
Dat'Art
Topic Modeling
NLP

# Sommaire


ATTENTIE
NIET AANRAKEN !
SV ONGEDIERTEBESTRIJDING
CARDS CORP
ULTRA SOUTH ON TOUR
SOTL
BANGKOK SOI DOGS
RUES NEST
VIDEO MARC DORCEL
SAVAGE
LAYING THE GRAVEL FOR YOU TO TRAVEL
REM JDI
NOK
There?
NOK
@plantpotheads
mauvais enfants
ANVA
ANVA

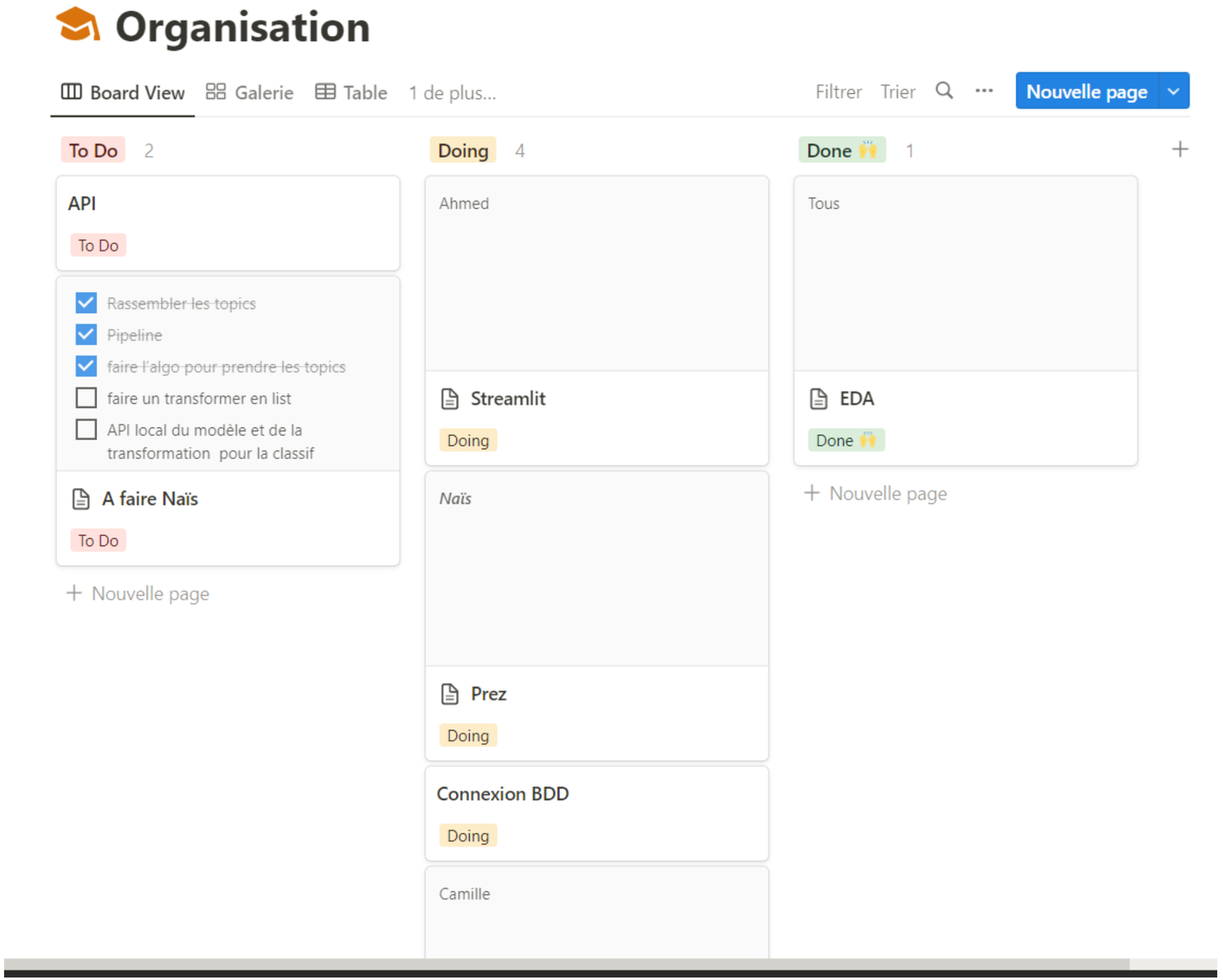
# Contexte

# Objectif :

- **Créer une interface streamlit pour prédire le sujet d'un article scientifique**
- **Déploiement d'une API sur Azure avec le modèle**
- **Connexion avec une base de données SQL**


# Organisation

# Organisation




Organisation


ATTENTIE
NIET AANRAKEN !
ONGEDIERTEBESTRIJDING
CARDS CORP
ULTRA SOUTH
ON TOUR
SOTL
BANGKOK SOI DOGS
RUES
NEST
VIDEO MARC DORCEL
SAVAGE
LAYING THE GRAVEL FOR YOU TO TRAVEL
REM JDI
NOK
There?
NOK
@plantpotheads
mauvais enfants
ANVIA
ANVIA
REAL DJ'S

# Demo


# Présentation du projet

**Base de données**

**API (contient le modèle)**

**Streamlit (interface)**

**En parallèle :**

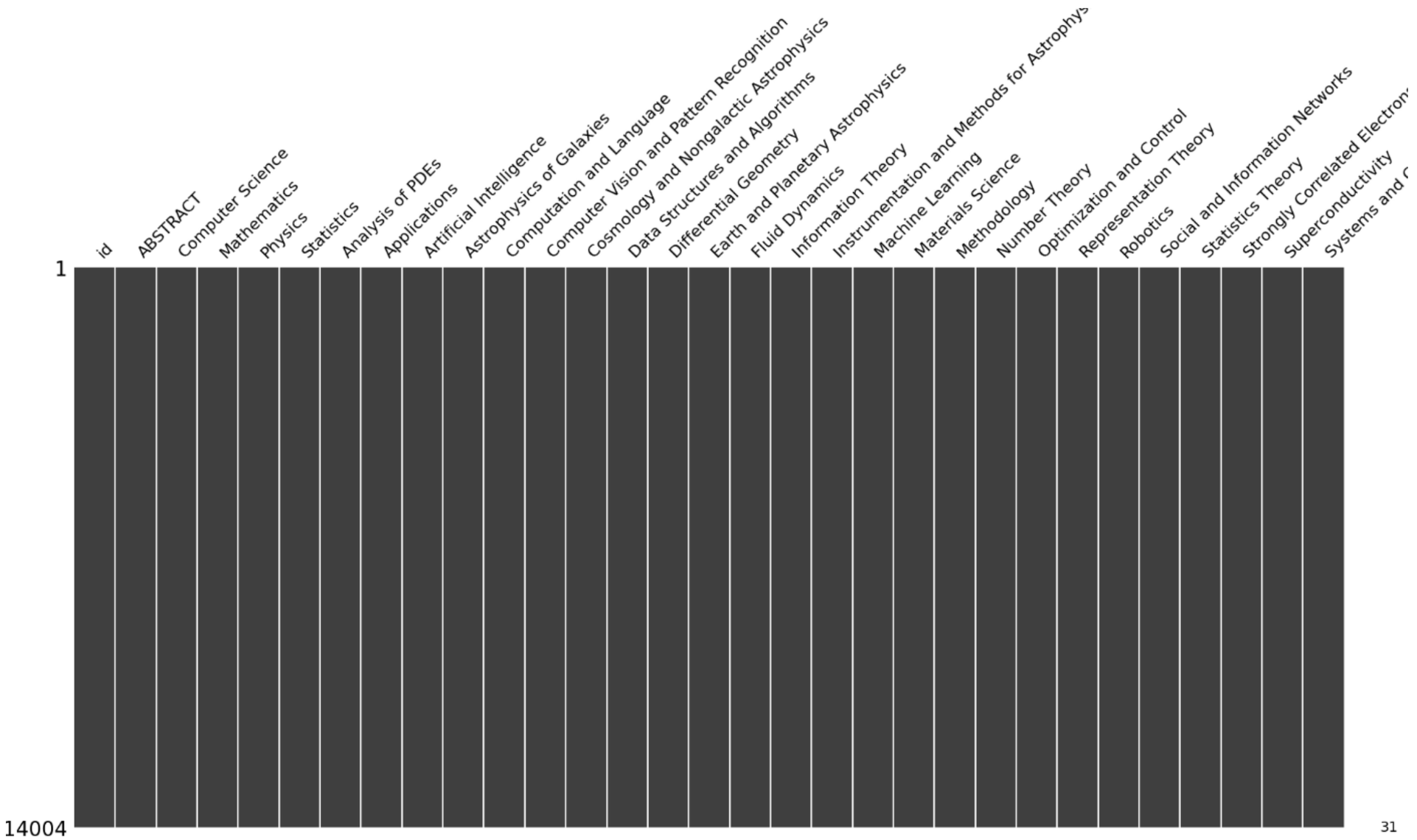
**Modèle de classification sur 29 topics**

**API pour aller requêter le modèle**


ATTENTIE
NIET AANRAKEN !
Deze voerdoos bevat een ratten- of muizenval.
Buiten bereik van kinderen en huisdieren houden.
SV ONGEDIERTEBESTRIJDING
CARDS CORP
ULTRA SOUTH
ON TOUR
SOUL
BANGKOK SOI DOGS
RUES
NEST
VIDEO MARC DORCEL
SAVAGE
LAYING THE GRAVEL FOR YOU TO TRAVEL
REM JDI
NOK
There?
NOK
@plantpotheads
mauvais enfants
ANNA
ANNA
REAL DJ'S

# Démarche

# Les données


id
ABSTRACT
Computer Science
Mathematics
Physics
Statistics
Analysis of PDEs
Applications
Artificial Intelligence
Astrophysics of Galaxies
Computation and Language
Computer Vision and Pattern Recognition
Cosmology and Nongalactic Astrophysics
Data Structures and Algorithms
Differential Geometry
Earth and Planetary Astrophysics
Fluid Dynamics
Information Theory
Instrumentation and Methods for Astrophys
Machine Learning
Materials Science
Methodology
Number Theory
Optimization and Control
Representation Theory
Robotics
Social and Information Networks
Statistics Theory
Strongly Correlated Electron
Superconductivity
Systems and
1
14004
31

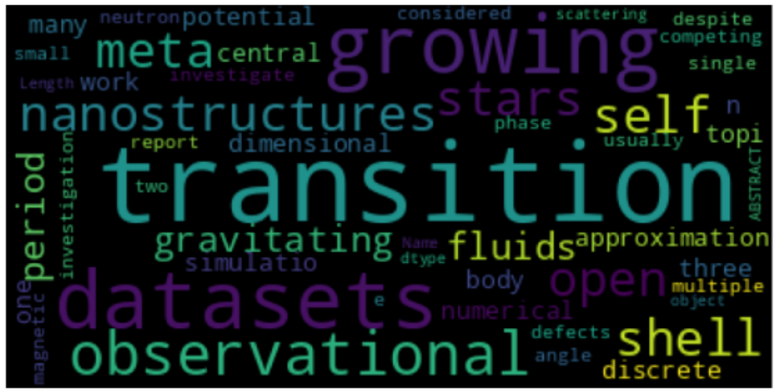
Démarche

# TOP WORDS FOR A GIVEN TOPIC

## COMPUTER_SCIENCE


## MATHEMATICS


## PHYSICS




## STATISTICS

# Modèle LDA

Soit un article traitant de du paiement mobil → _ Banque _ Mobil _ Paiements → Chacun de ces sujets ont des mots associés

Utilisation d'un modèle de probabilité basé sur la  loi de Dirichlet afin de déterminer en fonction des mots contenus dans le texte initial quel sera le sujet de ce dernier, car en fonction du topic certains mots ont plus de chance d'apparaître que d'autres.

# Classification

## 4 Preprocessing ¶

### 4.1 Remove punctuations

### 4.2 Remove stopwords

### 4.3 Remove frequents words

### 4.4 Remove rare words

### 4.5 Lemmatization

# Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.74 | 0.63 | 0.68 | 123 |
| 1 | 0.33 | 0.23 | 0.27 | 120 |
| 2 | 0.35 | 0.37 | 0.36 | 273 |
| 3 | 0.82 | 0.66 | 0.73 | 111 |
| 4 | 0.68 | 0.58 | 0.63 | 130 |
| 5 | 0.66 | 0.53 | 0.58 | 196 |
| 6 | 0.74 | 0.66 | 0.69 | 119 |
| 7 | 0.66 | 0.47 | 0.55 | 102 |
| 8 | 0.81 | 0.72 | 0.76 | 109 |
| 9 | 0.87 | 0.75 | 0.81 | 92 |
| 10 | 0.86 | 0.60 | 0.71 | 70 |
| 11 | 0.79 | 0.56 | 0.65 | 79 |
| 12 | 0.72 | 0.54 | 0.62 | 97 |
| 13 | 0.68 | 0.67 | 0.68 | 756 |
| 14 | 0.67 | 0.68 | 0.67 | 145 |
| 15 | 0.47 | 0.40 | 0.43 | 114 |
| 16 | 0.82 | 0.71 | 0.76 | 89 |
| 17 | 0.49 | 0.37 | 0.43 | 131 |
| 18 | 0.87 | 0.77 | 0.82 | 81 |
| 19 | 0.82 | 0.69 | 0.75 | 189 |
| 20 | 0.70 | 0.62 | 0.66 | 112 |
| 21 | 0.55 | 0.52 | 0.53 | 110 |
| 22 | 0.74 | 0.66 | 0.70 | 189 |
| 23 | 0.79 | 0.76 | 0.78 | 112 |
| 24 | 0.50 | 0.49 | 0.50 | 102 |
| micro avg | 0.66 | 0.59 | 0.62 | 3751 |
| macro avg | 0.69 | 0.59 | 0.63 | 3751 |

# Deep learning

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 200, 50)           3813750

 batch_normalization (BatchN  (None, 200, 50)          200
 ormalization)

 flatten (Flatten)           (None, 10000)             0

 Fully_Connected (Dense)     (None, 200)               2000200

 dropout (Dropout)           (None, 200)               0

 la_base (Dense)             (None, 200)               40200

 Output (Dense)              (None, 25)                5025

 dropout_1 (Dropout)         (None, 25)                0
```

```
Epoch 6/20
351/351 [==============================] - 27s 78ms/step - loss: 0.3107 - accuracy: 0.5402 - val_loss: 0.2137 - val_accuracy:
0.3920
```


# Axes d'amélioration

# Axes amelioration

- **Securiser l'API**
- **Finir le déploiement sur Azure**
- **Ameliorer le modèle**
- **Développer le Deep Learning => LSTM**

Axes d'amélioration

# Merci !

Camille, Naïs, Ahmed