Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique

Filière Génie Informatique et Digitalisation

Text Analytics

Dr. H. SEBBAQ *h.sebbaq@gmail.com*

Pr. N. EL FADDOULI *nfaddouli@gmail.com*

2023-2024

CC-BY NC SA

Découvrons quelques PRs

Document Reset → Supprime les annotations existantes

Tokeniser → Tokenise les annotations

Gazetter → Cree les annotations Lookup → Identifier des entités nommes en se basant sur des listes

Sentence Splitter → les annotations ‘Sentence Split’

POS tagger → Ajoute la propriété de catégorie grammaticale aux annotations du token

NE transducer → les annotations : Date, Person, Location, Organisation, Money et Percent → l'annotateur (tager) sémantique basé sur le langage JAPE

Orthomatcher → Ajoute la propriété ‘match’ aux annotations du NE transducer (le nombre d’occurrence de la même entité)

 CC-BY NC SA

Créer une nouvelle application

- ❑ Nettoyer GATE en supprimant toutes les ressources et applications (ou redémarrer GATE)
- ❑ Charger le corpus de textes d'actualité
- ❑ Créer une nouvelle application (pipeline de corpus) : Applications → Create New applications → Corpus Pipeline
- ❑ Sélectionner des PRs à ajouter parmi les PRs déjà chargés dans GATE.
- ❑ Nous allons commencer par le PR 'Document Reset', exécuter l'application et vérifier les annotations.
- ❑ Quel est le rôle de ce PR?
- ❑ Ajouter ces PRs un par un en exécutant à chaque fois l'application et vérifiant les résultats

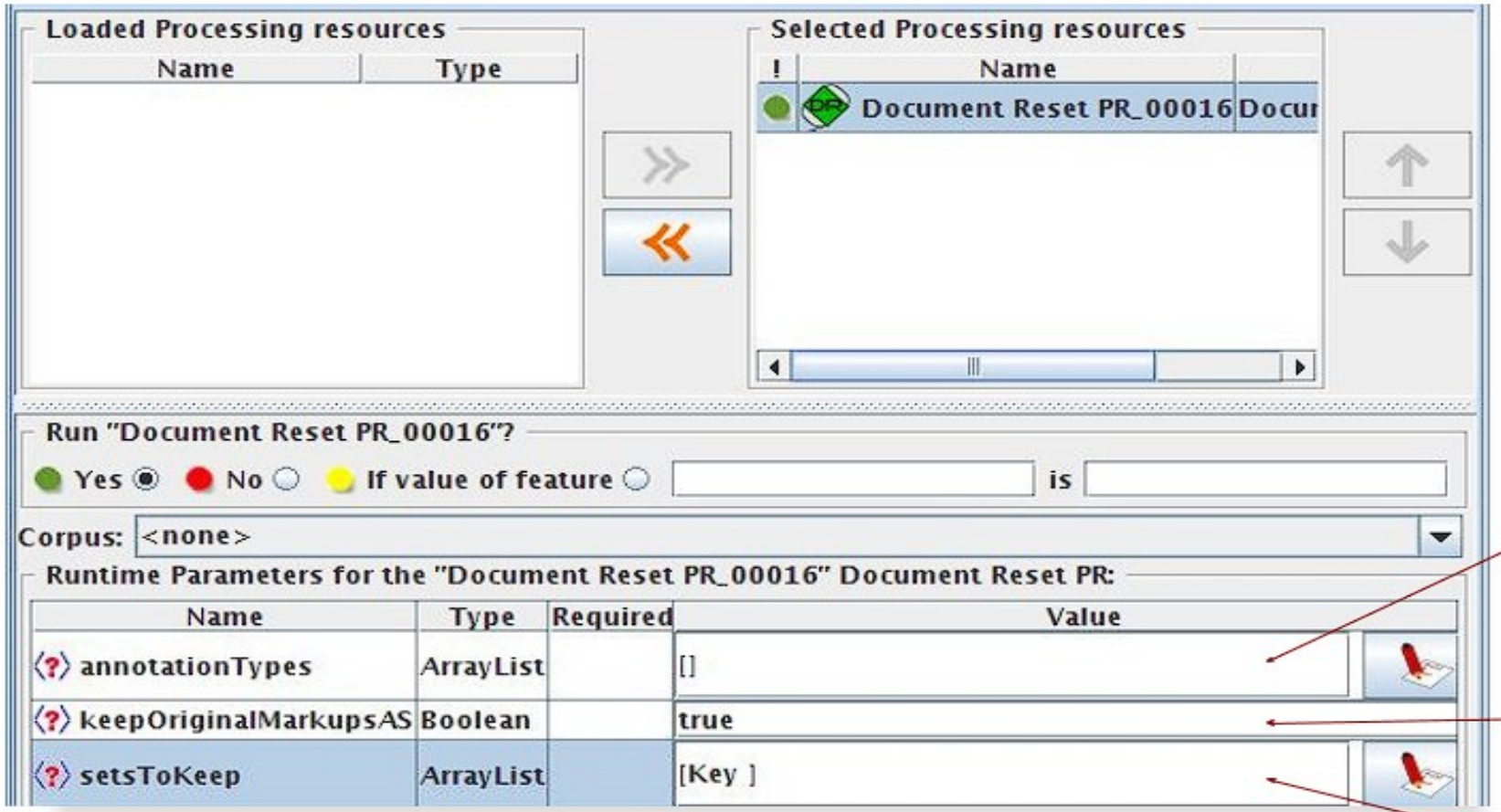
CC-BY NC SA

Document Reset

- Stockage séparé des annotations et du texte : GATE conserve annotations et texte séparément en format StandOff Markup.
- Aucune modification du texte initial : Aucune annotation ou information d'analyse n'est ajoutée au texte original dans GATE.
- Référencement faible et gestion des ressources :Utilisation d'un référencement faible dans la JVM nécessitant une suppression explicite des ressources pour éviter l'accumulation non désirée.
- Conséquences de la conservation des annotations en mémoire :
  - Anciennes annotations peuvent influencer le traitement des documents suivants.
  - Accumulation d'annotations anciennes en mémoire impactant les performances.
- Objectif de la réinitialisation des documents : La PR de réinitialisation vise à nettoyer les anciennes annotations pour libérer la mémoire, évitant l'accumulation inutile de données.

 CC-BY NC SA

Document Reset



Specify any specific annotations to remove. By default, remove all.

Keep Original Markups set

Keep Key set

 CC-BY NC SA

PR : ANNIE English Tokeniser

- Charger **ANNIE English Tokeniser**
- Ajoutez-le à l'application et exécutez-la sur le corpus.
- Voir les annotations Token et SpaceToken
- Quelles sont les différentes valeurs de la caractéristique "kind" que vous observez ?

CC-BY NC SA

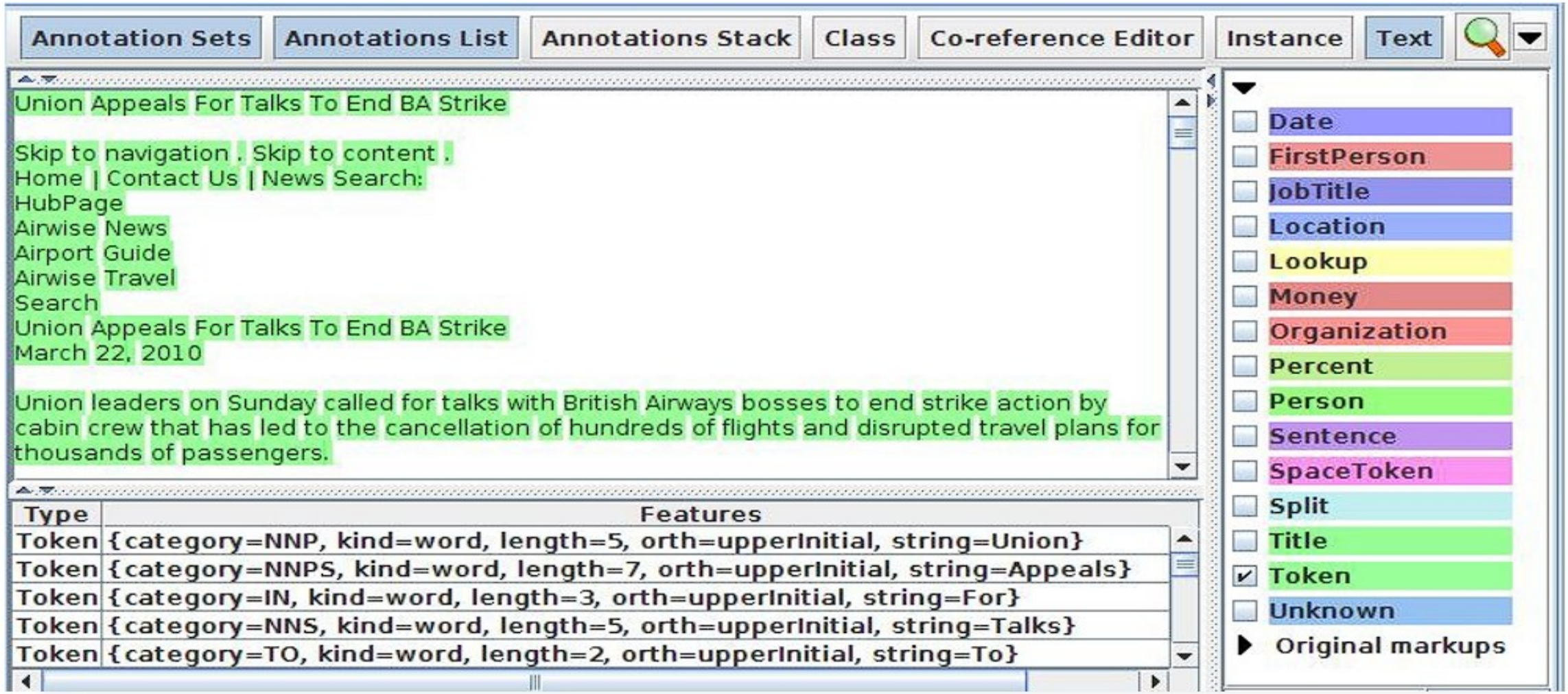
ANNIE English Tokeniser

Divise le texte en éléments très simples tels que des chiffres, des signes de ponctuation et des mots de types différents.

- Produit des annotations Token et SpaceToken avec des caractéristiques de type, d'orthographe, de longueur et de chaîne.
- Le type peut être : mot, nombre, symbole ou ponctuation.
- orth (orthographe) peut être :
  - upperInitial - la lettre initiale est une majuscule, les autres sont des minuscules
  - allCaps - toutes les lettres majuscules
  - lowerCase - toutes les lettres minuscules
  - mixedCaps - tout mélange de lettres majuscules et minuscules n'entrant pas dans les catégories ci-dessus.
- length : nombre de caractères dans le jeton
- string : texte du jeton

CC-BY NC SA

ANNIE English Tokeniser

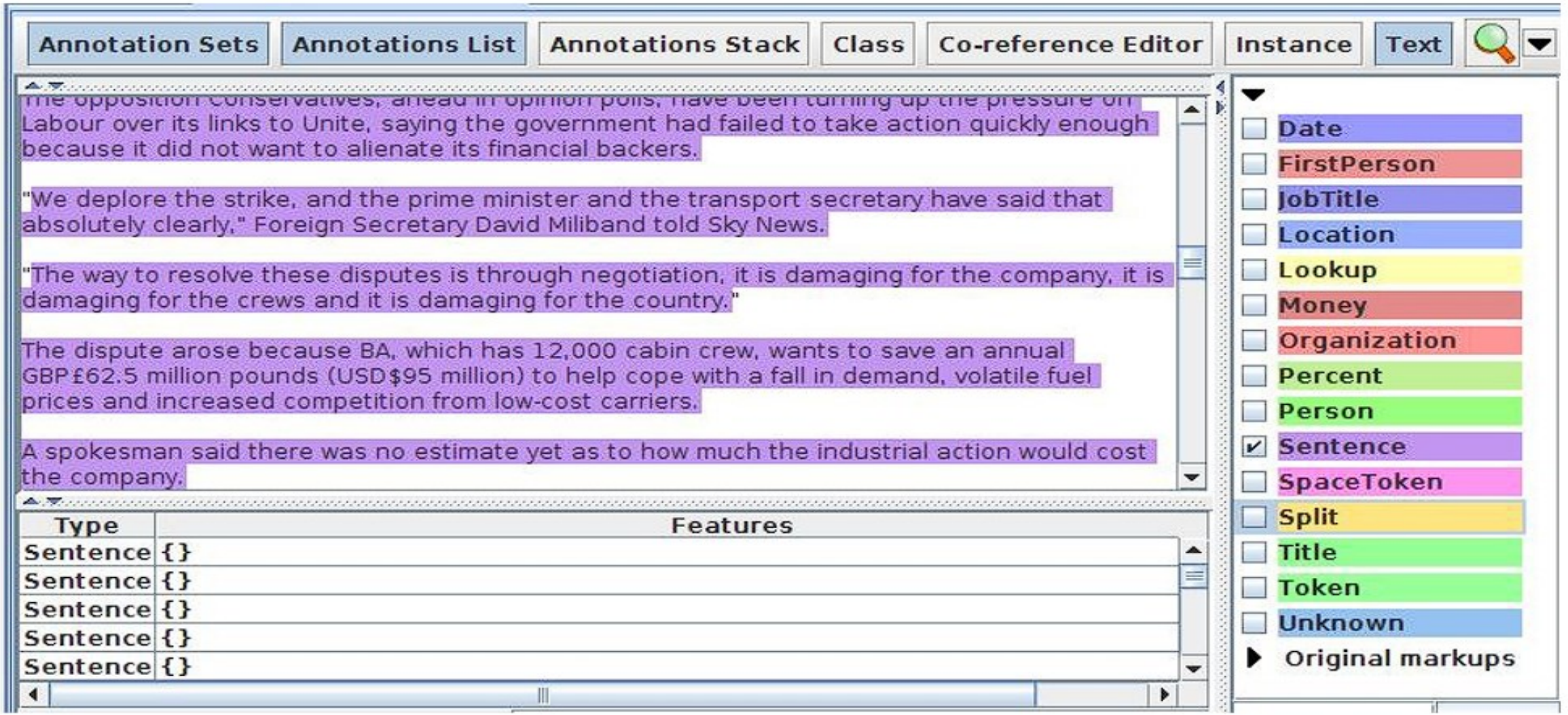


Sentence Splitter

- Le séparateur par défaut trouve les phrases en se basant sur les tokens.
- Crée des annotations de phrases et des annotations de divisions sur les délimiteurs de phrases.
- Utilise un répertoire d'abréviations, etc. et un ensemble de grammaires JAPE (A voir plus tard) qui trouvent les délimiteurs de phrases et annotent ensuite les phrases et les divisions.
- Charger un PR ANNIE Sentence Splitter et ajoutez-le à votre application (à la fin).
- Exécuter l'application et visualisez les résultats

CC-BY NC SA

Sentence Splitter



Analyse Syntaxique-Lexicale

- Ajoutez un ANNIE POS Tagger à votre application
- Ajoutez un analyseur morphologique GATE (**GATE Morphological Analyser)** après le POS Tagger (si ce PR n'est pas disponible, chargez d'abord le plugin Tools).
- Exécutez à nouveau votre application.
- Examinez les caractéristiques des annotations de tokens. → De nouvelles caractéristiques de catégorie et de racine ont été ajoutées.
- Ajouter un PR de stemming en le cherchant dans le store des plugins : Ajouter le **stemmer Snowball**
- Pouvez vous le placer tout seul? Sinon quels sont les PRs a ajouter en amont? Ajoutez les
- Apres l'ajout, ré-exécuter l'application et vérifier les résultats relatifs a ce PR

CC-BY NC SA

Shallow Parsing

- Ajouter par la suite les PRs ‘VP chunker’ et ‘Noun Phrase Chunker’
- Exécuter l’application après l’ajout de chaque PR. Vérifier les résultats
- Quel est le rôle de chaque PR? A quel niveau d’analyse appartiennent-ils? Préciser la technique

CC-BY NC SA

Parsing de constituance et de dépendance

- Explorer le store des plugins pour retrouver un PR permettant de faire du 'Dependendy parsing' → fait partie de l'application Stanford CoreNLP
- Charger l'application Stanford CoreNLP du store des plugins
- Choisir Applications -> Ready Made Applications -> Stanford parser -> English Dependency Parser
- Ajouter et exécuter
- Choisir annotation 'Dependency' dans la liste, Choisir 'annotation Stack' puis observer les résultats
- Choisir annotation 'syntexTreeNode' puis 'annotation stack' et double cliquer sur