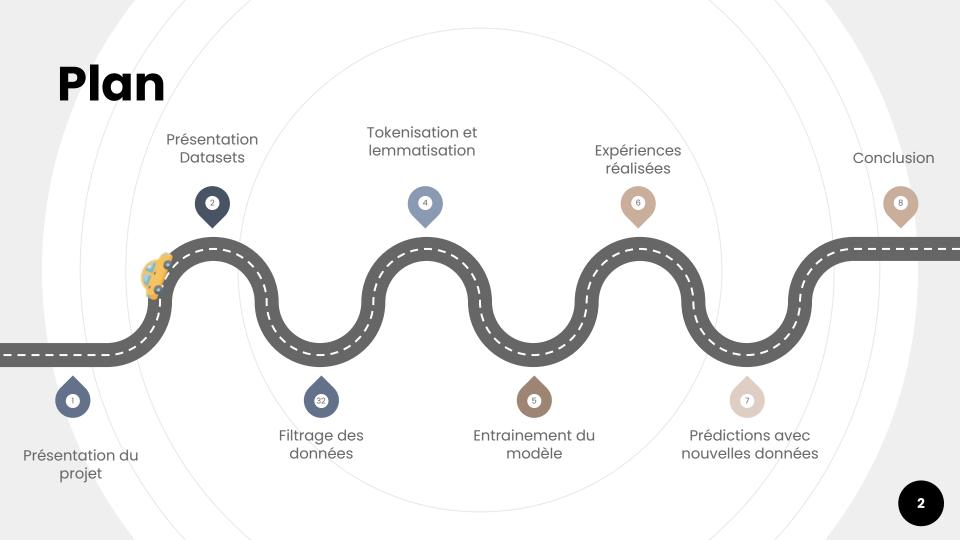


n automatiaue de aenres

Détection automatique de genres de film par IA

Benjamin VALLEIX, Augustin GIRAUDIER



Présentation

Deviner le genre des films à partir de leur synopsis

Objectif:

- Entrée : fournir une liste de synopsis
- Sortie : liste de genres (au départ

mono-label)



Présentation Dataset

Petit

- 5k données
- Champs: titre, synopsis, genres, poster, ...
- Nécessité de filtrer les données (plusieurs synopsis manquants)

Grand

- 50k données
- Champs: titre, synopsis, genres, poster, ...
- Nécessité de filtrer les données (7% données genres manquantes)



Filtrage des données

Filtrage des données

Suppression des lignes sans synopsis ou genres

```
data.dropna(subset=['overview', 'genre'], inplace=True)
```

Garder colonne intéressante

```
sorted_data = data[['original_title', 'overview', 'genre']]
filtered_data = sorted_data[sorted_data['overview'].str.len() >
10]
```

Tokenisation & Lemmatisation

Tokenisation & lemmatisation

- <u>Tokenisation</u>: division d'un texte en unités plus petites appelées "tokens"
- <u>Lemmatisation</u>: La lemmatisation est le processus de réduction des mots à leur forme de base, appelée "lemme".



Tokenisation

tokens = word_tokenize(text)

AVANT	APRES
<pre>['As youngs and naives recruits in Vietnam]</pre>	<pre>['As', 'youngs', 'and', 'naives', 'recruits', 'in', 'Vietnam']</pre>

Lemmatisation

lemmatized_tokens = [lemmatizer.lemmatize(token.lower()) for token in tokens]

AVANT	APRES
<pre>['Twin', 'sisters', 'Emma', 'and', 'Sam', 'come', 'up', 'with' 'a', 'scheme', 'to', 'switch', 'places']</pre>	<pre>['twin', 'sister', 'emma', 'and', 'sam', 'come', 'up', 'with', 'a', 'scheme', 'to', 'switch', 'place']</pre>

1. Divisions des données en ensemble d'entraînement et test

```
X_train, X_test, y_train, y_test =
train_test_split(filtered_data['overview'], filtered_data['genre'],
test_size=0.2, random_state=42)
```

2. Entraiment du modèle Word2Vec

```
model = Word2Vec(sentences=train_tokens, vector_size=100, window=5,
min_count=1, workers=4)
```

Train_tokens = ensemble de données qui est utilisée pour entraîner le modèle sur lequel on a appliqué le prétraitement (tokenisation et lemmatisation)

3. Transformation des données (entrainement + test) en embeddings

```
X_train_embeddings = np.array([np.mean([model.wv[word] for word in doc if
word in model.wv] or [np.zeros(model.vector_size)], axis=0) for doc in
train_tokens])
```

C'est ce qui va permettre ensuite aux modèles d'apprentissage de mieux comprendre la structure des données

C'est X_train_embeddings qui sera ensuite utilisé comme entrée pour le modèle

3. Transformation des données (entrainement + test) en embeddings

```
X_train_embeddings = np.array([np.mean([model.wv[word] for word in doc if
word in model.wv] or [np.zeros(model.vector_size)], axis=0) for doc in
train_tokens])
```

C'est ce qui va permettre ensuite aux modèles d'apprentissage de mieux comprendre la structure des données

C'est X_train_embeddings qui sera ensuite utilisé comme entrée pour le modèle

4. Entraîner le modèle de classification sur l'ensemble d'entraînement

```
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train_embeddings, y_train)
```

- Création d'une instance de RandomForestClassifier avec 100 arbres de décision et une graine aléatoire (42) -> Assurer la reproductibilité des résultats
- Entraînement du modèle (fit) avec les données d'entrainements sous formes d'embeddings

5. Prédictions sur ensemble de test

```
y_pred = clf.predict(X_test_embeddings)
```

Une fois le modèle entraîné (avec RandomForestClassifier), il tente de retrouver le genre des films du dataset Test :

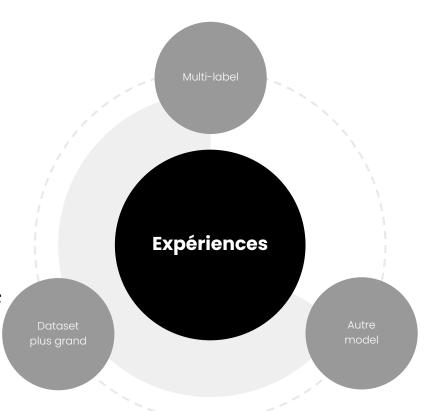
- Y_test : données réelles
- Y_pred : données prédites

C'est grâce à ceci que l'on peut étudier la performance du modèle. (rapport et matrice de confusion)



Expériences réalisées

 Utilisation de matrices de confusions pour modéliser les résultats



Expérience 1:

									М	atric	e de	Con	fusio	n									
	Action -	39	4	0	44	0	0	62	0	0	0	0	5	0	0	0	0	0	1	0	0	-	120
	Adventure -	20	2	0	23	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Animation -	4	0	0	8	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Comedy -	23	1	0	75	1	0	87	0	1	0	0	0	0	0	0	0	0	0	0	0	=	100
	Crime -	8	1	0	12	1	0	23	0	0	0	0	1	0	0	0	1	0	0	0	0		
	Documentary -	3	1	0	4	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Drama -	33	3	0	93	0	0	122	0	1	0	0	1	0	0	0	0	0	0	0	0		80
	Family -	4	0	0	8	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0		80
tes	Fantasy -	5	0	0	5	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0		
Vraies étiquettes	Foreign -	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
s éti	History -	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	=	60
/raie	Horror -	11	0	0	28	0	0	22	0	0	0	0	1	0	0	0	0	0	0	0	0		
	Music -	0	0	0	6	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Mystery -	1	0	0	6	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	-	40
	Romance -	3	0	0	14	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Science Fiction -	6	0	0	6	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0		
	TV Movie -	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	_	20
	Thriller -	2	0	0	12	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0		55
	War -	1	0	0	2	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Western -	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		- 00
		Action -	dventure -	nimation -	Comedy -	Crime -	mentary -	Drama -	Family -	Fantasy -	Foreign -	History -	Horror -	Music -	Mystery -	Romance -	e Fiction -	TV Movie -	Thriller -	War -	Western -	-	0

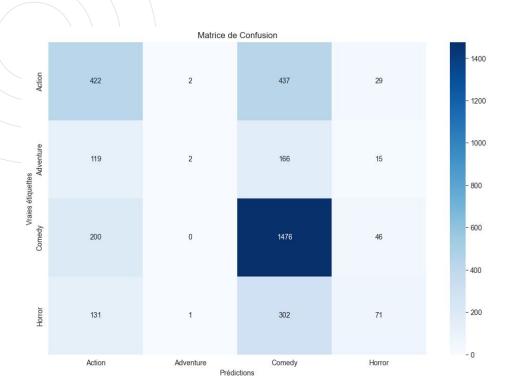
Rapport de class	ification :			
	precision	recall	f1-score	support
Action	0.24	0.25	0.25	155
Adventure	0.17	0.03	0.04	78
Animation	0.00	0.00	0.00	18
Comedy	0.21	0.40	0.28	188
Crime	0.33	0.02	0.04	47
Documentary	0.00	0.00	0.00	15
Drama	0.29	0.48	0.37	253
Family	0.00	0.00	0.00	16
Fantasy	0.00	0.00	0.00	19
Foreign	0.00	0.00	0.00	1
History	0.00	0.00	0.00	2
Horror	0.12	0.02	0.03	62
Music	0.00	0.00	0.00	8
Mystery	0.00	0.00	0.00	13
Romance	0.00	0.00	0.00	22
Science Fiction	0.00	0.00	0.00	18
TV Movie	0.00	0.00	0.00	1
Thriller	0.00	0.00	0.00	31
War	0.00	0.00	0.00	6
Western	0.00	0.00	0.00	2
accuracy			0.25	955
macro avg	0.07	0.06	0.05	955
weighted avg	0.20	0.25	0.20	955

Expérience 2 : Dataset plus grand

								М	atric	e de	Con	fusio	n								
Action -	329	4	0	177	0	37	355	0	0	0	0	16	0	0	0	0	0	0	0	0	- 1600
Adventure -	63	1	0	54	0	15	146	0	0	0	0	0	0	0	0	0	0	0	0	0	
Animation -	30	0	0	60	0	24	95	0	0	0	0	4	0	0	0	0	0	0	0	0	- 1400
Comedy -	82	0	0	696	0	85	851	0	0	0	0	5	0	0	0	0	0	0	0	0	1400
Crime -	80	0	0	68	0	11	192	0	0	0	0	6	0	0	0	0	0	0	0	0	
Documentary -	11	1	0	84	0	346	201	0	0	0	0	2	0	0	0	0	0	0	0	0	- 1200
Drama -	130	0	2	415	0	93	1635	0	0	0	0	17	0	0	0	0	0	1	0	0	
Family -	13	0	0	38	0	8	55	0	0	0	0	0	0	0	0	0	0	0	0	0	- 1000
Fantasy -	17	0	0	27	0	5	76	0	0	0	0	4	0	0	0	0	0	0	0	0	
Foreign -	1	0	0	7	0	0	10	0	0	0	0	1	0	0	0	0	0	0	0	0	
History -	9	0	0	6	0	21	32	0	0	0	0	0	0	0	0	0	0	0	0	0	- 800
Horror -	97	0	0	104	0	15	288	0	0	0	0	46	0	0	0	1	0	1	0	0	
Music -	3	0	0	26	0	31	47	0	0	0	0	0	0	0	0	0	0	0	0	0	- 600
Mystery -	8	0	0	21	0	3	67	0	0	0	0	3	0	0	0	0	0	0	0	0	
Romance -	8	0	0	66	0	5	156	0	0	0	0	3	0	0	0	0	0	0	0	0	
Science Fiction -	51	1	0	21	0	12	46	0	0	0	0	10	0	0	0	1	0	0	0	0	- 400
TV Movie -	12	0	0	19	0	6	44	0	0	0	0	2	0	0	0	0	0	0	0	0	
Thriller -	58	0	0	67	0	8	184	0	0	0	0	8	0	0	0	0	0	0	0	0	- 200
War -	19	0	0	6	0	9	37	0	0	0	0	0	0	0	0	0	0	0	0	0	
Western -	27	0	0	16	1	1	44	0	0	0	0	1	0	0	0	0	0	0	0	0	
	Action -	dventure -	nimation -	Comedy -	Crime -	mentary -	Drama -	Family -	Fantasy -	Foreign -	History -	Horror -	Music -	Mystery -	Romance -	ce Fiction -	TV Movie -	Thriller -	War -	Western -	- 0

Rapport de class	ification :				
	precision	recall	f1-score	support	
Action	0.31	0.36	0.33	918	
Adventure	0.14	0.00	0.01	279	
Animation	0.00	0.00	0.00	213	
Comedy	0.35	0.40	0.38	1719	
Crime	0.00	0.00	0.00	357	
Documentary	0.47	0.54	0.50	645	
Drama	0.36	0.71	0.48	2293	
Family	0.00	0.00	0.00	114	
Fantasy	0.00	0.00	0.00	129	
Foreign	0.00	0.00	0.00	19	
History	0.00	0.00	0.00	68	
Horror	0.36	0.08	0.14	552	
Music	0.00	0.00	0.00	107	
Mystery	0.00	0.00	0.00	102	
Romance	0.00	0.00	0.00	238	
Science Fiction	0.50	0.01	0.01	142	
TV Movie	0.00	0.00	0.00	83	
Thriller	0.00	0.00	0.00	325	
War	0.00	0.00	0.00	71	
Western	0.00	0.00	0.00	90	
accuracy			0.36	8464	
macro avg	0.12	0.11	0.09	8464	
weighted avg	0.28	0.36	0.29	8464	

Expérience 3 : exclusion des classes faibles



Rapport de cla	ssification :			
	precision	recall	f1-score	support
Action	0.48	0.47	0.48	890
Adventure	0.40	0.01	0.01	302
Comedy	0.62	0.86	0.72	1722
Horror	0.44	0.14	0.21	505
accuracy			0.58	3419
macro avg	0.49	0.37	0.36	3419
weighted avg	0.54	0.58	0.52	3419

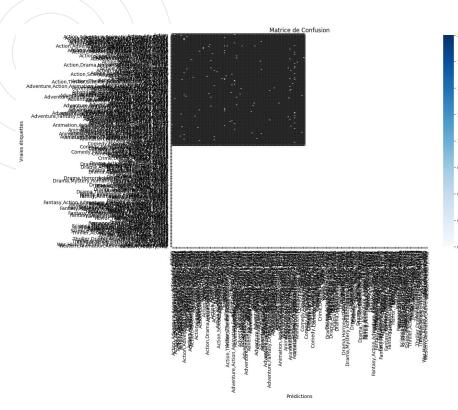
Expérience 4 : test du model FastText

									1	Matri	ce de	Con	fusio	n									40
	Action	232	1	0	200	1	38	434	0	0	0	0	12	0	0	0	0	0	0	0	0		- 16
	Adventure	46	1	0	65	0	14	149	0	0	0	0	4	0	0	0	0	0	0	0	0		
	Animation	26	0	0	72	0	25	89	0	0	0	0	1	0	0	0	0	0	0	0	0		- 14
	Comedy	77	0	0	656	1	95	885	0	0	0	0	5	0	0	0	0	0	0	0	0		
	Crime	35	0	0	77	1	9	232	0	0	0	0	3	0	0	0	0	0	0	0	0		
	Documentary	9	0	0	77	0	324	234	0	0	0	0	0	0	0	0	0	0	1	0	0		- 12
	Drama	111	1	0	431	1	135	1604	0	0	0	0	8	0	0	0	0	0	2	0	0		
	Family	9	0	0	42	1	5	57	0	0	0	0	0	0	0	0	0	0	0	0	0		- 10
Se	Fantasy	14	0	0	28	0	4	81	0	0	0	0	2	0	0	0	0	0	0	0	0		
Vraies étiquettes	Foreign	1	0	0	4	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0		
ss etic	History	9	0	0	5	0	18	36	0	0	0	0	0	0	0	0	0	0	0	0	0		- 80
Vraie	Horror	80	0	0	108	0	25	311	0	0	0	0	28	0	0	0	0	0	0	0	0		
	Music	1	1	0	31	0	26	48	0	0	0	0	0	0	0	0	0	0	0	0	0		- 6
	Mystery	6	0	0	20	0	5	71	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Romance	4	0	0	59	0	4	170	0	0	0	0	1	0	0	0	0	0	0	0	0		
	Science Fiction	34	0	0	34	0	11	58	0	0	0	0	4	0	0	0	1	0	0	0	0		- 4
	TV Movie	8	0	0	22	0	6	47	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Thriller	49	0	0	75	0	8	185	0	0	0	0	8	0	0	0	0	0	0	0	0		- 2
	War	11	0	0	8	0	7	45	0	0	0	0	0	0	0	0	0	0	0	0	0		
	Western	23	0	0	16	0	1	50	0	0	0	0	0	0	0	0	0	0	0	0	0		
		c	Ф	<u>_</u>	>	9	>	a	λ	λ.	_	>	70	2	>	Ф	U	e	-	ar	E		- 0
		Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	Family	Fantasy	Foreign	History	Horror	Music	Mystery	Romance	cience Fiction	TV Movie	Thriller	War	Western		

паррог с че стазз.	ification :		Ca.	
	precision	recall	f1-score	support
Action	0.30	0.25	0.27	918
Adventure	0.25	0.00	0.01	279
Animation	0.00	0.00	0.00	213
Comedy	0.32	0.38	0.35	1719
Crime	0.20	0.00	0.01	357
Documentary	0.43	0.50	0.46	645
Drama	0.33	0.70	0.45	2293
Family	0.00	0.00	0.00	114
Fantasy	0.00	0.00	0.00	129
Foreign	0.00	0.00	0.00	19
History	0.00	0.00	0.00	68
Horror	0.37	0.05	0.09	552
Music	0.00	0.00	0.00	107
Mystery	0.00	0.00	0.00	102
Romance	0.00	0.00	0.00	238
Science Fiction	1.00	0.01	0.01	142
TV Movie	0.00	0.00	0.00	83
Thriller	0.00	0.00	0.00	325
War	0.00	0.00	0.00	71
Western	0.00	0.00	0.00	90

Prédiction

Expérience 5 : tentative de multiclasse



Rapport de classification :				
	precision	recall	f1-score	support
Action, Adventure	0.00	0.00	0.00	
Action,Adventure,Animation,Family	0.00	0.00	0.00	0
Action, Adventure, Comedy, Drama, Mystery	0.00	0.00	0.00	
Action, Adventure, Comedy, Science Fiction	0.00	0.00	0.00	
Action, Adventure, Comedy, Science Fiction, Western	0.00	0.00	0.00	
Action,Adventure,Crime	0.00	0.00	0.00	
Action, Adventure, Crime, Fantasy, Science Fiction	0.00	0.00	0.00	
Action, Adventure, Drama	0.00	0.00	0.00	
Action, Adventure, Family, Fantasy	0.00	0.00	0.00	
Action,Adventure,Fantasy	0.33	0.50	0.40	
Action, Adventure, Fantasy, Science Fiction	0.00	0.00	0.00	
Action,Adventure,Fantasy,Thriller	0.00	0.00	0.00	
Action,Adventure,Science Fiction	0.00	0.00	0.00	
Action,Adventure,Thriller	0.00	0.00	0.00	
Action,Comedy,Drama,Thriller	0.00	0.00	0.00	
Action,Comedy,Science Fiction	0.00	0.00	0.00	
Action,Comedy,Thriller	0.00	0.00	0.00	
Action,Crime,Drama,Thriller	0.00	0.00	0.00	
Action, Crime, Fantasy	0.00	0.00	0.00	
Action,Crime,Science Fiction,Thriller	0.00	0.00	0.00	
Action, Drama, Adventure	0.00	0.00	0.00	0
Action, Drama, Mystery, Thriller	0.00	0.00	0.00	
,.,.,.,.,.,,,,				

Prédiction avec données utilisateur

"A spaceship travels to distant galaxies."	"A romantic drama set in Paris."
'Comedy'	'Drama'
['Adventure', 'Drama', 'Actions']	['Action', 'Adventure', 'Science Fiction']





1 Multi-label essentiel

2 Sous classes

3 Dataset plus important