



Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes  
Université Mohammed V - Rabat



# Unsupervised Learning Project

Arabic News Articles Clustering using  
K-means & Spectral Clustering

Présenté par :

EZZIDANI Ayoub

CHENNA Mohamed Reda

GRINI Anass

# Plan de présentation

## Introduction

I - Description du Dataset

II - Pré-traitement du text arabe

III - Les Méthodes du Clustering

IV - Résultats

## Perspectives et Conclusion

# Introduction (1)

سياسة اقتصاد ثقافة رياضة فن تكنولوجيا تراث ميدان ريادة البرامج المزيد

البيان الحي

## وسط إضراب لقضاة تونس.. الغنوشي يحذر من حرب أهلية وجبهة الخلاص تواصل تحركاتها لإسقاط قرارات سعيد

قال رئيس حركة النهضة ورئيس البرلمان التونسي راشد الغنوشي إن نهاية الاستبداد ليست بعيدة. بينما دعا رئيس جبهة الخلاص الوطني أحمد الشابي لحكومة إنشاء وطني تلتقي من حوار وطني جامع.

لوموند: هل بدأ العد التنازلي لمرحلة عالية المخاطر في تونس؟

زعيم جبهة الخلاص للجزيرة نت: قيس سعيد فقد مبررات وجوده كرئيس لكنه مستفيد من حياد الجيش والنظام اللخب

صورة أم ضرورية؟. التقسام تونس حول مصداقية لجان صياغة مشروع دستور جديد

سياسة اقتصاد ثقافة رياضة فن تكنولوجيا تراث ميدان ريادة

# Introduction (2)

## Objectif

Automatiser la tâche de diriger les nouveaux articles écrits en langue arabe vers leur rubrique correspondante .

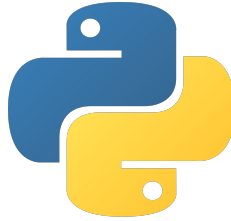
Screenshot of the website **الشعب برس** (Al-Sha'ab Press) showing a news article titled "بنكيران: بعض الحداثيين 'مرترقة' و'شياطين'.. ومستقبل الحزب بيد الله" (Benkirane: Some modernists 'meretrqa' and 'shaytan'.. The future of the party is in God's hands). The article is dated 24 ساعة (24 hours) and features a photo of Abdelhak Benkirane speaking at a podium. The website header includes navigation links: الرئيسية (Home), سياسة (Politics), جهات (Authorities), مجتمع (Society), اقتصاد (Economy), حوادث (Incidents), السلطة الرابعة (The Fourth Power), فن وثقافة (Art and Culture), تماريغت (Games), رياضة (Sports), صوت وصورة (Sound and Image), and خارج الحدود (Outside the Borders). The article text discusses the future of the party and the role of modernists.

الرئيسية سياسة جهات مجتمع اقتصاد حوادث السلطة الرابعة فن وثقافة تماريغت رياضة صوت وصورة خارج الحدود



# Description du Dataset

## 1- Scraping des données (1)



*Les Outils utilisés*



*Les Sources de Données*



# Description du Dataset

## *1- Scraping des données (2)*



culture



economy



politics



science



tamazight

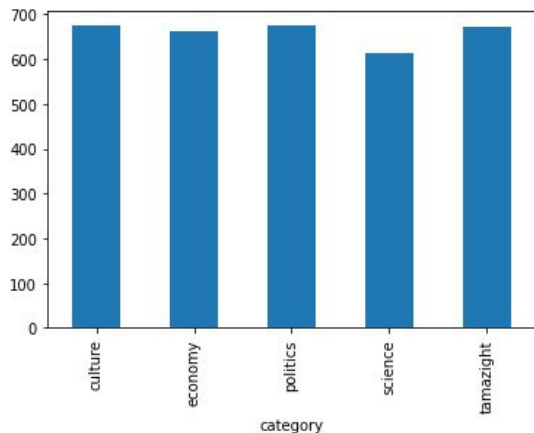
**Code source du scrapping :**

<https://github.com/Taylor-X01/News-Categories-Clustering/tree/master/scrapper>

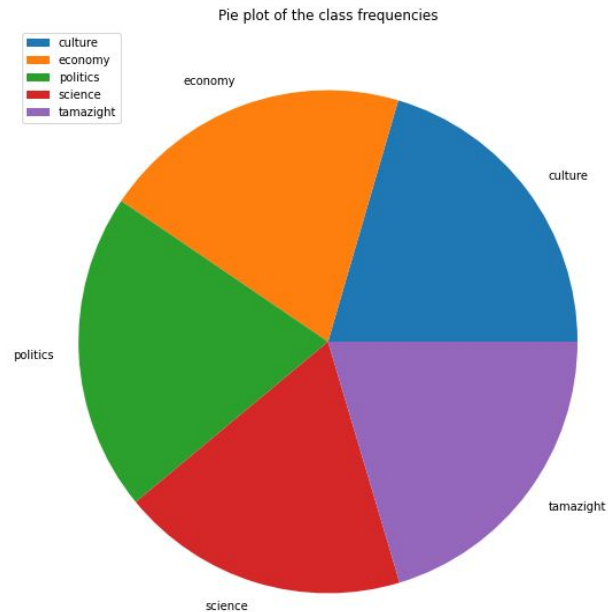


# Description du Dataset

## 2- Exploration des données et Visualisation



Fréquence des articles pour chaque thème



Répartition des articles sur les différentes thèmes

### 3 - Extraction des caractéristiques

<u>Label</u>	<u>LabelID</u>
<i>culture</i>	0
<i>politics</i>	1
<i>tamazight</i>	2
<i>science</i>	3
<i>economy</i>	4

## Labellisation des Classes

(3295, 2)

category	content
culture	... مهرجان "فيكام" بمكناس يسعى إلى خلق سوق مغربية
culture	... من إرنانديز ملقبي بشكلي بروج للثقافة الصحراوية
culture	... رحيل الشاعر العراقي حسب الشيخ جعفر عن 80 عاما
science	... الصديق المحتاج.. واصاب تكتف طريفة جديدة لحد"
science	... لحماية التصور.. عوغل تسمح لأهل بإزالة صور أبنيا
science	... حولت أحلام العديد من المبتكرين لواقع.. تكبر أ
science	... "توسيد تغلب على تسلا بشكل رسمي في مجال "المدى
tamazight	... بنشطاء يعولون على الانتخابات المقبلة لتصبح ال
politics	... مقتل 31 شخصا في تنافق قرب إحدى الكنائس جنوبى ن
politics	... بكيان يضع أخطاء الحركة الإسلامية تحت المجهر

### Extrait du jeu de données





# Pré-traitement du text arabe

## 1- Suppression de la ponctuation et des chiffres

```
arabic_punctuations = '``÷x-“”!|+!~{}`',.?"':/[-][%^&*()_<>!'`'  
english_punctuations = string.punctuation  
punctuations_list = arabic_punctuations + english_punctuations  
punctuations_list
```

executed in 5ms, finished 20:56:37 2022-06-05

```
'~{||}^[\]\]@?<=>;:/.-,+*()'\'&%$#"!-“”!|+!~{}`',.?"':/[-][%^&*()_<>!x÷`'
```

```
text = "".join([word for word in text if word not in string.punctuation])  
text = "".join([word for word in text if word.isdigit()==False])
```



# Pré-traitement du text arabe

## 2- Suppression des caractères spéciaux

```
def remove_tags(text):
    remove = re.compile(r'<.*?>')
    return re.sub(remove, '', text)

def remove_diacritics(text):
    arabic_diacritics = re.compile("""
        \s          | # Tashdid
        َ           | # Fatha
        ِ           | # Tanwin Fath
        ّ           | # Damma
        ٓ           | # Tanwin Damm
        ُ           | # Kasra
        ٌ           | # Tanwin Kasr
        ٔ           | # Sukun
        -           | # Tatwil/Kashida
    """, re.VERBOSE)
    text = re.sub(arabic_diacritics, '', str(text))
    return text
```

```
def remove_emoji(text):
    regex_pattern = re.compile(pattern = "["
        u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # symbols & pictographs
        u"\U0001F680-\U0001F6FF"  # transport & map symbols
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        "]" + "", flags = re.UNICODE)
    return regex_pattern.sub(r'', text)
```

# Pré-traitement du text arabe

## 3- Suppression des stop-words

```
def remove_stopwords(text):  
    stop_words = set(stopwords.words('arabic'))  
    stop_words.update({'سيكون', 'يأن', 'ذلك', 'وفي', 'أنه', 'قال', 'يمكن', 'وقد', 'لهذه', 'أيضا', 'خلال'})  
    words = word_tokenize(text)  
    return " ".join([i for i in words if i not in stop_words ])
```

بَ، و' كَلَامُهُا، و' سَاءَ، و' فَضْلًا، و' مَنذُ، و' هَلَا، و' لَكِنَّ، و' السَّمَّ، و' فِيهِ، و' لَبِثَ، و' حَيْثُمَا، و' شَرَعَ، و' بَكَمَ، و' بَعَثَ، و' سَتَ، و'  
ثَلَاثِينَ، و' حِينَ، و' هَذَيْنِ، و' بَشِ، و' ثَمْنِيَّةَ، و' أَيَّ، و' الْأَلَاءِ، و' تَخَذَ، و' تَسْعَمِيَّةَ، و' كَأَيْنَ، و' إِنْ، و' رِيثَ، و' عَلَيْكَ، و' رَابِعَ، و'  
رَايَ، و' أَنْ، و' بَاءَ، و' لَكِي، و' أَلْفِي، و' انْقَلَبَ، و' مِمَّ، و' غَيْنَ، و' ثَمَّةَ، و' أَنْتُمَا، و' حَبِيبَ، و' ثَمَانِي، و' تَلَكُمَ، و' نَحْنُ، و' أَفْرِيءَ  
فِيهَا، و' هَجَّ، و' إِيَاهُمَ، و' كَمَا، و' أَمْسَ، و' اللَّتِيَا، و' دَ، و' مَلِيمَ، و' رَجَعَ، و' كَلِمَا، و' يَنَابِرَ، و' رَاحَ، و' تَائِكَ، و' خَذَارَ، و' عَامَةً، و'  
بَ، و' صِرَاحَةً، و' مِنْهَا، و' بَضَعَ، و' لَا سِيَمَا، و' بِلَ، و' اللَّذَانِ، و' لَسْنَا، و' لَيْسَ، و' ضَ، و' آهَ، و' كَأَيَّ، و' ثَمَّ، و' زَ، و' لَهُ، و' ذَ  
و' لَعَلَّ، و' لَا، و' إِنْ، و' أَرْبَعَمِائَةَ، و' سَنَتِيمَ، و' إِلَّا، و' لَوْلَا، و' وَلَكِنْ، و' جَمِيعَ، و' لَاسِيَمَا، و' عُلِقَ، و' مِثْنَانِ، و' كَيْفَمَا، و' مَذَ، و'  
صَبْرًا، و' هَذِي، و' فَمِنْ، و' سِرْعَانَ، و' أَمَامَكَ، و' مَسَاءَ، و' ذِي، و' سِتَ، و' أَلْفَ، و' دَالٍ، و' دُونَ، و' ثَاءَ، و' دُونَكَ، و' نَحْجَ، و' إِيَا  
بَ، و' كَلَامُهُا، و' سَاءَ، و' فَضْلًا، و' مَنذُ، و' هَلَا، و' لَكِنَّ، و' السَّمَّ، و' فِيهِ، و' لَبِثَ، و' حَيْثُمَا، و' شَرَعَ، و' بَكَمَ، و' بَعَثَ، و' سَتَ، و'



# Pré-traitement du text arabe

## 4- tokenization du texte

"[مارسيل ' و 'خليفة' و 'يتمنى' و 'الأصل' و 'المغربي' و 'ويتمسك' و 'بالمقاومة' و 'عبر' و 'الموسيقى' و 'بحضور' و 'الموسيقى' و 'البارز' و 'مارسيل' و 'خليفة'، و 'استقبلت' و 'أكاديمية' و 'المملكة' و 'المغربية'، و 'اليوم' و 'الخميس' و 'بالرباط'، و 'حفلة' و 'نقاش'، و 'نظمها' و 'المعهد' و 'الأكاديمي' و 'للفنون'، و 'أسئلة' و 'الموسيقى' و 'حاضر' و 'المنطقة' و 'والعالم' و 'الحفلة' و 'النفاشية'، و 'أطرها' و 'مارسيل' و 'خليفة'، و 'وسيرها' و 'باسم' و 'المعهد' و 'الأكاديمي' و 'للفنون' و 'الأكاديمي' و 'محمد' و 'نور' و 'الدين' و 'أفاية'، و 'جمعت' و 'عددا' و 'الأسماء' و 'البحثية' و 'والأصوات' و 'الموسيقية' و 'المغربية'، و 'عبرت' و 'آرائها' و 'تاريخ' و 'وواقع' و 'الموسيقى'، و 'وتلقيها'، و 'والتربية' و 'الموسيقية'، و 'وأفاق' و 'توثيق' و 'الإبداعات' و 'المغربية' و 'وتتميتها' و 'أحدث' و 'المعهد' و 'الأكاديمي' و 'للفنون' و 'إطار' و 'أكاديمية' و 'المملكة' و 'المغربية'، و 'برسم' و 'قائد' ]"



# Pré-traitement du text arabe

## 4- Lemmatizing & Stemming

- La lemmatisation est le fait de prendre le lemme (forme canonique) des mots
- Le stemming fonctionne en coupant la fin ou le début du mot, en tenant compte d'une liste de préfixes et de suffixes courants que l'on peut trouver dans un mot infléchi.

```
def stem_word(text):  
    st = ISRIStemmer()  
    return " ".join([st.stem(word) for word in text])  
  
def lemmatize_(text):  
    lemmer = qalsadi.lemmatizer.Lemmatizer()  
    lemmas = lemmer.lemmatize_text(text)  
    return (lemmas)
```



# Pré-traitement du text arabe

## 5- Résultats du traitement

l'application des étapes de traitement de texte permet de transformer nos étapes en features qu'on peut exploiter ultérieurement dans l'application du modèle

'بركة يبسط خطة الحكومة لحماية المغاربة الوضعية المائية المقلقة أكد نزار بركة، وزير التجهيز والماء، الوضع المائي المغرب مقلق؛ وجاء أجوبته أسئلة الفرق البرلمانية، اليوم الاثنين، مجلس النواب وقال بركة "الوضعية المائية مقلقة لعدة اعتبارات، بينها المغرب مهدد بندرة المياه وسنوات الجفاف الشأن بالنسبة السنة، انعكاسات كبيرة الساكنة " وسجل الواردات المائية تراجعت بنسبة المائة مقارنة بسنة عادية، بفضل التساقطات الأخيرة تراجعت المائة، مشيرا تراجع الواردات المائية انطلق سنة وأوضح ارتفاع متوسط درجة الحرارة بدرجة السنوات الخمس الأخيرة يؤدي تبخر المياه ويتسبب إشكالا ت مستوى التربة والري، وانضاف أمر آخر تأثير الوضعية المائية المغرب الحرب الروسية الأوكرانية، أدت ارتفاع عدد المواد الأساسية تستعمل بناء السدود مقابل ذلك، أفاد بركة الحكومة وضعت خط ل لضمان تحقيق الأمن المائي تقوم تسريع وثيرة البرنامج وضعه الملك برسم ، إنجاز السدود، انطلقت عملية إنجاز سدا، مبرزا هدف الحكومة إنجاز سدا أفق سنة ، سيرفع حجم تخزين المياه مليار مة ر مكعب مليار متر مكعب، وهذا حد ذاته أمر مهم وشدد أنه " ظل قلة التساقطات المطرية، تكفي بالسدود فقط، الضروري نستعمل تحلية المياه، قمنا بوضع مخطط خاص لتحلية المياه بالنسبة للمناطق



## 6- Visualisation des WordClouds

Politics related words:

Culture related words:

Economy related words:



[illegible]





# Pré-traitement du text arabe

## 7- Vectorisation

La vectorisation du texte est le processus de conversion du texte en représentation numérique. 2 méthodes ont été utilisées dans ce projet:

TF-IDF

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t) = \ln\left(\frac{1 + n}{1 + df(t)}\right) + 1$$

Bag of words

- Vocabulary collection
- Vectorizer



# Les Méthodes du Clustering

## 1- K-means

*Après avoir initialisé des points étant les centroïdes en prenant des données au hasard dans le jeu de données, K-means alterne plusieurs fois ces deux étapes pour optimiser les centroïdes et leurs groupes :*

- 1. Regrouper chaque objet autour du centroïde le plus proche.***
- 2. Remplacer chaque centroïde selon la moyenne des descripteurs de son groupe.***

*Après quelques itérations, l'algorithme trouve un découpage stable du jeu de données : on dit que l'algorithme a convergé.*



# Les Méthodes du Clustering

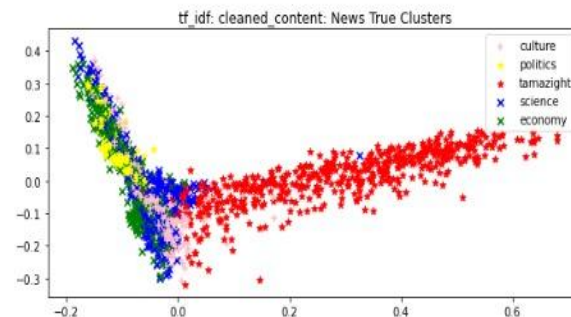
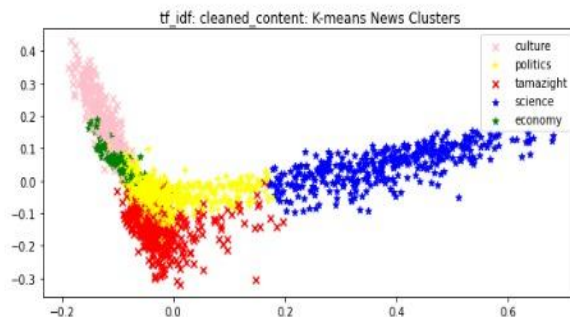
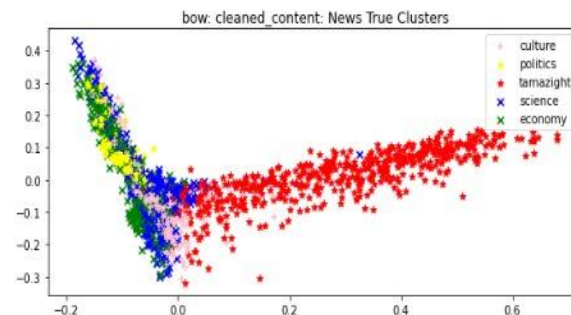
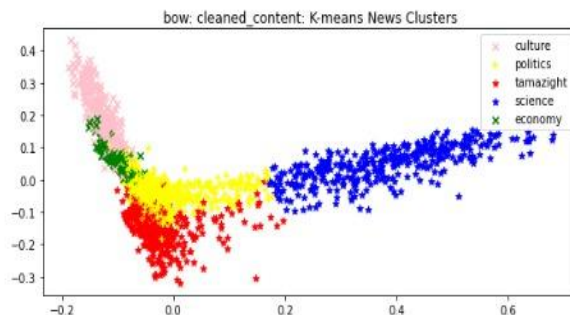
## 2- clustering spectrale

1. Construire un graphe de similarité et fixer le nombre  $k$  de clusters
2. Calculer  $W$  sa matrice d'adjacence pondérée.
3. Calculer la matrice de degré  $D$  et la matrice Laplacienne  $L = D - W$ .
4. Trouver les valeurs propres et les vecteurs propres de  $L$ .
5. Avec les vecteurs propres des  $k$  plus grandes valeurs propres calculées à l'étape précédente, former une matrice.
6. Normaliser les vecteurs.
7. Regrouper les points de données dans un espace à  $k$  dimensions.



# Résultats

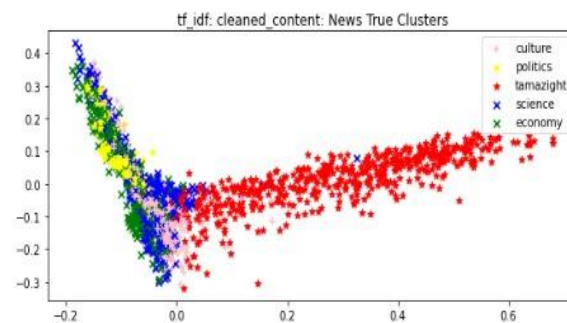
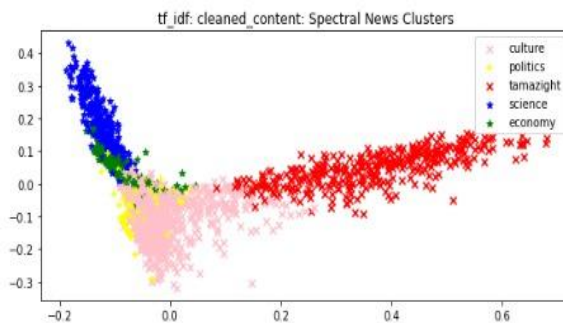
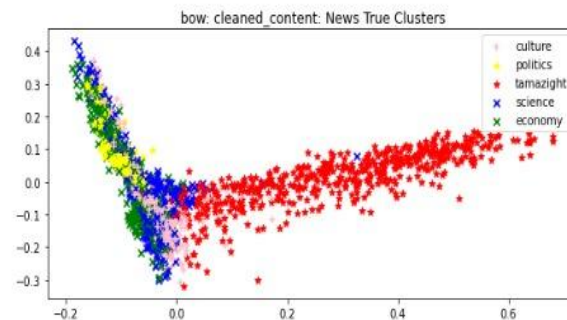
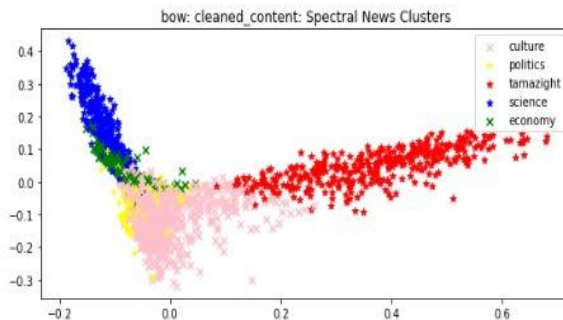
## 1- Avec k-means





# Résultats

## 2- Avec Spectrale clustering





# Résultats

## 1- Silhouette

Il se base sur la distance moyenne du point à son groupe :

$$a(i) = \frac{1}{|I_k|-1} \sum_{j \in I_k, j \neq i} d(x^i, x^j)$$

et la distance moyenne du point à son groupe voisin

$$b(i) = \min_{k' \neq k} \frac{1}{|I_{k'}|} \sum_{i' \in I_{k'}} d(x^i, x^{i'}).$$

Son expression est :

$$s_{sil}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$



# Résultats

## 1- *Davies bouldin*

L'indice (ou score) de Davies-Bouldin,  $S$  se base sur les points moyens de chaque groupe

$$\mu_k = \frac{1}{|I_k|} \sum_{i \in I_k} x^i$$

et la distance moyenne entre un point et le centre de son groupe

$$\bar{\delta}_k = \frac{1}{|I_k|} \sum_{i \in I_k} d(x^i, \mu_k).$$

Son expression est :

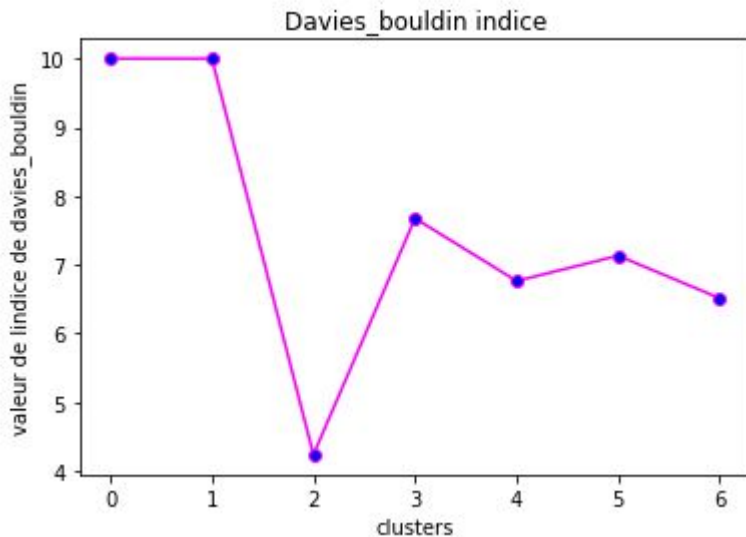
$$S_{DB} = \frac{1}{K} \sum_{k=1}^K \max_{k' \neq k} \left( \frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(\mu_k, \mu_{k'})} \right)$$



# Résultats

## 1- Validation de K-means

La variation de l'indice de davies-bouldin suivant le nombre de cluster se présente comme suit :



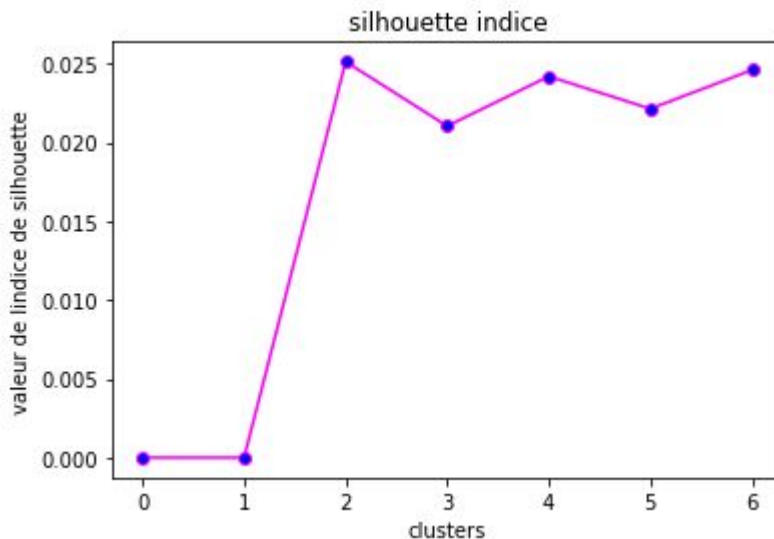




# Résultats

## 2- Validation de *K-means*

La variation de l'indice de silhouette suivant le nombre de cluster se présente comme suit :

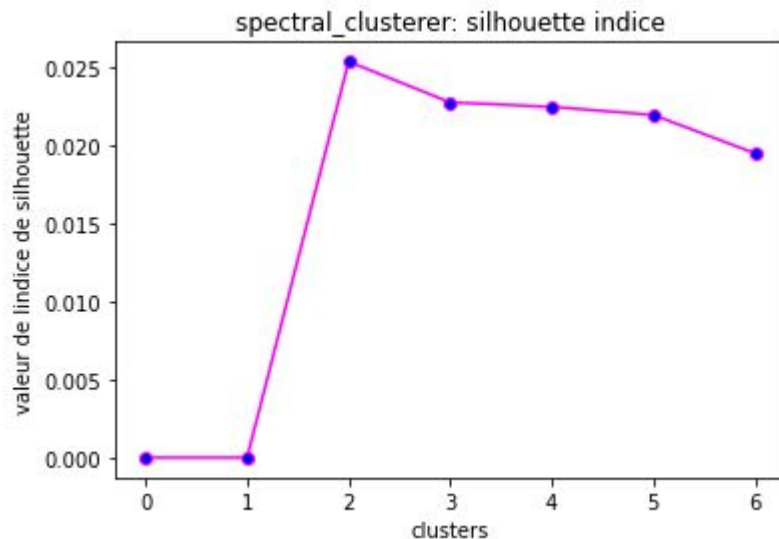




# Résultats

## 3- Validation de Clustering spectrale

La variation de l'indice de silhouette suivant le nombre de cluster se présente comme suit :

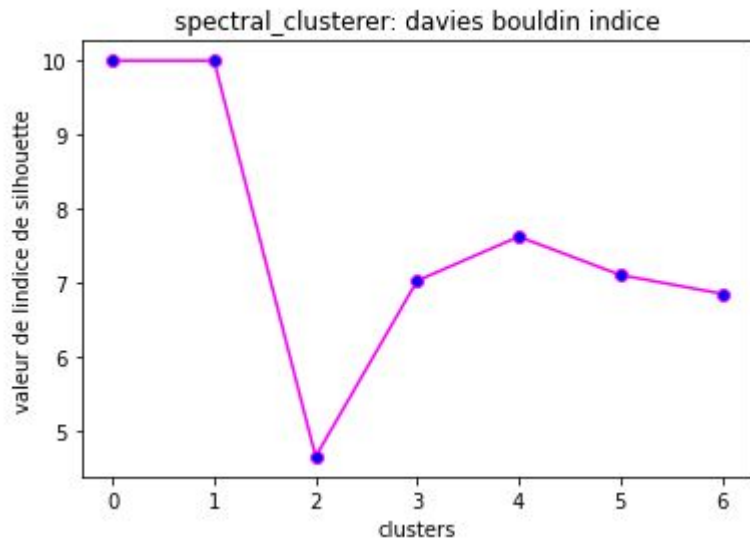




# Résultats

## 4- Validation de Clustering spectrale

La variation de l'indice de davies-bouldin suivant le nombre de cluster se présente comme suit :



# Perspectives du projet

- Augmenter la précision de clustering en testant d'autre algorithme
- Générer des titres ou des rubriques pour les nouveaux articles
- Appliquer la même procédure sur d'autre types de données
- Appliquer le même traitement pour les courriers électroniques

# Conclusion

*Les résultats de ce projet sont généralement applicables à n'importe quel domaine contenant des données textuelles et aident à traiter de grandes quantités de données. Cette approche peut être appliquée de manière pratique et étendue à des applications dans des domaines tels que l'analyse des réclamations d'assurance, le traitement automatique des fax, le filtrage des courriers électroniques, et bien d'autres applications de fouille de texte.*