
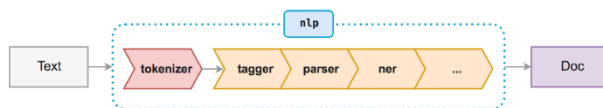


spaCy (v2.1.0)

	Outil de traitement automatique du langage naturel	
	:	Multiplateforme
	Développé par :	Explosion AI
Open source, MIT	Personne de contact :	katy.fokou@smals.be

Fonctionnalités

spaCy est un outil de traitement automatique du langage naturel (NLP) basé sur le langage de programmation Python. Il est construit pour une application pratique du NLP, notamment pour des tâches telles que l'[étiquetage morpho-syntaxique](#) (POS tag), la [reconnaissance d'entités](#) (NER) ou l'[analyse syntaxique](#) (parser).



spaCy supporte 8 langues (modèles) et pour chaque langue, spaCy propose des modèles en natif les fonctionnalités linguistiques élément de classification de textes.

suivantes : POS tagger, parser et NER. À

Chaque élément peut être facilement activé et désactivé selon les besoins. Le module statistique NER standard reconnaît le français (PER, ORG, MISC, LOC). En plus des modèles statistiques, spaCy propose des fonctionnalités de *matching* similaires aux expressions régulières et une fonction de calcul de similarité sémantique.

Les modèles natifs peuvent être enrichis de plusieurs manières, en modifiant le *tokenizer*, en ré-entraînant le *tagger*, *parser* et *NER* ou en ajoutant de nouvelles entités basées sur des règles prédéfinies. Il est aussi possible de développer un nouveau modèle « from scratch ».

Notons aussi que s de dépendances : displaCy.
Les résultats sont disponibles via un serveur web ou dans [Jupyter notebook](#) et bien à une application web.

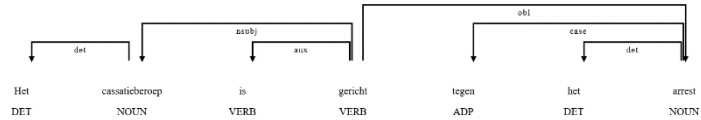
Conclusions & Recommandations

spaCy est une librairie simple avec le langage Python. Une des qualités de cet outil vient de sa modularité ; il est facile et/ou de modifier des composants du pipeline. un outil rapide et performant, orienté vers la production intègre parfaitement , on le recommande donc pour du NLP.

Tests et Résultats

La majorité des tests effectués ci-après portent sur des documents juridiques en néerlandais, une langue pour laquelle les modèles ne sont pas toujours disponibles ou alors sont peu précis.

Parsing et tagging : le *parsing* et le *tagging* se font de façon correcte et peuvent être visualisés avec displaCy.



Customisation du modèle - ajout d'un modèle de NER basé sur des règles prédéfinies : spaCy au composant NER existant ; de façon relativement simple et complémentaire au NER statistique en utilisant des règles ou *pattern* pour détecter les entités. Dans -dessous, une entité « KAMER » a été ajoutée au modèle existant. Pour cela on définit une règle qui sera appliquée dans le pipeline après le NER statistique.

gedeeltelijk vernietigde arrest. Houdt de kosten aan en laat de beslissing daaromtrent aan de feitenrechter over. Verwijst de aldus beperkte zaak naar het arbeidshof te **Brussel Loc** . 28 maart 2011 – 3^e kamer **KAMER** – **Voorzitter Loc** en verslaggever: **C. Storck PER** , voorzitter – **Gelijkkluidende PER** conclusie: **M. Palumbo PER** , advocaat-generaal met opdracht – **Advocaat PER** : **J. Oosterbosch PER** .

Customisation du modèle - ajout d'une nouvelle entité au modèle statistique : dans certaines situations, les règles sont difficiles à définir et on préfère utiliser du *machine learning* pour apprendre au modèle à reconnaître entité. Dans -dessous, un modèle NER est entraîné à extraire t juridique à partir de 25 données de *training*.

onderdeel 1. **Artikel 1017, eerste lid, Gerechtelijk Wetboek bepaalt dat, tenzij LAW** bijzondere wetten anders bepalen, ieder eindvonnis, zelfs ambtshalve, de in het ongelijk gestelde partij in de kosten verwijst, onverminderd de overeenkomst tussen partijen, die het eventueel bekrachtigt. Krachtens **artikel 1017, vierde lid, Gerechtelijk Wetboek kunnen de kosten worden omgeslagen zoals de rechter het raadzaam oordeelt, hetzij wanneer de partijen onderscheidenlijk omtrent enig geschilpunt in het ongelijk zijn gesteld, hetzij over echtgenoten, bloedverwanten in de opgaande lijn, broeders en zusters of aanverwanten in dezelfde graad. LAW** Krachtens **artikel 1018, eerste lid, 6°, Gerechtelijk Wetboek omvatten de kosten de rechtsplegingsvergoeding zoals bepaald in artikel 1022 Gerechtelijk Wetboek. LAW** Volgens **artikel 1022, eerste lid, Gerechtelijk Wetboek, is de rechtsplegingsvergoeding een forfaitaire tegemoetkoming in**

Le modèle détecte correctement les articles de loi présents dans le texte mais considère toute la phrase comme entité. Ceci est en parti du à la façon dont sont définies les données de training.

Customisation du modèle - ajout d'un classificateur : le modèle entraîné sur un set de seulement 300 textes avec le composant «TextCategorizer» de spaCy s est avéré moins performant qu SVM entraîné sur les même données. Ceci est probablement dû au fait que spaCy utilise le deep learning.

Calcul de similarité entre documents : le calcul de similarité se fait sur base de la mesure du cosinus entre les vecteurs. Pour certaines langues, spaCy propose des modèles avec vecteurs et sans vecteurs (ou plutôt avec des pseudo vecteurs). Les tests effectués sur des documents en français ont révélé que les mesures de similarité étaient plus cohérentes avec le modèle sans vecteurs vecteurs inclus.

Conditions d'utilisation & Budget

spaCy est une librairie gratuite et sous licence MIT, à utiliser dans un environnement Python