CentraleSupélec

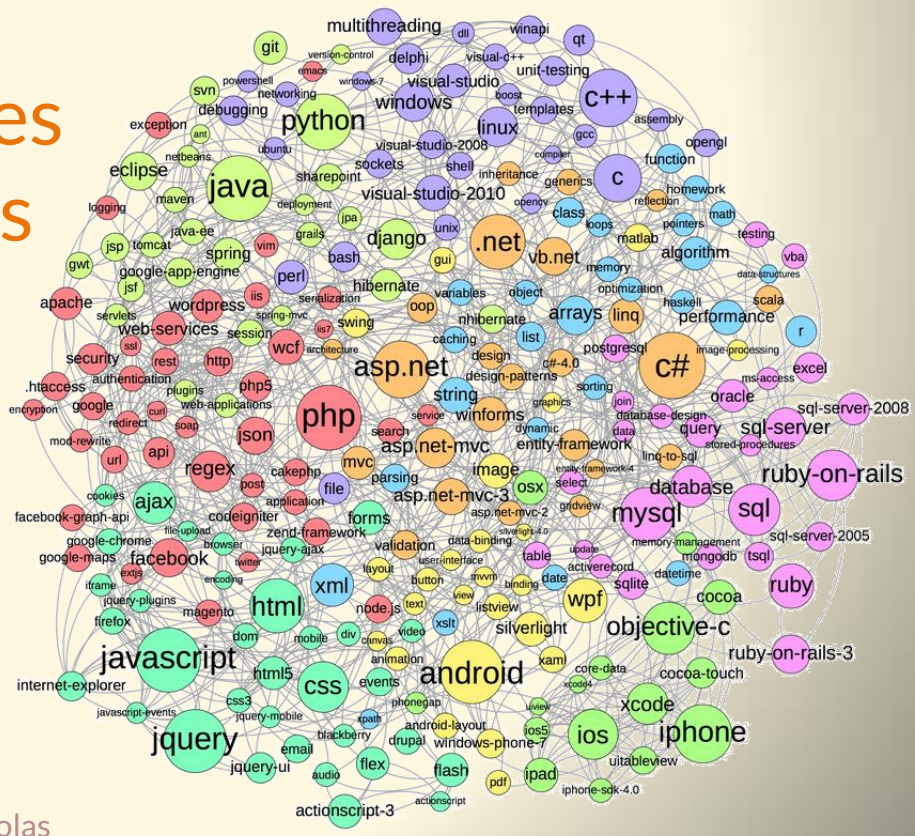
OPENCLASSROOMS

# Parcours Data Scientist

## Projet 6 :
## Proposition de Tags

# Sommaire

- Présentation et Objectifs
- Préparation des données
- Exploration
- Préparation des Matrices
- Modèles Non supervisés
- Modèles Supervisés
- API
- Pistes d'évolutions
- Conclusion



# Présentation et Objectifs

- Présentation
  - Basé sur divers questions existantes
  - Traitement de données textuelles
- Objectifs
  - Prédiction de Tags
    - Méthode non supervisée
    - Méthode supervisée
  - Mise en place d'une API

# Préparation des données

- Récupération du dataset
  - 2 x 50k Posts
    - Score > 3 (Pertinence des questions)
    - Type = 1 (Question uniquement)
    - Order By = Random (éviter les tendances)

```
SELECT Id, Title, Tags, Body
FROM Posts
WHERE PostTypeId = 1
AND Score > 3
ORDER BY RAND()
OFFSET 0 ROWS FETCH NEXT 50000 ROWS ONLY
```

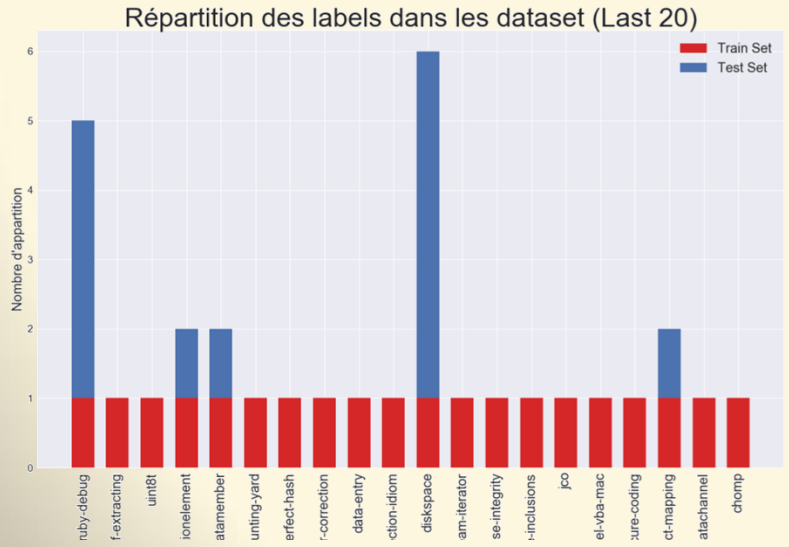
```
SELECT Id, Title, Tags, Body
FROM Posts
WHERE PostTypeId = 1
AND Score > 3
ORDER BY RAND()
OFFSET 0 ROWS FETCH NEXT 50000 ROWS ONLY
```

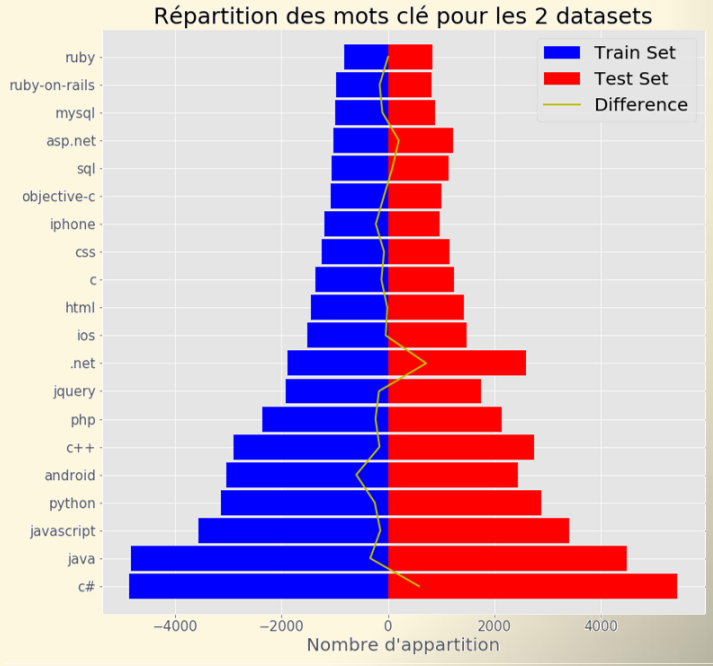
# Préparation des données

- Pré-traitement
  - Vérification Id de chaque dataset
  - Fusion Titre + Body
  - Suppression balises
  - Suppression <code>
  - Top 100 mots + StopWords (English)
  - Parsing tags
    - Liste de listes

# Exploration

- Tags
  - Réduction des tags (5000+ => 773)
    - Minimum 25 apparitions train Set
    - Réduction de 1k post (train) et 3k posts (test)
  - Vérification balance dataset



# Préparation des Matrices

- Matrice TF
  - Sans Lemmatisation (peu de dimensions en moins)
  - Sparse 48357 x 91349 avec 1.72 millions d'entiers
  - 1 elem sur 4000
- Matrice TF-IDF
  - Avec Lemmatisation
  - Sparse 48357 x 2764 avec 1,55 millions de float
  - 1 elem sur 86
- Réduction TF via LSA
  - 91349 => 3000 dimensions (88% de la variance)
  - Modèle très lourd (2go) + lent

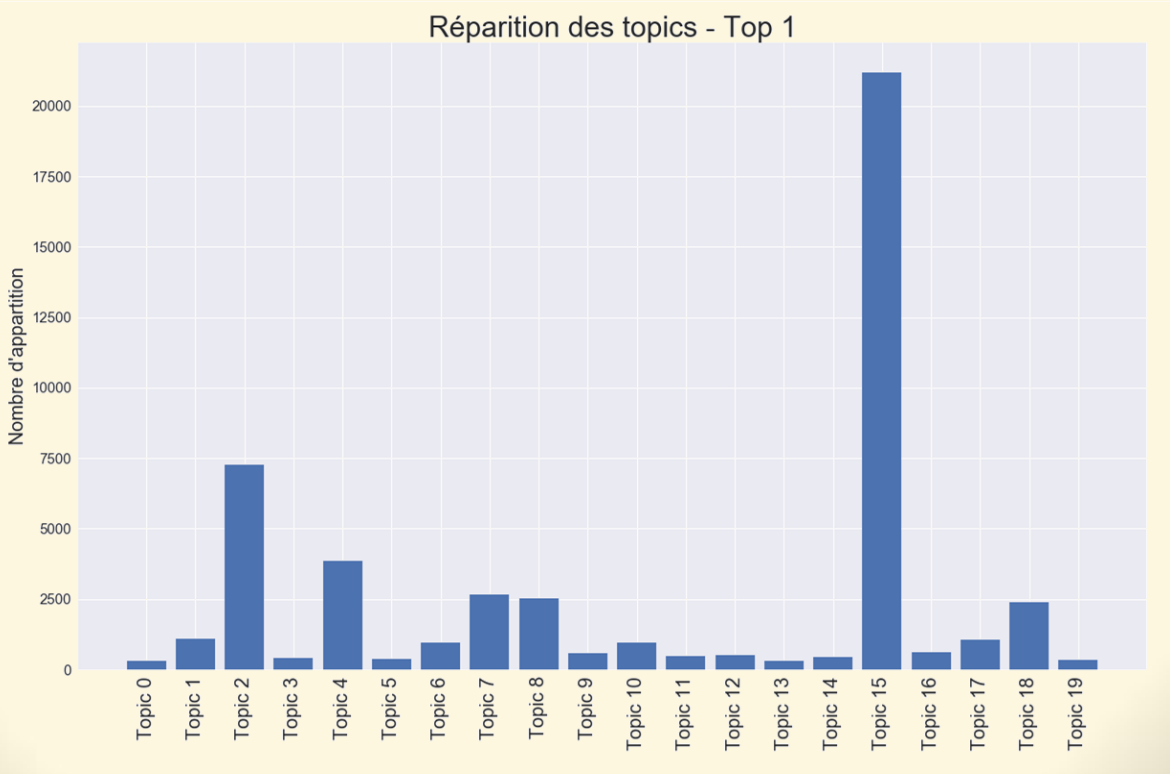
# Modèle Non supervisé

- Latent Dirichlet Allocation
  - Fait sur Matrice TF
  - 20 topics

| Topics | Mots clés | Analyse |
|---|---|---|
| 1 | javascript event js events node tag component tags form control | Gestion de formulaires HTML |
| 2 | string value number java memory values list variable two array | Types de données |
| 4 | text jquery element css html json button change set click | Mise en page site (CSS, js, forms) |
| 7 | project files android build version directory folder git studio eclipse | Gestion de projets/applications |
| 8 | table database sql query key mysql field column array id | Base de données |
| 10 | image images android size map points video draw plot matlab | Graphiques/images |
| 18 | page view net web http asp url controller request mvc | Fonctionnement site web |

# Modèle Non supervisé

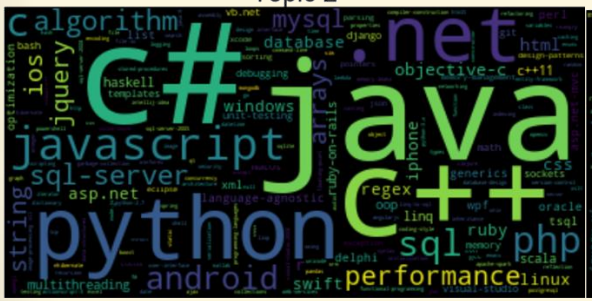
- Latent Dirichlet Allocation
  - Nombre de post par topics



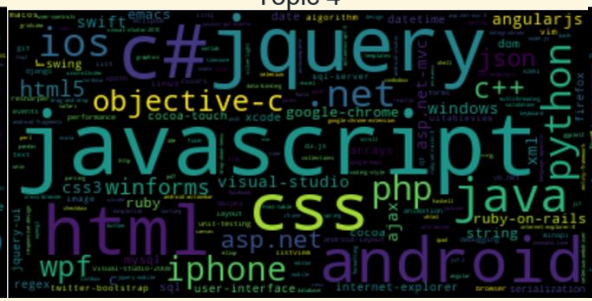
# Modèle Non supervisé

- Latent Dirichlet Allocation
  - Tags par Topics
  - Non Normalisé

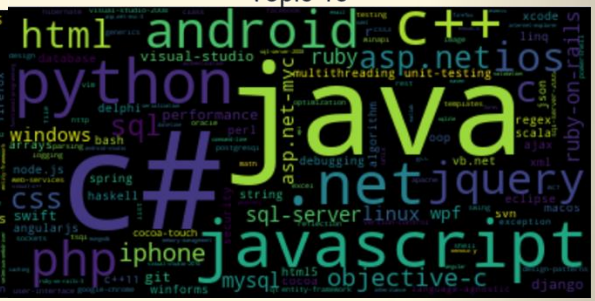
Topic 2




Topic 4

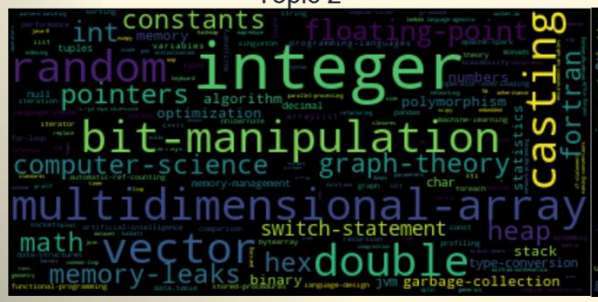



Topic 15

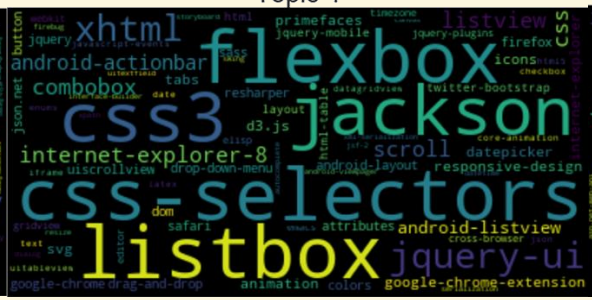



- Normalisé

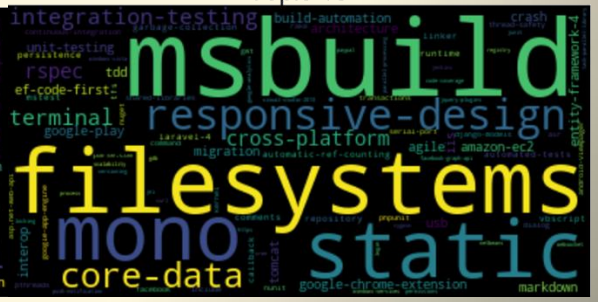
Topic 2




Topic 4




Topic 15



# Modèle Non supervisé

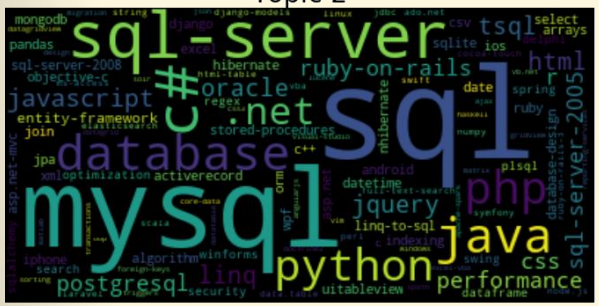
- Non-Negative Matrix Factorization
  - Fait sur Matrice TF-IDF
  - 20 topics

| Topics | Mots clés (lemmatisés) | Analyse |
|---|---|---|
| 1 | c compil program librari languag pointer b gcc declar dll | Compilation, librairie et C |
| 2 | tabl column queri row sql databas mysql select index field | Base de Données |
| 3 | server sql connect client databas servic web request send http | Requete Server |
| 10 | array element byte loop index numpi sort pointer size number | Type de données et Structures |
| 13 | php script mysql variabl 5 upload session page ini email | PHP |
| 15 | page jqueri html element javascript button click text event div | Mise en page web |
| 18 | valu return variabl key set null properti default field type | Type de données |

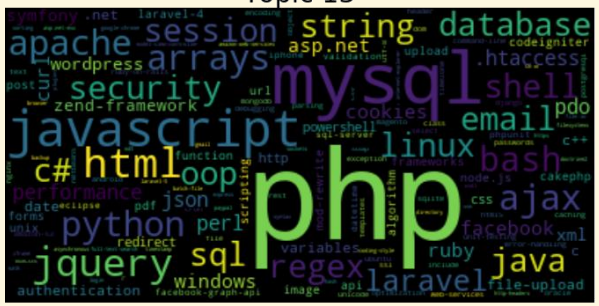
# Modèle Non supervisé

- Non-Negative Matrix Factorization
  - Tags par Topics
  - Non Normalisé

Topic 2




Topic 13

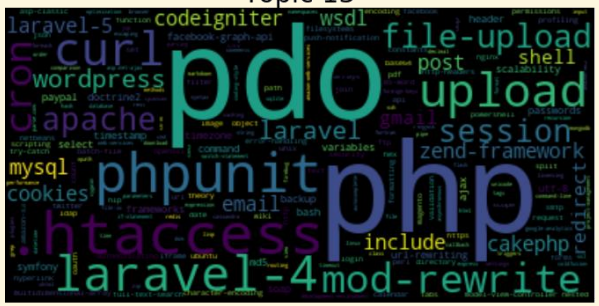



Topic 18

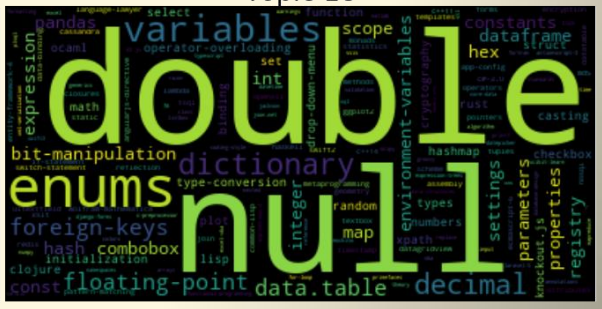

  - Normalisé

Topic 2


Topic 13

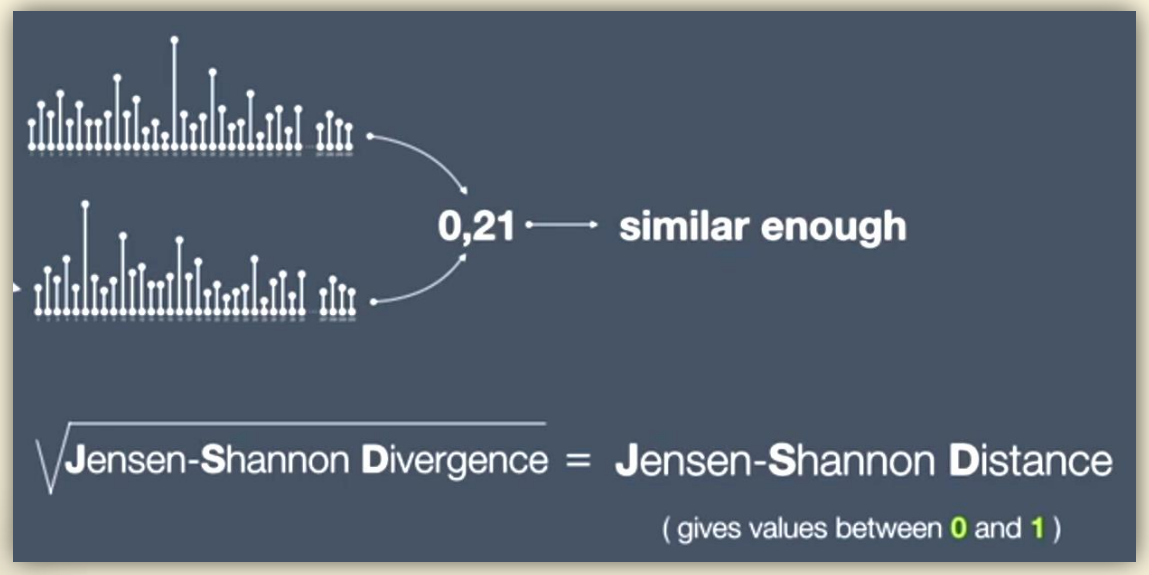



Topic 18



# Modèle Non supervisé

- Proposition de tags
  - Utilisation du LDA
  - Répartition des topics en %
  - KNN basé sur le Jensen Shannon Divergence



# Modèle Non supervisé

- Proposition de tags
  - Test sur un post au hasard
    - Tags : Php, html, apache, compression
    - Tags non normalisé : C#, javascript, .net, winforms, angular
    - Tags normalisé : Event-handling, include, android-intent, javascript-events, iframe
  - Peu de pertinence
  - Conservation uniquement du topic (HTML, PHP)
- Evaluation
  - 20 posts – 5 tags
  - Modèle non-normalisé : 37%
  - Modèle normalisé : 42%

# Modèle Supervisé

- Entrainement sur Matrice TF-IDF
  - Custom Score
    - Prédiction de probabilités par classe
    - Récupération du Top 5
    - % de Tags en commun avec le post
- Exemple:
  - Prédiction : Python, Algorithm, C++, Integer et Array
  - Post 1 : Python
    - 100 %
  - Post 2 : C++, Pointers, Compiler
    - 33 %

# Modèle Supervisé

- Entrainement sur Matrice TF-IDF
  - Fléau de la dimensions (Matrice TF)

| Type | Modèles | Résultats |
|---|---|---|
| MOC | SGDClassifier | Train : 79,3%<br>Test : 71,4% |
| OVR + Ensemble | AdaBoostClassifier<br>GradientBoostingClassifier | Out of time |
| Multi-label + Ensemble | ExtraTreesClassifier<br>RandomForestClassifier | Train : 37.5% - Test : 36.1%<br>Train : 47.1% - Test : 45.1% |
| Multilabel | KNeighborsClassifier | Memory Error (matrice non Creuse) |
| Multilabel | RidgeClassifierCV | Memory Error (inversion de trop grosses matrices) |
| Multilabel | MLPClassifier | Train : 82.0%<br>Test : 68.9% (overfitting malgré Early Stop) |

# Modèle Supervisé

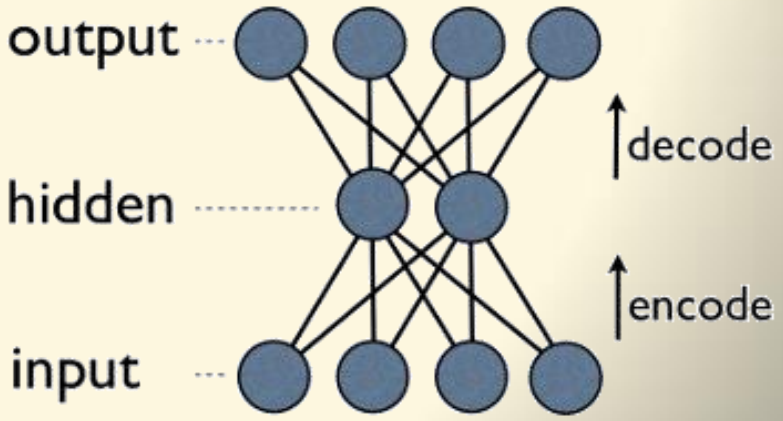
- Fine Tuning
  - SGDClassifier
  - Grid Search manuel (predict_proba + custom score)
  - Test de régularisation (trop d'overfitting)
  - Cross Validation sur train set

| Paramètres | Résultat Train Set | Résultat Test Set |
|---|---|---|
| alpha = 1e-6 – penalty = L1 | 98.8% | 64.9% |
| alpha = 1e-6 – penalty = L2 | 98.1% | 68.0% |
| alpha = 1e-5 – penalty = L1 | 82.9% | 73.1% |
| alpha = 1e-5 – penalty = L2 | 87.7% | 72.6% |
| alpha = 1e-4 – penalty = L1 | 66.5% | 66.0% |
| alpha = 1e-4 – penalty = L2 | 63.7% | 60.7% |

  - Résultat sur Test Set : 66,18%

# Modèle Supervisé

- Fine Tuning
  - Utilisation de Keras
    - MLPC => CPU
    - Keras => GPU
  - Comparaison MLPClassifier
    - 3000 inputs -> 1500 Dense (+Reg) -> 773 outputs (Sigmoïde)
    - BinaryCrossEntropy
    - L1 régularisation 1e-5
    - Score :
      - Train Set : 67,1 %
      - Test Set : 64,5 %



# Modèle Supervisé

- Prédiction
  - Faite sur un Topic au Hasard
  - Tags : python, iterator, iteration

Résultats (33%):
python (45,5%)
c++ (21,2%)
C (3,81%)
java (3,7%)
.net (2,32%)

Résultats (33%):
python (72%)
c++ (24%)
Performance (11%)
Optimisation (2%)
.net (2%)

  - Majoritairement langages

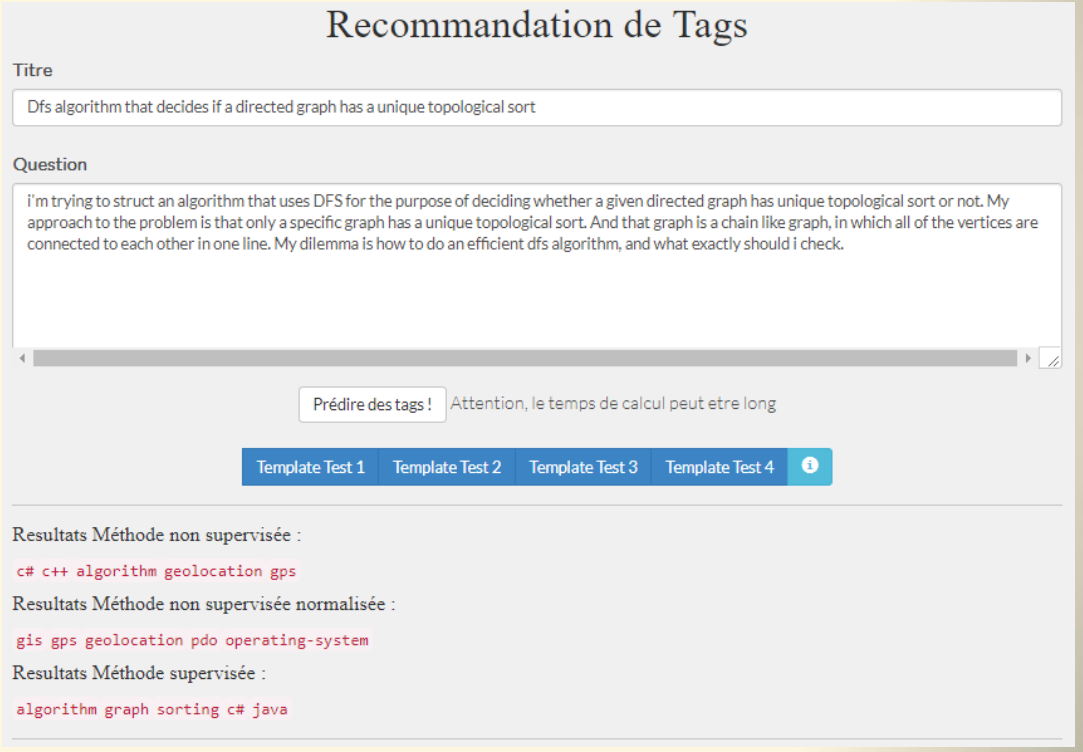
# Modèle Supervisé

- Analyse des mots-clés par Topics
  - Par modèles du MultiOutputClassifier
  - Poids les plus importants

| Tags | Mots clés |
|---|---|
| pandas | panda, datafram, seri, feedback, feed |
| dataset | dataset, zoom, feel, feed, featur |
| python | python, numpi, panda, django, matplotlib |
| machine-learning | zoom, fatal, feedback, feed, featur |
| git | git, commit, branch, repositori, repo |

# API

- Requête POST (Titre + Text)
- Non supervisé
  - TF Matrix
  - LDA
  - JSDivergence
  - Tags (train set)
  - Norm ou non
- Supervisé
  - TF-IDF
  - SGDC
  - Top 5 Classes
  - Conversion Colonne => Tag

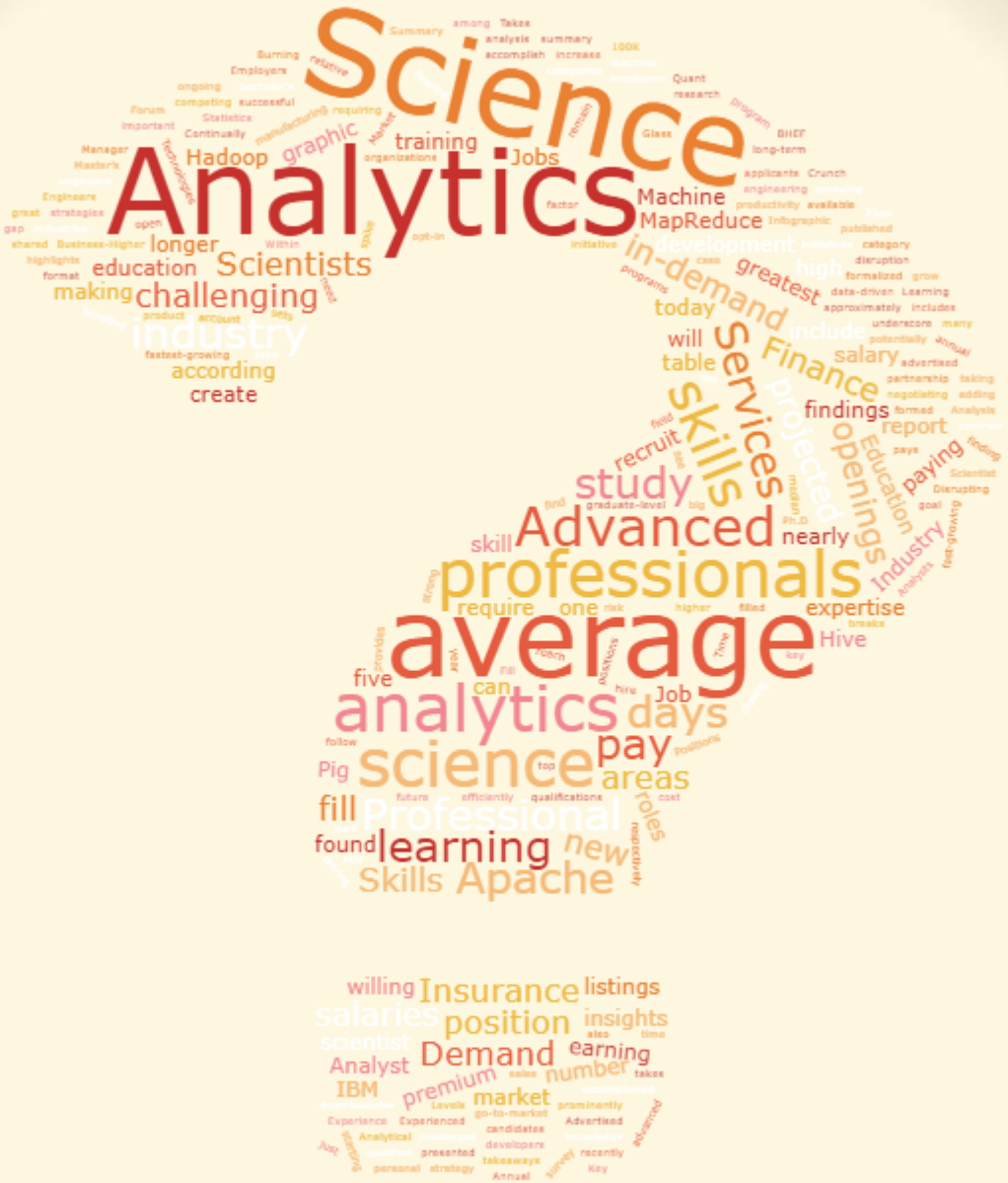


# Pistes d'évolutions

- Plus gros dataset (mémoire) ?
- Modèle Supervisé
  - Agrégation Topics + TF-IDF?
  - StopWords mots trop courants
- Modèle Non supervisé
  - Plus de StopWords (basé sur Topic 15)
    - Try, test, app, run , server
  - Ensemble NMF + LDA
- 2 modèles (Langage et Problématiques)
- Modèle évolutifs
- Prise en compte du contexte ?
  - Tokenizer multiple ou Word2Vec

# Conclusion

- Découverte de l'Analyse de Texte
- Pas d'analyse de contexte
- Résultats corrects en supervisé
- Résultats moyen en non supervisé (parfois incohérents)
- Diverses difficultés:
  - Taille des textes (petit comparé à des livres/articles)
  - Topics tous très proches (programmation)


Science
Analytics
training
Hadoop
graphic
Jobs
Machine
MapReduce
longer
education
Scientists
making
challenging
industry
according
create
development
in-demand
greatest
today
will
table
Services
Finance
salary
findings
report
skills
recruit
study
Advanced
skill
professionals
nearly
require
one
expertise
Hive
average
five
can
Job
analytics
days
pay
Pig
science
areas
fill
Professional
roles
found
learning
new
Skills
Apache
willing
Insurance
listings
salaries
position
insights
scientist
Demand
earning
Analyst
number
IBM
premium
market