

Rapport de stage

M1

Analyse des données pour la
création d'une aide à la
rédaction d'une offre d'emploi

Antoine SCHUFFENECKER, 23/08/2021

UFR mathématiques de Strasbourg



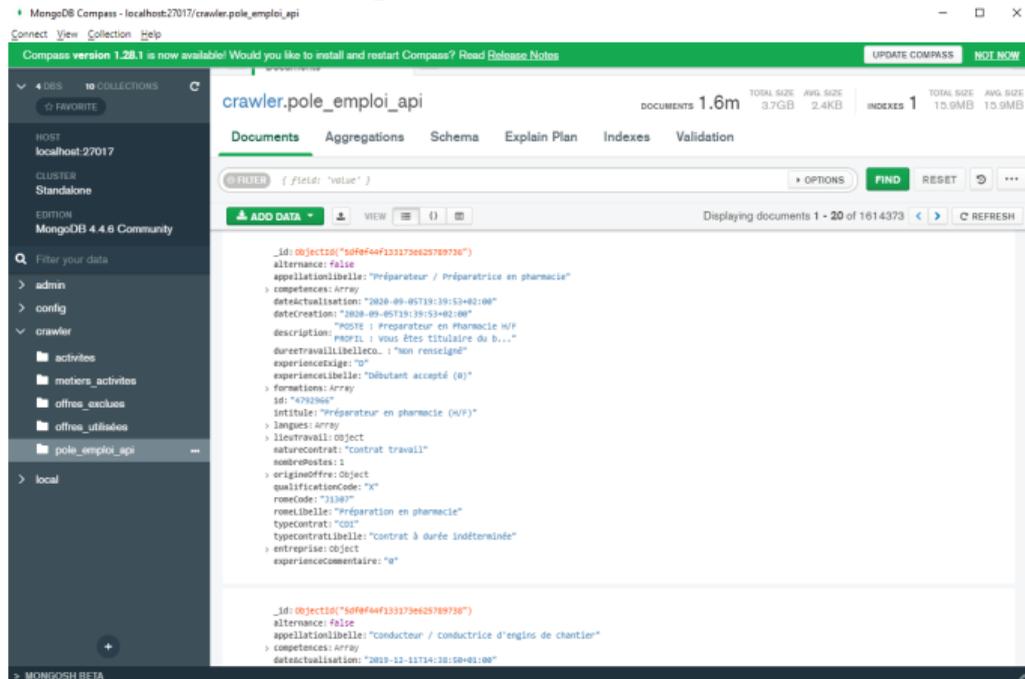
A L P H A B Y

Introduction

- Stage chez alphaby
- Présentation du problème: api pole emploi, aide au entreprises, préciser les recherches de profils.
- Buts: trouver des informations pertinentes sur un échantillon de 1,6 millions d'offres, trouver une manière efficace d'extraire des activités

Résultats préliminaires

- Utilisation de mongoDB compass



The screenshot shows the MongoDB Compass interface. The top bar indicates the version 1.28.1 and provides options to update or not update. The main area displays the database 'crawler.pole_emploi_api' with 1.6m documents and 1 index. The 'Documents' tab is active, showing a list of documents. The first document is expanded, showing a detailed JSON structure for a pharmacist.

```
{
  "_id": "50f9f44f333736e25789736",
  "alternance": false,
  "appellationlibelle": "Préparateur / Préparatrice en pharmacie",
  "competences": Array,
  "dateactualisation": "2020-09-05T19:30:53+02:00",
  "datecreation": "2020-09-05T19:30:53+02:00",
  "description": "PROFIL : Préparateur en Pharmacie H/F",
  "dureetravaillibelleco": "non renseigné",
  "experienceage": "0",
  "experiencelibelle": "Débutant accepté (0)",
  "formations": Array,
  "id": "4792966",
  "intitule": "Préparateur en pharmacie (H/F)",
  "langues": Array,
  "lieustravail": Object,
  "naturecontrat": "contrat travail",
  "numerosites": 1,
  "origineoffre": Object,
  "qualificationCode": "K",
  "romeCode": "J1307",
  "romeLibelle": "Préparation en pharmacie",
  "typecontrat": "CDD",
  "typecontratlibelle": "contrat à durée indéterminée",
  "entreprise": Object,
  "experiencecommentaire": ""
}
```

```
{
  "_id": "50f9f44f333736e25789736",
  "alternance": false,
  "appellationlibelle": "conducteur / conductrice d'engins de chantier",
  "competences": Array,
  "dateactualisation": "2020-12-17T14:30:50+01:00"
}
```

Résultats préliminaires

```
client=MongoClient()
db = client["local"]
collection = db["pole_emploi"]
cursor = collection.aggregate([{"$sample": {"size": 10000 } }])
entries=list(cursor)
df=pd.DataFrame(list(entries))
df.to_csv('Dataframe.csv')
```

```
# Pour afficher évolution en fonction du temps du nombres d'offres par secteur:
```

```
df=dfb.copy()
```

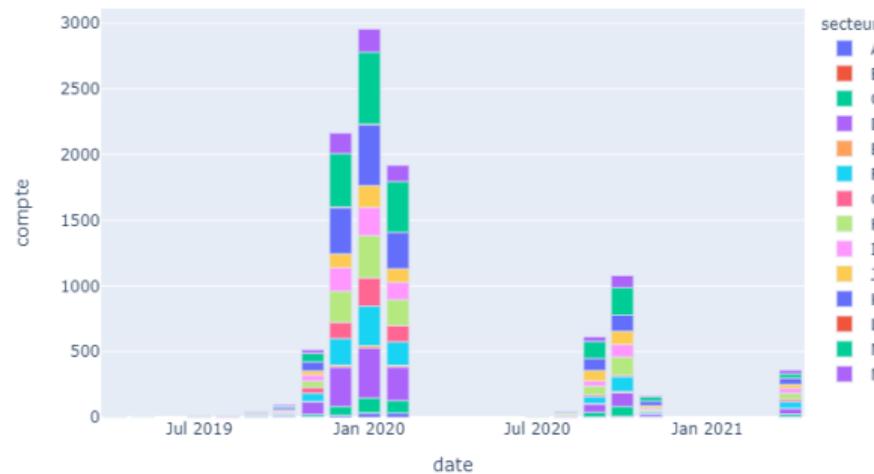
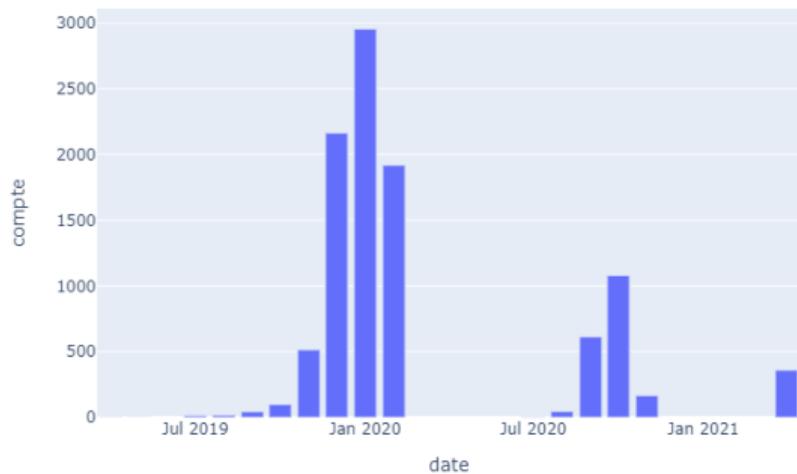
```
df['date']=df['dateCreation'].apply(lambda x : x[:7])
df['annee']=df['dateCreation'].apply(lambda x : x[:4])
df['secteurs']=df['romeCode'].apply(lambda x : x[0])
```

```
df=df[df['annee']>='2019']
df=df.groupby(['secteurs','date']).size().reset_index(name='compte')
```

```
fig = px.bar(df, x="date", y="compte",color='secteurs')
fig.show()
```

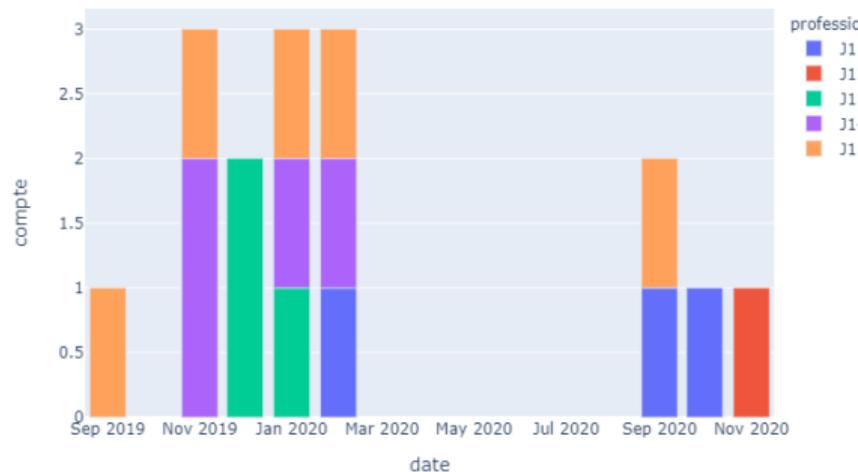
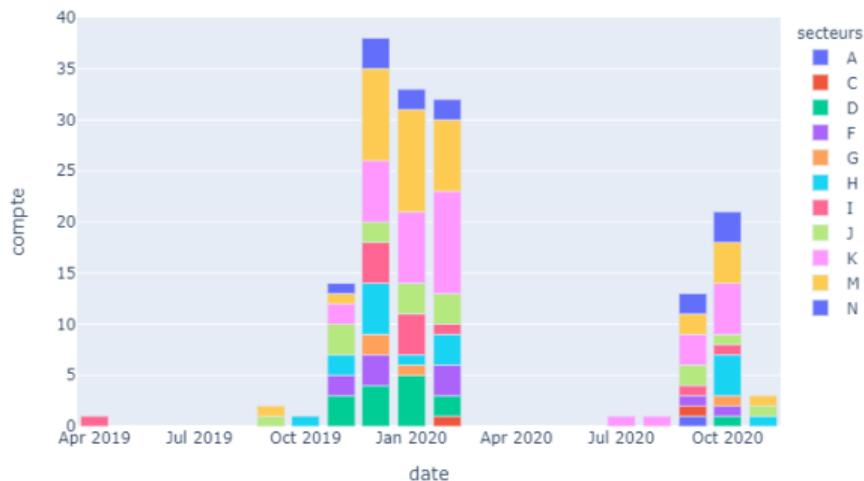
Résultats préliminaires

L'évolution du nombres d'offres sur un échantillon de 10000 offres



Résultats préliminaires

L'évolution du nombres d'offres dans un départements et l'évolution du nombres d'offres d'un secteur avec ses sous-secteurs



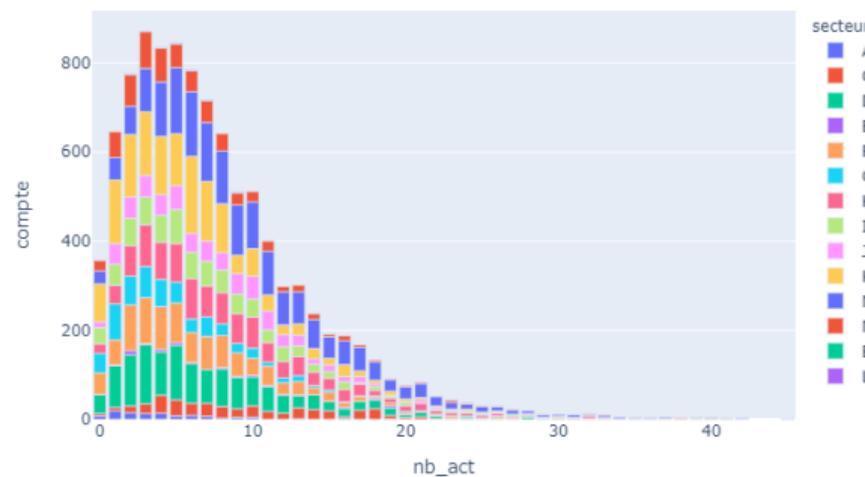
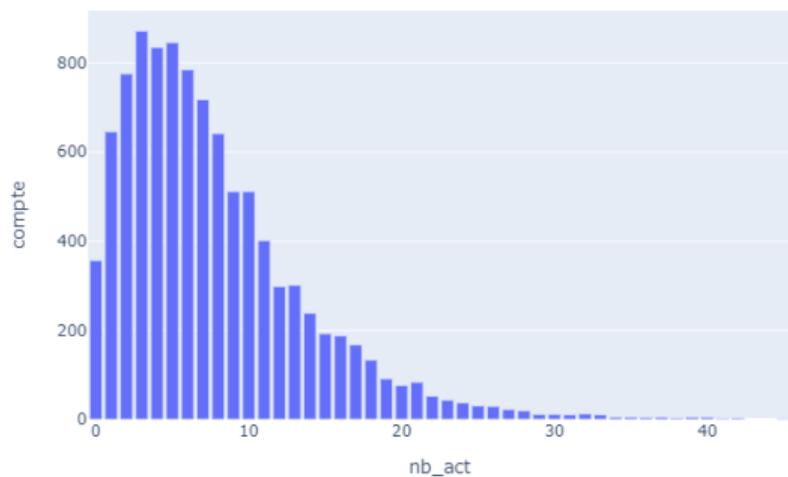
Extraction

- Principe "d'activités"
- Méthodologie primaire



Résultats

L'évolution du nombres d'activités par d'offres sur un échantillon de 10000 offres



Résultats

Création d'une nouvelle collection mongoDB

```
client=MongoClient()
db = client["local"]
#db_metiers = mongoClient.crawler.metiers_activites
collection = db["pole_emploi"]
# Recherche de toutes les offres dans mongo
offres_cursor = collection.find({}, {"description" : 1, "id": 1, "appellationlibelle": 1,"romeCode": 1})

# Sizing
offres_count = collection.estimated_document_count()
offres_batch_size = 10000
offres_nb_batches = offres_count//offres_batch_size

# Extraction activités + remplissage de la table db_metiers + calculs stats
i = 1
ratios_extraction = {}
for batch in tqdm(batched(offres_cursor, offres_batch_size), total=offres_nb_batches):
    print("[Batch "+str(i)+"]")
    df = pd.DataFrame(batch)
    df = add_activities(df)
    new_df=df
    new_df['secteurs']=df['romeCode'].apply(lambda x : x[0])
    new_df=new_df.explode('act')
    new_df=new_df.groupby(['act', 'secteurs', 'romeCode']).size().reset_index(name='occ')
    new_df.to_csv('Batch'+str(i)+'_csv')
    i += 1
```

Résultats

MongoDB Compass - localhost:27017/crawler.metiers_activites

Connect View Collection Help

Compass version 1.28.1 is now available! Would you like to install and restart Compass? [Read Release Notes](#) UPDATE COMPASS NOT NOW

4 DBS 10 COLLECTIONS

☆ FAVORITE

HOST
localhost:27017

CLUSTER
Standalone

EDITION
MongoDB 4.4.6 Community

Filter your data

- > admin
- > config
- ▼ crawler
 - activites
 - metiers_activites
 - offres_exclues
 - offres_utilisées
 - pole_emploi_api
- > local

crawler.metiers_activites

DOCUMENTS 141.6k TOTAL SIZE 26.1MB AVG. SIZE 193B INDEXES 1 TOTAL SIZE 1.3MB AVG. SIZE 1.3MB

Documents Aggregations Schema Explain Plan Indexes Validation

FILTER { field: 'value' } OPTIONS FIND RESET ↺ ⋮

ADD DATA ↓ ↓ VIEW [] [] []

Displaying documents 1 - 20 of 141616 < > REFRESH

```
{ "_id": ObjectId("60d3b6af081f408bb18570d4"),  
  "metier": "Boucher / Bouchère",  
  "activite": "abattage d'animaux",  
  "id": "4258275",  
  "romeCode": "D1101",  
  "count": 4  
}
```

```
{ "_id": ObjectId("60d3b6f9081f408bb1860e08"),  
  "metier": "Boucher industriel / Bouchère industrielle",  
  "activite": "abattage d'animaux",  
  "id": "4972155",  
  "romeCode": "H2101",  
  "count": 2  
}
```

```
{ "_id": ObjectId("60d3b972081f408bb18c99d1"),  
  "metier": "Élagueur-grimpeur / Élagieuse-grimpeuse",  
  "activite": "abattage d'arbres manuel",  
  "id": "09095CG",  
  "romeCode": "A1201",  
  "count": 2  
}
```

```
{ "_id": ObjectId("60d3b64e081f408bb1836e2f"),  
  "metier": "Bûcheron / Bûcheronne",  
  "activite": "abattage d'arbres de 30 à 40m de hauteur afin de rendre des terrains c...",  
  "id": "0909BST",  
  "count": 2  
}
```

Amélioration

Nouvelle philosophie pour extraire les activités :

- Plus seulement les premiers mots de chaque ligne
- Meilleure gestion avec mongo pour accélérer grandement la vitesse d'exécution

Amélioration de la précédente collection et création d'une nouvelle



Amélioration

MongoDB Compass - localhost:27017/crawler.activites

Connect View Collection Help

Local

4 DBS 10 COLLECTIONS

localhost:27017

Standalone

MongoDB 4.4.6 Community

admin

config

crawler

activites

metiers_activites

offres_exclues

offres_utilisees

pole_emploi_api

local

crawler.activites Documents

DOCUMENTS 365.7k TOTAL SIZE 51.0MB AVG. SIZE 146B INDEXES 1 TOTAL SIZE 19.7MB AVG. SIZE 19.7MB

Documents Aggregations Schema Explain Plan Indexes Validation

FILTER { field: 'value' }

ADD DATA

VIEW

Displaying documents 1 - 20 of 365691

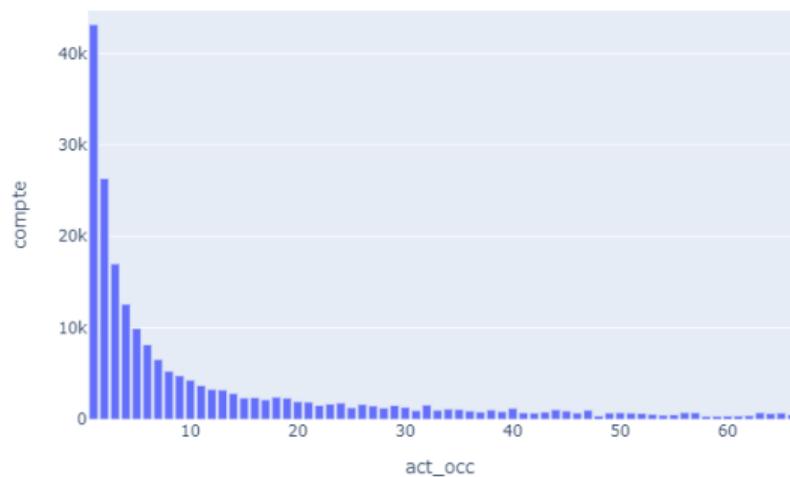
```
{
  "_id": "Object",
  "act": "preparation des viandes, de la charcuterie et des fromages, mise en pl...",
  "id_offre": "Array",
  "0": "0988WF1"
}
```

```
{
  "_id": "Object",
  "act": "veiller a la tracabilite et conformite des matieres premieres, des ins...",
  "id_offre": "Array",
  "0": "7390818"
}
```

```
{
  "_id": "Object",
  "act": "mettre en place le chantier",
  "id_offre": "Array",
  "0": "4894246",
  "1": "096CP3N",
  "2": "7293745",
  "3": "4384292",
  "4": "4087742",
  "5": "096G0XX",
  "6": "4087571",
  "7": "4897657",
  "8": "5146383",
  "9": "4987361",
  "10": "5548791",
  "11": "7085883",
  "12": "4354624"
}
```

Amélioration

Occurrences des activités dans l'ensemble de l'échantillon



[311872 rows x 1 columns]

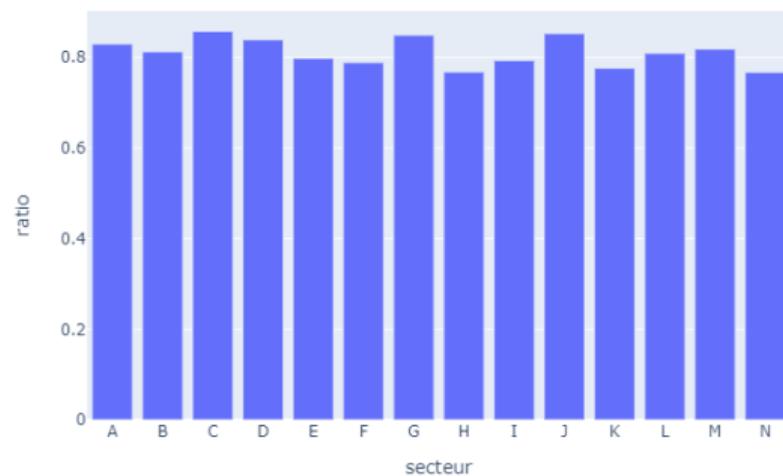
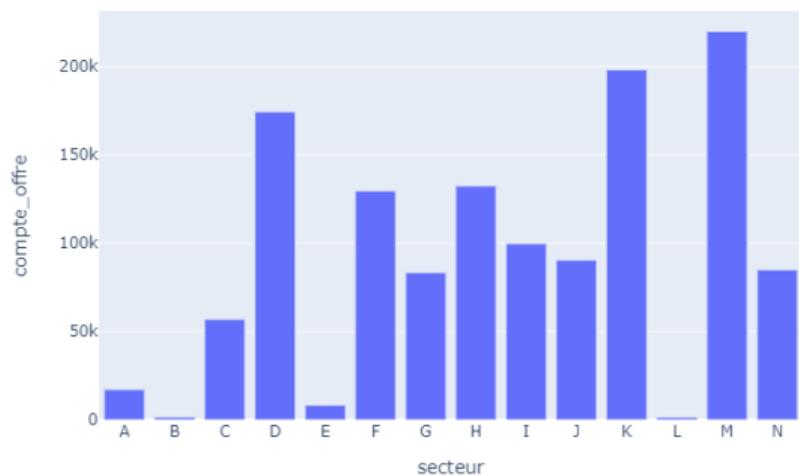
	act_occ	compte
0	1	43144
1	2	26327
2	3	16996
3	4	12577
4	5	9937
5	6	8152
6	7	6534
7	8	5257
8	9	4762
9	10	4284
10	11	3699
11	12	3277
12	13	3199
13	14	2826
14	15	2324
15	16	2377
16	17	2142
17	18	2423
18	19	2312
19	20	1920

Amélioration

Raison d'exclusion des offres ?

152.83517711584733

159.5896105537124



Conclusion

Travail de recherche

Réalité de situation concrète

Pertinence de l'utilisation des activités

Ouverture

Remerciement