Synthèse

Dan :

Source: A Simple Introduction to Natural Language Processing
https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

NLP : Natural Language Processing
Ou TAL Traitement Automatique des Langues

Techno d'aide à la compréhension du langage humain par un ordi, branche de l'IA.
Cheat Sheet
DataScience Simplified
Building API Infra
AI&NLP Workshop

Le NLP est utilisé pour:
Un traducteur comme Google Translate
Le correcteur orthographique de Word ou Grammarly qui utilise le NLP pour la précision grammaticale
Les app. d'assistants perso comme Ok Google, Siri, Cortana et Alexa

Le NLP est considéré comme un tâche compliquée en science des ordinateurs par la nature même du langage humain.
Certaines règles de haut niveau peuvent être abstraites à comprendre pour le NLP comme le langage sarcastique
Il y a les règles de bas niveau comme utilisé le "s" pour la pluralisation des mots.
Ces choses sont faciles à comprendre pour nous mais difficile à implémenter par le NLP pour un ordinateur

1) La syntaxe:

Techniques utilisées par le NLP:

La lemmatisation (Lemmatization): consiste à réduire les différentes significations d'un mot en une forme simplifiée (sans préfixe, suffixe par ex)

Segmentation morphologique (Morphological segmentation) : Elle consiste à diviser les mots en unités individuelles appelées morphèmes.

Segmentation des mots (Word segmentation): Diviser un texte en morceaux (mots ou groupe de mots)

Balisage de la partie du discours (Part-of-Speech tagging) : Associer la signification des mots au discours

Analyse (Parsing): Effectuer une analyse grammaticale de la phrase

Rupture de phrase (Sentence Breaking): Savoir délimiter une phrase dans un gros texte

Etymologie (Stemming): Couper les mots au niveau de leur racine

1) La sémantique :

Named entity recognition (NER): Il s'agit de déterminer les parties d'un texte qui peuvent être identifiées et classées dans des groupes prédéfinis. Les noms de personnes et les noms de lieux sont des exemples de ces groupes.

Word sense disambiguation: Il s'agit de donner un sens à un mot en fonction du contexte.

Natural language generation: Il s'agit d'utiliser des bases de données pour dériver des intentions sémantiques et les convertir en langage humain

Constant:

La NLP (Natural Language Processing) permet à l'ordinateur de reconnaître le langage humain beaucoup utilisé pour la traduction, la reconnaissance vocale, la reconnaissance de chat box ou la détection du spam exemple Siri, Google home, Alexa, dans gmail. Pour les textes, La NLP utilise 7 méthodes :

- sentence Tokenization (Segmentation d'une phrase en multiples éléments)
- word Tokenization (Découper un texte en token)
- Text Lemmatization and Stemming (c'est la représentation des mots sous forme canonique exemple un verbe sera son infinitif)
- Stop Words (c'est des mots filtrés avant et après le traitement d'un texte exemple comme : "et", "le" ...)
- Regex ( )
- Bag-of-Words
- et TF-IDF

NLTK est une bibliothèque pour créer des programmes python pour travailler avec des données en langage Humain. Il est open source

Caroline:

Le langage humain peut être trompeur notamment à cause de phrases à ne pas toujours interpréter au second degré ou encore l'emploi d'un ton cynique d'où l'importance de tenir compte du contexte dans lequel ça a été dit.

Voici les outils employés pour notre traitement du langage:

Bag of words ("sac de mots"):

Un modèle qui permet de compter tous les mots présents dans un extrait de texte. Il crée une matrice d'occurences sans tenir compte de la grammaire ou de l'ordre des mots. Ce qui fait que les articles du style "the" ou "a" (appelés stop words) peuvent ajouter un "bruit" dans l'analyse. Pour y remédier, on rééchelonne la fréquence des mots de sorte que les articles apparaissant trop souvent soient pénalisés. Ce rééechelonnage se nomme "Terme Frequency - Inverse Document Frequency" (TFIDF) qui améliore le "sac de mots" sans pour autant tenir compte du contexte.

	words	rain	a	paper	they	slip	the	universe	...
Words are flowing out like endless rain into a paper cup,	1	1	1	1	0	0	0	0	...
They slither while they pass, they slip away across the universe	0	0	0	0	3	1	1	1	...

Tokenization :

Il segmente le texte en phrases et mots. Les mots tirés du texte sont alors appelés tokens. Les ponctuations ne sont pas pris en compte dans cette segmentation.

Certains noms composés peuvent être séparés en 2 mots alors qu'il faudrait les compter comme un seul mot comme New York. De même les points présents dans les abbréviations sont éludés du fait que la tokenization ne tient pas compte des ponctutations ce qui peut poser poser problème également.

Topic Modeling :
Méthode qui consiste à découvrir des structures cachées dans des textes ou documents. Il classe les textes de sorte d'avoir un thème basé sur le contenu.