

WEM Groupe L

Manga Feelings



Table des matières

- Introduction
- Scraping
- Nettoyage des données
- Sentiment analysis
- Prédiction
- Visualisation
- Conclusion

Introduction

Objectif

Scraping des commentaires du site webtoons pour faire du sentiment analysis et essayer de prédire la note attribuée la note du webtoon



Comments 487

Please [log in](#) to leave a comment.

Post

0/500

✓ Top ✓ Newest

welovewithreason

TOP times like this make me wonder if wolf has CIPA (an extremely rare condition that prevents people from registering physical pain). edit: now that this is getting publicity, i just wanna clarify that i don't actually truly think wolf has CIPA. there have been plenty of times where we've seen him feel pain. i'm just highlighting how insanely durable he is.

May 18, 2023 | Report

Replies 26 ▼

👍 7670 🗨️ 17



Negative



Neutral



Positive



Introduction

Données

- Source
 - <https://www.webtoons.com/en/>
- Webtoon
 - Bande dessinée coréenne (manwha) publiée en ligne. S'applique aujourd'hui à toutes les bandes dessinées en ligne.
 - Hebdomadaire/saison
- TOS
 - Utilisation personnelle non commerciale
- Quantité de données
 - Un webtoon populaire par genre
 - 8 webtoons → 1472 épisode

série	épisodes
A Life Through Selfies	316
Goth Girl The Jock	32
How To Be A Dragon	36
My Husband Changes Every Night	5
Nerd And Jock	198
Power Pills	477
Seekers Log	168
Weakhero	240

Scraping

Technologies utilisées

- Site complexe
 - Javascript
 - Chargement dynamique du site
- Pas fonctionné
 - Scrapy
- Principales librairies utilisées



BeautifulSoup4



Scraping

Stratégie pour obtenir les informations des épisodes

1. Utilisation de Selenium

2. Renseigner url principale

3. Extraire

1. Episode name
2. Episode url
3. Episode likes
4. Episode Date

4. Page suivante

The screenshot displays the Webtoon interface for the series 'Weak Hero'. At the top, navigation tabs include ORIGINALS, GENRES, POPULAR, and CANVAS. The series title 'Weak Hero' is prominently displayed with the genre 'Action' and authors 'SEOPASS, RAZEN'. Below the title, a list of episodes is shown, including episode 243, 242, and 241, each with a thumbnail, title, release date, likes, and episode number. To the right, a sidebar provides statistics: 216.5M views, 1.4M followers, and a 9.84 rating. A description of the series is also present.

Episode	Date	Likes	Episode Number
(S3) Ep. 243	Jun 15, 2023	16,711	#244
(S3) Ep. 242	Jun 8, 2023	19,221	#243
(S3) Ep. 241	Jun 1, 2023	20,356	#242

```
1 def extract_episodes_info(driver, main_url, min_page, max_page):
2
3     episodes_info = []
4
5     for page_num in range(min_page, max_page + 1):
6         if page_num == min_page:
7             page_url = f"{main_url}&page={page_num}?lang=en"
8         else:
9             page_url = f"{main_url}&page={page_num}"
10
11         driver.get(page_url)
12
13         episode_items = driver.find_elements(By.CSS_SELECTOR, 'li._episodeItem')
14
15         for episode_item in episode_items:
16             episode_name = episode_item.find_element(By.CSS_SELECTOR, 'span.subj').text.strip()
17             episode_url = episode_item.find_element(By.CSS_SELECTOR, 'a').get_attribute('href')
18             likes_text = episode_item.find_element(By.CSS_SELECTOR, 'span.like_area').text.strip() # episode likes
19             likes_text = likes_text.replace(",", "") # Remove commas
20             episode_likes = int(re.search(r'\d+', likes_text).group())
21             episode_date = episode_item.find_element(By.CSS_SELECTOR, 'span.date').text.strip()
22
23             episodes_info.append({'name': episode_name, 'url': episode_url,
24                                   'likes': episode_likes, 'date': episode_date})
25
26     return episodes_info
```

Scraping

Stratégie pour obtenir les commentaires

1. Aller dans url épisode
2. Descendre bas de page
3. Ouvrir tous les «reply»
4. Télécharger html (bs4)
5. Parse commentaire/reply
 1. Date
 2. User
 3. ID Commentaire
 4. Parent ID (pour reply)
 5. Like/dislike
6. Fermer tous les «reply»
7. Page commentaires suivante

Thwack19

Ben and Gray are gonna have a fun little warm up mopping up all these low HP opponents.

May 25, 2023 | Report

Reply ▼

👍 99 🗨️ 0

Davy Jones

I'm telling y'all right now, Bens gonna take Donald on 1v1 while Gray has to solo the lieutenants.

May 25, 2023 | Report

Replies 4 ▼

👍 80 🗨️ 1

BLOODYPHANTOM 101

yah I'm still waiting for Gerrard to get up and murk Jimmy no way the author does garred dirty like that he ain't no punk bitch

May 25, 2023 | Report

Replies 4 ▼

👍 62 🗨️ 2

1 2 3 4 5 6 7 8 9 10 > »


```
1 soup = BeautifulSoup(driver.page_source, 'html.parser')
2 comments = soup.select('.u_cbox_comment')
3
4 comment_data = []
5
6 for comment in comments:
7     data_info = comment['data-info']
8     data_info = data_info.replace("'", "").replace(" ", "")
9     data_info_dict = dict(item.split(":") for item in data_info.split(",") if ":" in item)
10
11     comment_id = data_info_dict.get('commentNo')
12     reply_level = data_info_dict.get('replyLevel')
13     parent_comment_no = data_info_dict.get('parentCommentNo')
14     comment_text = comment.select_one('.u_cbox_contents').text
15     comment_date = comment.select_one('.u_cbox_date')['data-value']
16     comment_author = comment.select_one('.u_cbox_nick').text
17
18     # Extract likes and dislikes
19     likes = comment.select_one('.u_cbox_tool .u_cbox_cnt_recomm').text
20     if comment.select_one('.u_cbox_tool .u_cbox_cnt_recomm') else "0"
21     dislikes = comment.select_one('.u_cbox_tool .u_cbox_cnt_unrecomm').text
22     if comment.select_one('.u_cbox_tool .u_cbox_cnt_unrecomm') else "0"
23
24     comment_data.append([episode_url, episode_name, episode_likes, episode_date,\
25                           comment_id, reply_level, parent_comment_no, comment_text,\
26                           comment_date, comment_author, likes, dislikes])
```

Scraping

Résultat

- 1 csv par épisode
- 542 csv à traiter
- 280 Mo
- 370k commentaires
- 93k reply
- 433k lignes

Nettoyage des données

Principales étapes

- Utilisation de polars et pandas
- Nettoyage
 - Colonnes
 - Dates
 - Lignes dupliquées
- Ajout manuel
 - Genre
 - Vues
 - Subscribers
 - Rating
- Un seul csv avec toutes les données

Sentiment analysis

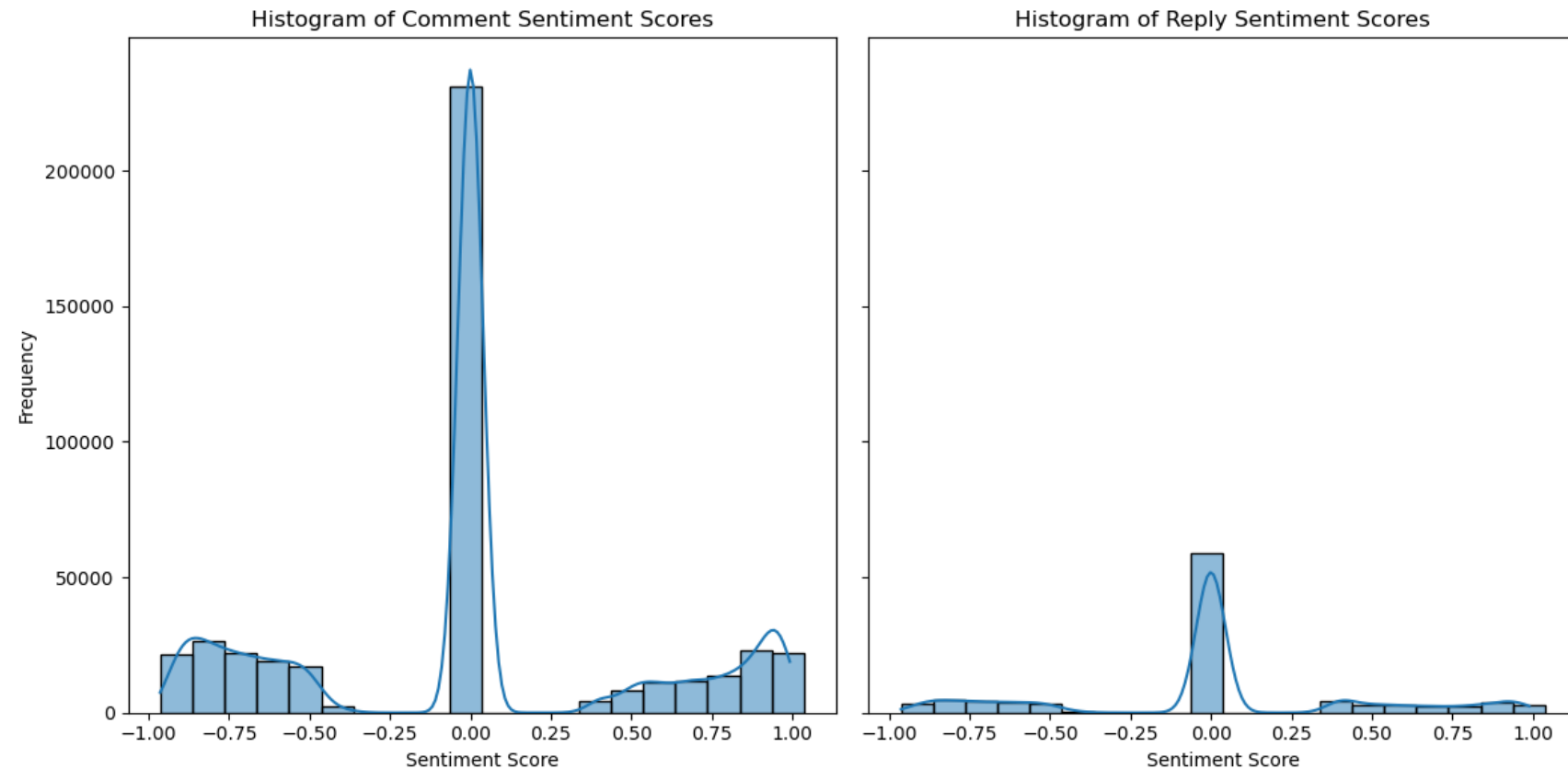
Principales étapes

- Utilisation de pandas, pytorch, transformers, nltk
- Préparation du texte
 - Tokenize
 - Supprimer mots trop longs (30 caractères)
 - Contractions
 - Supprimer stop-words
 - Lemmatize
- Sentiment analysis
 - Modèle «cardiffnlp/twitter-roberta-base-sentiment-latest» état de l'art en sentiment analysis sur tweets
 - Utilisation de GPU NVIDIA
 - Echelle (-1,1) → négatif/neutre/positif

Sentiment analysis

Résultats

- La plupart des commentaires sont neutres
- Le modèle n'a pas le contexte du commentaire
- Anglais mal écrit
- Sarcasme/ironie mal analysés



Prédiction

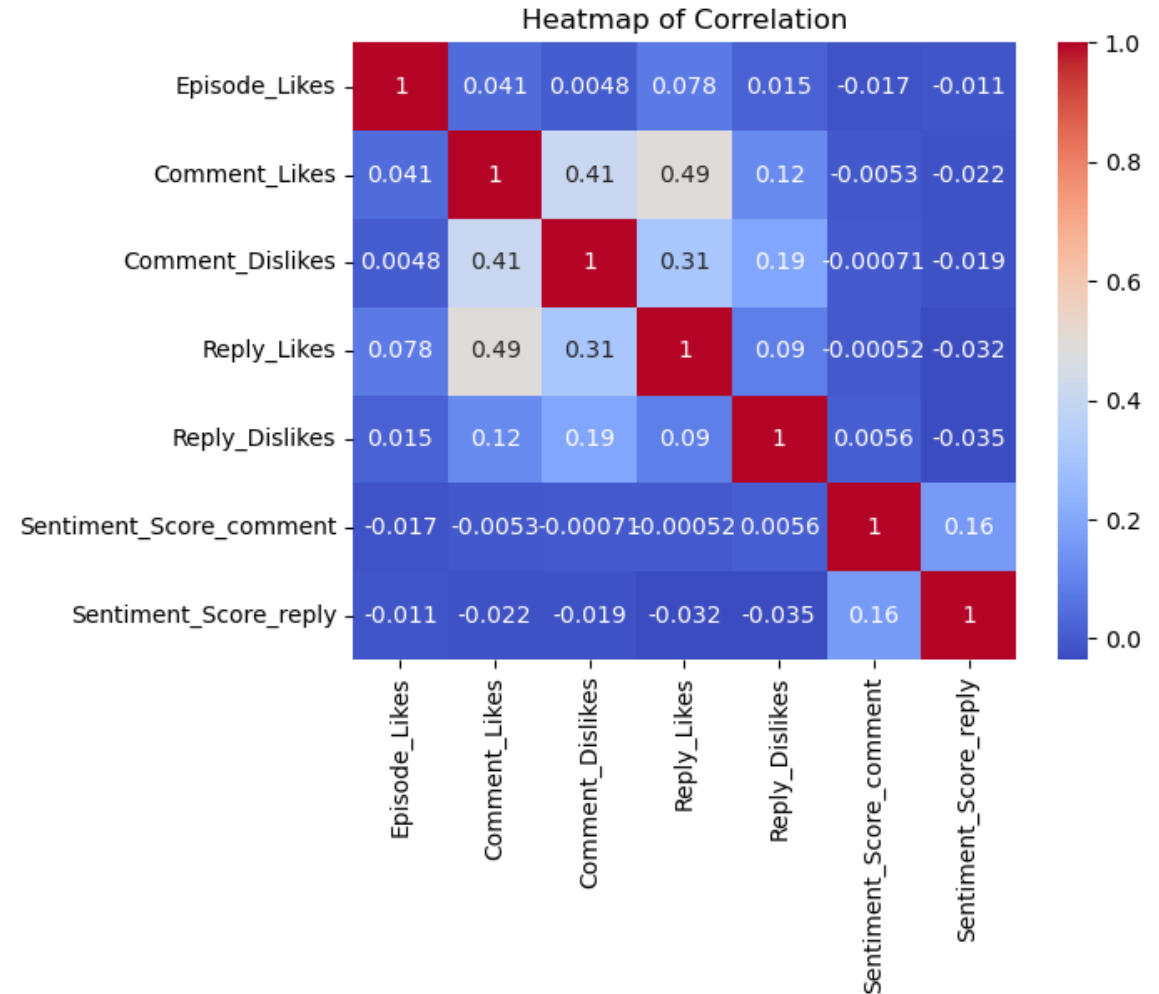
- Premier Model proposé
- Peu de donnée pour valider



Prédiction

Première approche

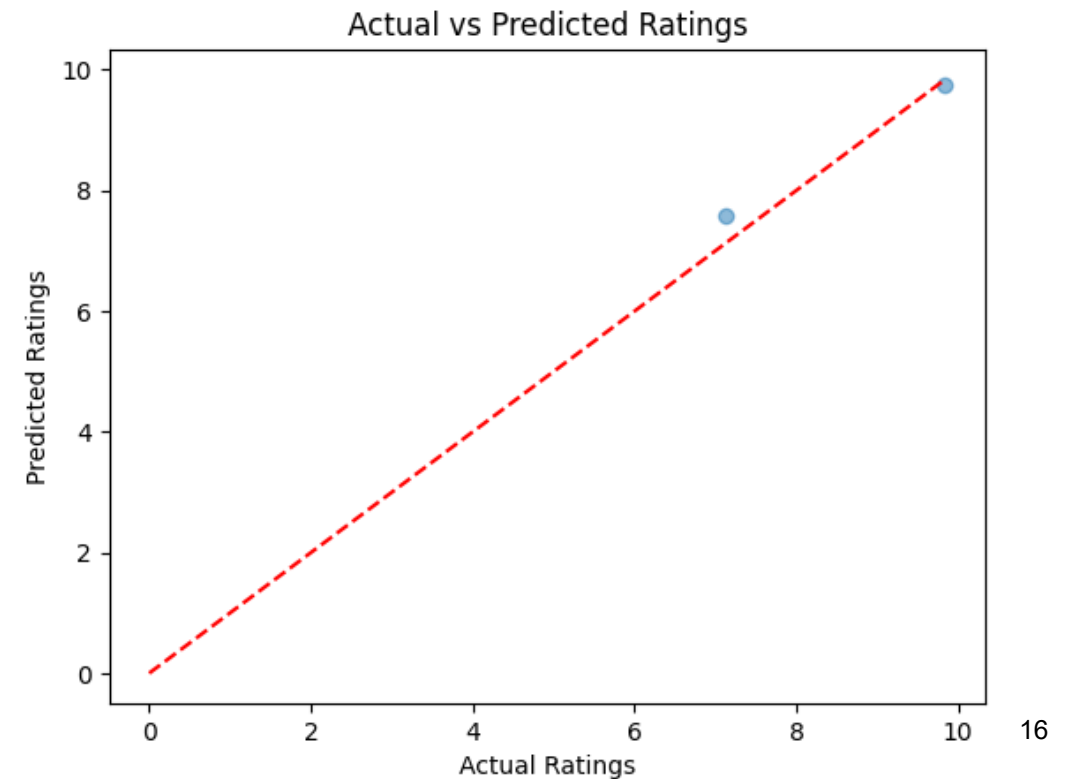
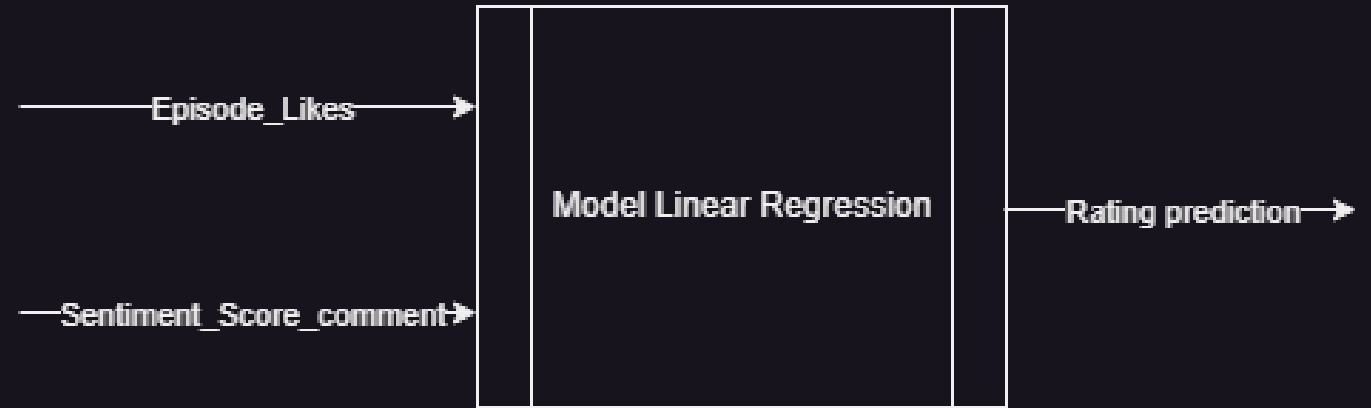
- Corrélation entre Comment_Likes, Comment_Dislikes, Reply_Likes
- Episode_Likes peu corrélée avec les autres inputs



Prédiction

Model linear

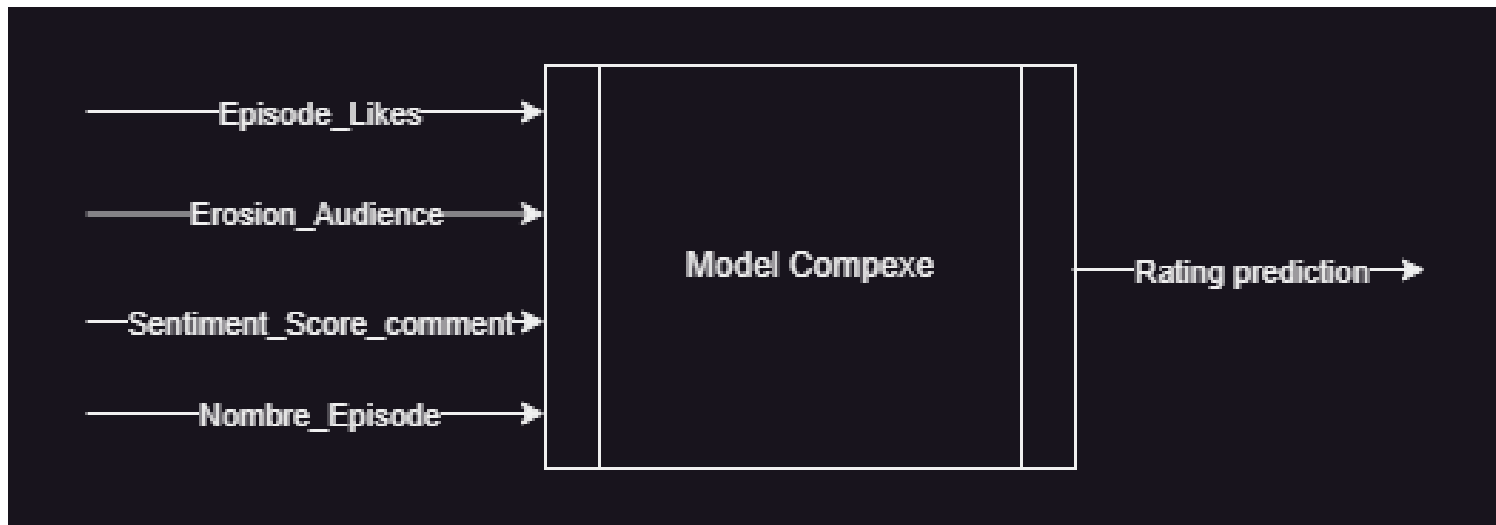
- Bon résultat
- Trop peu de données
- Faible fiabilité



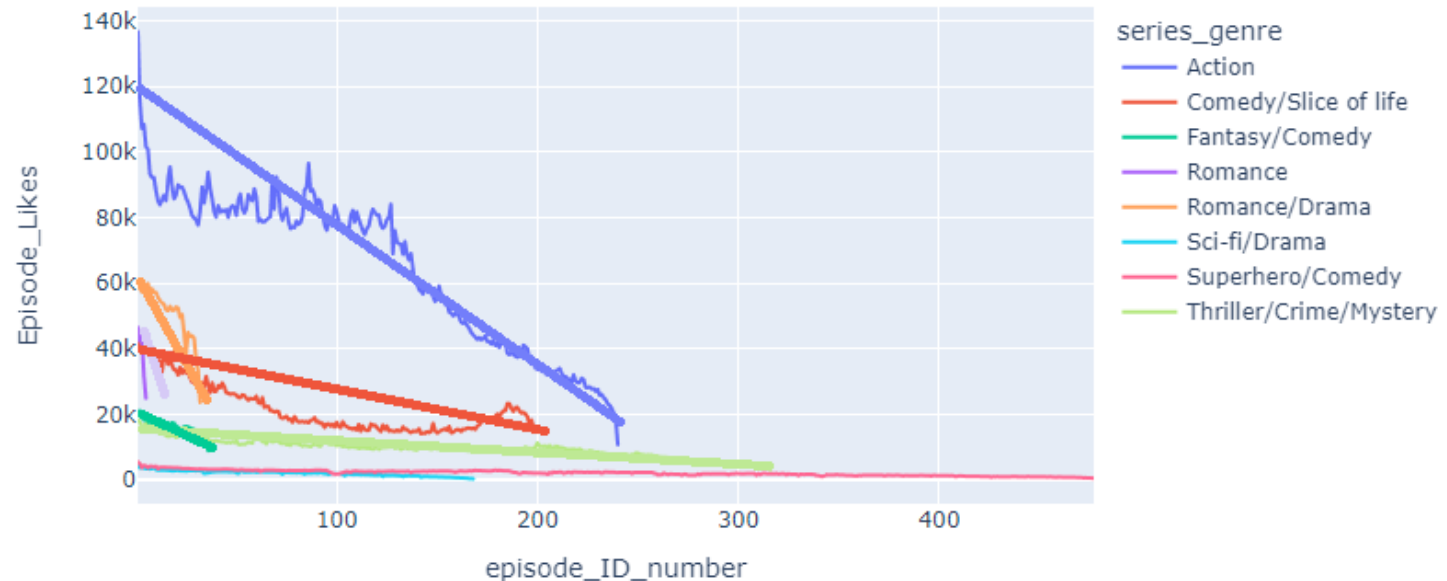
Prédiction

Model complexe

- Ajout de l'érosion de l'audience comme indication
- Validation des commentaires positif/négatif avec le nombre de like/dislike
- Possibilité d'extraire également un gradient de sentiment pour avoir l'évolution des utilisateurs
- Avec plus de donnée le genre vas devenir important



Likes per episode



Visualisation

Technologies utilisées

- Notebook pour tester les différents graphs
- Résultat final dans un script



Visualisation

Analyse des likes par épisode

- Isoler les épisodes

	series_genre	Episode_Date	Episode_Likes	episode_ID_number	Comment_Text
0	Action	2020-09-01	81653	100	I'm so glad Gray has grown and found peace for...
1	Action	2020-09-01	81653	100	I'm glad to see he's able to overcome his roof...
2	Action	2020-09-01	81653	100	Donald is playing 4D chess
3	Action	2020-09-01	81653	100	What is helmet up to these day Nothing much st...
4	Action	2020-09-01	81653	100	I really wish we can see Stephen again at some...
5	Action	2020-09-01	81653	100	Am I the only one who thinks the backpack is t...
6	Action	2020-09-01	81653	100	can we just take a moment and appreciate grays...
7	Action	2020-09-01	81653	100	I'm so happy that he's gotten over that fear a...
8	Action	2020-09-01	81653	100	My chest physically hurts again
9	Action	2020-09-01	81653	100	that mattress pin hiTS DEEP IN THE HEART BRO

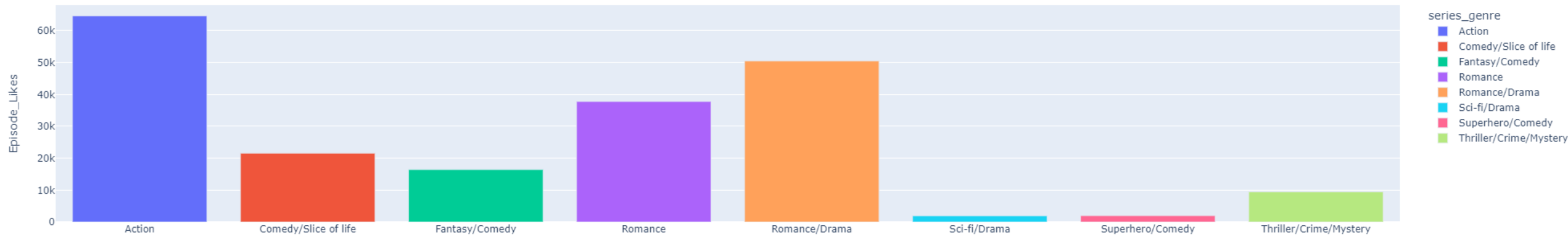
	series_genre	Episode_Date	Episode_Likes	episode_ID_number
0	Action	2019-09-10	136613	1
1	Action	2019-09-10	115644	2
2	Action	2019-09-10	107151	3
3	Action	2019-09-17	108406	4
4	Action	2019-09-24	101632	5
5	Action	2019-10-01	100795	6
6	Action	2019-10-08	93428	7
7	Action	2019-10-15	92176	8
8	Action	2019-10-17	92229	9
9	Action	2019-10-22	88114	10

Visualisation

Analyse des likes moyens par genre

- Grouper par genre, puis faire la moyenne

Average likes per genre

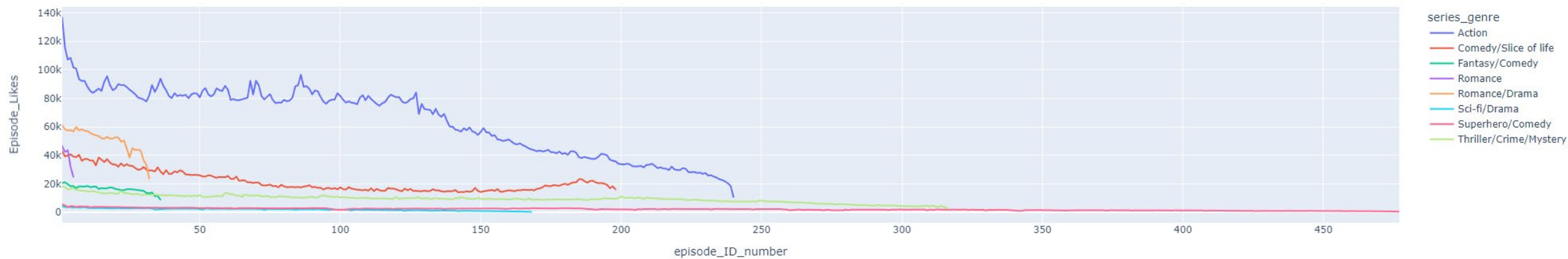


Visualisation

Analyse des likes par épisode

- Création du graphique à l'aide de plotly

Likes per episode



Visualisation

Analyse du déclin des likes par genre

- Calculer le pourcentage de like d'un épisode par rapport au premier épisode de la série

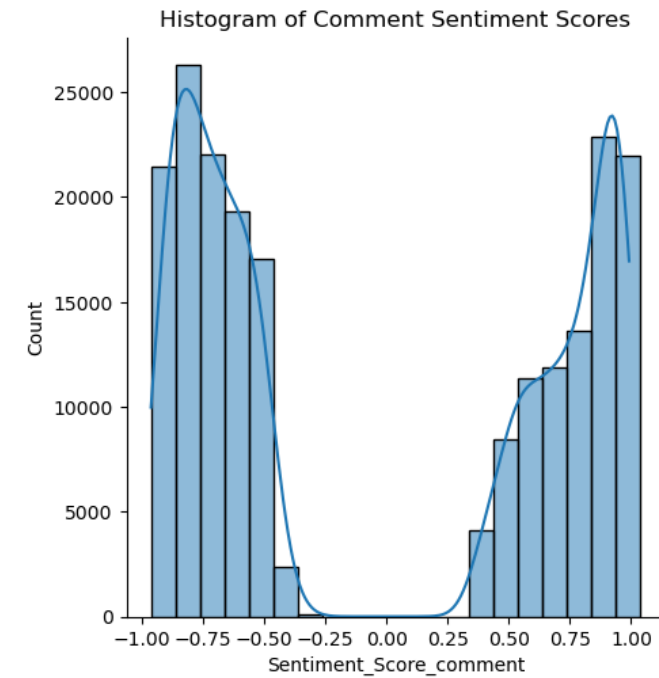
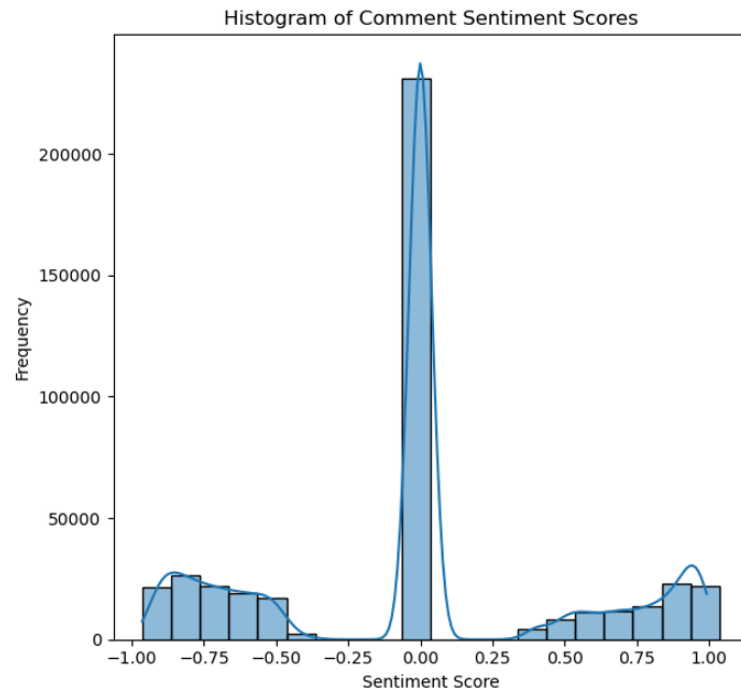
Percentage of likes compared to the first episode



Visualisation

Analyse du sentiment par genre

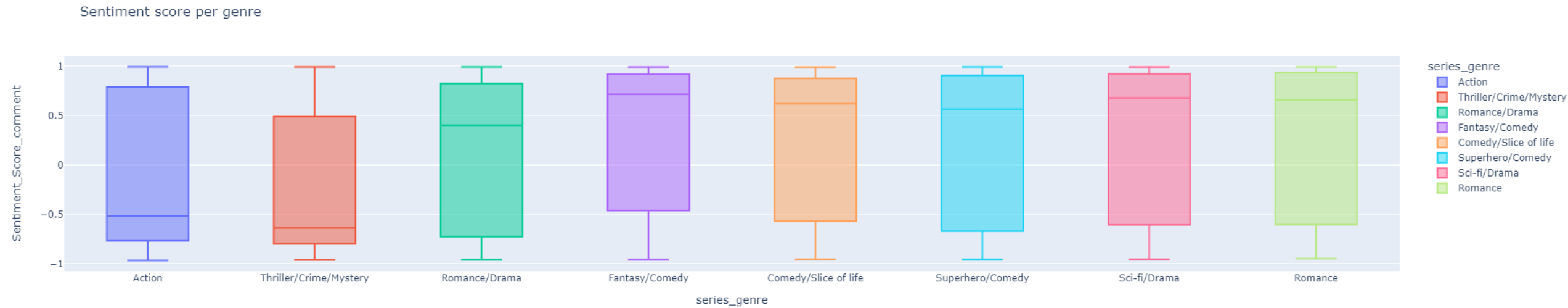
- Drop les commentaires qui ont un sentiment score de 0



Visualisation

Analyse du sentiment par genre

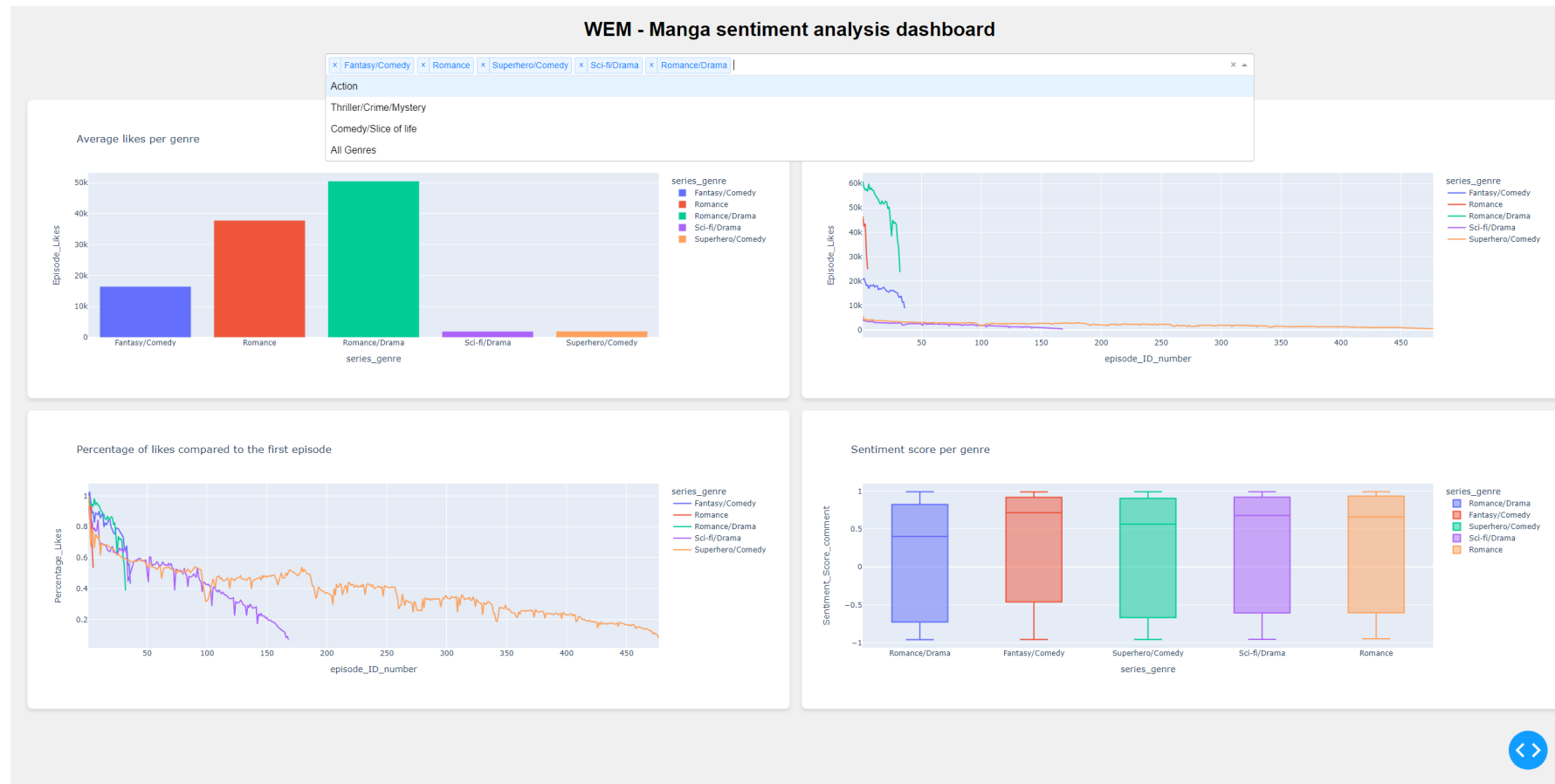
- Création de boxplots à l'aide de plotly



Visualisation

Création du dashboard

- Création de boxplots à l'aide de plotly



Conclusion

- Scraping
 - Très long
 - Code du site peut changer
- Nettoyage des données
 - Assez long d'ouvrir autant de csv avec pandas
- Sentiment analysis
 - Modèle «hugging face» facile à utiliser
 - Rapide avec un GPU
- Rating prédiction
- Visualisation
 - Très intéressant de tirer des graphs à partir d'un simple site web
 - Manque de données pour tirer de vraies conclusions

Pistes d'amélioration

- Général
 - Plus de séries, données obtenues insuffisantes
- Scraping
 - Code plus robuste et plus automatisé
 - Parallélisation du scraping
- Nettoyage de données
 - Utiliser exclusivement polars
- Sentiment analysis
 - Mieux évaluer influence pré-traitement texte
- Prédication
 - Complexe features (time séries)
- Visualisation
 - Créer un index pour quantifier le déclin des likes par épisode

Merci!