

From Bag of Words



Topic	Percentage
Software	30%
Data Science	60%
Business	5%
Self Help	0%
Career	5%
Humor	0%
Travel	0%

OPENCLASSROOMS

Table des Matières

1. Introduction.....3

1. Introduction

Dans cette étude j'ai appliqué le traitement automatique du langage naturel, 'natural language processing' (NLP) en anglais, au texte html de 'Stackoverflow', le plus grand site de questions-réponses sur Internet consacré à la programmation et au développement. Le but du projet était de développer un application pour classer automatiquement chaque question posé sur le site et l'assigner des mot clés (tags). Pour cette tache j'ai traité un jeux de données exporté par le site

<https://data.stackexchange.com/stackoverflow/query/new>

Sur ce site j'ai téléchargé un tableau des questions et respectifs mots clé avec une requête SQL. J'ai sélectionné les 50000 questions plus vues, favorites par les utilisateurs (Favoritecount) et pertinentes (Score), avec plus d'une une réponse. Le code est le suivant :

```
SELECT
    TOP 50000 ViewCount,
    CreationDate,
    Title,
    Body,
    Tags,
    Score,
    CommentCount,
    AnswerCount,
    FavoriteCount

FROM Posts
WHERE
    FavoriteCount > 10
    AND AnswerCount > 1
    AND Score > 100
ORDER BY Score DESC
```

Ce document est structuré comme suit. Dans la section 2 on trouve le nettoyage des documents pour obtenir un corpus prêt au traitement et une liste des mots clé . La section 3 présente une analyse exploratoire des données. La section 4 traite la représentation vectorielle du texte. La modélisation non-supervisée est présentée dans la section 5 et la modélisation supervisée dans la section 6. La section 7 présente la choix du modèle pour le traitement automatique et les conclusions.

```
StopWords = nltk.corpus.stopwords.words('english')  
  
print(len(StopWords)) #nltk  
print(len(STOP_WORDS)) #Spacy
```

179

326