



Présentation par : Orphila OURAGA, Assia BOUZNAD, Sonia CHIBOUB

Sources

- Site source : [Stanza](#) et [CoreNLP](#)
- [A Python Natural Language Processing Toolkit for Many Human Languages](#) (Google Scholar)
- Corpus (Web scraping) :
 - o [Ici-japon](#) (japonais)
 - o [github](#) (coréen)
 - o [Gutenberg](#) (anglais, espagnol, français)
 - o [Wikipédia](#) (afrikaans)

Sommaire

1. Stanza de CoreNLP , c'est quoi ?

- Présentation
- Documentation de l'outil
- Etat de l'art

2. Comment l'avons-nous utilisé ?

- Comprendre l'outil

3. Conclusion

- Avantages/Inconvénients

Stanza de CoreNLP, c'est quoi ?

Présentation

- Par Stanford NLP Group (Stanford University)
- Librairie python de **CoreNlp**
- Outil du TAL



Documentation de l'outil

- Installation guidée pour python
- Tutoriel pour l'utilisation et l'exploitation
- Accès au corpus d'entraînement

Language	Treebank	Icode	Default	Tokens	Sentences	Words	UPPERCASE
English	ESLSpok	en		100.00	92.81	100.00	98.28
English	EWT	en		99.17	89.16	98.92	96.20
English	GUM	en		99.78	95.34	99.69	97.59
English	LinES	en		99.96	88.63	99.96	97.42
English	ParTUT	en		99.72	100.00	99.63	96.70
Erzya	JR	myv	✓	99.63	96.92	99.08	84.84
Estonian	EDT	et	✓	99.97	92.88	99.97	97.19
Estonian	EWT	et		98.79	79.57	98.79	93.14
Faroese	FarPaHC	fo	✓	99.65	93.94	99.64	96.51
Finnish	FTB	fi		100.00	90.13	99.98	96.88
Finnish	TDT	fi	✓	99.73	90.70	99.70	97.51
French	GSD	fr	✓	99.71	95.03	99.50	97.63
French	ParTUT	fr		99.88	100.00	99.42	96.92
French	ParisStories	fr		99.76	92.80	99.56	97.08
French	Rhapsodie	fr		99.81	99.82	99.52	97.31
French	Sequoia	fr		99.88	87.46	99.64	98.80
Galician	CTG	gl	✓	99.91	99.30	99.37	97.11

<https://stanfordnlp.github.io/stanza/performance.html>

État de l'art

[A Python Natural Language Processing Toolkit for Many Human Languages, Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning, Stanford University \(2003\)](#) : L'article démontre une présentation détaillée de l'**outil Stanza de Corenlp** et en effet, il est avantageux car il comporte une **couverture linguistique vaste** et une **précision élevée** (anglais et chinois) grâce à son architecture entièrement neurale. Les chercheurs nous expliquent qu'inclure **une réduction des modèles en taille** pourrait être efficace et l'ajout de **nouvelles fonctionnalités** pour enrichir encore plus l'**analyse textuelle**.

Comment l'avons-nous utilisé ?

Comprendre l'outil

Processeurs (Neural Pipeline)

- Tokenize, Lemma : Tokenisation et lemmatisation
- POS : part of speech (Partie du discours)
- Constituency (Groupes syntaxiques)
- Depparse : Dependency parsing (Relation de dépendances syntaxiques)
- NER : Named-entity recognition (Entités nommées)
- Mwt : Multi-word-Token (Mots composés)
- Sentiment (positifs/négatifs)

Anglais

Processor	Package
tokenize	combined
mwt	combined
pos	combined_charlm
lemma	combined_nocharlm
constituency	ptb3-revised_charlm
depparse	combined_charlm
sentiment	sstplus_charlm
ner	ontonotes-ww-multi_charlm

Afrikaans

Processor	Package
tokenize	afribooms
pos	afribooms_charlm
lemma	afribooms_nocharlm
depparse	afribooms_charlm
ner	nchlt

Français

Processor	Package
tokenize	combined
mwt	combined
pos	combined_charlm
lemma	combined_nocharlm
depparse	combined_charlm
ner	wikiner

Espagnol

Processor	Package
tokenize	ancora
mwt	ancora
pos	ancora_charlm
lemma	ancora_nocharlm
constituency	combined_charlm
depparse	ancora_charlm
sentiment	tass2020_charlm
ner	conll02

Coréen

Processor	Package
tokenize	kaist
pos	kaist_nocharlm
lemma	kaist_nocharlm
depparse	kaist_nocharlm

```
2024-05-09 18:06:18 INFO: Using device: cpu
2024-05-09 18:06:18 INFO: Loading: tokenize
2024-05-09 18:06:18 INFO: Loading: pos
2024-05-09 18:06:19 INFO: Loading: lemma
2024-05-09 18:06:19 INFO: Loading: depparse
2024-05-09 18:06:19 INFO: Done loading processors!
```

Japonais

Processor	Package
tokenize	gsd
pos	gsd_charlm
lemma	gsd_nocharlm
constituency	alt_charlm
depparse	gsd_charlm
ner	gsd

```
2024-05-09 18:18:53 INFO: Using device: cpu
2024-05-09 18:18:53 INFO: Loading: tokenize
2024-05-09 18:18:54 INFO: Loading: pos
2024-05-09 18:18:54 INFO: Loading: lemma
2024-05-09 18:18:54 INFO: Loading: constituency
2024-05-09 18:18:54 INFO: Loading: depparse
2024-05-09 18:18:55 INFO: Loading: ner
2024-05-09 18:18:56 INFO: Done loading processors!
```

Conclusion

Avantages

- Installation
- Documentation
- Compatible avec plusieurs langages de programmation
- Analyse textuelle approfondie
- Langues Naturelles (+70 langues)

Inconvénients

- Processeurs différentes selon les langues
- Anglais > autres langues
- Ambiguïté syntaxique
- Nécessite une connexion Internet
- Prend beaucoup d'espace

Merci de votre attention !

DÉMONSTRATION

EN LIGNE

<http://stanza.run/>

Exemples :

- « Le petit garçon mange une pomme. »
- « La petite porte le voile. »

SUR PYTHON

<https://we.tl/t-ysrVobjMCI>