

## **M2SOLO34 Corpus, ressources et linguistique outillée**

### **TD1 : Introduction aux Corpus et à l'IA**

Wiem TAKROUNI

Sorbonne Université

Master « Langue et Informatique »

UFR Sociologie et Informatique pour les Sciences Humaines

Semestre 2, 2024-2025, le 28 mars 2025

#### **Objectifs du TD**

- 1. Installation et préparation de l'environnement**
- 2. Annotation automatique d'un texte avec SpaCy et Stanza**
- 3. Comparaison des résultats et évaluation de la qualité**
- 4. Utilisation d'un modèle pré-entraîné pour améliorer l'annotation**

#### **Problème Global : Annotation et Analyse d'un Corpus de Tweets**

Vous travaillez sur un projet de TAL qui vise à **analyser et annoter automatiquement un corpus de tweets en français**. Votre objectif est de comparer différentes méthodes d'annotation et d'évaluer leur qualité.

Installation et Préparation de l'Environnement

**Créez un environnement virtuel** (si ce n'est pas déjà fait) et activez-le.

```
python -m venv mon_env  
mon_env\Scripts\activate # Sous Windows  
source mon_env/bin/activate # Sous Linux/Mac
```

**Installez les bibliothèques nécessaires :**

```
pip install spacy stanza datasets transformers torch
```

**Téléchargez les modèles linguistiques :**

Pour **SpaCy**

```
python -m spacy download fr_core_news_sm
```

Pour **Stanza**

```
import stanza
stanza.download('fr')
```

**Téléchargez un petit corpus de tweets en français :**

```
from datasets import load_dataset
dataset = load_dataset("tweet_eval", "sentiment") # Dataset annoté en sentiments
print(dataset['train'][0]) # Afficher un exemple de tweet
```

**Question 1:** Quelles différences remarquez-vous entre les modèles téléchargés de SpaCy et de Stanza en termes de fichiers et de taille ?

### EXERCICE 1 : Annotation Automatique d'un Tweet

Annoter un tweet avec **SpaCy** et **Stanza** et comparer les résultats.

**Chargez un tweet du corpus précédemment téléchargé.**

```
tweet = dataset['train'][0]['text']
print("Tweet :", tweet)
```

**Annoter le tweet avec SpaCy**

```
import spacy
nlp_spacy = spacy.load("fr_core_news_sm")
doc_spacy = nlp_spacy(tweet)

for token in doc_spacy:
    print(token.text, token.pos_, token.ent_type_)
```

### Annoter le tweet avec Stanza

```
import stanza
nlp_stanza = stanza.Pipeline('fr')
doc_stanza = nlp_stanza(tweet)


for sentence in doc_stanza.sentences:
    for word in sentence.words:
        print(word.text, word.upos, word.ner)
```

**Question 1 :** Quelle est la différence entre les types d'étiquettes retournées par SpaCy et Stanza ?

**Question 2 :** Quel modèle donne le plus d'informations ?

**Exercice 2 :** Comparer les résultats de l'annotation et mesurer leur qualité.

1. Affichez et comparez les annotations obtenues par SpaCy et Stanza.
2. Quels sont les mots qui ont des différences d'annotation ?
3. Mesurez l'accord entre les deux outils en comparant leur sortie.

 **Indication :** Pour mesurer la similarité, on peut compter le nombre de mots ayant la même étiquette entre SpaCy et Stanza.

**Exercice 3 :** Utiliser un modèle BERT pré-entraîné pour générer des annotations plus précises.

Installez et chargez **CamemBERT** pour l'étiquetage des entités nommées (NER).

```
from transformers import pipeline

ner_pipeline = pipeline("ner", model="Jean-Baptiste/camembert-ner",
                        tokenizer="Jean-Baptiste/camembert-ner")
annotations = ner_pipeline(tweet)
print(annotations)
```

**Question 1 :** Comparez les annotations obtenues avec celles de SpaCy et Stanza.

**Question 2 :** BERT améliore-t-il l'annotation ?

**Question 3 :** Quels sont les avantages et limites de cette approche ?