Projet 5

**CATEGORISEZ AUTOMATIQUEMENT DES QUESTIONS**

****

Contents

[Présentation de la problématique 3](#_Toc25159480)

[Cleaning 3](#_Toc25159481)

[Feature engineering 3](#_Toc25159482)

[Analyse explaratoire 3](#_Toc25159483)

[Modélisation non-supervisée 3](#_Toc25159484)

[Modélisation supervisee 3](#_Toc25159485)

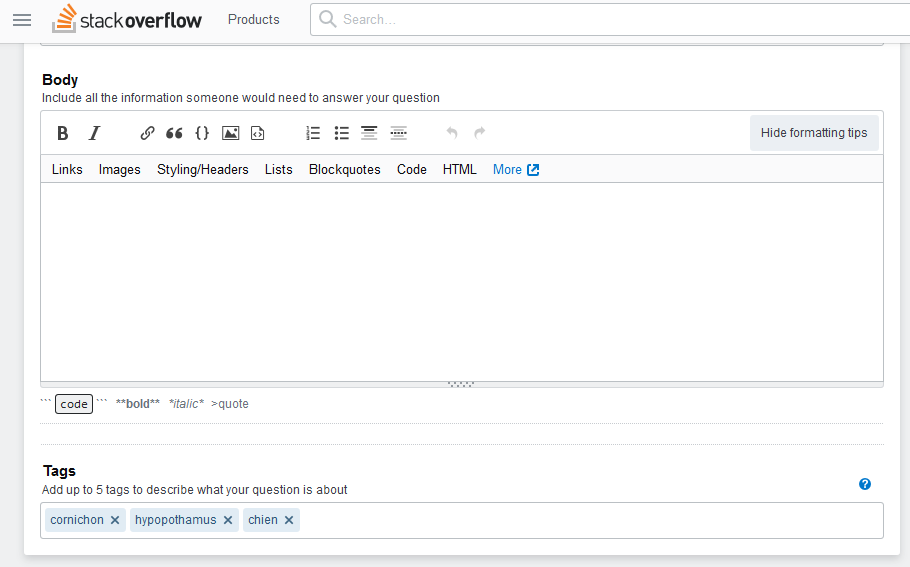
[Conclusion 3](#_Toc25159486)

# Présentation de la problématique

Stack overflow est une des plus grands forums de questions – réponses pour les développeurs. Si nous avons besoin d’aide avec un code qui ne marche pas, il suffit souvent d’aller chercher une question qui existe déjà et qui était déjà résolu par les autres utilisateurs de la plateforme.

Aujourd’hui, la base de données de question – réponses devient immense et il est de plus en plus difficile de trouver des questions pertinentes. Afin de simplifier la recherche, les utilisateurs sont censés de rentrer jusqu’à 5 tags qui correspondent à la question. Les tags sont un champ de texte libre, même s’il existe un système d’aide qui nous propose des tags les plus fréquents quand nous commençons à éditer les tags, nous pouvons quand même renseigner tout et n’importe quoi.

Figure - Edition d'une question Stack Overflow



Développer un système qui peut prédire un ou des tags à partir de notre question peut être utile à :

* Mieux classifier des questions
* Proposer des questions déjà existantes plus pertinentes
* Eviter des questions en doublons
* Fidéliser les utilisateurs en proposant un fonctionnement simple, rapide et efficace

Aujourd’hui, la base de données et questions contiennent presque 47 millions de questions, dont 18,5 millions de questions complètes, c’est à dire qui contient et le titre, le corps et les tags. Le monde informatique est constamment en évolution, donc ce qui nous intéresse ce sont des questions les plus récentes. Seulement en 2019, les utilisateurs ont posté presque 1,9 millions de questions complètes.

Notre but dans ce projet est trouver un algorithme de machine learning approprié qui nous servira à proposer les tags automatiquement aux utilisateurs. Les prédictions de tags se feront à partir de titre et de corps de question saisi par l’utilisateur. L’algorithme est censé de proposer un ou plusieurs tags en temps réel, donc le calcul de la prédiction devrait être fait très rapidement.

# Description de données

Notre projet sera basé sur un échantillon de questions complètes qui contient et les tags, le titre et le corps de question, posté en 2019. Nous avons téléchargé un jeu de données de 50 000 questions pour les analyses et la modélisation et un autre échantillon de 500 000 questions pour création d’un modèle word2vect.

# Cleaning

## Features

Le feature est composé de deux variables : le titre et le corps de question. La particularité de corps est qu’il s’agit d’un texte qui contient des balises HTML et parfois aussi le code entouré de balises <code> et </code>. Hormis le traitement de texte standard, nous allons également enlever le code, ainsi que les balises HTML. Le schéma suivant représente le workflow de cleaning :

## Cible

# Feature engineering

# Analyse exploratoire

# Modélisation non-supervisée

# Modélisation supervisee

# Conclusion