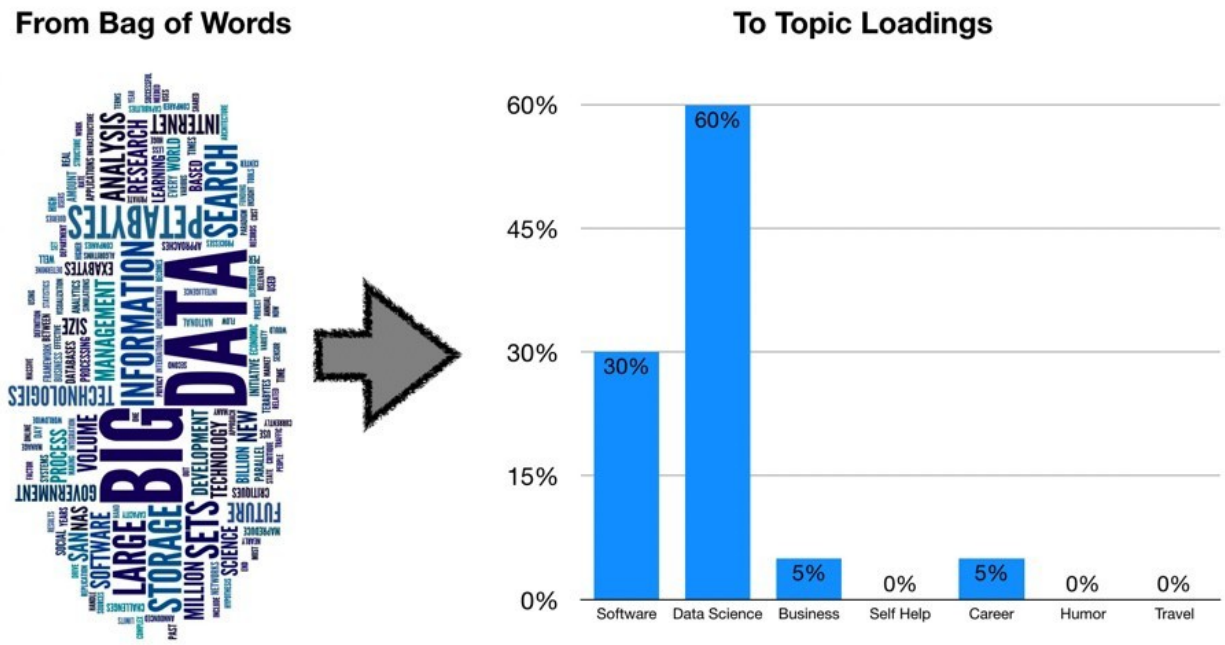
Catégorisation automatique des questions posées sur Stackoverflow



Marco Berta

Table des Matières

1. Introduction

Dans cette étude j'ai appliqué le traitement automatique du langage naturel, 'natural langage 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 télécharge 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
```

J'ai obtenu un ficher csv de 50MB, prêt pour être importé avec Python comme Pandas dataframe (Tableau 1).

```
[4]: print(df_raw.shape)
     df_raw.head()
```

```
(37175, 9)
```

[4]:

	ViewCount	CreationDate	Title	Body	Tags	Score	CommentCount	AnswerCount	FavoriteCount
0	1483128	2012-06-27 13:51:36	Why is processing a sorted array faster than p...	<p>Here is a piece of C++ code that shows some...	<java><c++><performance><optimization><branch-...	24320	22	26	10983
1	8547399	2009-05-29 18:09:14	How do I undo the most recent local commits in...	<p>I accidentally committed the wrong files to...	<git><version-control><git-commit><undo><pre-c...	20895	13	83	6776
2	8115583	2010-01-05 01:12:15	How do I delete a Git branch locally and remot...	<p>I want to delete a branch both locally and ...	<git><version-control><git-branch><git-push><g...	16826	6	40	5357
3	2782271	2008-11-15 09:51:09	What is the difference between 'git pull' and ...	<p>What are the differences between <code>git ...	<git><version-control><git-pull><git-fetch>	11833	9	35	2333
4	2783219	2009-01-25 15:25:19	What is the correct JSON content type?	<p>I've been messing around with <a href="http...	<json><http-headers><content-type>	10204	0	36	1446

Tableau 1. Données importées dans Pandas à partir du fichier obtenu par la requête SQL.

Ce document est structuré comme suit. Dans la section 2 on trouve le nettoyage et le pré-traitement des documents pour obtenir un corpus prêt à l'analyse 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.

2. Pré-traitement des données

L'objectif de cette étape est de standardiser le texte pour rendre son traitement facile par les algorithmes d'apprentissage automatique. Dans la colonne des mots clé, Tags, (Tableau 2) ça suffit d'enlever les parenthésés.

Title	Body	Tags
Why is processing a sorted array faster than p...	Here is a piece of C++ code that shows some...	<java><c++><performance><optimization><branch-...
How do I undo the most recent local commits in...	I accidentally committed the wrong files to...	<git><version-control><git-commit><undo><pre-c...
How do I delete a Git branch locally and remot...	I want to delete a branch both locally and ...	<git><version-control><git-branch><git-push><g...
What is the difference between 'git pull' and ...	What are the differences between <code>git ...	<git><version-control><git-pull><git-fetch>
What is the correct JSON content type?	I've been messing around with <a href="http...	<json><http-headers><content-type>

Tableau 2. Les colonnes 'Body' et 'Tags' dans la forme originelle des fichiers html.

Par contre, le pré-traitement du corpus du texte (colonne 'Body') est plus complexe. Le texte de chaque question est dans la forme du langage html, par exemple dans la première ligne on lit :

'' Here is a piece of C++ code that shows some very peculiar behavior. For some strange reason, sorting the data miraculously makes the code almost six times faster: \n\n<pre class="lang-cpp prettyprint-override"><code>#include <algorithm>\n#include <ctime>\n#include <iostream>\n\nint main()\n{\n   // Generate data\n   const unsigned arraySize = 32768;\n ''

Enlever les caractères html à la main serait très long, et pour çà j'ai utilisé le paquet ''Beautiful soup'' . Après l'application de la commande :

```
df_raw['Body_text'] = [BeautifulSoup(text).get_text() for text in df_raw['Body'] ] # get text from html body
```

Le même texte est devenu

‘‘ Here is a piece of C++ code that shows some very peculiar behavior. For some strange reason, sorting the data miraculously makes the code almost six times faster:\n#include <algorithm>\ n#include <ctime>\n#include <iostream>\n\nint main()\n{\n   // Generate data\n   const unsigned arraySize = 32768;\n ’’

Mieux, mais pas encore prêt au traitement automatique. Après la concaténation avec les mots du titre, plusieurs étapes de nettoyage étaient nécessaires.

- traitement des contractions plus couramment utilisées en anglais, ex. changer “it's“ en “it is“.
- enlever les chiffres
- enlever les caractères spéciaux comme @, #,%, <
- enlever la ponctuation

```
StopWords = nltk.corpus.stopwords.words('english')

print(len(StopWords)) #nltk
print(len(STOP_WORDS)) #Spacy
```

```
179
326
```