Thursday',
'NNP'), ('morning', 'NN'),
('Arthur', 'NNP'), ('did',
'VBD'), ("n't", 'RB'),
('feel', 'VB'), ('very',
'RB'), ('good', 'JJ'),
('.', '.')]
```

Création arbre, et reconnaissance EN

```
from nltk import Tree
from nltk.draw.tree import TreeView
```

```
sentence = """At eight o'clock on Thursday morning Arthur didn't feel very good."""
tokens = nltk.word_tokenize(sentence)
print(tokens)
tagged = nltk.pos_tag(tokens)
print(tagged)
entities = nltk.chunk.ne_chunk(tagged)
print("entities", entities)
#Permet de visualiser l'arbre tree.tree.Tree (entities dans le cas CI PRESENT)
TreeView(entities)._cframe.print_to_file("test.ps")
```

Visualisation arbre




```
(S
  At/IN
  eight/CD
  o'clock/NN
  on/IN
  Thursday/NNP
  morning/NN
  (PERSON Arthur/NNP)
  did/VBD
  n't/RB
  feel/VB
  very/RB
  good/JJ
  ./.)
```

Stemmatisation

```
from nltk.stem.snowball import FrenchStemmer
```

```
mots_exemples = ["désespoir","désespérément","désespérant","désespérer","donner","don"]
stemmer = FrenchStemmer()
for mot in mots_exemples:
    print(stemmer.stem(mot))
```

```
désespoir
désesper
désesper
désesper
don
don
```

Concordancier

```
nltk.download('nps_chat')
nltk.download('webtext')
""" pour travailler sur les corpus"""
from nltk.book import *
```

```
text1.concordance("man")
```

```
Displaying 25 of 527 matches:
 Civitas ) which is but an artificial man ." -- OPENING SENTENCE OF HOBBES ' S
y of that sort that was killed by any man , such is his fierceness and swiftnes
in his deepest reveries -- stand that man on his legs , set his feet a - going
it ? The urbane activity with which a man receives money is really marvellous ,
, and that on no account can a monied man enter heaven . Ah ! how cheerfully we
ure truly , enough to drive a nervous man distracted . Yet was there a sort of
ss needle sojourning in the body of a man , travelled full forty feet , and at
 him ), bustles a little withered old man , who , for their money , dearly sell
ld put up with the half of any decent man ' s blanket . " I thought so . All ri
king as much noise as the rest . This man interested me at once ; and since the
 . I have seldom seen such brawn in a man . His face was deeply brown and burnt
ions had mounted to its height , this man slipped away unobserved , and I saw n
us to the entrance of the seamen . No man prefers to sleep two in a bed . In fa
 an uncomfortable feeling towards the man whom you design for my bedfellow -- a
lord , that harpooneer is a dangerous man ." " He pays reg ' lar ," was the rej
nd a papered fireboard representing a man striking a whale . Of things not prop
 me . I remembered a story of a white man -- a whaleman too -- who , falling am
ter all ! It ' s only his outside ; a man can be honest in any sort of skin . B
eard of a hot sun ' s tanning a white man into a purplish yellow one . However
 you sabbee me , I sabbee -- you this man sleepe you -- you sabbee ?" " Me sabb
```