



compte rendu

Mini-Projet

sur le projet de web scraping et data
visualisation avec power bi

encadré par :

Pr Imade Benelallam

- Nabigh Mohamed
 - DS
- Aissame Kaouch
 - DS

Web scraping : définition

Le Web scraping (de l'anglais *scraping* = « gratter/raceler ») consiste à extraire des données de sites Internet et à les enregistrer afin de les analyser ou de les utiliser de toute autre façon. Le scraping permet de collecter des informations de nature bien différente. Il peut par ex. s'agir de **coordonnées** comme des adresses e-mail ou des numéros de téléphone, mais aussi de **mots-clés individuels ou d'URL**. Ces informations sont alors rassemblées dans des bases de données locales ou des tableaux.

Comment fonctionne le Web scraping ?

Le scraping comprend différentes fonctionnalités, mais on opère généralement une distinction entre le scraping manuel et automatique. Le **scraping manuel** désigne le fait de copier et insérer manuellement des informations et des données. On peut le comparer avec le fait de découper et rassembler des articles de journaux. Le scraping manuel est uniquement effectué lorsque l'on souhaite trouver et enregistrer des informations de façon sporadique. Il s'agit d'un **processus très laborieux** qui est rarement appliqué pour de grandes quantités de données.

Dans le cas du **scraping automatique**, on utilise un logiciel ou un algorithme qui explore plusieurs sites Internet afin d'extraire des informations. Un logiciel spécifique est utilisé en fonction de la nature du site Internet et du contenu. Dans le scraping automatique, on distingue différentes **méthodes** :

- **Les analyseurs syntaxiques** : un analyseur syntaxique est utilisé pour convertir le texte en une nouvelle structure. Dans le cas de l'analyse d'un HTML par exemple, le logiciel lit le document HTML et enregistre les informations. L'analyse d'un DOM utilise l'affichage des contenus dans le navigateur côté client pour extraire les données.
- **Les robots** : un robot est un logiciel réalisant des tâches spécifiques et les automatisant. Dans le Web harvesting, les robots sont utilisés pour explorer automatiquement des sites Internet et collecter des données.
- **Le texte** : les personnes sachant utiliser la Command Line peuvent utiliser les instructions Unix grep pour explorer le Web à la recherche de certains termes dans

Python ou Perl. Il s'agit d'une méthode très simple pour obtenir des données qui requiert toutefois davantage de travail que lorsqu'on utilise un logiciel.

Scraping Target

- [challenge.ma](#)



Challenge, anciennement Challenge Hebdo, est un hebdomadaire économique marocain francophone créé en 2004.

Le premier numéro de Challenge Hebdo, d'environ quarante pages et incluant un dossier sur la Chine, est sorti le 2 avril 2004. En juillet 2009, une fusion a été réalisée avec un autre titre de presse, La Gazette du Maroc, de son groupe d'appartenance : Les Éditions de la Gazette dont dépend aussi VH Magazine (mensuel masculin francophone) et Lalla Fatéma (mensuel féminin arabophone).

- [www.huffpostmaghreb](#)



Le Huffington Post est un journal d'information gratuit d'origine américaine publié exclusivement sur Internet. Cofondé en 2005 par Arianna Huffington, qui en a été la rédactrice en chef jusqu'au 11 août 2016, Kenneth Lerer (en), Jonah Peretti (en) et Andrew Breitbart, il fait appel à de nombreuses collaborations et sources externes. Il emploie ainsi près de 20 000 blogueurs qu'il ne rémunère pas.

Le site originel The Huffington Post est racheté aux États-Unis en février 2011 par AOL pour 315 millions de dollars américains, puis connaît une expansion internationale grâce à des versions développées le plus souvent avec des partenaires locaux, dans un premier temps (2011) en anglais (Royaume-Uni et Canada), puis, en 2012, en français avec Le Huffington Post (fondé en partenariat avec le groupe Le Monde) et Le Huffington Post Québec.

- www.lavieeco.com



Créé en 1957 par Marcel Herzog, La Vie Eco est un hebdomadaire économique et financier. En 1994, Jean-Louis Servan Schreiber s'installe à Casablanca avec son épouse marocaine. Il rachète le journal et en devient propriétaire et directeur. Il contribue à faire de la Vie Eco un des journaux les plus dynamiques du pays. Son départ du Maroc est précipité en 1997 par Driss Basri qui ne l'aime pas. En 1997, La Vie Eco est achetée par le groupe Caractères, détenue par Aziz Akhannouch, alors réputé proche de Basri. Après avoir été dirigé par Saad Benmansour, Nabila Fathi est nommée directrice de la rédaction en 2021.

- [leconomiste](http://leconomiste.com)



L'Économiste est un journal marocain francophone basé à Casablanca fondé en 1991. Il traite principalement des informations économiques, financières et boursières du Maroc. Il fait partie du groupe Eco-Medias.

- LeMatin



Le Matin, anciennement Le Matin du Sahara et du Maghreb, est un quotidien marocain publié en français, présentant des actualités nationales et internationales ainsi que des informations pratiques. Édité par la société Maroc Soir, il est considéré comme le journal quasi officiel du Palais Royal. Le groupe Le Matin est leader sur la scène médiatique marocaine. Il compte plusieurs titres papier et digitaux. Sur le volet papier, il édite aussi le quotidien Assahra Al Maghribia ainsi que plusieurs spéciaux thématiques et ouvrages dont Maroc RSE de la série Maroc Business Intelligence. Le groupe compte aussi un pôle industriel d'impression presse, dépliants livres scolaires et travaux numériques petit et grand format. Le groupe le Matin se distingue aussi par un réseau unique de treize agences régionales qui couvrent l'ensemble du territoire national.

Résultat de scraping:

en utilisant le code que nous avons développé, nous avons réussi à récupérer le titre, le lien et le contenu de chaque article de ces sites Web

quelques exemples

- code pour huffingtonpost

```
from bs4 import BeautifulSoup
import requests

#creating empty list
urls=[]
#function created
def scrape_huff(site):

    urls=[]
    #getting the request from url
    r=requests.get(site)

    #converting the text
    s=BeautifulSoup(r.text,"html.parser")

    link=[]
    head=[]
    content=[]

    for i in s.find_all("a",class_="card__headline card__headline--long"):

        head.append(i.text)
        link.append(i.attrs['href'])

    for i in link:
        r=requests.get(i)

        #converting the text
        s=BeautifulSoup(r.text,"html.parser")

        for i in s.find_all("div",class_="content-list-component"):

            content.append(i.text)

    return [head,link,content]
```

- résultats:

In [8]: huff

- on stock dans une pandas sata frame

	title	link	contenu
0	Une enquête ouverte après la mort d'une Française dans l'Algérie... Yacéf Saadi, héros de l'indépendance de l'Algérie...	https://www.huffingtonpost.fr/entry/maroc-une-mort-de-yac...	MAROC - La police marocaine a ouvert ce samedi...
1	Plus de 700 hectares brûlés dans des incendies... Incendies en Algérie: les Canadair français dé... La France envoie des moyens aériens à l'Algérie...	https://www.huffingtonpost.fr/entry/maroc-plus-incendies... https://www.huffingtonpost.fr/entry/la-france-envoie-des-moyens-aerien...	ALGÉRIE - Yacéf Saadi, héros de la lutte pour ... À voir également sur le HuffPost: À Alger, une... MAROC - La mobilisation continue ce lundi 16 a...
2	En Algérie, le bilan des incendies monte encore... Le témoignage d'un étudiant "torturé" soulève ... BLOG - Madame Bendouda, la richesse culturelle...	https://www.huffingtonpost.fr/entry/en-algerie-les-incendies... https://www.huffingtonpost.fr/entry/le-temoignage-d-un-etudiant-torture... https://www.huffingtonpost.fr/entry/madame-ben...	49 degrésUne vague de chaleur traverse le Maroc... ALGÉRIE - L'aide humaine et matérielle promise... FR 🇲🇦 Les 2 #Canadair et le Beech sont arrivés... Ce soutien aérien français est très précieux p...
3	3 recettes de couscous qui vont réchauffer vos... Le couscous du Maghreb au patrimoine immatériel...	https://www.huffingtonpost.fr/entry/recettes-couscous... https://www.huffingtonpost.fr/entry/le-couscou...	INCENDIES - Paris va envoyer deux Canadair et ... Face aux drames auxquels sont confrontés les a... L'annonce a été rapidement appuyée par une déo...
4	Qui en tête au 1er tour? Une majorité pour Mac... "J'ai même lu qu'il mangeait les enfants": Vér...	https://www.huffingtonpost.fr/entry/tous-les-soutiens... https://www.huffingtonpost.fr/entry/france-202...	INCENDIES - Paris va envoyer deux Canadair et ... Face aux drames auxquels sont confrontés les a... L'annonce a été rapidement appuyée par une déo...
5	Donald Trump placé au centre d'un "complot" pa... Covid-19: L'OMS n'écarte plus la thèse d'une f... Sarkozy prend parti dans la guerre Estrosi-Cio...	https://www.huffingtonpost.fr/entry/donald-trump-place... https://www.huffingtonpost.fr/entry/covid-19-oms... https://www.huffingtonpost.fr/entry/sarkozy-pr...	INCENDIES - Paris va envoyer deux Canadair et ... Face aux drames auxquels sont confrontés les a... L'annonce a été rapidement appuyée par une déo...
6	Stéphane Rotenberg ouvre un restaurant avec to...	https://www.huffingtonpost.fr/entry/top-chef-s...	#Solidarité🇩🇿 FR Dans le cadre du Mécanisme de...

- code pour lavieeco

```
from bs4 import BeautifulSoup
import requests

#creating empty list
urls=[]
#function created
def scrape_lavieeco(site):

    r=requests.get(site)

    #converting the text
    s=BeautifulSoup(r.text,"html.parser")

    link=[]
    head=[]
    content=[]

    for i in s.find_all("a",class_="post-title post-url"):

        head.append(i.text)
        link.append(i.attrs['href'])

    for i in link:

        r=requests.get(i)

        #converting the text
        s=BeautifulSoup(r.text,"html.parser")

        c=''
        for i in s.find_all("p"):
            c+=i.text
            c = c.replace("\n", "")
        content.append(c)

    return [head,link,content]
```

- résultats:
- on stock dans une pandas sata frame

	title	link	contenu
0	Loi de finances 2023 : La CGEM dépose ses prop...	https://lematin.ma/express/2022/loi-finances-2...	Ces incendies ont débuté lundi soir en Kabylie...
1	La Bourse de Casablanca renouvelle sa certific...	https://lematin.ma/express/2022/bourse-casabla...	
2	Barid Al-Maghrib : émission d'un timbre-poste ...	https://lematin.ma/express/2022/barid-al-maghri...	ALGÉRIE - Des pompiers soutenus par des milita...
3	Ithmar Capital réunit les fonds souverains afr...	https://lematin.ma/express/2022/ithmar-capital...	
4	Bourse : BMCE capital recommande d'accumuler l...	https://lematin.ma/express/2022/tablons-amelio...	ALGÉRIE - Le témoignage d'un jeune étudiant al...
5	Sound Energy lève 4 millions de livres sterlin...	https://lematin.ma/express/2022/sound-energy-l...	
6	Bourse : BMCE capital recommande de cumuler le...	https://lematin.ma/express/2022/bcp-enregistre...	PATRIMOINE - Depuis la nuit des temps, le comb...
7	Maroc Telecom maintient sa position de premièr...	https://lematin.ma/express/2022/maroc-telecom-...	
8	Un nouveau départ pour le Conseil d'affaires M...	https://lematin.ma/express/2022/reactivation-c...	SONDAGES - Il était temps. La campagne a fini ...
9	IDE: Avec près de 83 milliards de dollars, l'A...	https://lematin.ma/express/2022/investissement...	
10	Les acteurs de l'écosystème mobilisés pour réu...	https://lematin.ma/express/2022/reussir-virage...	Si l'on a assisté à une baisse importante de l...
11	Bourse : BMCE capital recommande l'allégement ...	https://lematin.ma/express/2022/hps-connaître-...	
12	Blé tendre : le Maroc, 3e client hors UE de la...	https://lematin.ma/express/2022/ble-tendre-mar...	Cette majorité absolue, c'est Ensemble! qui en...
13	Tanger Med : TMPA sur un projet de gestion sma...	https://lematin.ma/express/2022/tanger-med-tmp...	
14	Bourse : BMCE capital recommande l'allégement ...	https://lematin.ma/express/2022/difficultés-op...	Le compilateur de sondages du HuffPost, commen...
15	Banque mondiale : la croissance ralentirait au...	https://lematin.ma/express/2022/banque-mondial...	

- code pour lematin

```

from bs4 import BeautifulSoup
import requests

#creating empty list
urls=[]
#function created
def scrape_lematin(site):

    r=requests.get(site)

    #converting the text
    s=BeautifulSoup(r.text,"html.parser")

    link=[]
    head=[]
    content=[]

    for i in s.find_all("div",class_="card h-100"):
        link.append(i.a.attrs['href'])

    for j in link:
        r=requests.get(j)

        #converting the text
        s=BeautifulSoup(r.text,"html.parser")

        for i in s.find_all("article",class_="card p-1 mb-2"):
            head.append(i.h1.text)

        for i in s.find_all("div",class_='card-body p-2'):
            content.append(i.text)

    return [head,link,content]

```

- résultats:
- on stock dans une pandas sata frame

	title	link	contenu
0	\nOpération Marhaba 2022 : La RAM propose 6 mi...	https://www.lavieeco.com/actualite-maroc/opera...	LÉGISLATIVES - Pouvoir d'achat, retraites, édu...
1	\nRyanair annonce son programme d'hiver pour l...	https://www.lavieeco.com/actualite-maroc/ryana...	🔴 🚨 « J'ai lu dans un journal d'extrême-droite ...
2	\nHôtellerie : Le Maroc, un hub stratégique ve...	https://www.lavieeco.com/actualite-maroc/hotel...	"Je trouve ça scandaleux""J'ai même lu dans un...
3	\nTourisme : Fatim-Zahra Ammor active l'octroi...	https://www.lavieeco.com/actualite-maroc/touri...	Pap Ndiaye est un militant racialiste et anti-...
4	\nTransition numérique : Le tout digital c'est...	https://www.lavieeco.com/economie/transition-n...	À voir également sur Le HuffPost: Vous ne comp...
5	\nAgriculture urbaine : Un projet pilote de 53...	https://www.lavieeco.com/economie/agriculture-...	
6	\nRetraite-réforme : Le double défi d'éviter l...	https://www.lavieeco.com/economie/retraite-ref...	ÉTATS-UNIS - Les images de l'assaut du Capitole...
7	\nKissa'n relance le verre soufflé artisanal	https://www.lavieeco.com/economie/kissan-relan...	Thread le @January8thCmte vient de montrer une...
8	\nVers un nouveau record de recettes en 2022 !	https://www.lavieeco.com/economie/vers-un-nouv...	"Le 6 janvier a été la culmination d'une tenta...
9	\nSourcing : Managem Group fournit Renault en ...	https://www.lavieeco.com/economie/sourcing-man...	part 4 de la vidéo chrono assaut du Capitole d...

Out[2]:

	address	awards/0/name	awards/0/year	awards/1/name	awards/1/year	awards/2/name	awards/2/year	awards/3/name	awards/3/year	awards/4/name
0	3 rue Aristide Bruant, 75018 Paris France	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	55 rue Monge, 75005 Paris France	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	Certificate of Excellence 2019	2019.0	Certificate of Excellence 2018	2018.0	NaN
2	4 boulevard Malesherbes, 75008 Paris France	Travelers' Choice 2021	2021.0	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	NaN	NaN	NaN
3	10 rue d Ormesson, 75004 Paris France	Travelers' Choice 2021	2021.0	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	Certificate of Excellence 2019	2019.0	Certificate of Excellence 2018
4	52 rue François 1er, 75008 Paris France	Travelers' Choice 2021	2021.0	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	Certificate of Excellence 2019	2019.0	Certificate of Excellence 2018
...
95	9 rue Roy, 75008 Paris France	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	Certificate of Excellence 2019	2019.0	Certificate of Excellence 2018	2018.0	Certificate of Excellence 2017
96	111 avenue de la Bourdonnais, 75007 Paris France	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	Certificate of Excellence 2019	2019.0	Certificate of Excellence 2018	2018.0	Certificate of Excellence 2017
97	22 rue Saint Sulpice, 75006 Paris France	Certificate of Excellence 2021	2021.0	Certificate of Excellence 2020	2020.0	Certificate of Excellence 2019	2019.0	Certificate of Excellence 2018	2018.0	Certificate of Excellence 2017

tokenizer

on doit pour chaque article

1. appliquer le word tokenizer du NLTK
2. charger les stop words unique du français
3. filtrer les stop words
4. extraire les bigrammes
5. detecter les bigrammes uniques
6. compter les bigrammes uniques
7. définir la fonction qui calculr le TF
8. définir la fonction qui calcule le IDF
9. définir la fonction qui calcule le TF-IDF
10. calculer le tf IDF

alors on définit une fonction pour nous aider :

```
In [50]: def freq(df):  
  
    text=''  
    for i in df['contenu']:  
        text+=str(i)  
    text=text.lower()  
  
    tokens = WordPunctTokenizer().tokenize(text)  
  
    stop_words=stopwords.words('french')  
  
    ponct="!@#$%^&*()_+=-[]{}';:/?.>,<\|^~"  
  
    tknz=[]  
    for i in tokens:  
        if (i not in stop_words) and (i not in ponct) :  
            tknz.append(i)  
  
    # Instantiate a stemmer  
    ps = PorterStemmer()  
    # and stem  
    stems = [ps.stem(tk) for tk in tknz]  
  
    return stems
```

```
In [84]: def clean_freq(df):  
  
    text=''  
    for i in df['contenu']:  
        text+=str(i)  
    text=text.lower()  
  
    tokens = WordPunctTokenizer().tokenize(text)  
  
    stop_words=stopwords.words('french')  
  
    ponct="!@#$%^&*()_+=-[""]{}';:/?.>,<\|^~"  
  
    tknz=[]  
    for i in tokens:  
        if (i not in stop_words) and (i not in ponct) :  
            tknz.append(i)  
  
    # Instantiate a stemmer  
    ps = PorterStemmer()  
    # and stem  
    stems = [ps.stem(tk) for tk in tknz]  
  
    count=[i[1] for i in Counter(stems).most_common()]  
    terms=[i[0] for i in Counter(stems).most_common()]  
  
    return [terms,count]
```

```

: def uni_bg_freq(df):

    stems_l=freq(df)
    stems=stems_l
    bigrams = [w for w in ngrams(stems,n=2)]

    unique_bigram=[]
    for i in bigrams:

        c=i[0] + ' ' + i[1]
        unique_bigram.append(c)

    Counter(unique_bigram).most_common()

    count=[i[1] for i in Counter(unique_bigram).most_common()]
    term=[i[0] for i in Counter(unique_bigram).most_common()]

    return [term,count]

```

```

def tf_freq(df):

    text=[clean(i, no_emoji=True) for i in df['contenu']]

    vectorizer = TfidfVectorizer()
    vectorizer.fit(text)

    idf=list(vectorizer.idf_)
    dicti=vectorizer.vocabulary_

    mot=list(dicti.keys())
    index=list(dicti.values())

    return [idf,mot,index]

```

- le but de ces fonctions est de répondre au 10 étape mentionné ou chaque fonction produit une des étapes pour un entrée

sortie des fonction:

	word	count		bigram	count
0	a	75	0	suit aprè	31
1	aprè	43	1	aprè cett	31
2	cett	40	2	cett publicité	31
3	suit	33	3	", a	12
4	publicité	31	4	voir également	10
...
1569	mégaphon	1	3221	vidéo chrono	1
1570	délire	1	3222	capitol january6thcmt	1
1571	part	1	3223	january6thcmt january6thcommitteehearingsp	1
1572	4	1	3224	com upqxo7lux7	1
1573	upqxo7lux7	1	3225	upqxo7lux7 —	1

1574 rows × 2 columns

3226 rows × 2 columns

	index	word	tf
0	1014	maroc	3.662588
1	931	la	3.374906
2	1240	police	4.068053
3	1016	marocaine	4.068053
4	1160	ouvert	3.151762
...
1736	961	lisant	4.068053
1737	1033	megaphone	4.068053
1738	518	delire	3.662588
1739	1179	part	4.068053
1740	1670	upqxo7lux7	4.068053

Airflow



La plateforme Apache Airflow permet de créer, de planifier et de surveiller des workflows (flux de travail) par le biais de la programmation informatique. Il s'agit d'une solution totalement open source, très utile pour l'architecture et l'orchestration de pipelines de données complexes et le lancer de

tâches. Elle présente plusieurs avantages. Il s'agit tout d'abord d'une plateforme dynamique, puisque tout ce qui peut être fait avec le code Python peut être fait sur Airflow. Elle est aussi extensible, grâce à de nombreux plugins permettant l'interaction avec la plupart des systèmes externes les plus communs. Il est aussi possible de créer de nouveaux plugins pour répondre à des besoins spécifiques. En outre, Airflow apporte une elasticité. Les équipes de Data Engineers peuvent l'utiliser pour exécuter des milliers de tâches différentes chaque jour. Les workflows sont architecturés et exprimés sous forme de Directed Acyclic Graphs (DAGs) :Graphe orienté acyclique en français , dont chaque nœud représente une tâche spécifique. Airflow est conçue comme une plateforme code-first , permettant d'altérer très rapidement sur les workflows. Cette philosophie offre un haut degré d'extensibilité par rapport à d'autres outils de pipeline.

- Docker

Pour activer l' airflow on a besoin d'un container, dans notre cas on va utiliser Docker.



Le mot « Docker » fait référence à plusieurs choses, y compris un projet communautaire open source ; outils du projet open source ; Docker Inc., la société qui soutient principalement ce projet ; et les outils que l'entreprise prend officiellement en charge. Le fait que les technologies et l'entreprise partagent le même nom peut prêter à confusion.La technologie Docker utilise le noyau Linux et des fonctions de ce noyau, telles que les groupes de contrôle cgroups et les espaces de noms, pour séparer les processus afin qu'ils puissent s'exécuter de façon indépendante. Cette indépendance reflète l'objectif des conteneurs : exécuter plusieurs processus et applications séparément les uns des autres afin d'optimiser l'utilisation de votre infrastructure

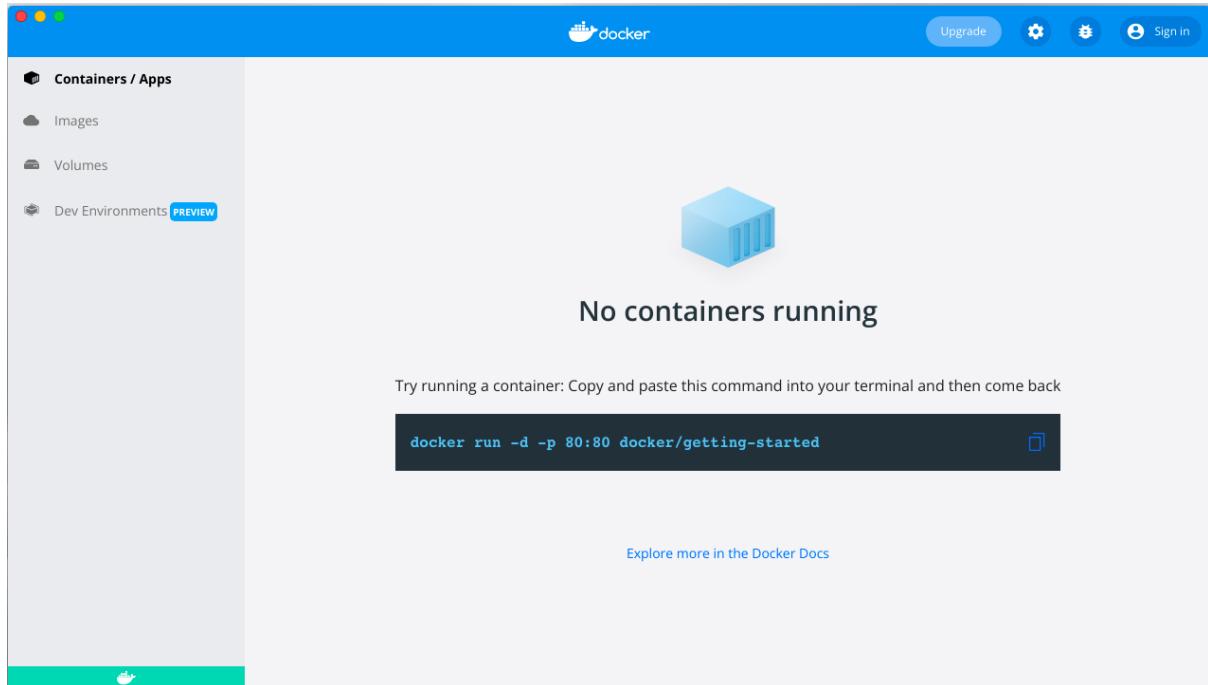
tout en bénéficiant du même niveau de sécurité que celui des systèmes distincts. Les outils de conteneurs, y compris Docker, sont associés à un modèle de déploiement basé sur une image. Il est ainsi plus simple de partager une application ou un ensemble de services, avec toutes leurs dépendances, entre plusieurs environnements. Docker permet aussi d'automatiser le déploiement des applications (ou d'ensembles de processus combinés qui forment une application) au sein d'un environnement de conteneurs. Ces outils conçus sur des conteneurs Linux (d'où leur convivialité et leur singularité) offrent aux utilisateurs un accès sans précédent aux applications, la capacité d'accélérer le déploiement, ainsi qu'un contrôle des versions et de l'attribution des versions.

Voici un bref explicatif :

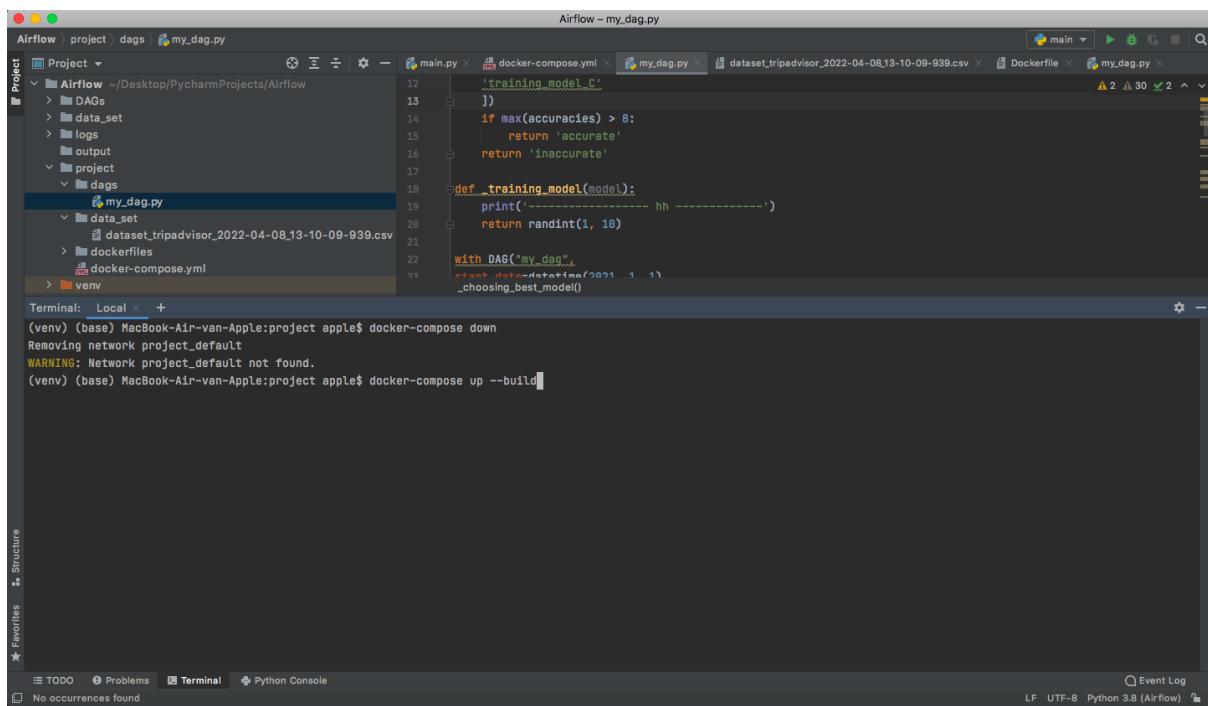
- Le logiciel informatique "Docker" est une technologie de conteneurisation qui permet la création et l'utilisation de conteneurs Linux®.
- La communauté Docker open source travaille à l'amélioration de ces technologies au profit de tous les utilisateurs.
- La société, Docker Inc., s'appuie sur le travail de la communauté Docker, la rend plus sûre et partage ces avancées avec la communauté au sens large. Il prend ensuite en charge les technologies améliorées et renforcées pour les entreprises clientes.

Avec Docker, vous pouvez traiter les conteneurs comme des machines virtuelles modulaires extrêmement légères. Et vous bénéficiez de la flexibilité avec ces conteneurs - vous pouvez les créer, les déployer, les copier et les déplacer d'un environnement à l'autre, ce qui permet d'optimiser vos applications pour le cloud.

Alors on active Docker:



il faut maintenant ouvrir et connecter le serveur avec airflow :



The screenshot shows the PyCharm IDE interface. The code editor displays Python code for a DAG named 'my_dag.py'. The terminal window shows logs from a webserver process, indicating the upgrade of the alembic.runtime.migration table and the start of the scheduler. The event log at the bottom right shows the Python version used.

```

Terminal: Local + 
webserver_1 | INFO  [alembic.runtime.migration] Running upgrade fe461863935f -> 7939bcff74ba, Add DagTags table
webserver_1 | Done.
webserver_1 | [2022-04-10 14:46:22,127] {{settings.py:253}} INFO - settings.configure_orm(): Using pool settings. pool_size=5, max_overflow=10, pool_recycle=1800, pid=1
webserver_1 | [2022-04-10 14:46:22,174] {{settings.py:253}} INFO - settings.configure_orm(): Using pool settings. pool_size=5, max_overflow=10, pool_recycle=1800, pid=20
webserver_1 | -----
webserver_1 | [2022-04-10 14:46:26,134] {{__init__.py:51}} INFO - Using executor LocalExecutor
webserver_1 | [2022-04-10 14:46:26,171] {{scheduler_job.py:1344}} INFO - Starting the scheduler
webserver_1 | [2022-04-10 14:46:26,181] {{scheduler_job.py:1352}} INFO - Running execute loop for -1 seconds
webserver_1 | [2022-04-10 14:46:26,195] {{scheduler_job.py:1353}} INFO - Processing each file at most -1 times
webserver_1 | [2022-04-10 14:46:26,199] {{scheduler_job.py:1356}} INFO - Searching for files in /usr/local/airflow/dags
webserver_1 | [2022-04-10 14:46:26,249] {{scheduler_job.py:1358}} INFO - There are 1 files in /usr/local/airflow/dags
Event Log
LF  UTF-8  Python 3.8 (Airflow)  9m

```

The screenshot shows the Airflow web interface. The top navigation bar includes links for Airflow, DAGs, Data Profiling, Browse, Admin, Docs, and About. The timestamp in the top right corner is 2022-04-10 14:47:23 UTC. The main content area is titled 'DAGs' and contains a table listing one DAG entry:

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
		my_dag	@daily	airflow				

Below the table, a message indicates 'Showing 1 to 1 of 1 entries'. A link 'Hide Paused DAGs' is visible.

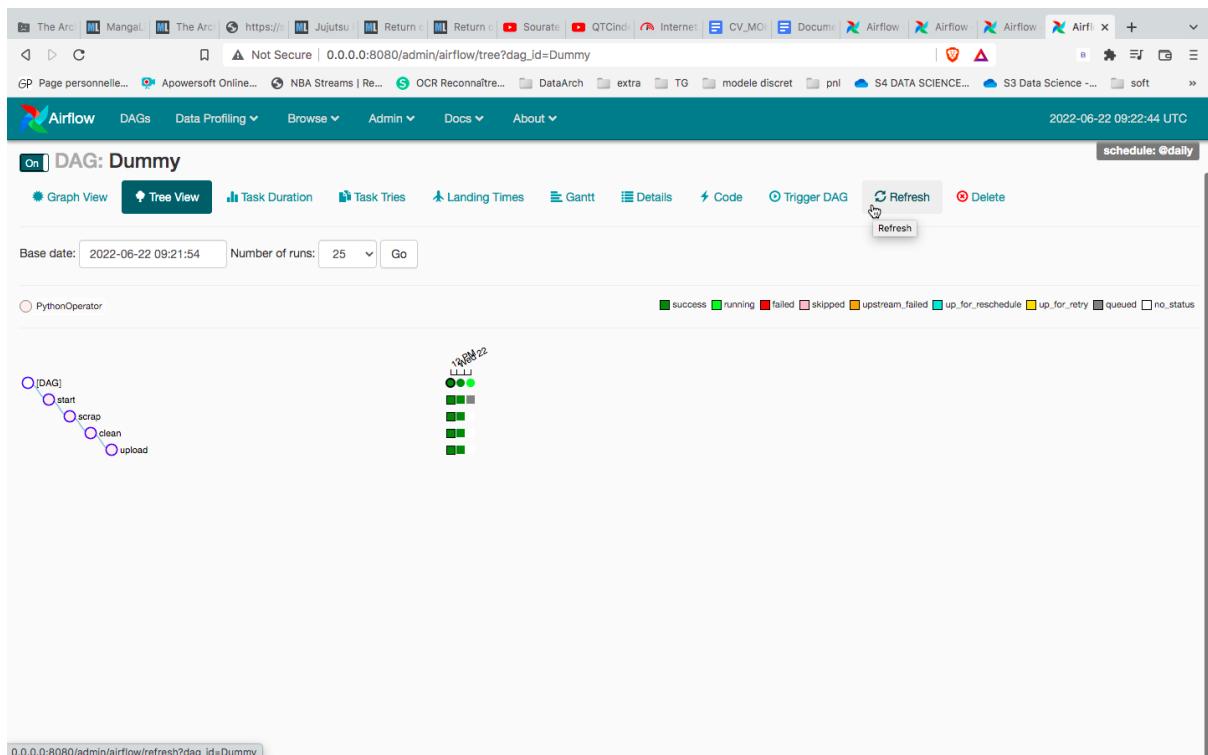
Airflow Dags Data Profiling Browse Admin Docs About 2022-04-10 15:12:15 UTC

Triggered my_dag, it should start any moment now. X

DAGs

Search:

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
		my_dag	@daily	airflow	<img alt="Task 120			



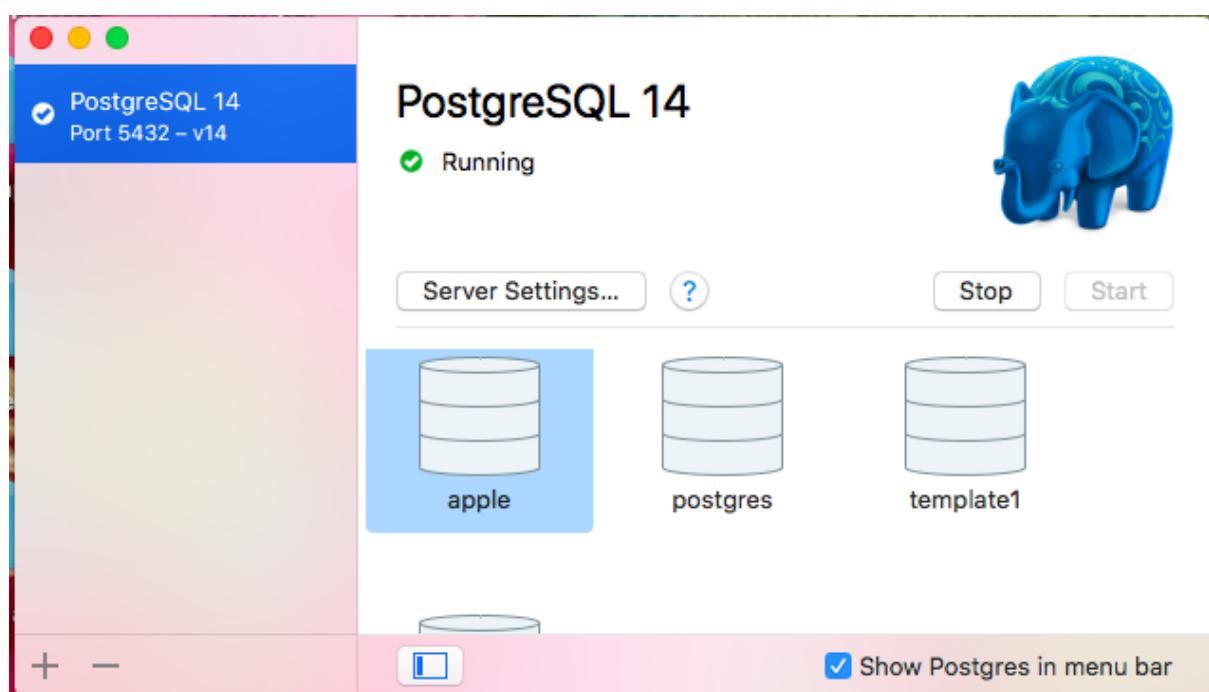
Database :

- Postgres :



PostgreSQL est un système de gestion de base de données relationnelle orienté objet puissant et open source qui est capable de prendre en charge en toute sécurité les charges de travail de données les plus complexes. Alors que MySQL donne la priorité à l'évolutivité et aux performances, Postgres donne la priorité à la conformité et à l'extensibilité SQL. Les entreprises qui souhaitent maintenir un haut niveau d'intégrité

et de personnalisation de leurs données choisissent généralement Postgres en raison de sa fiabilité, l'intégrité de ses données, la robustesse de ses fonctionnalités, et parce qu'il fournit des solutions toujours performantes et innovantes. PostgreSQL fonctionne sur tous les principaux systèmes d'exploitation et est conforme à ACID depuis 2001. Postgres peut être téléchargé gratuitement et déployé sur du matériel standard, ou peut être exécuté dans le Cloud par le biais d'une variété de fournisseurs. Bien que Postgres soit riche en fonctionnalités et adapté aux charges de travail OLAP, les performances de Postgres ont tendance à atteindre une limite lorsque les volumes de données dépassent plusieurs téraoctets.



```

apple — more - psql -p5432 postgres — 96x26
(base) MacBook-Air-van-Apple:~ apple$ /Applications/Postgres.app/Contents/Versions/14/bin/psql -p5432 "postgres"
psql (14.2)
Type "help" for help.

[postgres=# \list
               List of databases
   Name | Owner | Encoding | Collate | Ctype | Access privileges
---+-----+-----+-----+-----+-----+
apple | apple | UTF8 | en_US.UTF-8 | en_US.UTF-8 |
postgres | postgres | UTF8 | en_US.UTF-8 | en_US.UTF-8 |
template0 | postgres | UTF8 | en_US.UTF-8 | en_US.UTF-8 | =c/postgres      +
| postgres=CTc/postgres
template1 | postgres | UTF8 | en_US.UTF-8 | en_US.UTF-8 | =c/postgres      +
| postgres=CTc/postgres
testdata | user1 | UTF8 | en_US.UTF-8 | en_US.UTF-8 |
(5 rows)

(END)
               List of databases
   Name | Owner | Encoding | Collate | Ctype | Access privileges
---+-----+-----+-----+-----+
apple | apple | UTF8 | en_US.UTF-8 | en_US.UTF-8 |
postgres | postgres | UTF8 | en_US.UTF-8 | en_US.UTF-8 |
template0 | postgres | UTF8 | en_US.UTF-8 | en_US.UTF-8 | =c/postgres      +
| postgres=CTc/postgres

```

- Pgadmin :

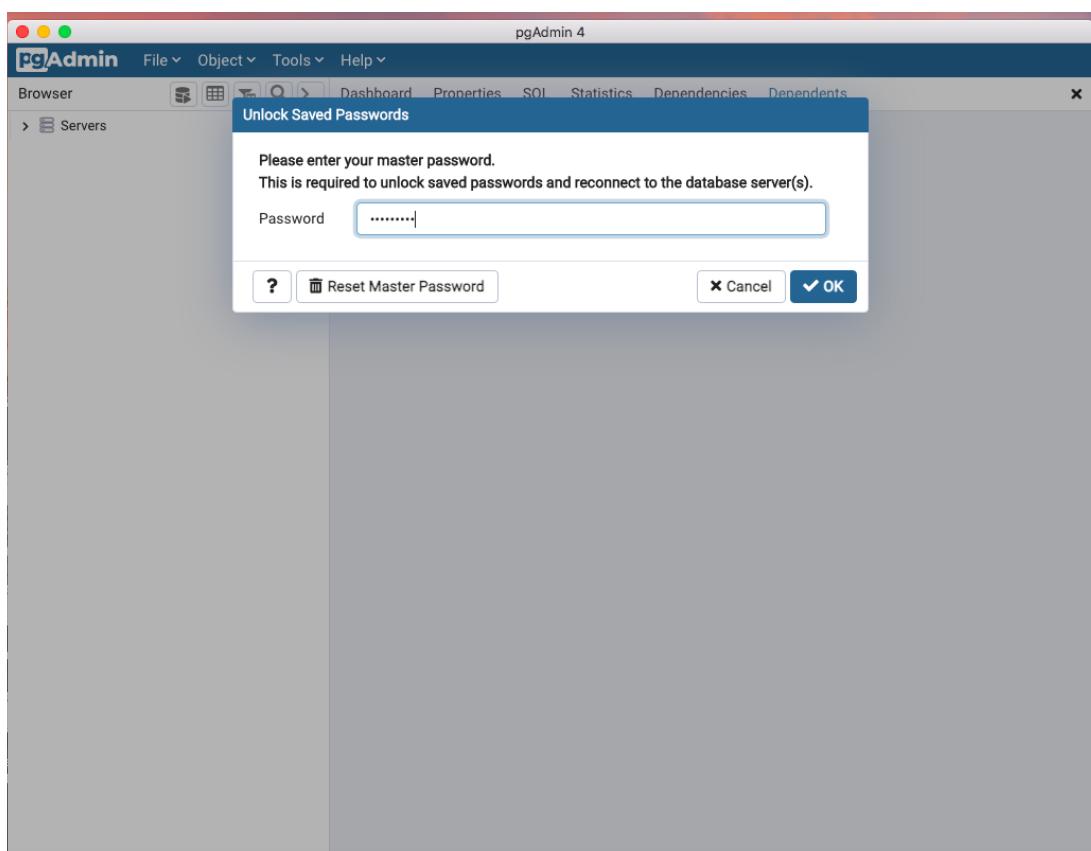
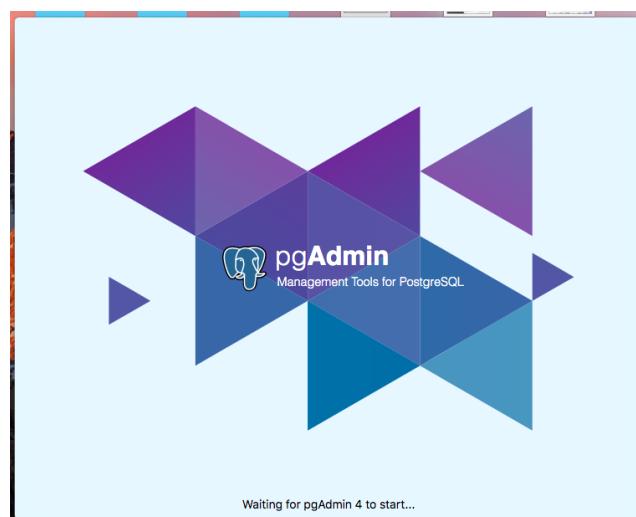


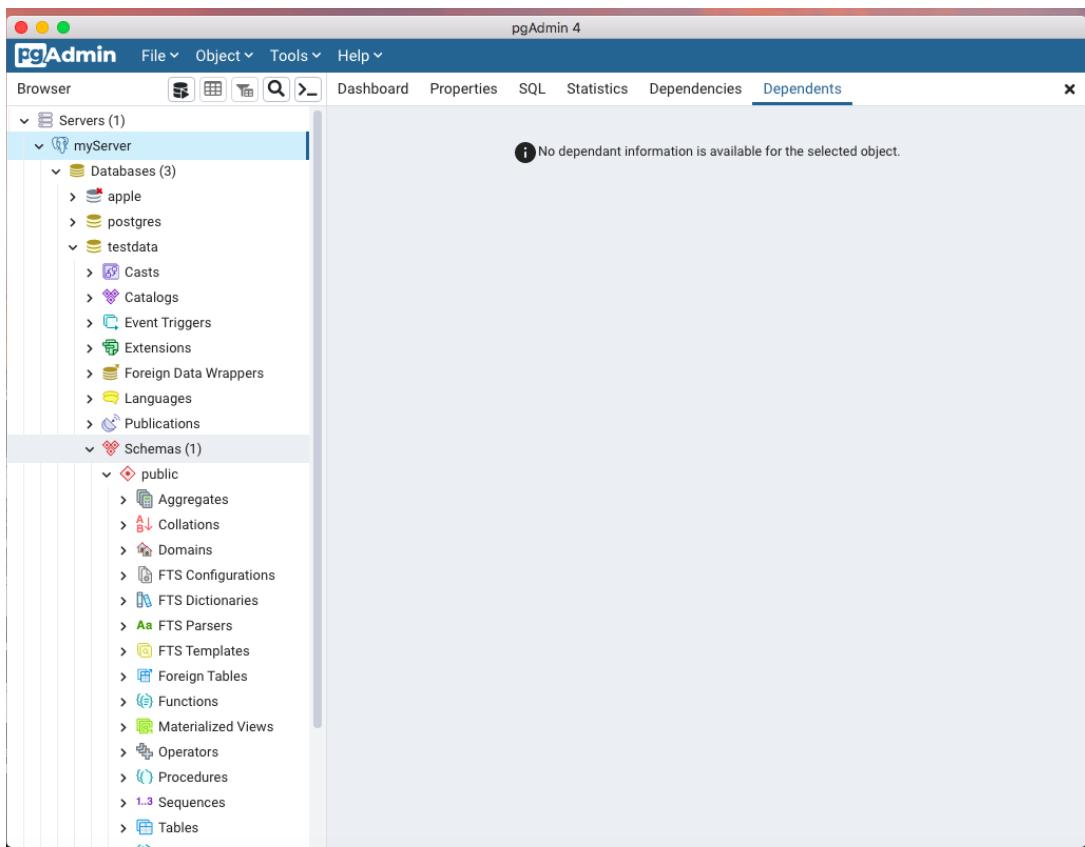
pgAdmin est un outil d'administration graphique pour PostgreSQL distribué selon les termes de la licence PostgreSQL. Il peut être utilisé sur les plateformes Linux, FreeBSD, Solaris, Mac OS X et Windows pour gérer les différentes moutures de PostgreSQL à partir de la version 7.3, mais également les versions commerciales et dérivées de PostgreSQL telles que Postgres Plus Advanced Server et Greenplum Database.

pgAdmin est conçu pour répondre aux besoins des utilisateurs, de l'écriture de requêtes SQL simples au développement de bases de données complexes. L'interface graphique supporte toutes les fonctionnalités de PostgreSQL et facilite l'administration.

Après plus de 10 000 heures de développement, l'équipe de développement de pgAdmin annonce la disponibilité de pgAdmin 4 v1 bêta 1. pgAdmin 4 est en fait une réécriture complète de pgAdmin (la version actuelle étant pgAdmin 3 v1.22.1), ce qui justifie tout le temps qui a été investi dans cette nouvelle version.

pgAdmin 4 peut être installé en mode application web (écrit en Python et JavaScript / jQuery) ou desktop (en Python et C++ à travers Qt).





- Connection avec jupyter :

on doit stocker notre data frame dans une database pour le visualiser
on déclare les configuration de notre Database server

```
In [11]: #Some imports for work ...
import pandas as pd
import pandas.io.sql as sqlio
import psycopg2 as ps#with postgresql database(to connect!)

In [12]: dbname="postgres"
user="postgres"
password="123456789"
host="localhost"
port="5432"
```

vérifier si la connexion a été établie

```
In [13]: def connect_to_db(host, dbname, port, user, password):
    try:
        conn = ps.connect(host=host, database=dbname, user=user, password=password, port=port)

    except ps.OperationalError as e:
        raise e
    else:
        print('Connected!')
        return conn

In [14]: con=connect_to_db(host, dbname, port, user, password)
Connected!
```

par suite on va charger notre data frame

- on commence par la création du base de donnée

```
In [6]: create_table_command = ("""CREATE TABLE IF NOT EXISTS words (
    word TEXT NOT NULL,
    count INTEGER NOT NULL
)""")

curr.execute(create_table_command)
con.commit()

In [70]: create_table_command = ("""CREATE TABLE IF NOT EXISTS bigram (
    bigram TEXT NOT NULL,
    count INTEGER NOT NULL
)""")

curr.execute(create_table_command)
con.commit()

In [67]: create_table_command = ("""CREATE TABLE IF NOT EXISTS tf (
    word TEXT NOT NULL,
    count numeric NOT NULL,
    index numeric NOT NULL
)""")

curr.execute(create_table_command)
con.commit()
```

- charger notre data frame

```
In [72]: for i in range(len(list_count)):  
    insert = """INSERT INTO bigram (bigram, count) VALUES(%s,%s);"""  
    row_to_insert = (list_words[i], list_count[i])  
    curr.execute(insert, row_to_insert)  
    curr = con.cursor()  
    con.commit()
```

```
In [69]: for i in range(len(list_count)):  
    insert = """INSERT INTO tf (word,count, index ) VALUES(%s,%s,%s);"""  
    row_to_insert = (list_words[i], list_count[i],list_index[i])  
    curr.execute(insert, row_to_insert)  
    curr = con.cursor()  
    con.commit()
```

```
In [10]: for i in range(len(list_count)):  
    insert = """INSERT INTO words (word, count) VALUES(%s,%s);"""  
    row_to_insert = (list_words[i], list_count[i])  
    curr.execute(insert, row_to_insert)  
    curr = con.cursor()  
    con.commit()
```

The screenshot shows a database query editor interface. At the top, there is a toolbar with various icons for file operations like download, save, search, and filter, followed by a "No limit" button. Below the toolbar, there are two tabs: "Query Editor" (which is selected) and "Query History".

In the main area, a query is displayed:

```
1 select * from bigram;
```

Below the query, there are four tabs: "Data Output" (selected), "Explain", "Messages", and "Notifications".

The "Data Output" tab displays a table with the following data:

	bigram	count
1	suit aprè	31
2	aprè cett	31
3	cett publicité	31
4	", a	12
5	voir également	10
6	également huffpost	10
7	tizi ouzou	10
8	' algéri	9
9	' extrêm	8
10	extrêm droit	8
11	"'	7
12	' alger	7
13	deux canadair	7
14	iinsuu'	6

The screenshot shows a PostgreSQL Query Editor interface. At the top, there is a toolbar with various icons for file operations like download, search, and filter. Below the toolbar, the title bar displays "Query Editor" and "Query History". The main area contains a query editor window with the following SQL command:

```
1 select * from words;
```

Below the query editor, there are tabs for "Data Output", "Explain", "Messages", and "Notifications". The "Data Output" tab is selected, displaying a table with two columns: "word" (text) and "count" (integer). The data is as follows:

	word	count
	text	integer
1	a	75
2	aprè	43
3	cett	40
4	suit	33
5	publicité	31
6	",	27
7	incendi	26
8	"	24
9	plu	21
10	fait	19
11	selon	19
12	algéri	19
13	deux	18
14	"	13

Screenshot of a PostgreSQL database management interface showing a query result.

The interface includes a top navigation bar with links: Dashboard, Properties, SQL, Statistics, Dependencies, Dependents, and a connection status: postgres/postgres@PostgreSQL 14*.

The toolbar below the navigation bar contains various icons for database operations like creating tables, files, and indexes; searching; and managing data.

The main area shows two tabs: Query Editor (selected) and Query History.

A query has been run:

```
1 select * from tf;
```

The results are displayed in a table under the Data Output tab:

	word	count	index
	text	numeric	numeric
1	maroc	3.662587827025453	1014
2	la	3.374905754573672	931
3	police	4.0680529351336165	1240
4	marocaine	4.0680529351336165	1016
5	ouvert	3.151762203259462	1160
6	ce	3.374905754573672	339
7	samedi	4.0680529351336165	1477
8	15	4.0680529351336165	9
9	janvier	4.0680529351336165	902
10	une	3.662587827025453	1664
11	enquete	3.662587827025453	645
12	sur	4.0680529351336165	1566
13	mort	4.0680529351336165	1086
14	francaise	4.0680529351336165	764

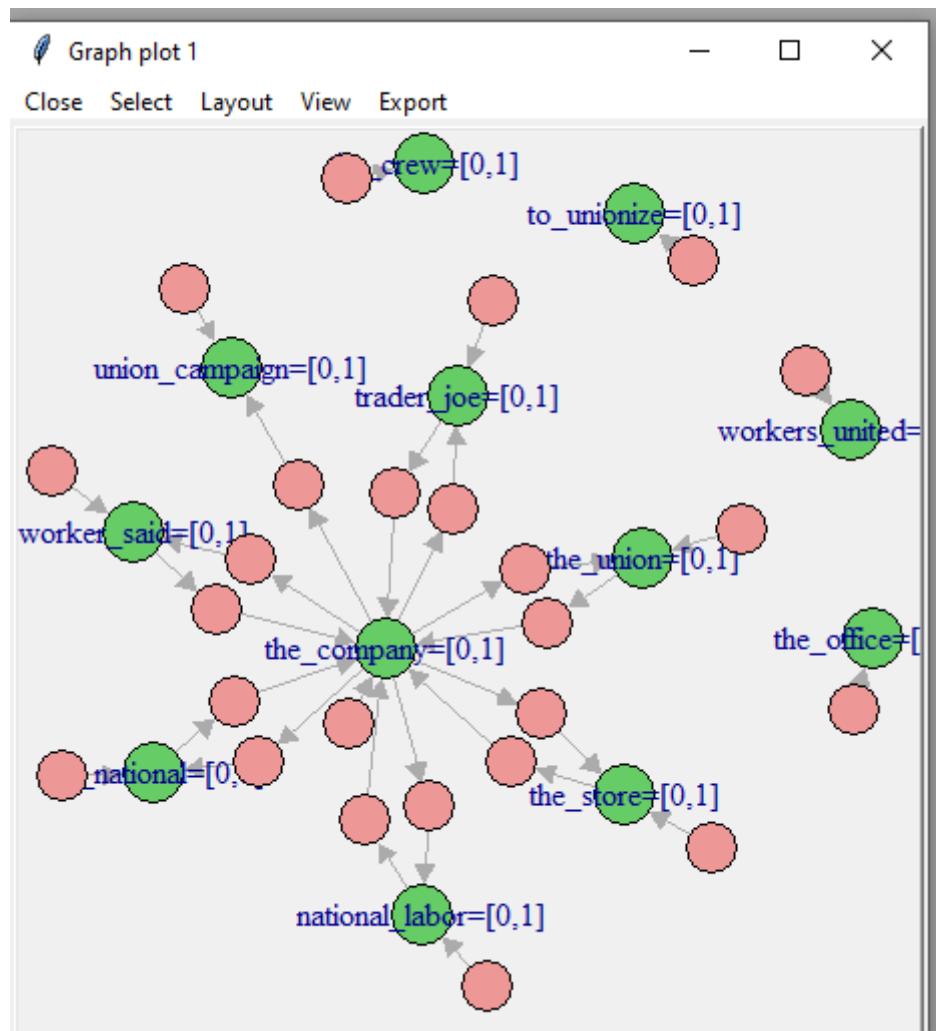
- Identification des Bigrammes pertinents qui apparaissent au même temps dans un article

	support	itemsets	length
0	0.089514	(the_company)	1
1	0.084399	(trader_joe)	1
2	0.007673	(worker_â€™)	1
3	0.010230	(worker_said)	1
4	0.007673	(the_national)	1
5	0.007673	(national_labor)	1
6	0.007673	(relations_board)	1
7	0.010230	(workers_are)	1
8	0.023018	(the_union)	1
9	0.010230	(that_musk)	1
10	0.020460	(the_store)	1
11	0.020460	(according_to)	1
12	0.017903	(have_been)	1
13	0.010230	(for_comment)	1
14	0.020460	(company_â€™)	1
15	0.020460	(activision_blizzard)	1
16	0.017903	(the_nrb)	1
17	0.012788	(union_campaign)	1
18	0.012788	(against_the)	1
19	0.012788	(when_they)	1

- Identification des règles pertinentes

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedent conviction
1	(workers_are)	(the_company)	0.010230	0.089514	0.002558	0.250000	2.792857	0.001642	1.213981	
2	(the_company)	(the_union)	0.089514	0.023018	0.007673	0.085714	3.723810	0.005612	1.068574	
3	(the_union)	(the_company)	0.023018	0.089514	0.007673	0.333333	3.723810	0.005612	1.365729	
4	(the_company)	(that_musk)	0.089514	0.010230	0.002558	0.028571	2.792857	0.001642	1.018881	
5	(that_musk)	(the_company)	0.010230	0.089514	0.002558	0.250000	2.792857	0.001642	1.213981	
6	(the_company)	(the_store)	0.089514	0.020460	0.002558	0.028571	1.396429	0.000726	1.008350	
7	(the_store)	(the_company)	0.020460	0.089514	0.002558	0.125000	1.396429	0.000726	1.040555	
8	(the_company)	(according_to)	0.089514	0.020460	0.005115	0.057143	2.792857	0.003284	1.038906	
9	(according_to)	(the_company)	0.020460	0.089514	0.005115	0.250000	2.792857	0.003284	1.213981	
10	(the_company)	(have_been)	0.089514	0.017903	0.002558	0.028571	1.595918	0.000955	1.010982	
11	(have_been)	(the_company)	0.017903	0.089514	0.002558	0.142857	1.595918	0.000955	1.062234	
12	(the_company)	(company_â€™)	0.089514	0.020460	0.020460	0.228571	11.171429	0.018629	1.269774	
13	(company_â€™)	(the_company)	0.020460	0.089514	0.020460	1.000000	11.171429	0.018629	inf	
14	(the_company)	(activision_blizzard)	0.089514	0.020460	0.005115	0.057143	2.792857	0.003284	1.038906	
15	(activision_blizzard)	(the_company)	0.020460	0.089514	0.005115	0.250000	2.792857	0.003284	1.213981	
16	(the_company)	(the_nlrb)	0.089514	0.017903	0.002558	0.028571	1.595918	0.000955	1.010982	
17	(the_nlrb)	(the_company)	0.017903	0.089514	0.002558	0.142857	1.595918	0.000955	1.062234	
18	(the_company)	(union_campaign)	0.089514	0.012788	0.005115	0.057143	4.468571	0.003970	1.047043	
19	(union_campaign)	(the_company)	0.012788	0.089514	0.005115	0.400000	4.468571	0.003970	1.517477	

- Implémentation d'un graphe orienté de corrélation entre les bigrammes



Pour visualiser ces données il faut connecter notre base de données par notre plateforme de power BI :

Power bi

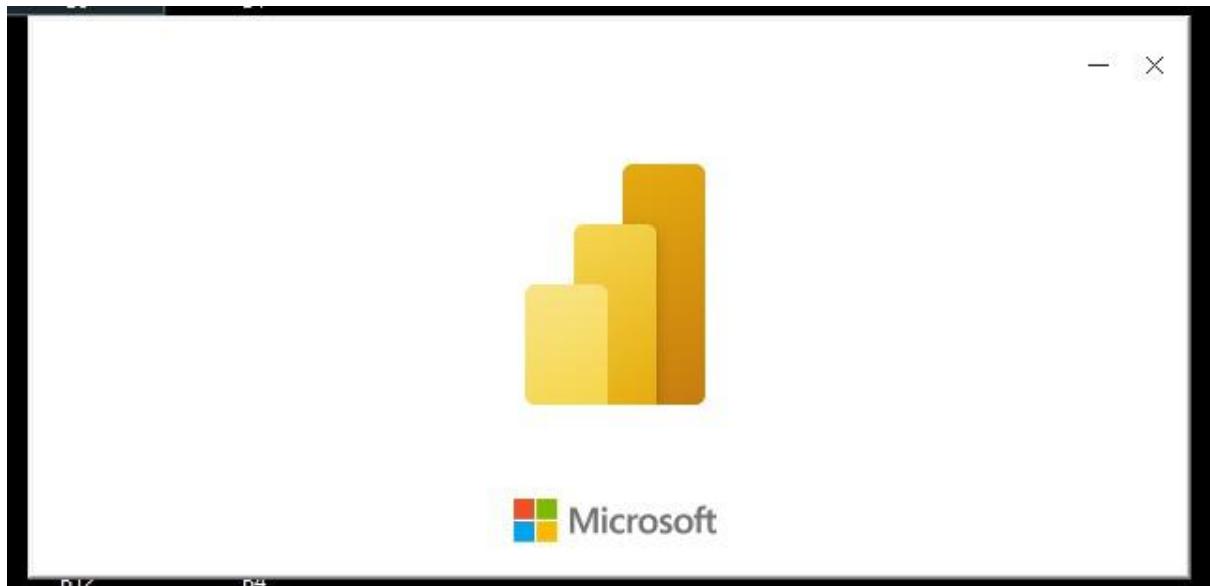


Mieux qu'une simple solution de Data Visualization, il permet aux Data Analysts chargés de fournir des rapports et des analyses à leurs entreprises d'augmenter leur productivité et leur créativité. L'outil Power BI Desktop permet notamment à tout moment de combiner des données en provenance de nombreuses sources sur site ou sur le Cloud : bases de données, fichiers, services web. En utilisant l'outil Gateways, il est possible de se connecter aux bases de données SQL. Cela concerne aussi les modèles Microsoft Analysis Services et bien d'autres sources de données vers un même tableau de bord.

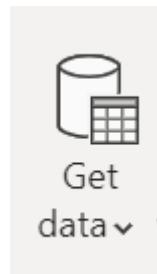
Des d'outils visuels facilitent la compréhension des données. Ils améliorent automatiquement leur qualité (Data Qualité) et de résoudre les éventuels problèmes de formats. De base, Microsoft a intégré plus d'une vingtaine de visuels. Grâce à une communauté d'utilisateurs très impliqués, on compte aussi un grand nombre de Dataviz customisées, permettant de créer des rapports encore plus efficaces.

Les tableaux de bord personnalisés offrent une vue à 360 degrés, se mettent à jour en temps réel, et permettent à tous les utilisateurs (même sans compétences techniques) de comprendre et d'exploiter les données de l'entreprise. Un seul clic suffit pour explorer les données à l'aide d'outils intuitifs à la portée du plus grand nombre et d'en dégager des informations exploitables.

- Visualisation de données et tableau de bord



On a connecter par ODBC méthode :





Common data sources

Excel workbook

Power BI datasets

Power BI dataflows

Dataverse

SQL Server

Analysis Services

Text/CSV

Web

OData feed

Blank query

Power BI Template Apps

More...

Get Data

X

Search

All

File

Database

Power Platform

Azure

Online Services

Other

Database

-  SQL Server Analysis Services database
-  Oracle database
-  IBM Db2 database
-  IBM Informix database (Beta)
-  IBM Netezza
-  MySQL database
-  PostgreSQL database
-  Sybase database
Import data from a PostgreSQL database.
-  Teradata database
-  SAP HANA database
-  SAP Business Warehouse Application Server
-  SAP Business Warehouse Message Server
-  Amazon Redshift
-  Impala
-  Google BigQuery
-  Vertica

Certified Connectors

Template Apps

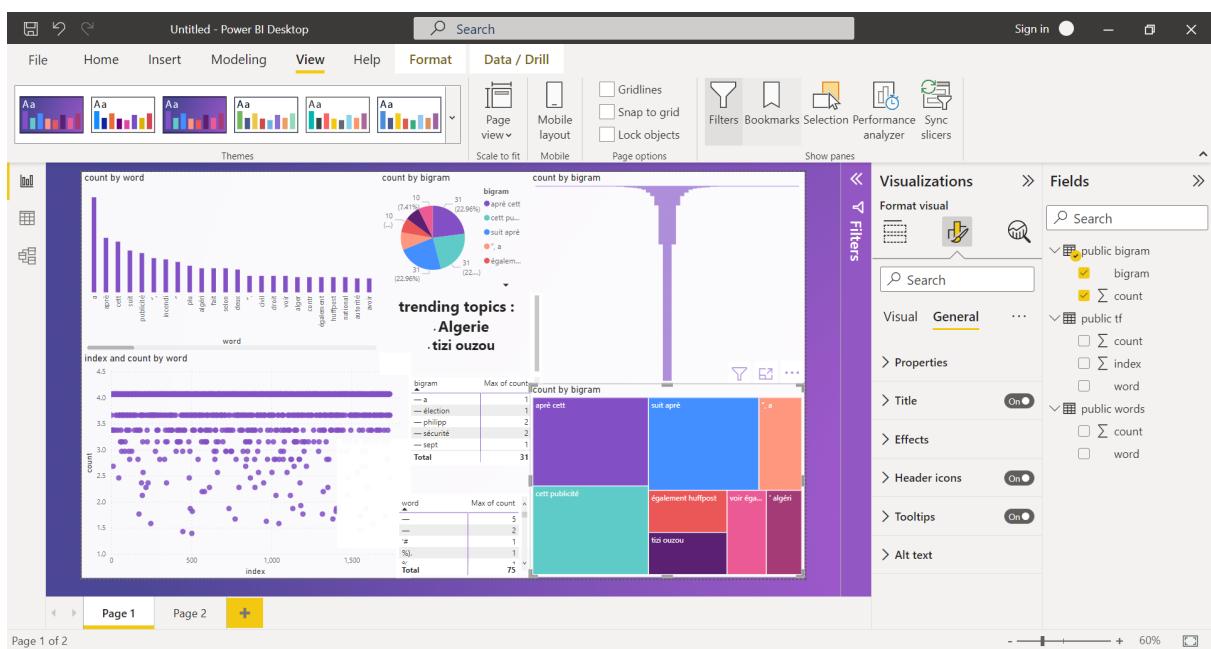
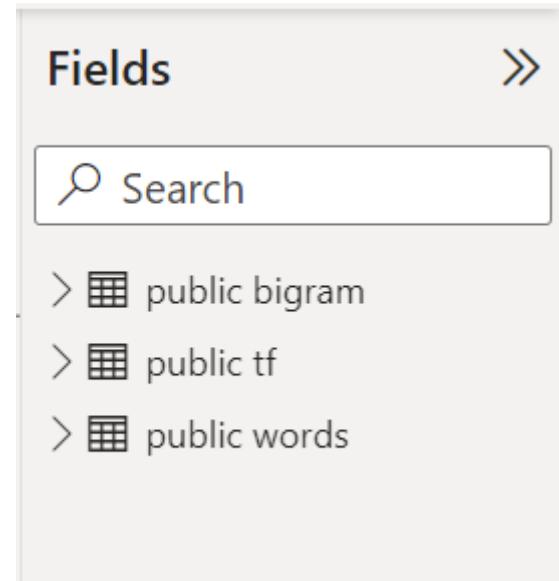
Connect

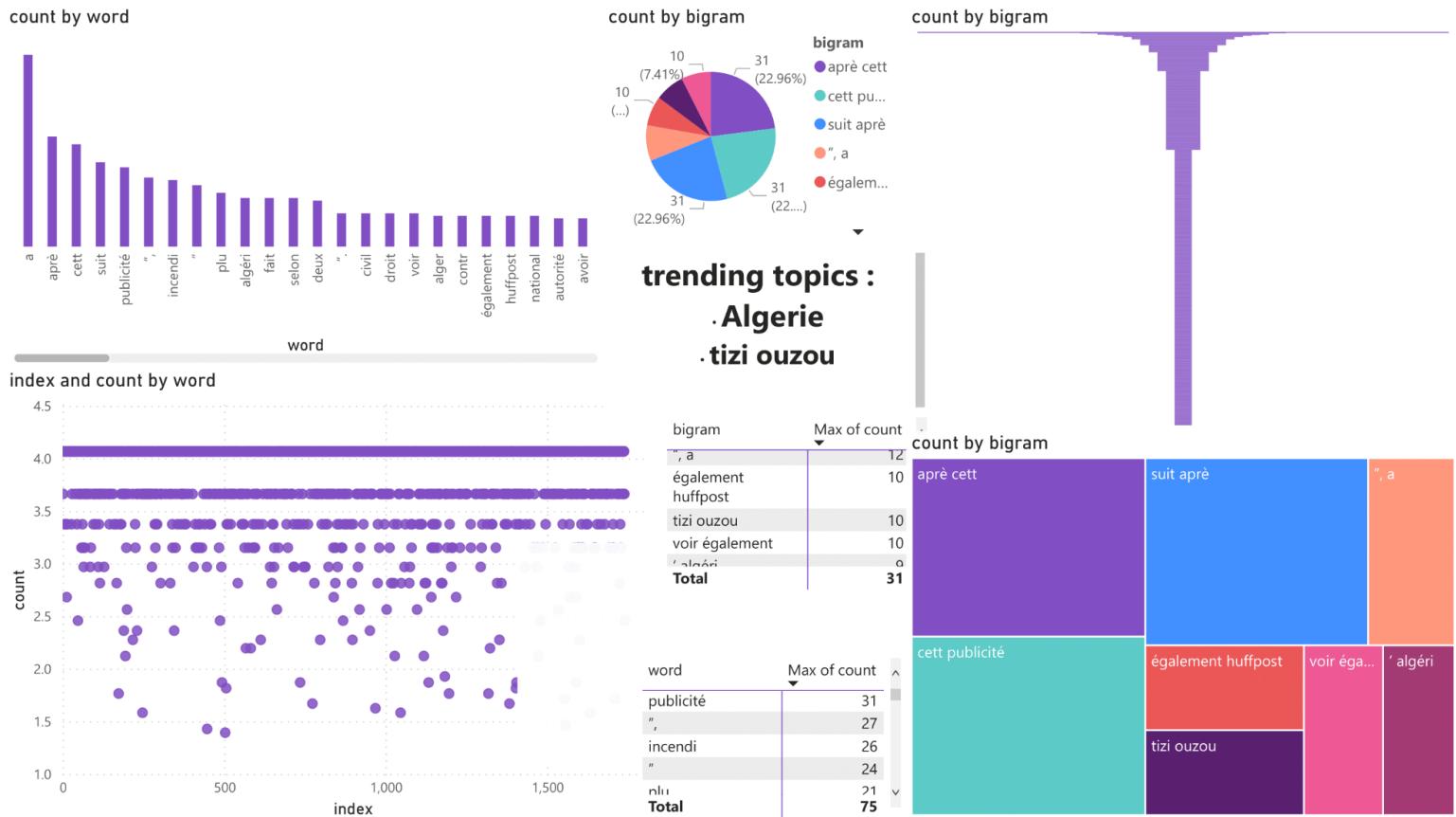
Cancel



The Navigator window title is "Navigator". It features a search bar and a "Display Options" dropdown. The left pane shows a tree view of the database structure under "localhost:5432: postgres [5]". The right pane displays the message "No items selected for preview". At the bottom are buttons for "Select Related Tables", "Load", "Transform Data", and "Cancel".

- localhost:5432: postgres [5]
 - public.bigram
 - public.scraped
 - public.screaped
 - public.tf
 - public.words





à partir de ce tableau de bord on peut observer plusieurs informations importantes à la fois, par exemple on voit que "trending topics" dans le secteur d'économie est l'Algérie et Tizi Ouzou .