

Système de suggestion de tag pour le site Stackoverflow



Outline

- **Introduction**
 - Objectif de l'étude
 - Jeux de données
- **Nettoyage**
 - Extraction du texte à partir des données html
 - Numéros, ponctuation, caractères spéciaux
 - Tokenisation et lemmatisation
- **Analyse exploratoire**
 - Les tags plus fréquentes
 - Les mots du corpus plus fréquentes
 - Les mots du corpus associés aux 3 tags principaux
- **Modélisation du corpus**
 - Matrice des termes (TF, TF-IDF)
 - Modèles non-supervisés – optimisation et performance
 - Modèles supervisés
 - Réduction dimensionnelle par ACP
- **Conclusions**
 - Choix des meilleurs algorithmes de prédiction pour une API

Modalités de la soutenance

5 min - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.

5 min - Présentation du cleaning effectué, du feature engineering et de l'exploration.

10 min - Présentation des différentes pistes de modélisation effectuées.

5 min - Présentation du modèle final sélectionné et résultats.

5 à 10 minutes de questions-réponses.

Objectif du modèle

Dans cette étude avec un apprentissage supervisé ou non, on essaye de trouver le meilleur modèle pour

- Prédire les mots-clés d'une question posée sur le site Stackoverflow
- Appliquer ce modèle à une API

Données du départ

outil d'export de données avec une requête SQL:

"stackexchange explorer"

```
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
```

- Le 50000 questions les plus consultées
- Bien notées (note >100)
- Avec au moins une réponse



Title	Body	Tags
Why is processing a sorted array faster than p...	<p>Here is a piece of C++ code that shows some...	<java><c++><performance><optimization><bran...
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...
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...
What is the difference between 'git pull' and ...	<p>What are the differences between <code>git ...	<git><version-control><git-pull><git-fetch>
What is the correct JSON content type?	<p>I've been messing around with <a href="http...	<json><http-headers><content-type>

Corpus

Mots-clés

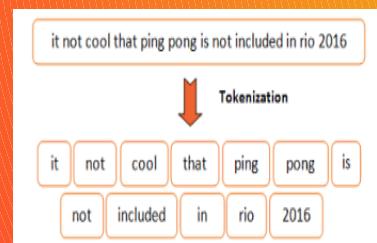
Traitement du texte non-structuré



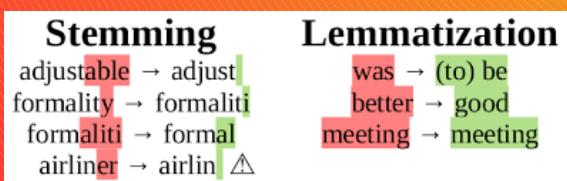
- **BeautifulSoup :**
extraction du texte de chaque document/question dans le « Body » à partir du contenu en html (« soup »)

- Jointure des colonnes « **Body** » et « **Title** »

- **Pre-traitement :**
filtrage des chiffres, ponctuation, caractères spéciaux, mots non-pertinents (stopwords)



- **Tokenisation :** décomposition du texte en mots
- **Lemmatisation :** remplacement de chaque mot par sa forme canonique



Exploration des données

Conclusions

-