Rapport de stage M1

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

Antoine SCHUFFENECKER, 23/08/2021

UFR mathématiques de Strasbourg

ALPHABY

Introduction

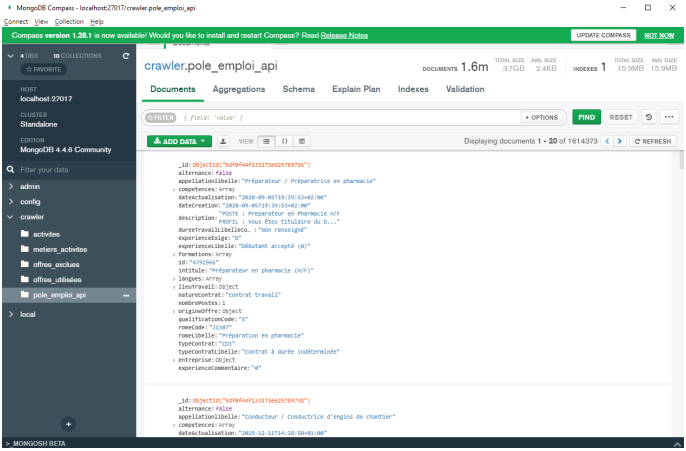
- Stage chez alphaby

- 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



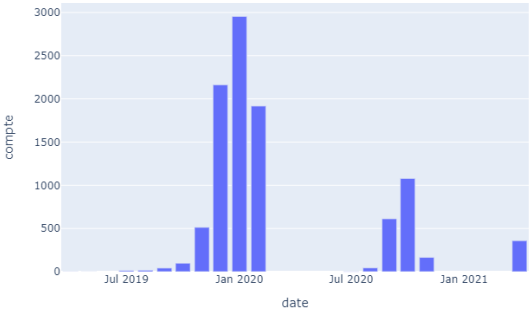
Résultats préliminaires

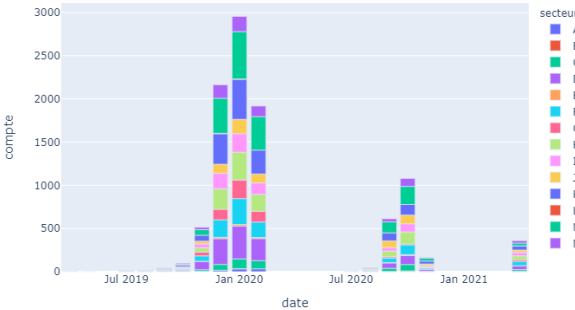
```
client=MongoClient()
db = client["local"]
collection = db["pole_emploi"]
cursor = collection.aggregate([{ "$sample": { "size": 10000 } }])
entries=list(cursor)
df=pd.DataFrame(list(entries))
df.to_csv('Dataframe.csv')
```

```
# 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

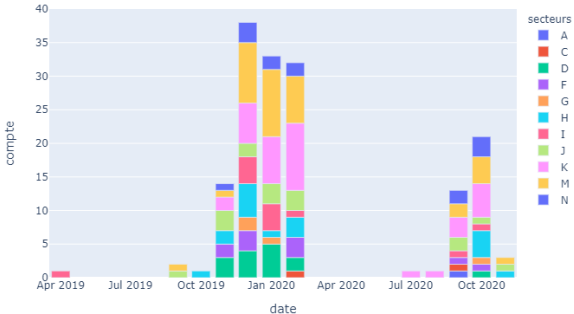
L'évolution du nombres d'offres sur un échantillon de 10000 offres

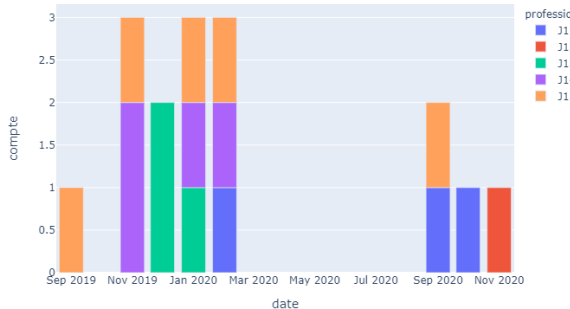




Résultats préliminaires

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





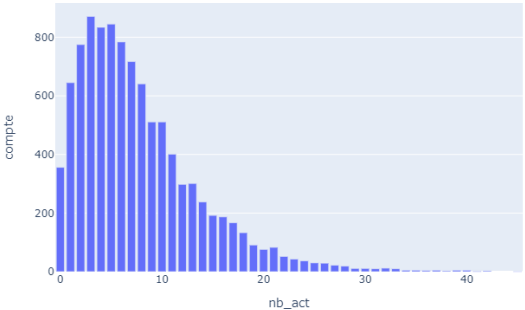
Extraction

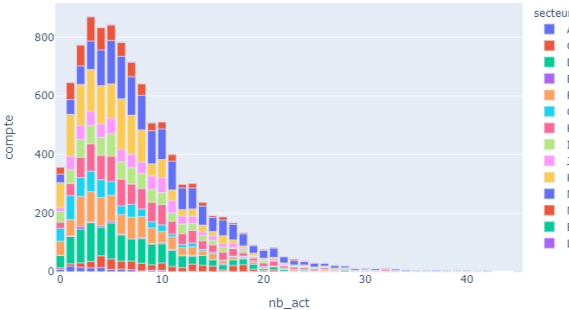
- Principe "d'activités"

- Méthodologie primaire

Résultats

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



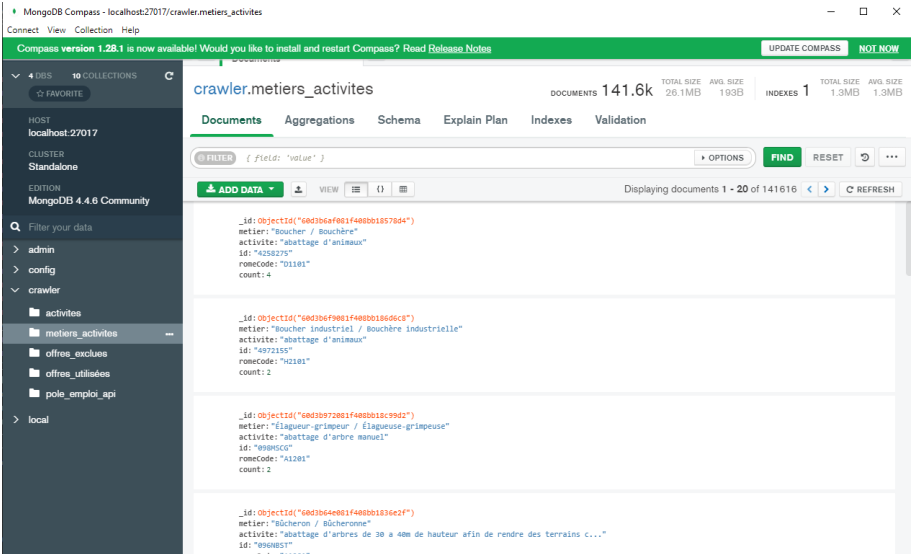


Résultats

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("60d3b6af081f408bb1857d4")
metier: "Boucher / Bouchère"
activite: "abattage d'animaux"
id: "4258275"
romeCode: "D1101"
count: 4
_id: ObjectId("60d3b6f9081f408bb186d6c8")
metier: "Boucher industriel / Bouchère industrielle"
activite: "abattage d'animaux"
id: "4972155"
romeCode: "H2101"
count: 2
_id: ObjectId("60d3b972081f408bb18c99d2")
metier: "Élagueur-grimpeur / Élagueuse-grimpeuse"
activite: "abattage d'arbre manuel"
id: "098MSCG"
romeCode: "A1201"
count: 2
_id: ObjectId("60d3b64e081f408bb1836e2f")
metier: "Bûcheron / Bûcheronne"
activite: "abattage d'arbres de 30 a 40m de hauteur afin de rendre des terrains c..."
id: "096NBS1"

Amélioration

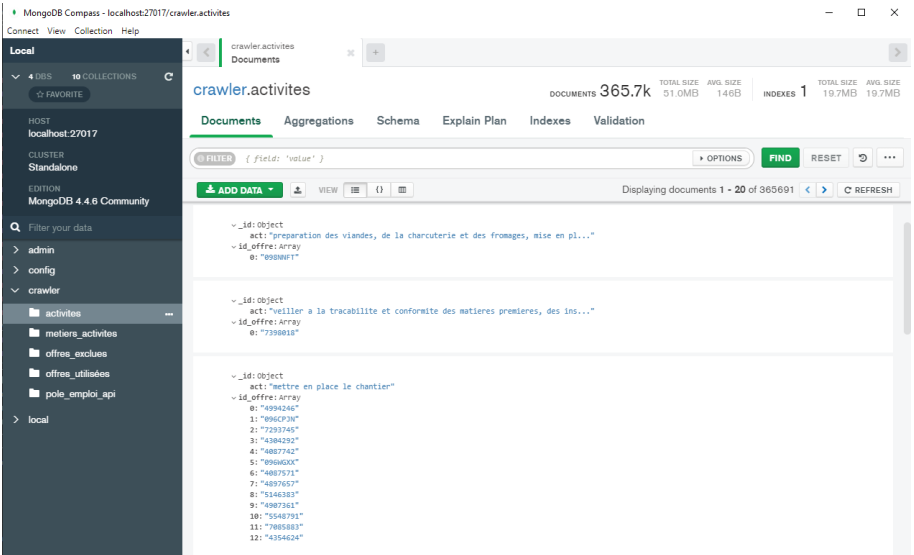
Nouvelle philosophie pour extraire les activités :

- Plus seulement les premiers mots de chaqe ligne

- Meilleur gestion avec mongo pour accélerer grandement la vitesse d'exécution

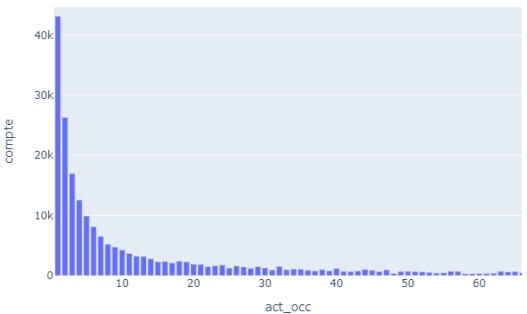
Amélioration de la précendente collection et création d'une nouvelle

Amélioration



Amélioration

Occurences des activitités dans l'ensemble de l'échantillon



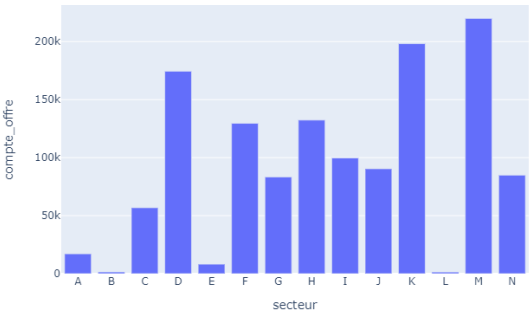
[311872 rows x 1 columns]

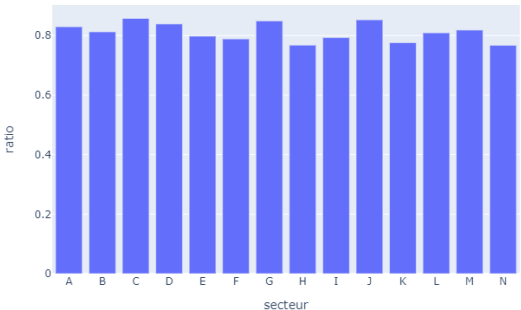
	act_occ	compte
0	1	43144
1	2	26327
2	3	16996
3	4	12577
4	5	9937
5	6	8152
6	7	6534
7	8	5257
8	9	4762
9	10	4284
10	11	3699
11	12	3277
12	13	3199
13	14	2826
14	15	2324
15	16	2377
16	17	2142
17	18	2423
18	19	2312
19	20	1920

Amélioration

Raison d'exclusion des offres ?

```
152.83517711584733
159.5896105537124
```





Conclusion

Travail de recherche

Réalité de situation concrète

Pertinence de l'utilisation des activitités

Ouverture

Remerciement