SpaCy Tuto

Cours de Documents Structurés

Qu’est-ce que Spacy ?

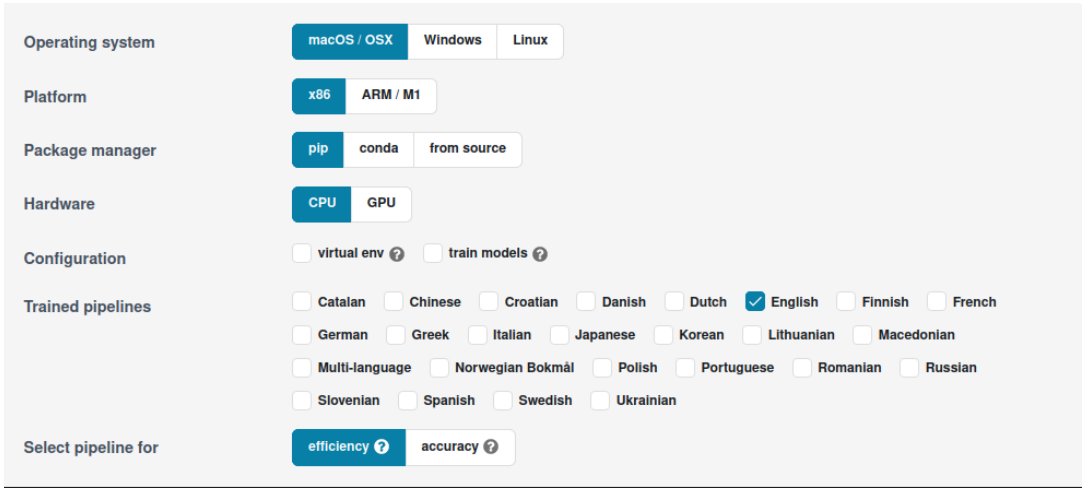
- Une bibliothèque open-source crée en 2015 par Matthew Honnibal and Ines Montani (société Explosion)

- développé en Python et Cython

A quoi ça sert ? (liste non exhaustive)

- Tokenisation
- Pos tagging
- lemmatization
- Détection d’entités nommées
- Analyse en dépendance
- Calcul de similarité entre documents

Comment installer Spacy ?



Avec pip:
pip install -U setuptools
wheel pip install -U spacy

Avec conda:
conda install -c conda-forge
spacy

https://spacy.io/usage

Comment installer Spacy ?

Avec pip

```
anne@anne-VivoBook-ASUSLaptop-X515JAB-X515JA:~$ pip install -U pip setuptools wheel
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pip in /usr/lib/python3/dist-packages (22.0.2)
Collecting pip
  Downloading pip-23.2.1-py3-none-any.whl (2.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 8.7 MB/s eta 0:00:00
Requirement already satisfied: setuptools in /usr/lib/python3/dist-packages (59.6.0)
Collecting setuptools
  Downloading setuptools-68.2.2-py3-none-any.whl (807 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 807.9/807.9 KB 7.2 MB/s eta 0:00:00
Requirement already satisfied: wheel in /usr/lib/python3/dist-packages (0.37.1)
Collecting wheel
  Downloading wheel-0.41.2-py3-none-any.whl (64 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.8/64.8 KB 7.5 MB/s eta 0:00:00
Installing collected packages: wheel, setuptools, pip
Successfully installed pip-23.2.1 setuptools-68.2.2 wheel-0.41.2
```

```
anne@anne-VivoBook-ASUSLaptop-X515JAB-X515JA:~$ pip install -U spacy
Defaulting to user installation because normal site-packages is not writeable
Collecting spacy
  Obtaining dependency information for spacy from https://files.pythonhosted.org/packages/58/93
manylinux2014_x86_64.whl.metadata
  Downloading spacy-3.6.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (
Collecting spacy-legacy<3.1.0,>=3.0.11 (from spacy)
  Using cached spacy_legacy-3.0.12-py2.py3-none-any.whl (29 kB)
Collecting spacy-loggers<2.0.0,>=1.0.0 (from spacy)
  Obtaining dependency information for spacy-loggers<2.0.0,>=1.0.0 from https://files.pythonhos
none-any.whl.metadata
  Downloading spacy_loggers-1.0.5-py3-none-any.whl.metadata (23 kB)
Collecting murmurhash<1.1.0,>=0.28.0 (from spacy)
  Obtaining dependency information for murmurhash<1.1.0,>=0.28.0 from https://files.pythonhoste
310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metad
  Downloading murmurhash-1.0.10-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_
Collecting cymem<2.1.0,>=2.0.2 (from spacy)
```

Choix de la langue : installation du package

Package français:
python -m spacy download fr_core_news_sm

Package anglais:
python -m spacy download en_core_web_sm

Autres packages:
https://spacy.io/models/fr

fr_core_news_sm

RELEASE DETAILS
Latest: 3.6.0

French pipeline optimized for CPU. Components: tok2vec, morphologizer, parser, senter, ner, attribute_ruler, lemmatizer.

LANGUAGE	FR French
TYPE	CORE Vocabulary, syntax, entities
GENRE	NEWS written text (news, media)
SIZE	SM 15 MB
COMPONENTS	tok2vec, morphologizer, parser, senter, attribute_ruler, lemmatizer, ner
PIPELINE	tok2vec, morphologizer, parser, attribute_ruler, lemmatizer, ner
VECTORS	0 keys, 0 unique vectors (0 dimensions)
DOWNLOAD LINK	fr_core_news_sm-3.6.0-py3-none-any.whl
SOURCES	UD French Sequoia v2.8 (Candito, Marie; Seddah, Djamé; Perrier, Guy; Guillaume, Bruno) WikiNER (Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, James R Curran) spaCy lookups data (Explosion)
AUTHOR	Explosion
LICENSE	LGPL-LR

Tokenisation, lemmatisation, pos tagging

```
['Bienvenu', 'Bienvenu', 'PROPN']
['dans', 'dans', 'ADP']
['le', 'le', 'DET']
['cours', 'cours', 'NOUN']
['de', 'de', 'ADP']
['Documents', 'document', 'NOUN']
['Structurés', 'structurer', 'ADJ']
['!', '!', 'PUNCT']
['Il', 'il', 'PRON']
["s'", 'se', 'PRON']
['agit', 'agir', 'VERB']
["d'", 'de', 'ADP']
['un', 'un', 'DET']
['cours', 'cours', 'NOUN']
['enseigné', 'enseigner', 'VERB']
['par', 'par', 'ADP']
['M.', 'm.', 'NOUN']
['Pierre', 'Pierre', 'PROPN']
['Magistry', 'Magistry', 'PROPN']
['.', '.', 'PUNCT']
```

```
text = ("Bienvenu dans le cours de Documents Structurés !
Il s'agit d'un cours enseigné par M. Pierre Magistry.")
doc = nlp(text)

 for token in doc:
            analyse = [token.text, token.lemma_, token.pos_, token.dep_]
            print(analyse)
```

Reconnaissance d’entités nommées

```
text = ("Bienvenu dans le cours de Documents Structurés !
Il s'agit d'un cours enseigné par M. Pierre Magistry.")
doc = nlp(text)


for entity in doc.ents:
   print(entity.text, entity.label_)
```

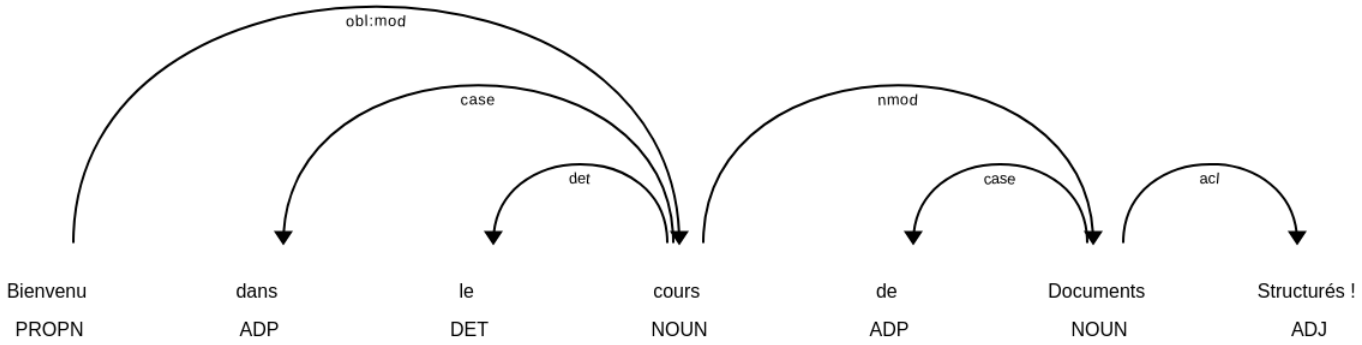
Output:

Documents Structurés ! **MISC**
M. Pierre Magistry **PER**

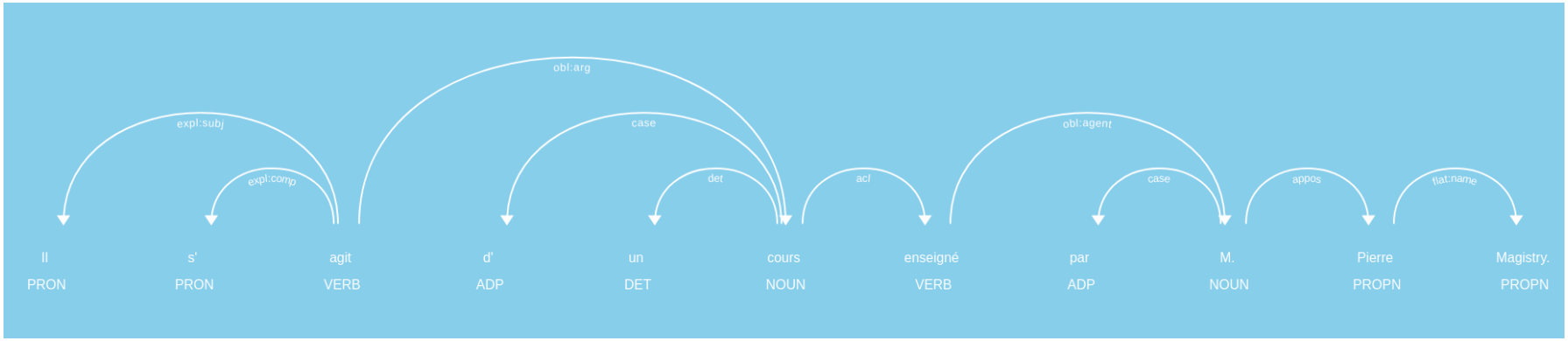
Entitée non catégorisée

Personne

Analyse en dépendance



Style de l’arbre




```
f = open("tuto_spacy.html", "w")
   options = {"color": "#ffffff", "bg": "#87CEEB"}
   html = displacy.render(doc, style='dep', options=options, page=True)
   f.write(html)
```

Calcul de similarité

```
phrase1 = "Je suis étudiante à l'université Sorbonne Nouvelle"
   phrase2 = "J'étudie à la faculté P3"
   doc1 = nlp(phrase1)
   doc2 = nlp(phrase2)
   similarity_score = doc1.similarity(doc2)
   print(phrase1, "<->", phrase2, similarity_score)
```

Remarque : Utiliser un web package plutôt qu'un small (sm)

```
/home/anne/Documents/M2/classes/Docs_structures/demo_spacy.py:33: UserWarning: [W007] The model you're using has no word vectors loaded, so the result of the Doc.similarity method will be based on the
gger, parser and NER, which may not give useful similarity judgements. This may happen if you're using one of the small models, e.g. `en_core_web_sm`, which don't ship with word vectors and only use co
ext-sensitive tensors. You can always add your own word vectors, or use one of the larger models instead if available.
  similarity_score = doc1.similarity(doc2)
Je suis étudiante à l'université Sorbonne Nouvelle <-> J'étudie à la faculté P3 0.5751566573710091
```