


# Catégoriser automatiquement des questions

...

9 juin 2023  
Yoann Poupart

*Problématique: développer un système de suggestion pour Stack Overflow, qui propose 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
  - ➔ ASCII
- Nettoyage du texte
  - ➔ Retrait de la ponctuation
  - ➔ Retrait des mots “inutiles”
- Lemmatisation
  - ➔ Regroupement sémantique



# Exploration



# Vocabulaire

Text (body)

## Code (body)

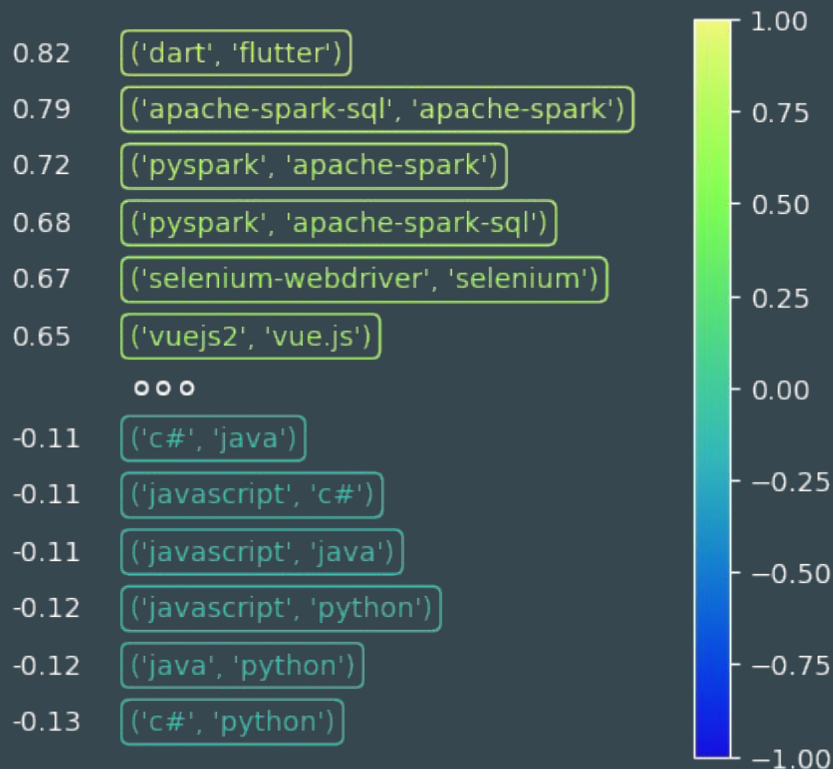




# Corrélations des tags

## Analyse

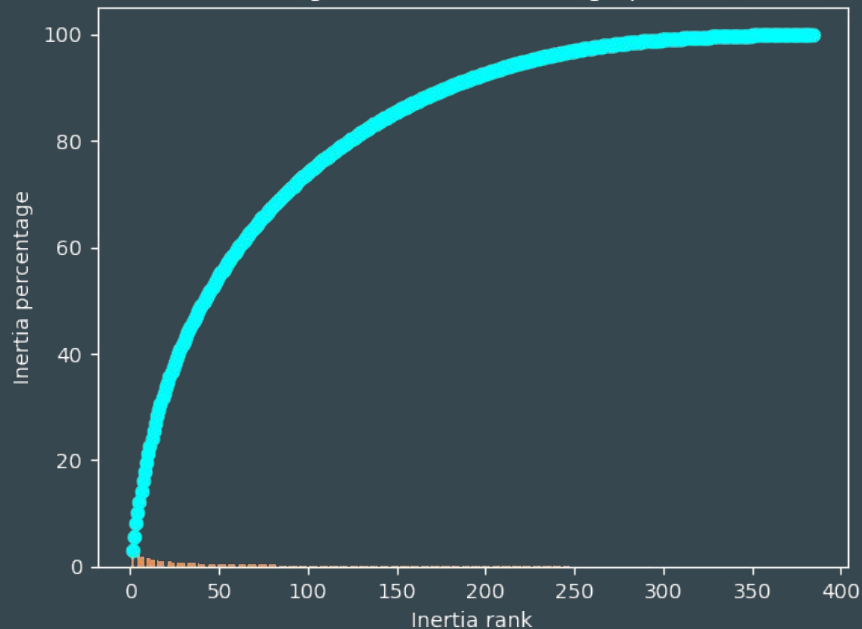
- Corrélations
  - ➡ Similarité/connexion entre tags
  - ➡ Regroupement de tags
- Anti-corrélations
  - ➡ Langages de programmation distinct
  - ➡ Tags potentiellement polarisants



# Réduction de dimensions

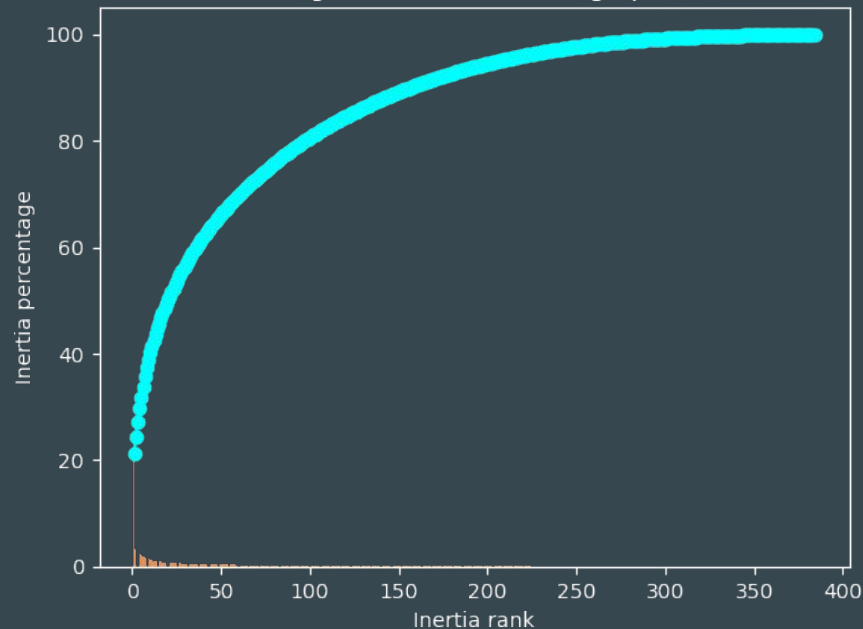
Titre

Eigen value cumulative graph



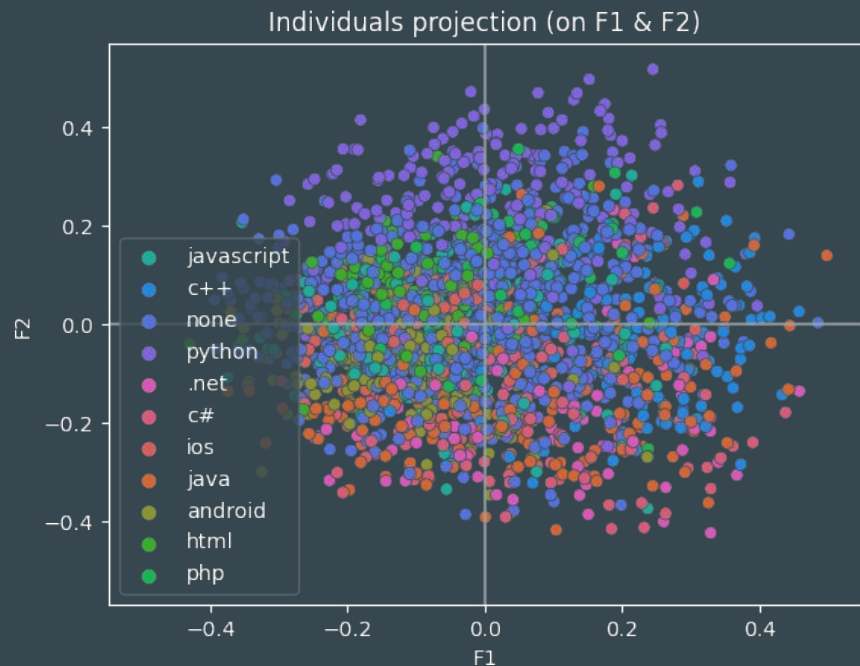
Code

Eigen value cumulative graph

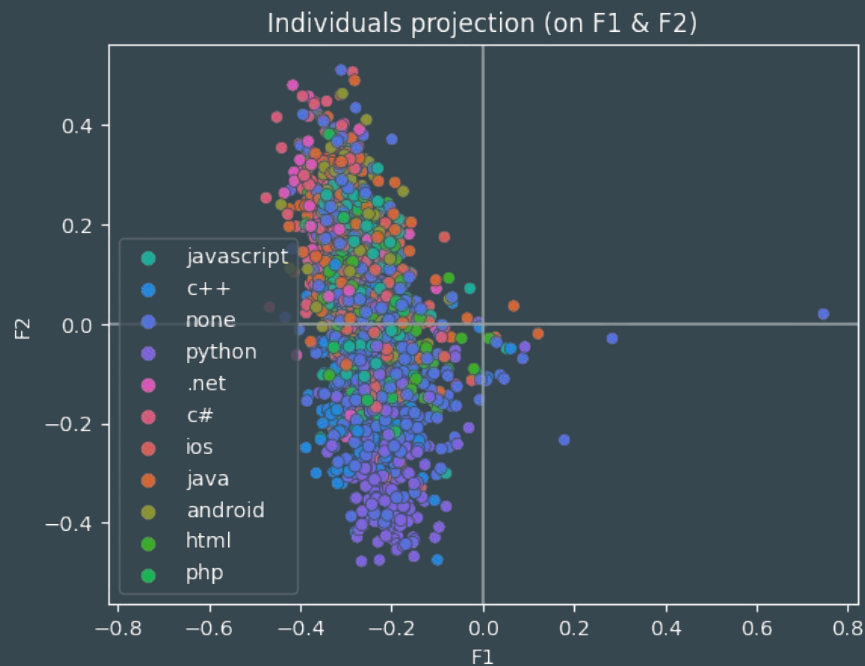


# Réduction de dimensions

Titre

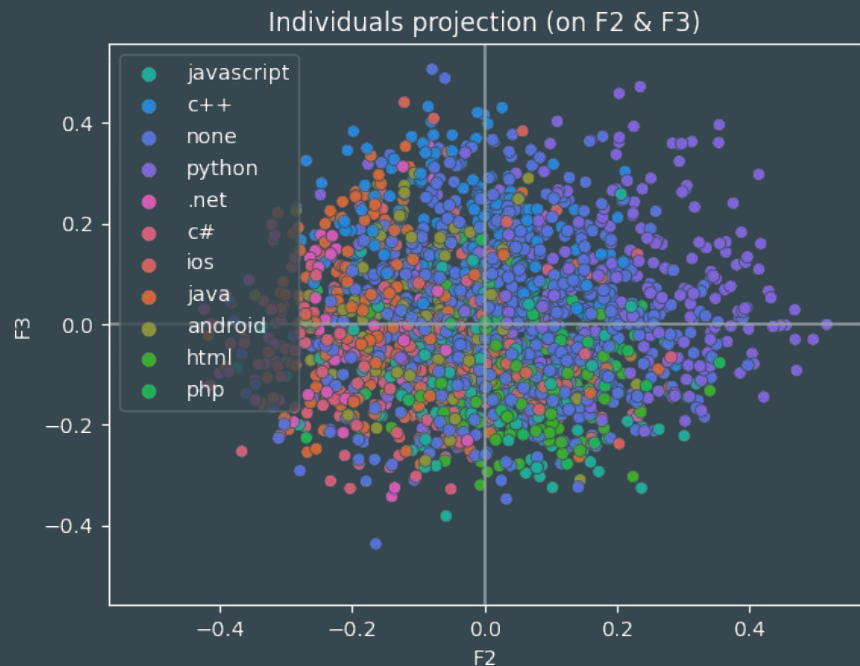


Code

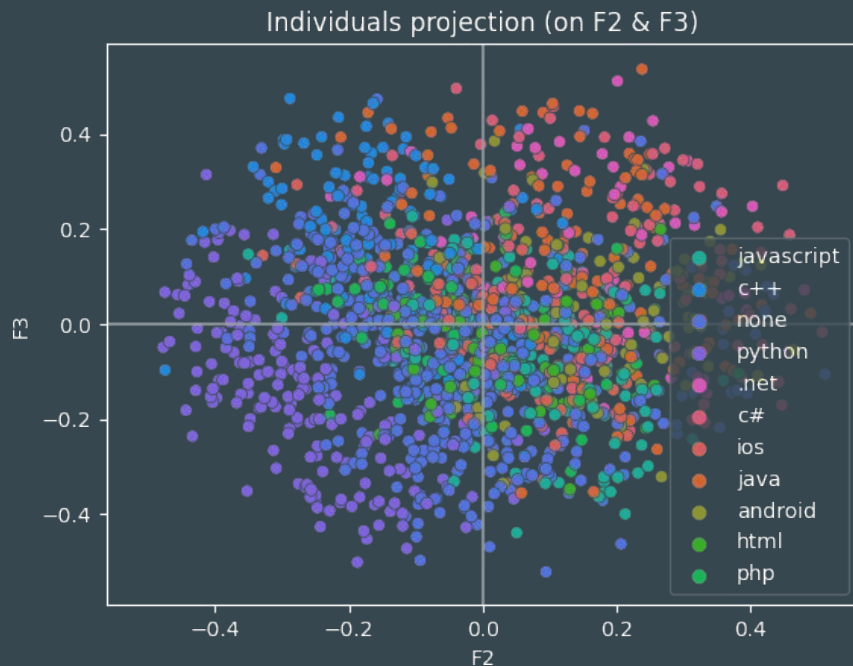


# Réduction de dimensions

Titre

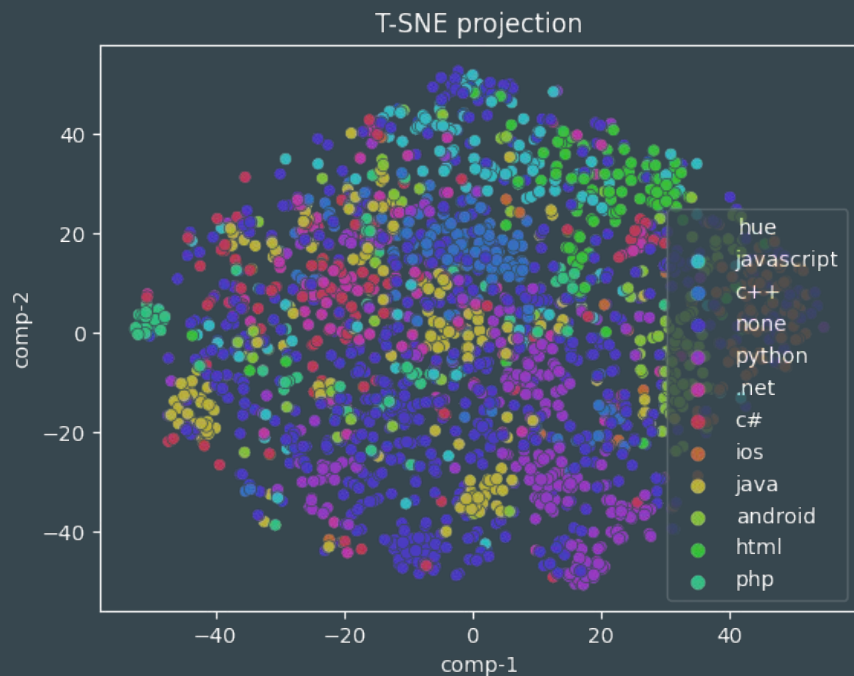


Code

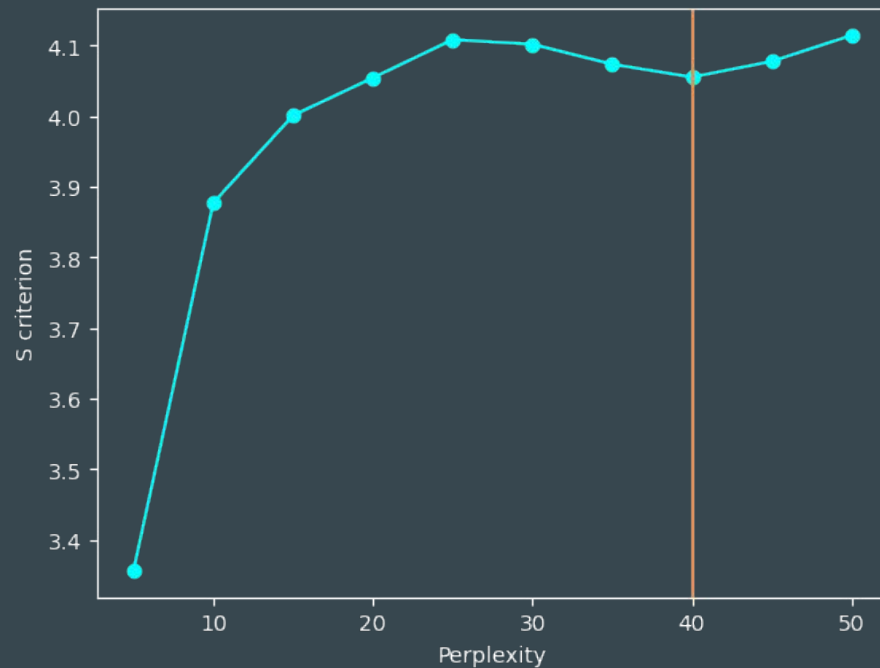


# Réduction de dimensions

Titre



Perplexité





# **Modélisation non-supervisée**

# Clustering

- Développement d'applications web
- Calculs et graphiques



# Modélisation de sujets

- Regroupement par thème
- Regroupement par langage

```
Topic 0:
157 access entity method 13 change end head max Bar
Topic 1:
value dispatch span ForeignKey tab ini const module os line
Topic 2:
SQL log init migration title password redirect reference exist extend
Topic 3:
use open command extern redirect client platform int64 compile table
Topic 4:
go conf project IList language request system reference plot expect
Topic 5:
float window timeout recipe header plt datum npm state button
Topic 6:
dir vector NotePage export constraint alert android retrieve 50 Red
Topic 7:
test work function login bucket admin paramiko sum parse IList
Topic 8:
component auto df insert project browser output apply header Add
```

```
Topic 0:
public class foo int String void return Foo bar extend
Topic 1:
Android Studio float load Java head format location override image
Topic 2:
use directory interface Microsoft Studio notification status css display good
Topic 3:
difference select const std key insert NET argument vector range
Topic 4:
function var std return this console code JavaScript remove error
Topic 5:
ios run app error key device build miss unable io
Topic 6:
Windows cursor window application convert set language log console format
Topic 7:
android image id room extend 157 Java shape title center
Topic 8:
Python find version python import pd list print install 10
```





# **Modélisation semi-supervisée**

# ZS prompt

- Description du cadre
- Description de la tâche
- Formatage des données

## Prompt

```
PROMPT_TEMPLATE = """
```

```
You will be provided with the following information:
```

1. A Stack Overflow question. The question is delimited with triple backticks
2. List of tags the question can be assigned to. The tags in the list are

```
Perform the following tasks:
```

1. Identify to which tags the provided question belongs to with the highest probability
2. Assign the question to any tags based on the probabilities. If no tag is assigned, return an empty list
3. Provide your response in a JSON format containing a single key 'label'

```
List of tags: {labels}
```

```
Stack Overflow question:
```

```
"""
```

```
[title]
```

```
{title}
```

```
[body_text]
```

```
{body_text}
```

```
[body_code]
```

```
{body_code}
```

```
"""
```

```
Your JSON response:
```

```
"""
```

# ZS résultats

Évaluation:  
20 exemples aléatoire

	Scores (20)	Scores (500)
gpt-3.5-turbo (20)	A: 0.60 - R: 0.76	-
gpt-4 (20)	A: 0.50 - R: 0.63	-
gpt-3.5-turbo (500)	A: 0.40 - R: 0.75	A: 0.05 - R: 0.73
gpt-4 (500)	A: 0.55 - R: 0.73	A: 0.10 - R: 0.62



# **Modélisation supervisée**

# Vectorisation

Données:  
Titre (1000)

Sans réduction  
de dimension

Classifieur:  
Régression Logistique

	Train	Test
BOW	A: 0.51 - R: 0.38	A: 0.28±0.05 - R: 0.03±0.01
TF-IDF	A: 0.27 - R: 0.03	A: 0.27±0.05 - R: 0.002±0.003
BERT	A: 1.00 - R: 1.00	A: 0.36±0.04 - R: 0.27±0.03
USE	A: 0.29 - R: 0.07	A: 0.29±0.06 - R: 0.03±0.02
all-MiniLM	A: 0.39 - R: 0.20	A: 0.36±0.05 - R: 0.13±0.02
all-MPNet	A: 0.40 - R: 0.21	A: 0.37±0.05 - R: 0.15±0.02

# Données pertinentes

Vectorisation:  
all-MPNet

Classifieur:  
Régression Logistique

	Train	Test
Titre	A: 0.40 - R: 0.21	A: 0.37±0.05 - R: 0.15±0.02
Code	A: 0.37 - R: 0.18	A: 0.34±0.05 - R: 0.12±0.02
Titre + Code	A: 0.42 - R: 0.25	A: 0.40±0.04 - R: 0.18±0.01
Titre ^ Code	A: 0.50 - R: 0.37	A: 0.40±0.05 - R: 0.20±0.04

# Analyse des résultats

- Rapide et pas forcément besoin de réduction de dimension
- Embedding à base de Transformer plus performants
- Combinaison titre et code pour de meilleures performances



# Démonstration



# API

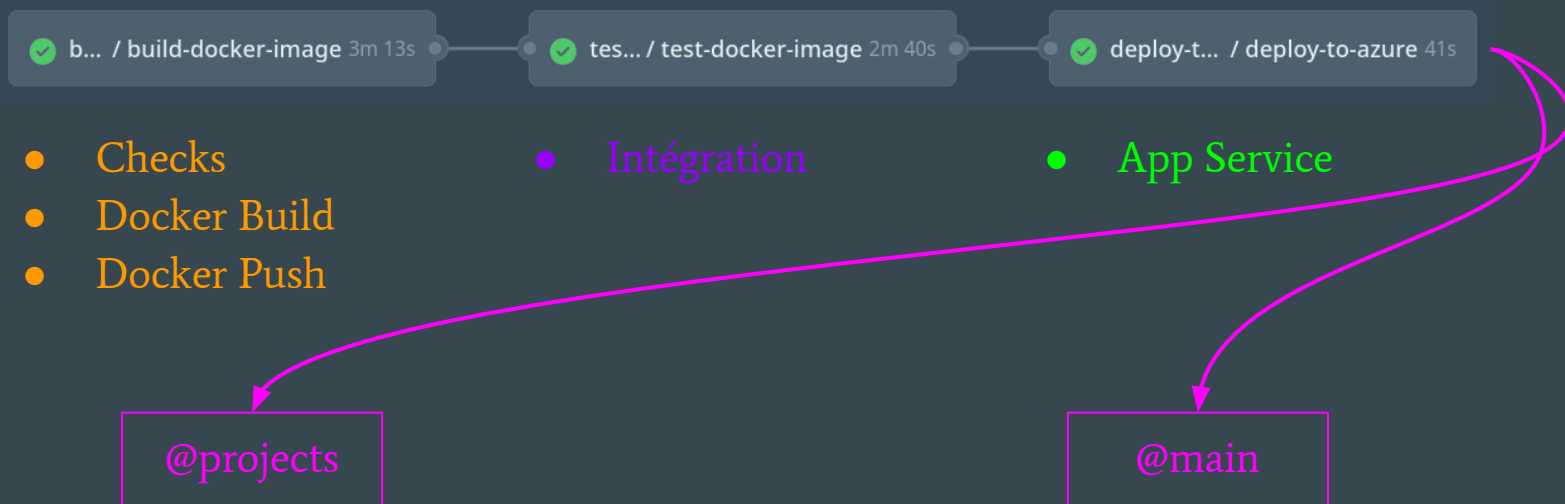
## FastAPI

- Rapidité
  - ⇒ Développement agile
- User-friendly
  - ⇒ Swagger UI
- Modularité
  - ⇒ Multi modèles
  - ⇒ Multi modules

## Déploiement sur Azure

- App service
  - ⇒ Facilité / scalabilité / maintenabilité
  - ⇒ Déploiement automatique
- Activation
  - ⇒ Recrée l'image à chaque activation
- Mémoire
  - ⇒ Image Docker limitée

# CI-CD



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

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



# Conclusion

# Conclusion générale

## Traitement des données

- Feature engineering
  - ➔ Agrégation de tags
  - ➔ Moins bonnes questions
- Exploration non-supervisée
  - ➔ Pour différents pré-traitements
- API
  - ➔ Pré-traitement

## Modélisation

- Supervisée
  - ➔ Explorer d'autres modèles
  - ➔ Analyse par tag
- Semi-supervisée
  - ➔ Coûts et temps élevés
  - ➔ Scalable
- Versionnage des modèles
  - ➔ DVC / MLFlow

# Merci de votre attention.

...

Des questions ?