


Catégoriser automatiquement des questions

...

9 juin 2023
Yoann Poupart

Problématique: développer un système de suggestion pour le site Stack Overflow, capable de proposer plusieurs tags pertinents pour une question.

- Feature engineering
- Exploration des données
- Modélisation non-supervisée
- Modélisation supervisée
- Démonstration
- Conclusion



Feature engineering

Récolte des données

Entraînement d'un modèle

- Réponse
 - ➔ Éviter les outliers
- Score et vues
 - ➔ Pertinence de la question
- Classement
 - ➔ Meilleures questions

```
SELECT TOP 50000 Title, Body, Tags, Id, Score, ViewCount, AnswerCount
FROM Posts
WHERE (
    PostTypeId = 1 AND AcceptedAnswerId IS NOT NULL
    AND (LEN(Tags) - LEN(REPLACE(Tags, '<', '')) >= 4)
) AND (
    Score > 20 AND ViewCount > 1000
)
ORDER BY Score DESC
```

Extraction et traitement

Features

- Title
 - ⇒ Non filtré
- Texte
 - ⇒ Balises titres et paragraphes
- Code
 - ⇒ Séparé du texte

Traitement

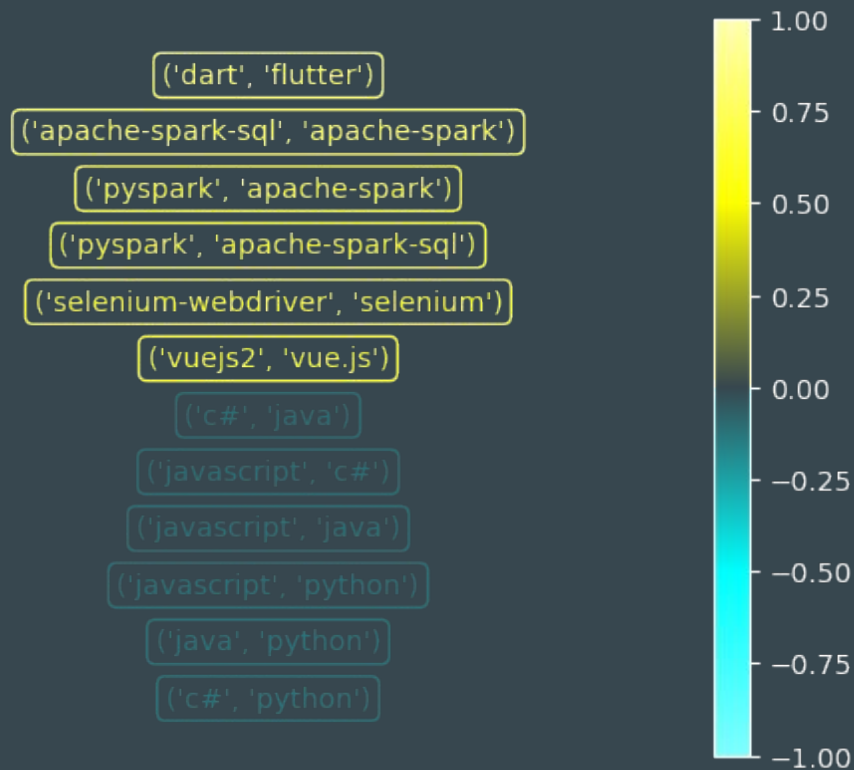
- Normalisation
 - ⇒ Minuscule
- Nettoyage du texte
 - ⇒ Retrait de la ponctuation
 - ⇒ Retrait des mots “inutiles”
- Lemmatisation
 - ⇒ Regroupement sémantique



Exploration

Vocabulaire

Corrélations des tags



Réduction de dimensions



Modélisation non-supervisée

Mots-clés

Modélisation de sujets



Modélisation supervisée

Classification multi-output

Embeddings

Résultats



Démonstration

API

URL: <https://oc-nlp.azurewebsites.net>

FastAPI

- Rapidité
 - ➡ Développement agile
- Modularité
 - ➡ Multi modèles
 - ➡ Versionning basique

Déploiement sur Azure

- App service
 - ➡ Facilité / scalabilité / maintenabilité
 - ➡ Versionnage automatique
 - ➡ Déploiement automatique
- Activation
 - ➡ Recréé l'image à chaque activation (après inactivité)



Conclusion

Conclusion générale

Modélisation

- Non-supervisée
➡
- Supervisée
➡ Agrégation de tags
- Récolte des données
➡ Moins bonnes questions

Axes d'amélioration

- Feature engineering
➡ Méthodes avancées d'Embedding
- Modélisation non-supervisée
➡
- Versionnage
➡ DVC (plus complet)

Merci de votre attention.

...

Des questions ?