Note sur le projet: Zero Shot Learning for Automatic Topic Classification for URL Conte

(Apprentissage à zéro coup pour automatisé la classification des sujets pour les URL)

1. Quelques définitions

-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.

-Apprentissage non supervisé :

L’apprentissage non supervisé est une tache d’apprentissage automatique sans utilisée de fonction de prédiction, ainsi elles ne se basent sur aucun exemple choisi au préalable.

-Natural language processing (NLP):

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 (ex : google home, traducteur). 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).

1. Notre objectif

Le 29 mai 2020, le papier de recherche Zero Shot Topic Classification est sorti. Il nous présente un outil qui permet d’associer à un texte donnée une liste de sujet (fournie par l’utilisateur). Mais ici, le modèle utilisé n’a pas besoin d’entraînement (et donc c’est une apprentissage non supervisé) et il se base donc sur les similarité sémantique entre le texte et la liste d’éléments composants la liste de sujet.

Petite étude de l’outil mis a disposition : (http://35.208.71.201:8000/)

Prenons l’exemple “IMBD Avengers review”

Lorsqu’on lit le texte, on peut voir que c’est une review d’un des films Marvels exprimant ici son enchantement grâce à un “mash up” de super-héros. Il remercie les cinéastes et les accessoiristes d’avoir rendu ce film plus qu’un simple film d’action avec des cascades et super-héros mais d’avoir inclus la profondeur de chaque personnage.

On peut donc voir en lisant la review qu’on parle de film d’action avec des superheros.

Lorsque l’on utilise l’outil de Zero Shot Topic Classification, on peut voir directement que les sujets abordé dans ce texte sont principalement les films puis superheros/action. On peut surtout voir qu’on ne parle pas de livres.

J’ai rajouté les sujets “story” et “dream” pour voir l’impact et on voit que l’outil fonctionne bien car on parle bien d’histoire dans cette review (27.4%), par contre le sujet “dream” Apparaît a 94.6%

1. Comprendre les différents méthodes de NLP puis prendre en main la librairie “Hugging Face”
2. Premier programme qui permet d’extraire les mots clés d’un texte (en anglais)
3. Deuxième programme qui prendra un texte en entré et qui ressort une liste de mots-clés ainsi que leur probabilité de bien être dans le texte (on fixera un seuil minimal)
4. Dernier programme qui permet de prendre un URL en entrée et de ressortir la liste des sujets associés a cet URL. Donc cette fonction prend l’URL en entrée et utilise les deux premiers programmes pour analyser le texte de l’URL et en ressortir les mots-clés important avec leur probabilité.
5. Quelques recherches réalisé

Le NLP (Traitement automatique du langage naturel) est une technologie dont l’objectif est la communication entre l’ordinateur et l’être humain sur un même niveau. Le NLP allie les connaissances issues de la linguistique et les méthodes les plus récentes de l’informatique et de l’intelligence artificielle.

Afin que le Natural Language Processing fonctionne, il est d’abord nécessaire de travailler la reconnaissance vocale (google home, bixby (samsung), siri (apple)). Le NLP est considérée comme une technologie prometteuse dans le domaine du IHM pour la gestion d’équipements ou d’applications web. Par exemple, le travail des chatbots ou des assistants linguistiques numériques se base sur ce principe.

**Contexte**

Le développement du NLP débute dans les années 1950 à l’époque où le scientifique Alan Turing publie un article titré *Computing Machinery and Intelligence*. Il y présente une méthode de mesure de l’intelligence artificielle. Le "Turing Test" existe encore de nos jours.

Les chercheurs avaient déjà réussi dès 1954 à traduire à l’aide d’une machine une soixantaine de phrases du russe à l’anglais. Dans l’euphorie des débuts, beaucoup de chercheurs informatiques pensaient que la traduction automatique n’était plus qu’une question de temps. Il a cependant fallu attendre jusque dans les années 1980 pour voir se développer les premiers systèmes de traduction automatique basés sur les statistiques. Entre temps, quelques idées furent trouvées pour traduire les informations du monde "réel" en langue informatique.

Un grand pas fut fait à la fin des années 1980, à l’époque où l’apprentissage automatique est devenu populaire. Les algorithmes, alliés à la puissance informatique de plus en plus importante des ordinateurs, purent alors être utilisés pour le NLP. L’un des pionniers de ce domaine fut et est encore aujourd’hui le linguiste Noam Chomsky. L’entreprise de logiciels IBM a elle aussi joué un rôle dans le développement du Natural Language Processing.

Les programmes informatiques basés sur le NPL ne sont aujourd’hui plus en mesure de profiter des fichiers de données collectés manuellement, mais sont cependant capables d’analyser directement et tout seuls des corpus de textes comme des sites web ou de la langue parlée.

**Conditions**

Le NLP est fondé sur l’idée de base que chaque forme de langue, parlée ou écrite, doit tout d’abord être reconnue. La langue est cependant un système très complexe de signes. Ce qui importe n’est pas seulement le mot en lui-même mais le rapport qui existe avec d’autres mots, des phrases entières ou encore les faits.

Pour comprendre ce que les êtres humains apprennent naturellement dès la naissance, les ordinateurs doivent utiliser des algorithmes. Si l’Homme peut tirer des réponses de son expérience de vie, l’ordinateur doit lui avoir recours à des expériences créées de manière artificielle. Le défi pour le traitement automatisé du langage naturel est finalement de comprendre la langue plus que de la produire.

## Fonctionnalité

Le NLP moderne se base sur des algorithmes fondés eux-mêmes sur des statistiques du machine learning. La particularité ici est que les ordinateurs, de cette manière, sont non seulement capables d’apprendre des problèmes résolus autrefois, mais également de les reconnaître de façon autonome et de résoudre des secteurs problématiques en se basant sur d’importants corpus de documents. Les ordinateurs n’apprennent pas de cette manière à trouver une solution pour chaque problème mais plutôt des tendances générales grâce auxquelles ils pourront traiter des questions individuelles. Cela fait du NLP un précurseur de l'intelligence artificielle.

**Utilisation**

-Récapitulatif automatique

-Analyse du discours

-Traduction automatique

-Segmentation morphologique

-Conversion en langue humaine

-Reconnaissance