

Création d'un modèle NLP avec Spacy

Table des matières

1.	C'est quoi le modèle Spacy :	2
2.	Quelques caractéristiques clés de SpaCy :	2
3.	Pourquoi le modèle Spacy ?	2
4.	Etapes du modèle NLP :	3



1. C'est quoi le modèle Spacy :

Spacy est une bibliothèque open-source en Python utilisée principalement pour le traitement automatique du langage naturel (NLP). Elle offre des fonctionnalités avancées pour le traitement du texte, telles que l'analyse syntaxique, la reconnaissance d'entités nommées, la lemmatisation et d'autres tâches liées au NLP.

2. Quelques caractéristiques clés de SpaCy :

- ✓ **Analyse syntaxique** : SpaCy analyse la structure grammaticale des phrases, en identifiant les parties du discours, les dépendances syntaxiques et d'autres informations linguistiques.
- ✓ **Reconnaissance d'entités nommées** : SpaCy est capable de détecter et classer des entités nommées telles que les personnes, les lieux, les organisations, etc.
- ✓ **Lemmatisation** : SpaCy fournit une lemmatisation efficace pour obtenir les formes de base des mots.
- ✓ **Traitement rapide** : Il est conçu pour être rapide et efficace, ce qui en fait un choix populaire pour le traitement de grands ensembles de données textuelles.
- ✓ **Intégration de modèles pré-entraînés** : SpaCy propose des modèles pré-entraînés pour différentes langues, permettant une utilisation immédiate pour certaines tâches sans nécessiter un entraînement supplémentaire.

3. Pourquoi le modèle Spacy ?

Le choix d'un modèle pour le traitement du langage naturel (NLP) dépend souvent des besoins spécifiques du projet, des langues prises en charge, de la performance, de la facilité d'utilisation et d'autres critères.

Nous avons identifié deux différents modèles pour le traitement du langage naturel (NLP) : **BERT** et **SPACY**.

	BERT	SPACY
Architecture et Objectif	Basé sur l'architecture Transformer. Conçu pour comprendre le contexte des mots dans une phrase en utilisant une approche bidirectionnelle, capturant ainsi les dépendances sémantiques complexes.	Propose l'analyse syntaxique, la reconnaissance d'entités nommées, la lemmatisation, etc. Elle utilise une architecture basée sur les règles et des modèles statistiques.
Tâches	Excelle dans une variété de tâches NLP, y compris la classification de texte, la traduction, la réponse aux questions, etc.	Couvre un large éventail de tâches NLP, mais chaque modèle SpaCy est généralement spécialisé dans une tâche spécifique (par exemple, l'analyse syntaxique)

		ou la reconnaissance d'entités nommées).
Facilité d'utilisation	BERT peut être plus complexe à mettre en œuvre en raison de sa structure sophistiquée. L'utilisation de modèles pré-entraînés peut simplifier le processus.	SpaCy est réputé pour sa facilité d'utilisation. Son API est bien documentée et son intégration dans les projets est généralement plus simple.
Taille des modèles	Le modèle BERT peut être relativement grands en taille, nécessitant souvent des ressources informatiques plus importantes.	SpaCy propose des modèles de taille plus modeste, ce qui peut être plus adapté pour des applications avec des contraintes de ressources.

En résumé, BERT est souvent utilisé lorsque des tâches NLP complexes nécessitent une compréhension approfondie du contexte, tandis que SpaCy est une bibliothèque plus polyvalente et conviviale, adaptée à une gamme de tâches NLP standard.

4. Etapes du modèle NLP :

- **Obtention des données :** développement d'un programme python qui permet de générer les données.
- **Nettoyage des données :** Nettoyer les données en supprimer la ponctuation, adverbes, verbes, nom...etc.
- **Ajouter les flags <DEPART> et <ARRIVEE> avec l'indexation :** exemple ('albert à avelsches de saint-lambert-et-mont-de-jeux', {'entities': [(22, 51, 'DEPART'), (9, 18, 'ARRIVEE')]})
- **Entraînement du modèle.**
- **Evaluation du modèle :** Precision: 95.66%, Accuracy: 94.75%.
- **Tester le modèle.**