Un assistant de productivité qui allie la puissance de l'intelligence collective et de l'intelligence artificielle

Projet data science Jedha

Clement.sirvente@gmail.com / 06.06.70.42.18

Présentation de holi.io

Holi.io est un assistant intelligent basé sur un graphe de connaissance

Nous offrons aux cabinets de conseil une solution de knowledge management qui leur permet de centraliser et mettre à disposition des consultants les ressources essentielles pour leur travail.

L'assistant s'intègre directement dans le travail quotidien des consultants avec :

- un smartbot Slack qui permet de partager et de suggérer des contenus pertinents dans les discussions
- un extension Chrome qui permet de contextualiser les recherches Google avec les ressources de l'entreprise
- un hub de connaissance pour centraliser et explorer les contenus à la manière d'une carte de connaissance boosté par des algorithmes de graphe et des visualisations.

Projet 1: Topic modeling

Problème: Aider les utilisateurs à qualifier plus facilement les contenus grâce à des suggestions de mots clés.

Solution: Développer un algorithme d'extraction de mots clés de contenus (articles de presse). Pour chaque article sortir les mots clés avec le score de pertinence associé.

Understanding: C'est un problème qui fait appel au NLP où un des enjeux est de trouver l'équibre en l'exhaustivité et la pertinence.

Technical requirement: Code de préférence en python ou javascript. Je suis ouvert à des API cloud et à l'utilisation de modèles deep learning.

Identification: [url, title, text (content), tag]

Highly scalable:
1-10k rows for the démo and ideally millions for the prod ;)
<5s per response.

Projet 1: Topic modeling – Approche

Jeu de donnnées suggestion:
1/ MIND – Microsoft News Dataset https://msnews.github.io/
Les jeux de données sur disponible sur le drive.
Il faut utiliser le fichier « news.tsv » qui contient ID, Catégorie, Sous-catégorie, Titre, Résumé, URL, Mots clés du titre, Mots clés du résumé (seul ID, Titre, Url, Mots clés sont utiles) et le fichier « msn-article_content.json » qui contient le texte des articles.
Le jeu de données ne contient pas le contenu des articles je les ai téléchargé avec le crawler. Autres liens utiles : doc Microsoft et command line
2/ Medium - plusieurs jeux de données sont disponible sur Kaggle (à voir)
3/ 20 Newsgroups https://www.kaggle.com/crawford/20-newsgroups

Algorithmes suggestion:
On s'appuie d'abord sur des prétraitement sur les stop words, et des algorithmes de base « TF-IDF », puis sur des algorithmes de ML non supervisés. Algorithmes LSA/PCA gensim word2vec

Projet 1: Topic modeling – bonus entitées nommées

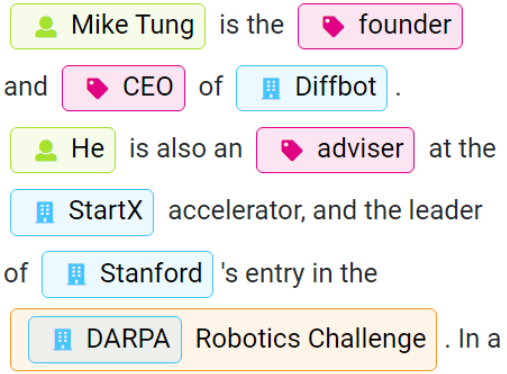
Problème: Aider les utilisateurs à lier les concepts d'un article à des concepts de wikipedia pour facilité la compréhension et la désambiguïsation.

Solution: Développer un algorithme d'extraction d'entités nommées (concept, personne, entreprise) de contenus (articles de presse). Pour chaque article sortir les mots clés avec l'entité associé si pertinent.

Understanding : L'objectif est de faire de la "Wikification" en s'appuyant sur des algorithmes déjà existants.

Algorithmes existants:
Facebook BLINK
Neo4j Graph data science book github

Exemple avec Diffbot



Projet 2: Classification

Problème: Aider les utilisateurs à qualifier rattacher les contenus à un ensemble de thématique défini.

Solution: Développer un algorithme de classification de contenus (articles de presse). Pour chaque article définir si on peut l'associer à une thématique avec le score de pertinence associé.

Understanding: C'est un problème de classification

Technical requirement: Code de préférence en python ou javascript. Je suis ouvert à des API cloud et à l'utilisation de modèles deep learning.

Identification: [url, title, text (content), thématique]

Highly scalable:
1-10k rows for the démo and ideally millions for the prod ;)
<5s per response.

Projet 2: Classification – Approche

Jeu de donnnées suggestion:
1/ MIND – Microsoft News Dataset https://msnews.github.io/
Les jeux de données sur disponible sur le drive.
Il faut utiliser le fichier « news.tsv » qui contient ID, Catégorie, Sous-catégorie, Titre, Résumé, URL, Mots clés du titre, Mots clés du résumé (seul ID, Catégorie, Sous-catégorie, Titre, Url, Mots clés sont utiles) et le fichier « msn-article_content.json » qui contient le texte des articles.
Le jeu de données ne contient pas le contenu des articles je les ai téléchargé avec le crawler. Autres liens utiles : doc Microsoft et command line
2/ Medium - plusieurs jeux de données sont disponible sur Kaggle (à voir)
3/ 20 Newsgroups https://www.kaggle.com/crawford/20-newsgroups

Algorithmes suggestion:
Voir scikit learn https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html
Les méthodes peuvent être supervisé (avec les catégories) ou non supervisé (clustering)

Contacts

clement.sirvente@holi.io
0606704218

Clément Sirvente

Fondateur