Projet ISN - Zero Shot Learning for Automatic Topic Classification for URL Content

1. Quelques définition avant de présenter le projet

Zero-Shot-Learning (ZSL): le ZSL fait référence à un cas d'utilisation spécifique du Machine Learning (et du Deep Learning) où vous voulez que le modèle classifie les données sur la base de très peu ou même d'aucun exemple étiqueté, ce qui signifie classifier à la volée.

Topic Classification: Topic Classification est une forme d'apprentissage non supervisé, de sorte que l'ensemble des sujets possibles sont inconnus a priori.

NLP: NLP est d'abord l'initiale de Natural Language Processing (traitement automatique du langage naturel). Il s'agit donc de l'ensemble des techniques qui permettent à une interface machine d'analyser et traiter automatiquement les propos écrits ou oraux d'un individu et d'exprimer les réponses nécessaires. Les procédures NLP sont notamment utilisées pour traiter les requêtes des moteurs de recherche, pour la traduction automatique et pour faire fonctionner les agents conversationnels (chatbot / voicebot). Dans le cas des voicebots ou assistants vocaux, les capacités de NLP sont associées à des fonctionnalités de synthèse vocale. Le

traitement automatique du langage naturel repose en grande partie sur les techniques de mise en oeuvre de l'intelligence artificielle.

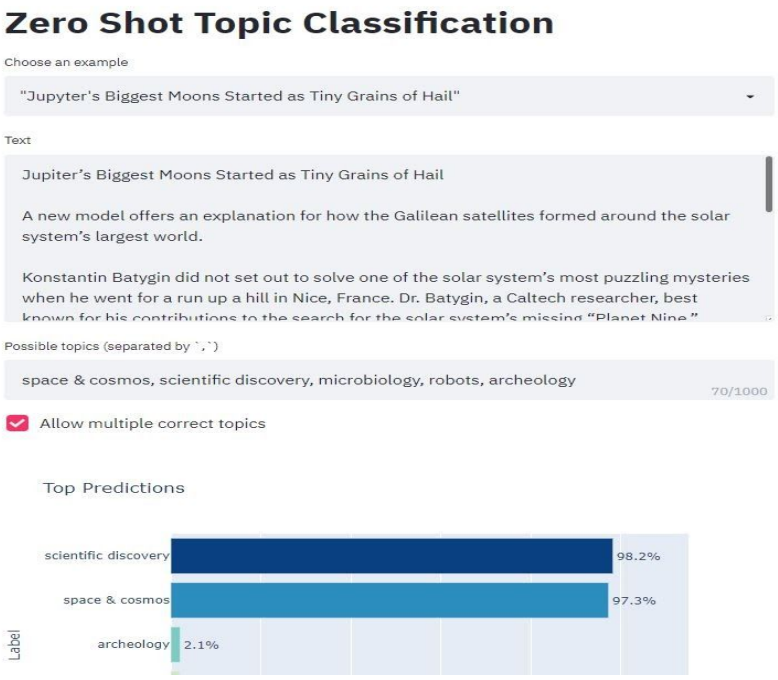
Scraping: Le scraping est une technique d'extraction du contenu de sites Web, via un script ou un programme, dans le but de le transformer pour permettre son utilisation dans un autre contexte, par exemple le référencement. Il peut être utilisé pour de l'extraction de texte mais aussi de tableaux (données structurées).

2. Le Projet

a. Etats des lieux

Le 29 mai 2020, le papier de recherche **Zero Shot Topic Classification** est sorti avec une démonstration de l'outil ici: http://35.208.71.201:8000/. Cette méthode permet d'associer à une liste de sujets fournie par l'utilisateur une probabilité permettant de quantifier le lien entre chaque éléments de cette liste et un texte donnée. Cela permet d'associer un Label à un texte. Cependant le modèle utilisé n'a pas besoin d'entraînement et se base sur une similarité sémantique entre le texte et la liste d'éléments.

Voici un exemple pour illustrer l'outil:



Dans cette exemple nous avons une liste **définie par l'utilisateur et un texte**. La méthode de Zero Shot Learning a permis d'associer "scientific discovery" et "space & cosmos" au texte sans entraînement au préalable.

b. Automatic & for Content URL

i. La première étape de ce projet sera de bien comprendre les différentes méthodes de NLP. Vous serez guidé afin de monter en compétence la dessus et d'avoir une vision globale du NLP et des différentes méthodes de vectorisation. Suite à cette première étape, il sera nécessaire de prendre en main la Library **Hugging Face** sur lequel votre travail sera basé.

ii. Le premier livrable se tournera vers l'extraction de mots clés d'un texte donnée. Cette méthode sera dans un premier temps réalisé sur de l'anglais. Il existe de nombreuses approches pour la partie Keywords Extraction. Ici il sera demandé de créer une fonction permettant de prendre en entrée un texte en anglais et de ressortir un liste de mots clés.

iii. La seconde partie sera l'assemblage de la première partie et de la partie Zero Shot Learning exposé ci dessus. Cela permettra de passer d'une méthode manuelle à une méthode automatique. Le second livrable sera une fonction prenant en entrée une texte et récupérant en sortie une liste de mots-clés avec une probabilité associée. Un seuil sera défini afin de ne garder que les probabilité au dessus de ce seuil.

iv. La dernière partie consistera à réaliser le scraping d'URL, d'en extraire le texte pertinent afin qu'il puisse aliment les deux blocs définis dans les étapes précédentes. Cela permettra ainsi d'associer à une URL un Topic. Le livrable finale sera une fonction permettant de prendre entrée une URL et de ressortir la liste des Topics associés à celle ci.

v. **Bonus**: Réaliser une interface avec **Streamlit**

3. Conclusion

Une liste de ressources vous sera fourni afin de vous permettre de trouver les bons modèles, de vous aider dans le projet et de réaliser l'interface en bonus.