Le Monde Covid

Analyse de documents textuels grâce à l'intelligence artificielle

Topic Modeling

François LABASTIE

Décembre 2020

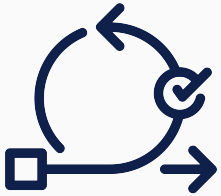
Titre professionnel "Développeur en intelligence artificielle" de niveau 6 enregistré au RNCP sous le n°34757

Passage par la voie de la formation - parcours de 19 mois achevé le 20 octobre 2020

Concept du projet & méthodologie

Topic modeling · Modélisation de thèmes
à partir de documents du média lemonde.fr

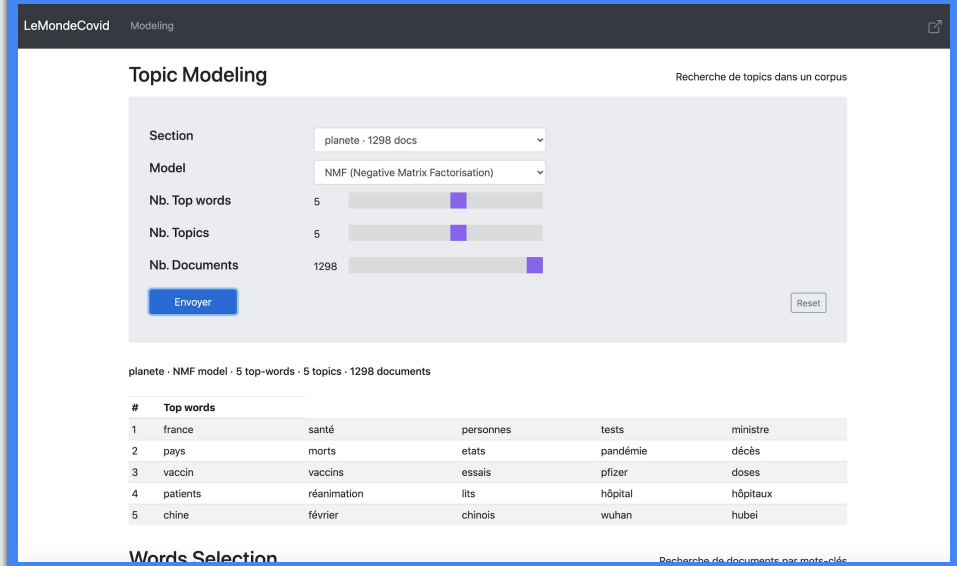
- Scraping de la page “recherche” du Monde
- Méthodologie → SCRUM & UML

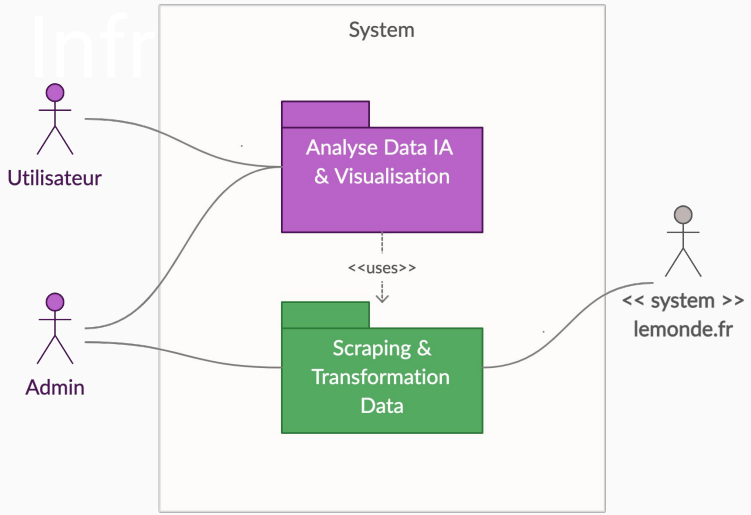


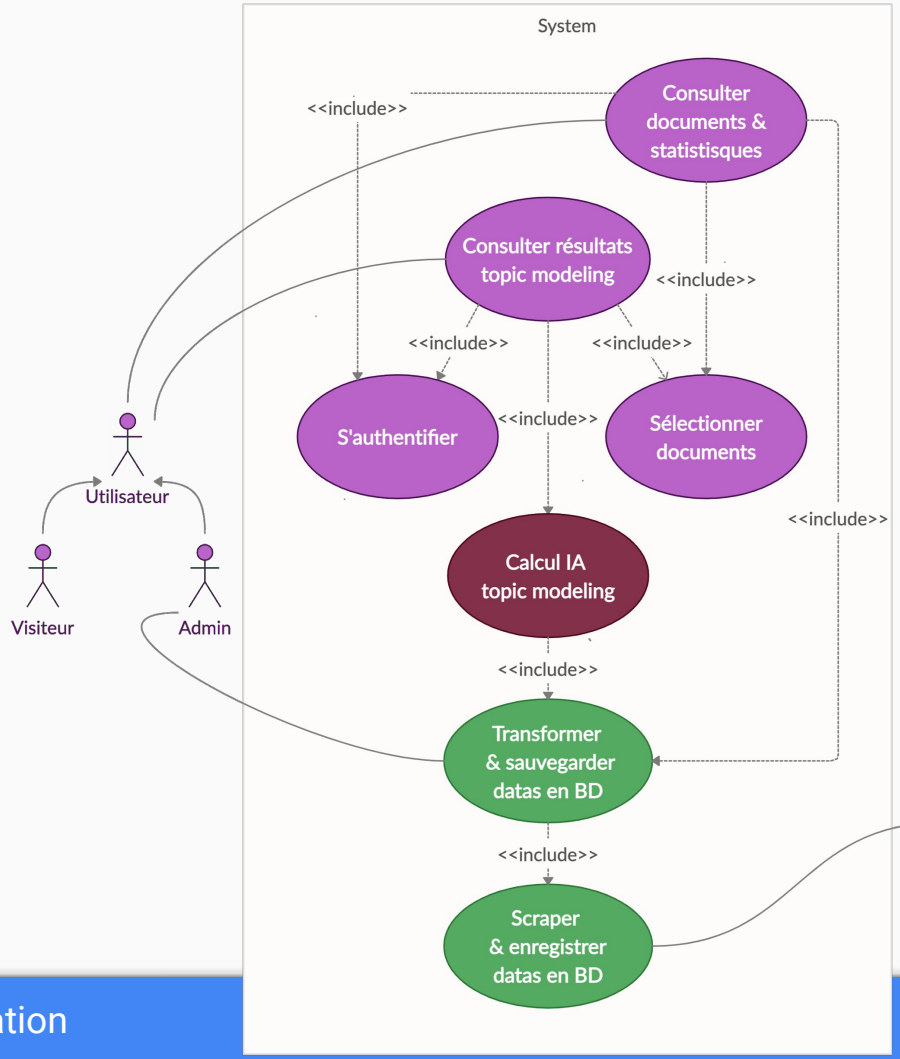
DEMO

Du point de vue **utilisateur**

...

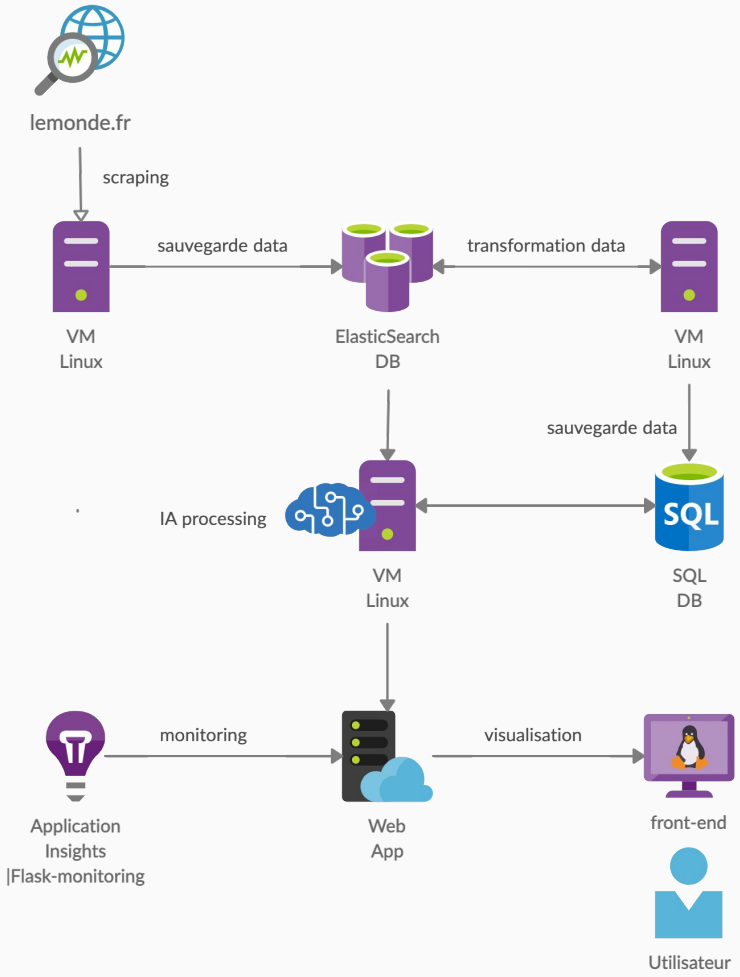



System
Utilisateur
Admin
Analyse Data IA
& Visualisation
<<uses>>
Scraping &
Transformation
Data
<< system >>
lemonde.fr


System
Consulter
documents &
statistisques
<<include>>
Consulter résultats
topic modeling
<<include>>
<<include>>
<<include>>
S'authentifier
<<include>>
Sélectionner
documents
<<include>>
Calcul IA
topic modeling
<<include>>
Transformer
& sauvegarder
datas en BD
<<include>>
Scraper
& enregistrer
datas en BD
Utilisateur
Visiteur
Admin
<< system >>
lemonde.fr

Infrastructure

- Scraping → scrapy
- Elasticsearch
- Azure SQL database
- VM sur Azure ML Studio
- Azure Web Apps
- Flask + bootstrap
- Flask-monitoring



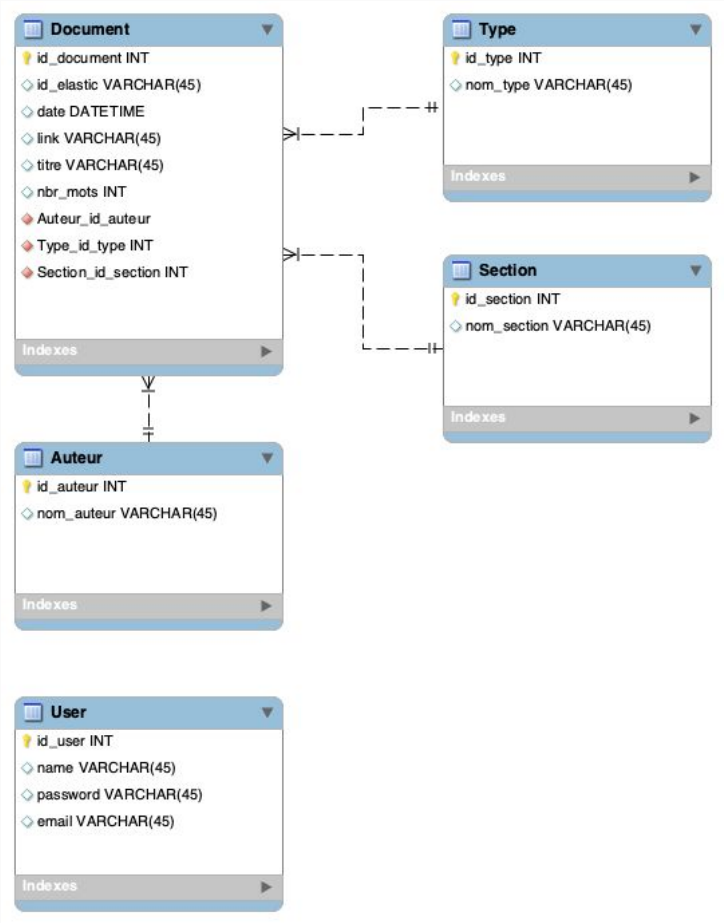
Recueil et traitement des données

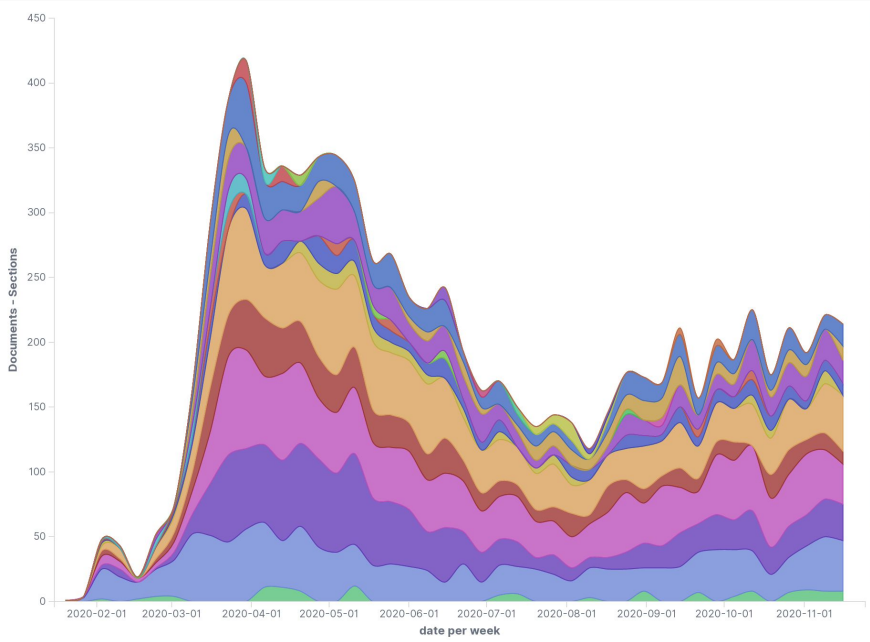
- Scraping
- Insertion data → bd Elasticsearch (tokenization, stopwords, etc.)
- Insertion data → bd Azure SQL database

- package Analyse → https://github.com/flabastie/news-analysis
- package Scraping → https://github.com/flabastie/news-analysis-data

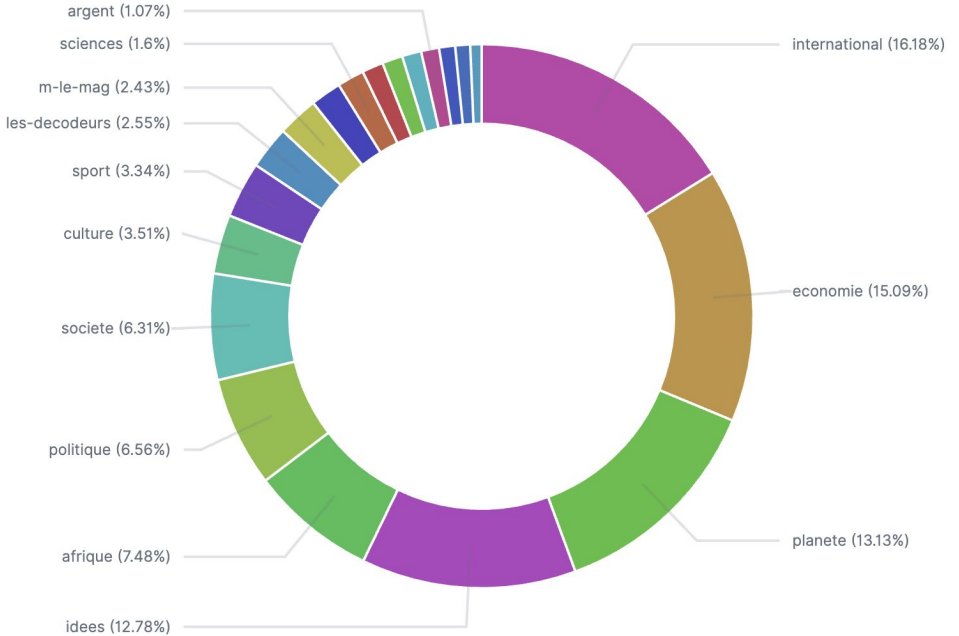
Dev tools Elastic → GET /news_analysis/_mapping


```
{
  "news_analysis" : {
    "mappings" : {
      "properties" : {
        "author" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
        "content_all" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
        "content_html" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
```


Documents - Sections
450
400
350
300
250
200
150
100
50
0
2020-02-01
2020-03-01
2020-04-01
2020-05-01
2020-06-01
2020-07-01
2020-08-01
2020-09-01
2020-10-01
2020-11-01
date per week


les-decode
planete
idees
international
afrique
economie
m-le-mag
culture
pixels
sante
campus
politique
sciences
education
sport
societe
m-perso
disparitions
argent
smart-cities
livres
m-styles
series-d-eti
football
emploi


argent (1.07%)
sciences (1.6%)
m-le-mag (2.43%)
les-decodeurs (2.55%)
sport (3.34%)
culture (3.51%)
societe (6.31%)
politique (6.56%)
afrique (7.48%)
idees (12.78%)
international (16.18%)
economie (15.09%)
planete (13.13%)

Code Scraping

- scrapy_projects
- transform_projects

- lemonde_covid/spiders/lm_covid.py
- lemonde_covid/pipelines.py

- transform_projects/LeMonde_stop_words.ipynb
- transform_projects/elastic-to-sql.py

Code IA

- TopicModelingLDA
- TopicModelingNMF

- **project/main.py**
- **project/queries/selection.py**
- **project/processing/modeling.py**

Model Evaluation

- *Subjectivité ...*
- Topic coherence
- Coherence score

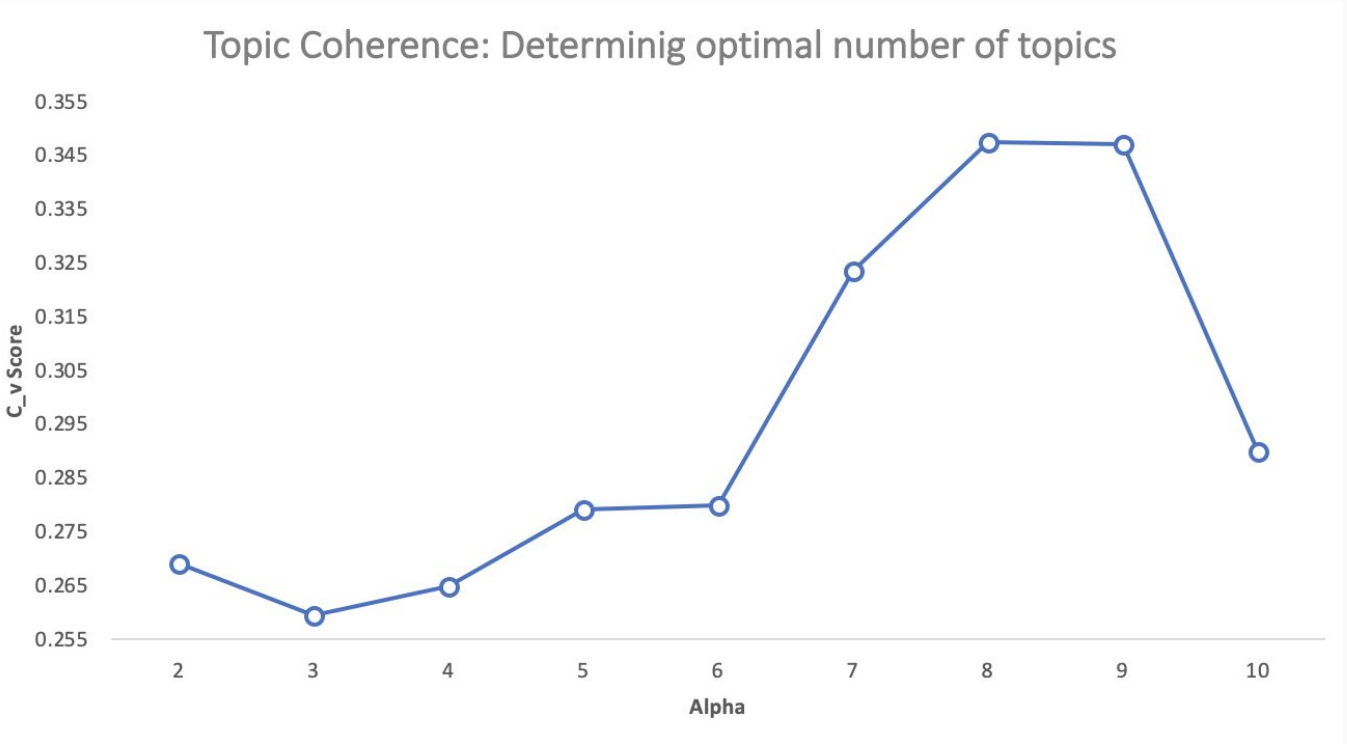
Topic coherence

Attribuent un score à chaque topic en mesurant le degré de similitude sémantique entre les mots les mieux notés du topic.

Mesures qui aident à distinguer entre les topics qui sont des sujets sémantiquement interprétables et les topics qui sont des artefacts d'inférence statistique.

Coherence score C_v

La mesure C_v est basée sur une fenêtre glissante, une segmentation en un seul ensemble des premiers mots et une mesure de confirmation indirecte qui utilise des informations mutuelles ponctuelles normalisées (NPMI) et la similitude cosinus.



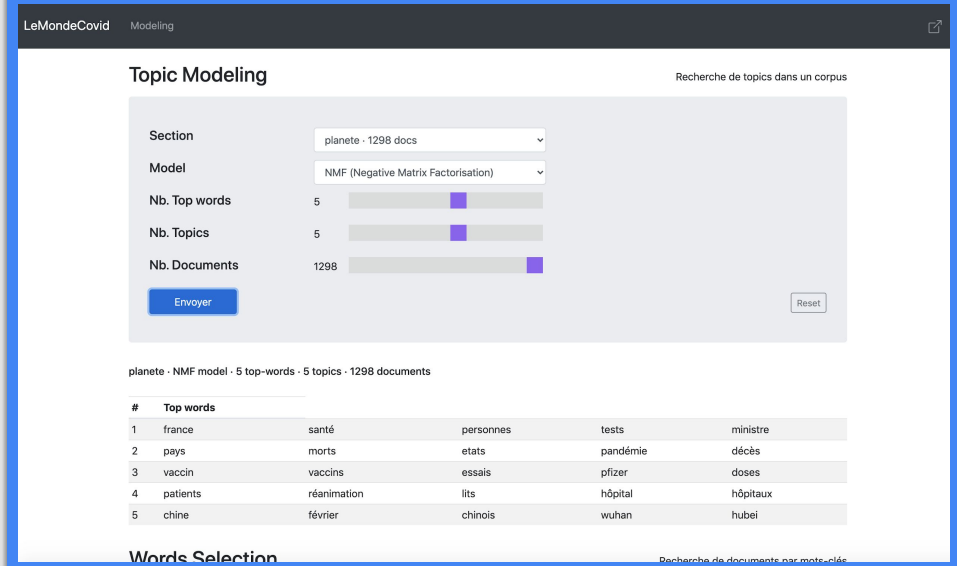
Exemple de scoring avec gensim - C_vScore sur nombre optimal de topics

DEMO

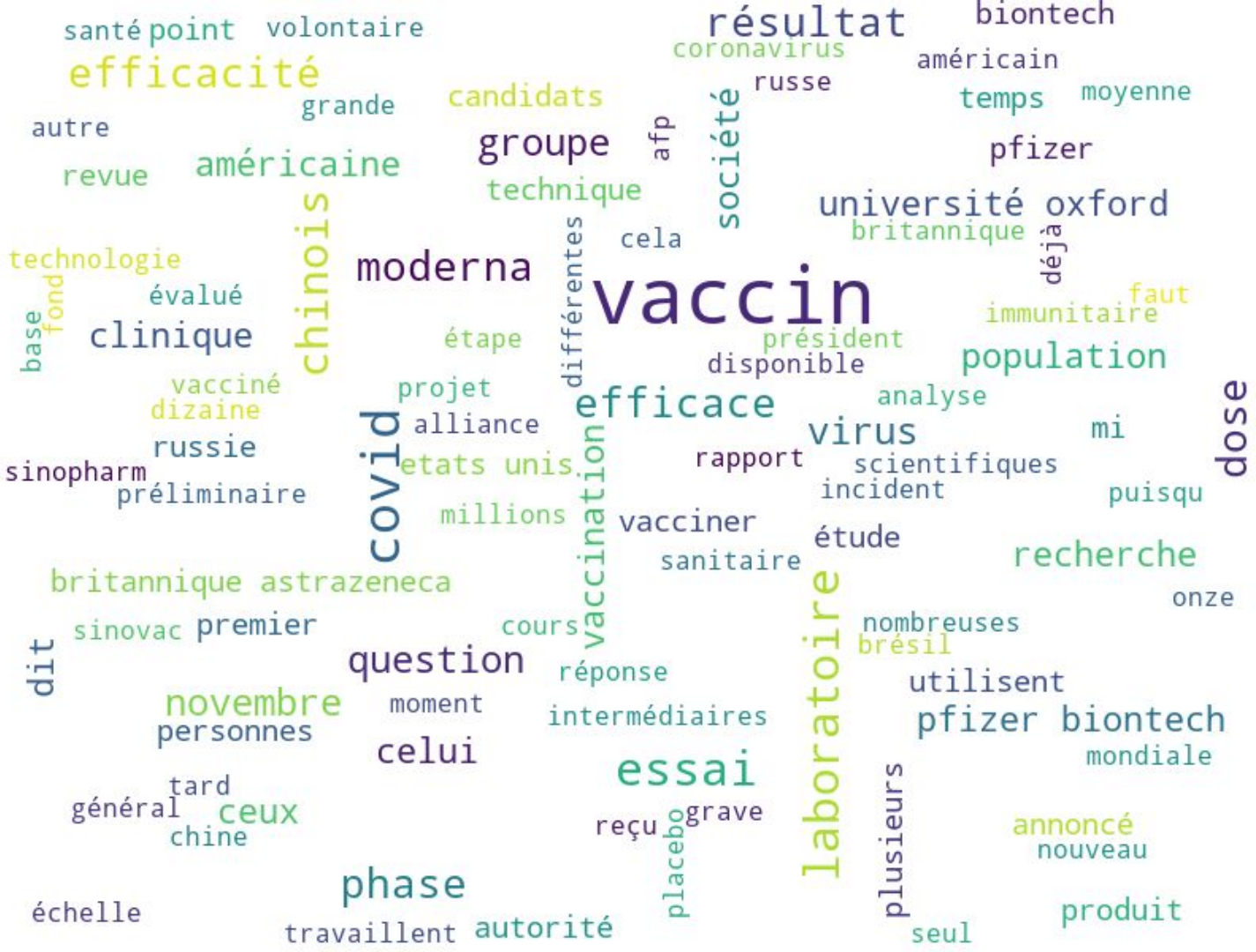
Du point de vue **développeur**

...

\+ monitoring



Merci !


résultat
biontech
santé
point
volontaire
coronavirus
américain
efficacité
russe
temps
moyenne
grande
candidats
autre
groupe
afp
société
pfizer
revue
américaine
technique
université oxford
britannique
déjà
chinois
cela
technologie
moderna
vaccin
différentes
fond
évalué
faut
immunitaire
base
clinique
étape
président
population
vacciné
projet
disponible
dizaine
efficace
analyse
covid
alliance
virus
mi
dose
russie
vaccination
rapport
scientifiques
sinopharm
etats unis
incident
puisqu
préliminaire
millions
vacciner
étude
recherche
britannique astrazeneca
sanitaire
onze
sinovac
premier
cours
nombreuses
brésil
dit
question
réponse
utilisent
novembre
moment
intermédiaires
pfizer biontech
personnes
laboratoire
mondiale
celui
essai
tard
général
ceux
reçu
grave
annoncé
chine
placebo
plusieurs
nouveau
phase
produit
échelle
travaillent
autorité
seul