

# Textual Big Data : Application du "Deep Learning" au traitement des langues naturelles

Projet Option INFO - Ecole Centrale de Lyon  
Année 2018/2019

ADIM Mehdi

BABINET Pierre

DARTIGUENAVE Aliénor

KUOCH Thomas

# Plan

1

Equipe, Contexte, Objectifs, Livrables

2

Projet et Evaluation

3

Perspectives

4

Conclusion

# Organisation de l'équipe

---



Alexandre SAIDI  
Tuteur



Thomas KUOCH  
Chef de projet



ADIM Mehdi  
Responsable Données



BABINET Pierre  
Responsable Traitement  
de données



DARTIGUENAVE Aliénor  
Responsable Traitement  
de données et Livrables



Réunion d'équipe tous les jeudis



Points d'avancements tous les 15 jours avec M.Saidi

# Contexte et objectifs

---

Massification des données

Masse de documents numériques

Clustering, Classification ...

**L'objectif : utiliser des techniques du Text Mining et implémenter des approches de traitement automatique du langage naturel pour finalement classifier et arpenter des données textuelles.**

# Livrables

---



Dépôt GitHub:  
Codes, Données, README.md

# Plan

1

Equipe, Contexte, Objectifs, Livrables

2

Projet et Evaluation

3

Perspectives

4

Conclusion

# Organisation du projet



**Choisir les  
données**



Traiter et nettoyer  
les données



Créer un modèle de  
prédiction

# Enjeux du projet

Sur IMDb, la catégorisation des films selon le genre se fait manuellement en se basant sur les suggestions des utilisateurs.



- Manque d'expérience humaine.
- L'erreur humaine.
- Attente de plusieurs suggestions pour conclure.
- Difficulté de catégoriser les nouveaux films.



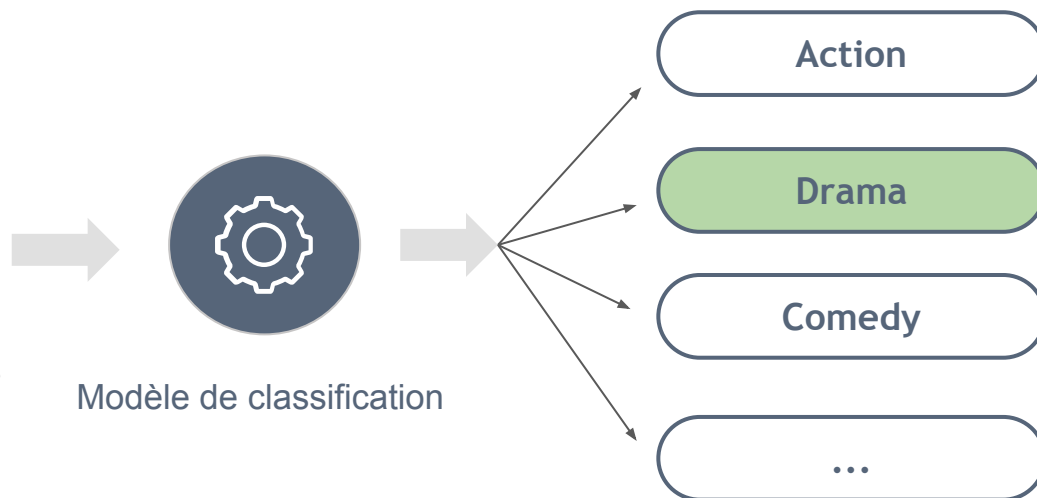
# Application choisie

## Classification des films par genre d'après leur synopsis

### Film : 'Germinal'

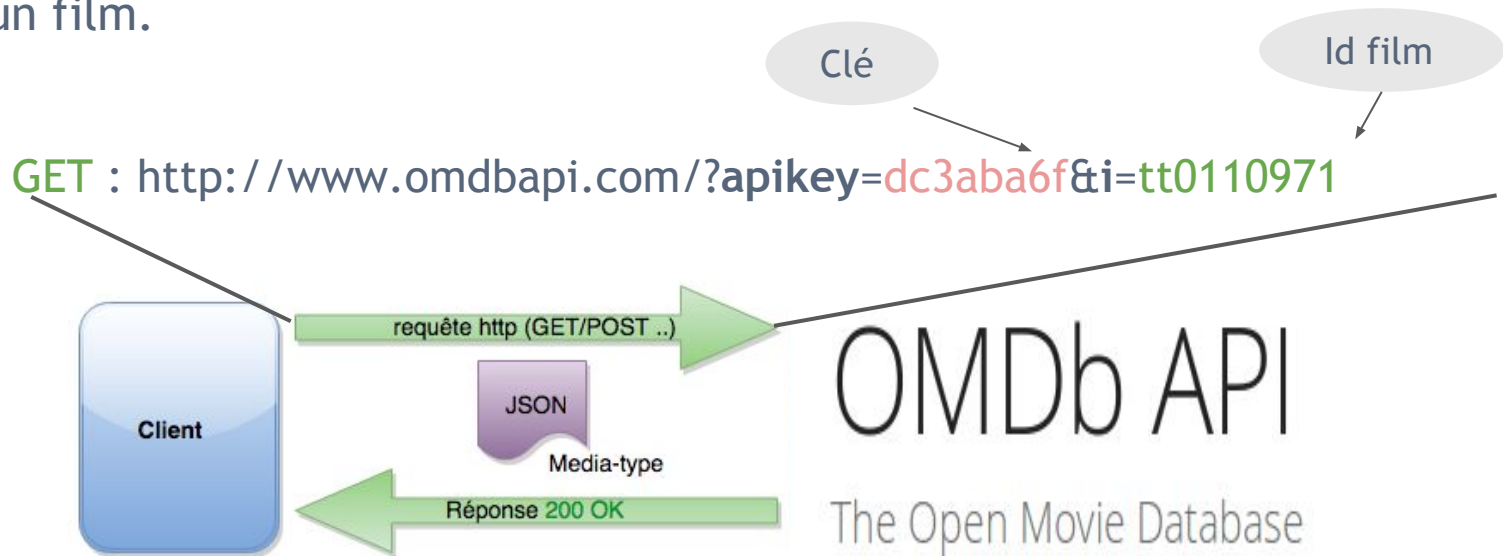
#### *Synopsis*

It's mid 19th century, north of France. The story of a coal miner's town. They are exploited by the mine's owner. One day they decide to go on strike, and then the authorities repress them.



# Collection de données

Utilisation d'une API Restful pour récupérer toutes les informations disponibles sur un film.



# Objet JSON retourné par l'API

| JSON        | Raw Data  | Headers |
|-------------|---|---------|
| Save        | Copy  |         |
| Title:      | "Renaissance Man"   |         |
| Year:       | "1994"  |         |
| Rated:      | "PG-13"   |         |
| Released:   | "03 Jun 1994"   |         |
| Runtime:    | "128 min"   |         |
| Genre:      | "Comedy, Drama"   |         |
| Director:   | "Penny Marshall"  |         |
| Writer:     | "Jim Burnstein"   |         |
| ▶ Actors:   | "Danny DeVito, Gregory Hines, Remar, Ed Begley Jr."   |         |
| ▼ Plot:     | "A failed businessman is hired by the army to teach a group of underachieving recruits in order to help them pass basic training."  |         |
| Language:   | "English"   |         |
| Country:    | "USA"   |         |
| Awards:     | "N/A"   |         |
| ▶ Poster:   | " <a href="https://m.media-amazon.com/images/M/MV5BMTUwNjY0MTUwOV5BMl5BanBnXkFtZTcwMjA3MDA@._V1_SX300.jpg">https://m.media-amazon.com/images/M/MV5BMTUwNjY0MTUwOV5BMl5BanBnXkFtZTcwMjA3MDA@._V1_SX300.jpg</a> " |         |
| ▶ Ratings:  | [...]   |         |
| Metascore:  | "44"  |         |
| imdbRating: | "6.2"   |         |
| imdbVotes:  | "15,587"  |         |
| imdbID:     | "tt0110971"   |         |
| Type:       | "movie"   |         |
| DVD:        | "01 Jul 2003"   |         |
| BoxOffice:  | "N/A"   |         |
| Production: | "Buena Vista"   |         |
| Website:    | "N/A"   |         |


# Sélection et Fusion

- Sélection des attributs nécessaires (Synopsis, Genre)
- Fusion des données dans un seul fichier csv dont la structure est la suivante.

4000  
synopsis

19  
genres

## Caractéristiques sélectionnées :



| Title                     | Synopsis  | Genre1    | Genre2 | Genre3  |
|---------------------------|---|-----------|--------|---------|
| Renaissance Man           | A down-on-his-luck businessman desperately takes the only job he can get          | Comedy    | Drama  |         |
| Rising Sun                | At the offices of a Japanese corporation, during a party, a woman is killed       | Action    | Crime  | Drama   |
| The Road to Wellville     | A story about the ins and outs of one unusual health facility in the 1950s        | Comedy    | Drama  |         |
| RoboCop 3                 | The mega corporation Omni Consumer Products is still bent on controlling the city | Action    | Crime  | Sci-Fi  |
| Robin Hood: Men in Yellow | The standard story of Robin Hood: Evil Prince John is oppressing the people       | Adventure | Comedy | Musical |
| Romeo Is Bleeding         | Detective Jack Grimaldi (Gary Oldman) takes us through his last case              | Action    | Crime  | Drama   |
| Romper Stomper            | Nazi skinheads in Melbourne take out their anger on local Vietnamese              | Crime     | Drama  |         |

# Nettoyage de données

---



Suppression des films qui ne contiennent pas de genre.



Suppression des tags HTML (BeautifulSoup library).



Suppression des nombres, ponctuation, et caractères spéciaux (regex).

# Organisation du projet



**Choisir les  
données**

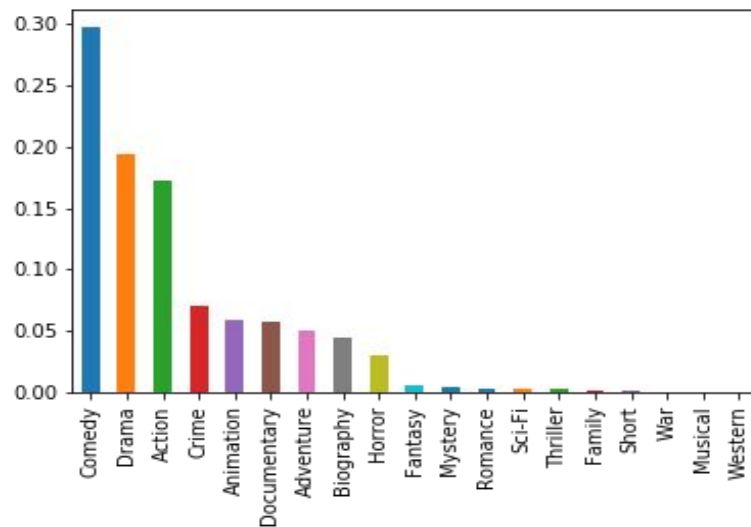


Traiter et nettoyer  
les données



Créer un modèle de  
prédiction

# Sélection des catégories (1/2)



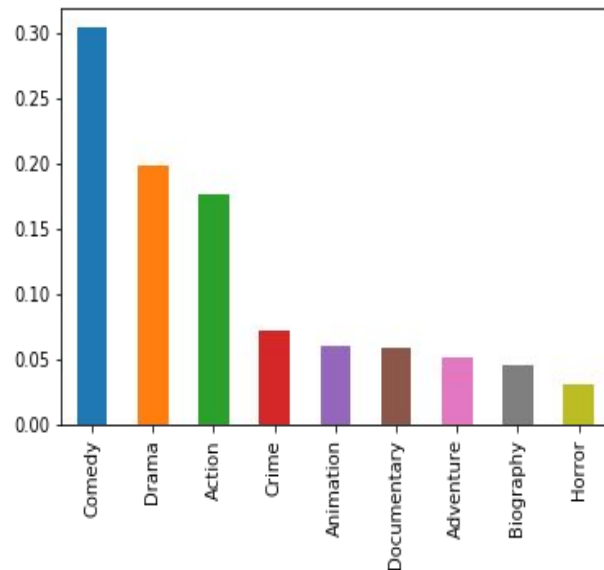
**Données déséquilibrées** : 3 genres dominants (Comedy, Drama et Action)

**Choix** : Ne garder que les catégories dont le nombre d'éléments est supérieur à 100.

|             |      |
|-------------|------|
| Comedy      | 1189 |
| Drama       | 774  |
| Action      | 688  |
| Crime       | 282  |
| Animation   | 237  |
| Documentary | 229  |
| Adventure   | 203  |
| Biography   | 179  |
| Horror      | 121  |
| Fantasy     | 24   |
| Mystery     | 20   |
| Romance     | 14   |
| Sci-Fi      | 12   |
| Thriller    | 9    |
| Family      | 8    |
| Short       | 7    |
| War         | 2    |
| Musical     | 1    |
| Western     | 1    |

Name: Genre1, dtype: int64

## Sélection des catégories (2/2)



|             |          |
|-------------|----------|
| Comedy      | 0.304716 |
| Drama       | 0.198360 |
| Action      | 0.176320 |
| Crime       | 0.072271 |
| Animation   | 0.060738 |
| Documentary | 0.058688 |
| Adventure   | 0.052025 |
| Biography   | 0.045874 |
| Horror      | 0.031010 |

Après sélection, il nous reste **3902 films**, ce qui correspond à **97%** du dataset  
La classe majoritaire est "Comedy" avec 30% des films: l'accuracy de **0.30** sera le **score 0R** pour nos **modèles de prédiction**



# Nettoyer les données

---

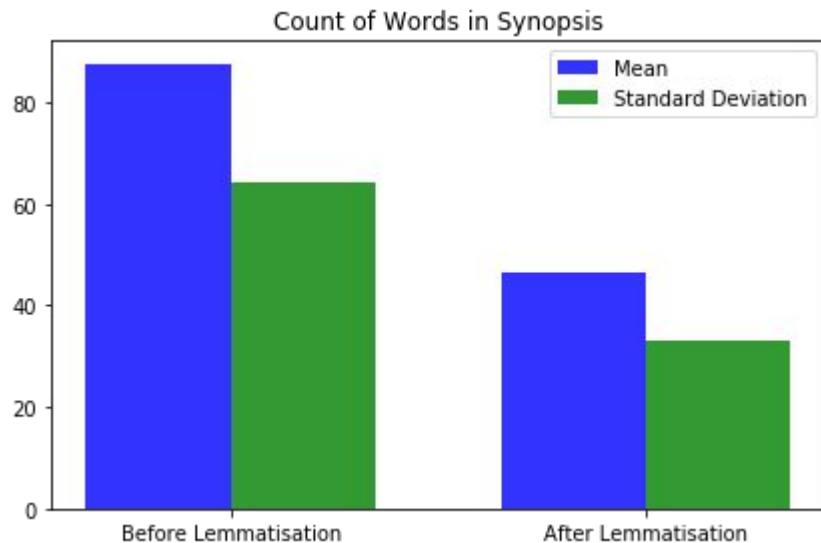
Pour chaque synopsis, on effectue le traitement suivant, avec utilisation de la **librairie Python NLTK**:

*Exemple : “Le chien mange un os et les oiseaux mangent des vers”*

- 1 On enlève les “**stopwords**” de la langue : *“Chien mange os oiseaux mangent vers”*
- 2 On **lemmatise** : *“Chien manger os oiseaux manger vers”*

- ⇒ Réduction du nombre de mots du corpus
- ⇒ Conservation des mots significatifs

# Nettoyer les données



Division par 2 du nombre de mots en moyenne

# Word embedding par le **Lexicon**

## Création de la matrice d'apprentissage

|         | Mot 1 | Mot 2 | Mot 3 | ... |  |  | Mot N |
|---------|-------|-------|-------|-----|--|--|-------|
| Genre 1 | 0     | 2     | 1     |     |  |  | 1     |
| Genre 2 | 2     | 1     | 0     |     |  |  | 1     |
| Genre 3 |       |       |       |     |  |  |       |

Synopsis 1 : “ Mot 1 , Mot 2 , Mot 1 , Mot N “ de Genre 2

Synopsis 2 : “ Mot 2 , Mot 2 , Mot 3 , Mot N “ de Genre 1

# Word embedding par le **Lexicon**

## Classification d'un synopsis

Nouveau Synopsis : "Mot 1, Mot 3, Mot N, Mot 3"

|         | Mot 1 | Mot 2 | Mot 3 | ... |  |  | Mot N | Somme | Somme normalisée |
|---------|-------|-------|-------|-----|--|--|-------|-------|------------------|
| Genre 1 | 156   | 32    | 1     |     |  |  | 12    | 170   | 0.84             |
| Genre 2 | 53    | 128   | 31    |     |  |  | 63    | 143   | 0.52             |
| Genre 3 | 42    | 138   | 78    |     |  |  | 79    | 278   | 0.82             |

⇒ Conclusion : Le Synopsis 1 correspond à un film de genre 3 (somme) ou 1 (somme normalisée)

# Word Embedding par TF-IDF

Objectif : Sélection des mots significatifs du texte

Corpus (tiré d'œuvres de Friedrich Gottlieb Klopstock)<sup>2</sup>

| Document 1  | Document 2   | Document 3  |
|---|--|---|
| Son nom est célébré par le bocage <b>qui</b> frémit, et par le ruisseau <b>qui</b> murmure, les vents l'emportent jusqu'à l'arc céleste, l'arc de grâce et de consolation que sa main tendit dans les nuages. | À peine distinguait-on deux butts à l'extrémité de la carrière : des chênes ombrageaient l'un, autour de l'autre des palmiers se dessinaient dans l'éclat du soir. | Ah ! le beau temps de mes travaux poétiques ! les beaux jours que j'ai passés près de toi ! Les premiers, inépuisables de joie, de paix et de liberté ; les derniers, empreints d'une mélancolie <b>qui</b> eut bien aussi ses charmes. |

|        |  |                  |
|--------|--|------------------|
| TF     | $\frac{\text{Occurrence du mot}}{\text{Nombre de mots dans le document}}$              | $= \frac{2}{38}$ |
| IDF    | $\frac{\text{Nombre de Documents du Corpus}}{\text{Nb de document comportant le mot}}$ | $= \frac{3}{2}$  |
| TF-IDF | $\text{TF} * \log(\text{IDF})$   | $= 0.0092$       |



On obtient pour chaque mot une **importance** dans chaque document

# Word Embedding par TF-IDF

Matrice d'apprentissage :

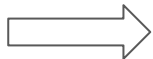
|    | 22    | 23    | 24    | 25    | 26    |
|----|-------|-------|-------|-------|-------|
| 15 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 16 | 0.000 | 0.000 | 0.000 | 0.421 | 0.000 |
| 17 | 0.000 | 0.000 | 0.000 | 0.000 | 0.281 |
| 18 | 0.288 | 0.000 | 0.260 | 0.000 | 0.000 |
| 19 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Dracula : ['help', 'vampire', 'count', 'meet', 'castle', 'dracula', 'mentor', 'guard', 'long',  
'young']

# Word embedding par Spacy

## Synopsis d'Hercules lemmatisé

hercul son greek god zeus turn  
half god half mortal evil hade  
god underworld plan overthrow  
zeus hercule raise earth retain  
god like strength discover  
immortal heritage zeus tell  
return mount olympus true [...]



## Document Embedding par Spacy (vecteurs de taille 128)

|        |     |
|--------|-----|
| 1.345  | 128 |
| -2.450 |     |
| 0.435  |     |
| 0.189  |     |
| ...    |     |

# Organisation du projet



**Choisir les  
données**



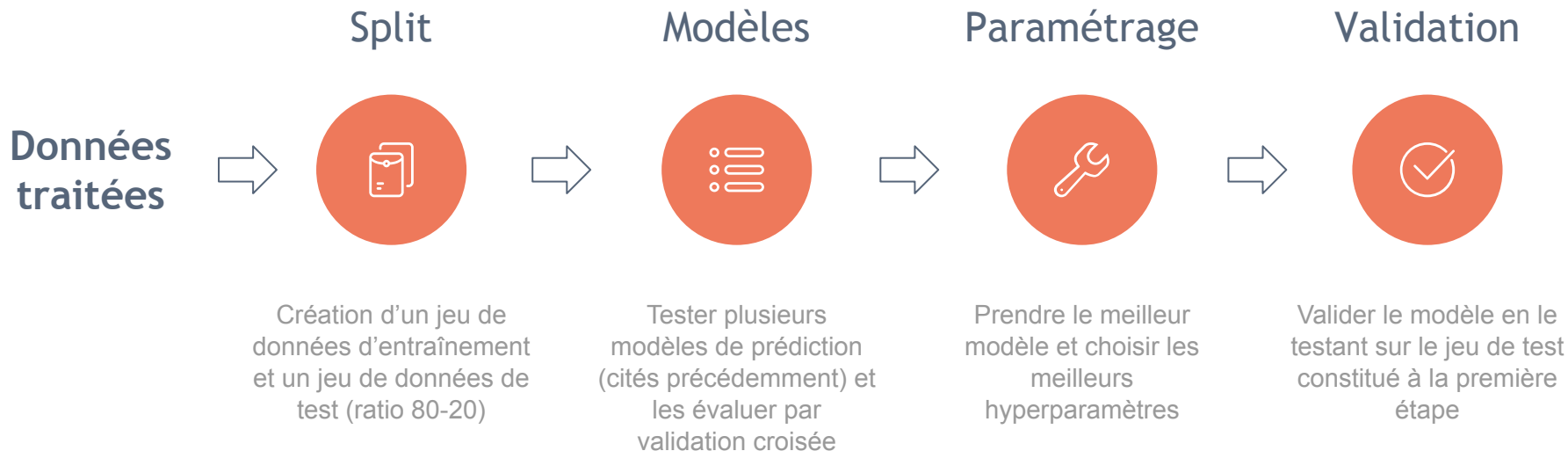
Traiter et nettoyer  
les données



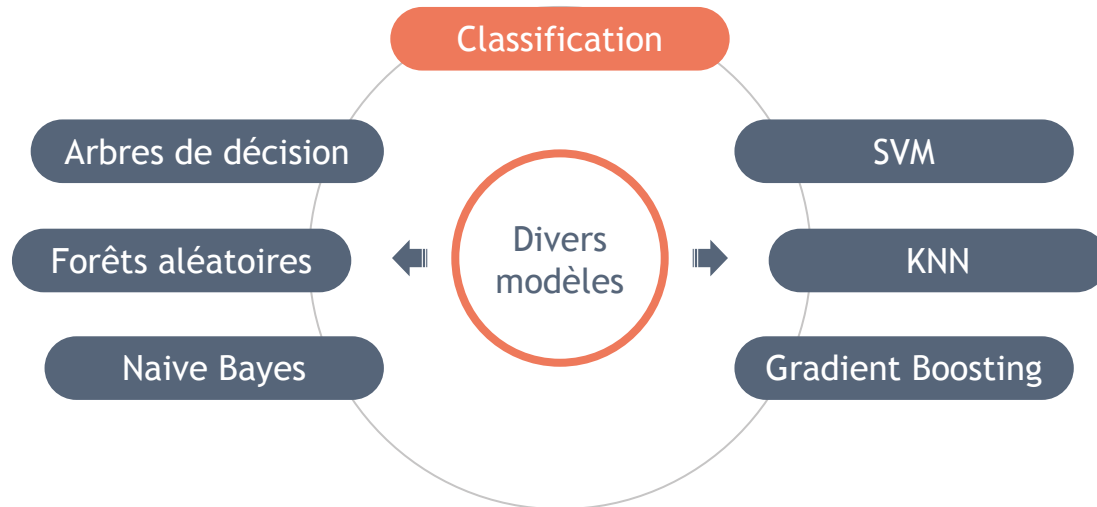
Créer un modèle de  
prédiction



# Entraînement du modèle de prédiction

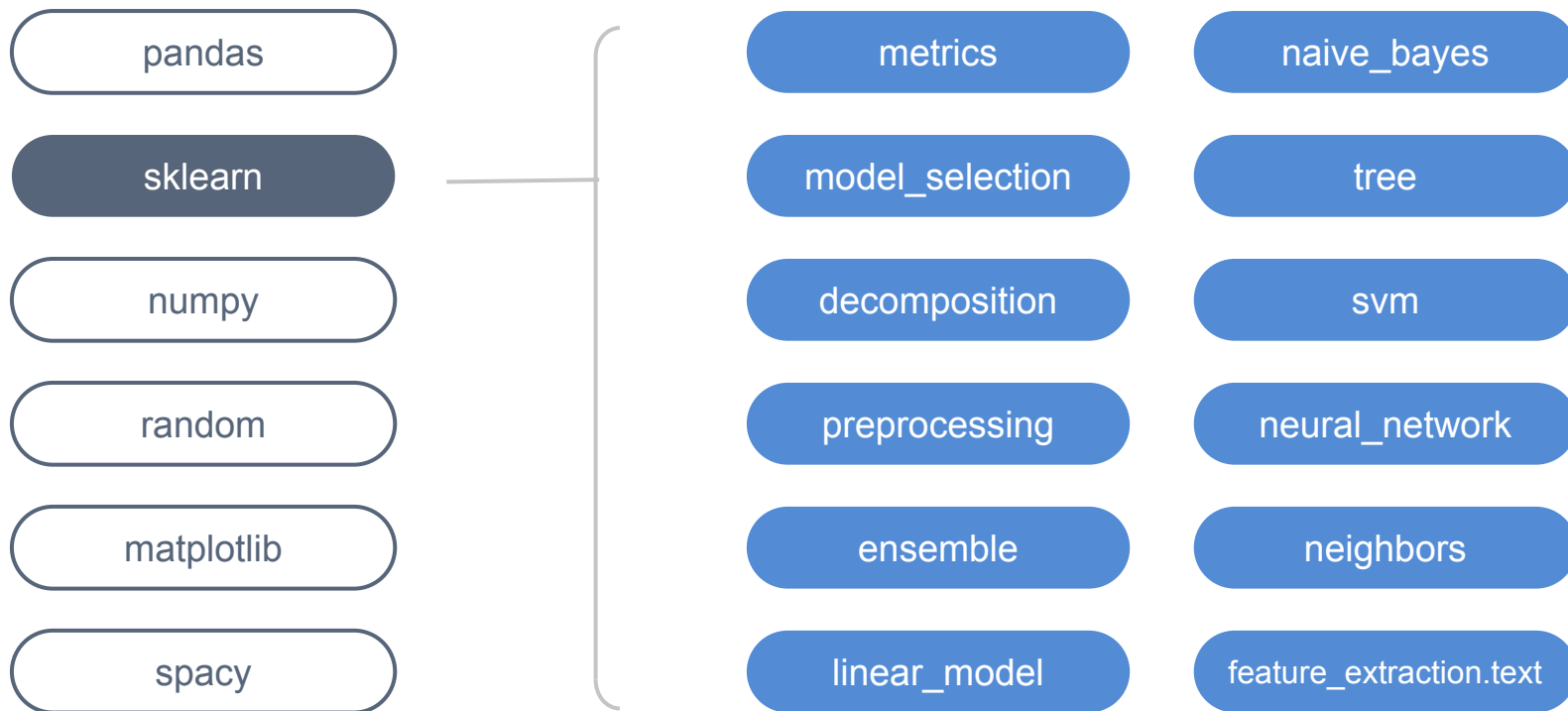


# Modèles de classification

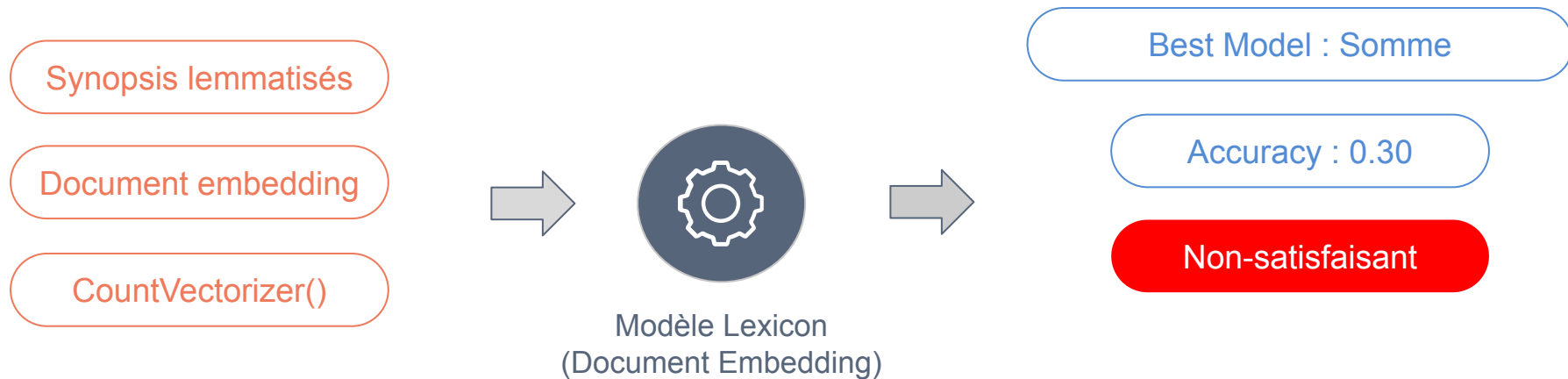


- Les données sont numérisées
- On va tenir compte de la disproportion des catégories en donnant le dictionnaire de poids au modèle
- Pour évaluer le modèle; on regardera la précision pour chacune des catégories et l'accuracy globale

# Utilisation de **scikit-learn**

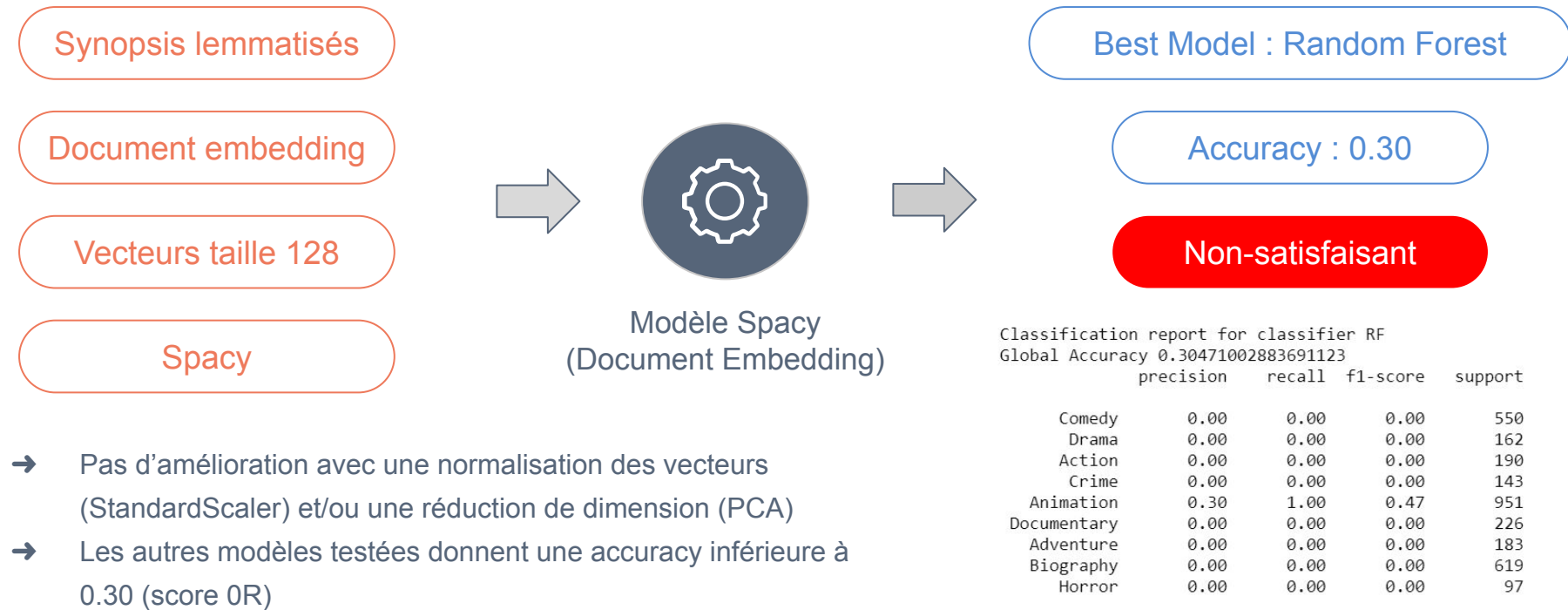


# Modèle **Lexicon**



→ **Remarque:** Résultats initiaux fortement biaisés par la distribution de la base de données

# Modèle Spacy



# Modèle TfIdf

Synopsis lemmatisés

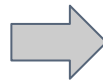
Matrice TfIdf

17300 mots

TfidfVectorizer()



Modèle TfIdf



Best Model : LinearSVM

Accuracy : 0.49

Le plus satisfaisant

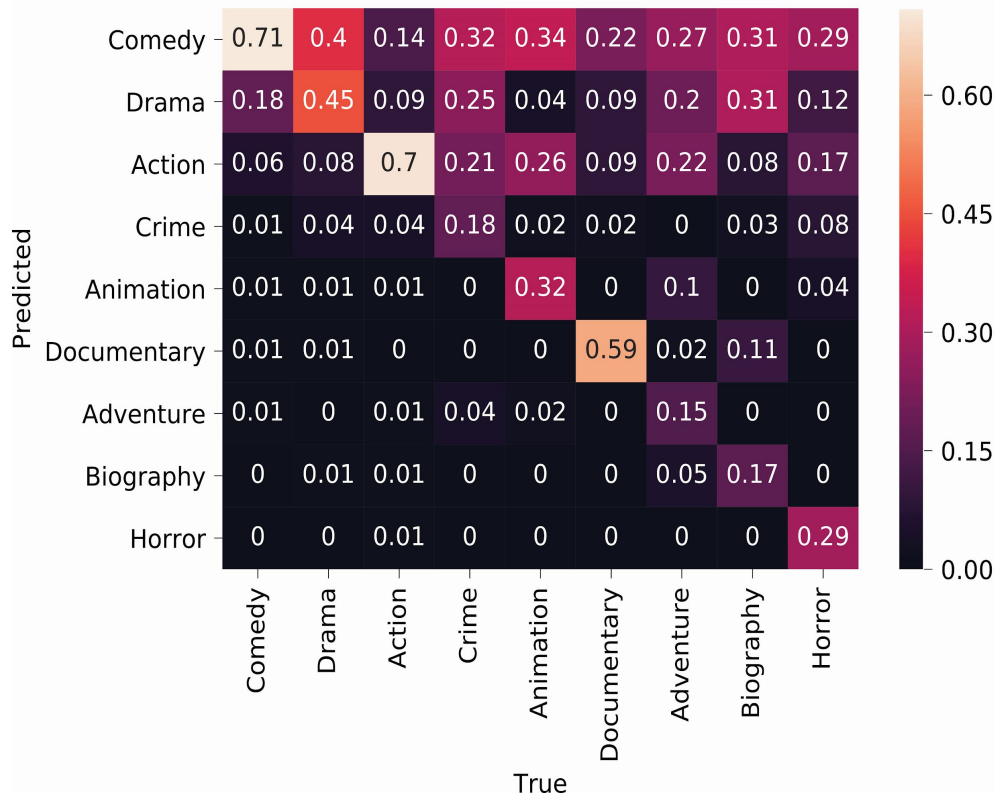
→ C'est le résultat le plus satisfaisant obtenu, même après recherche des meilleurs paramètres

Classification report for LinearSVM

Global Accuracy 0.48958667093880165

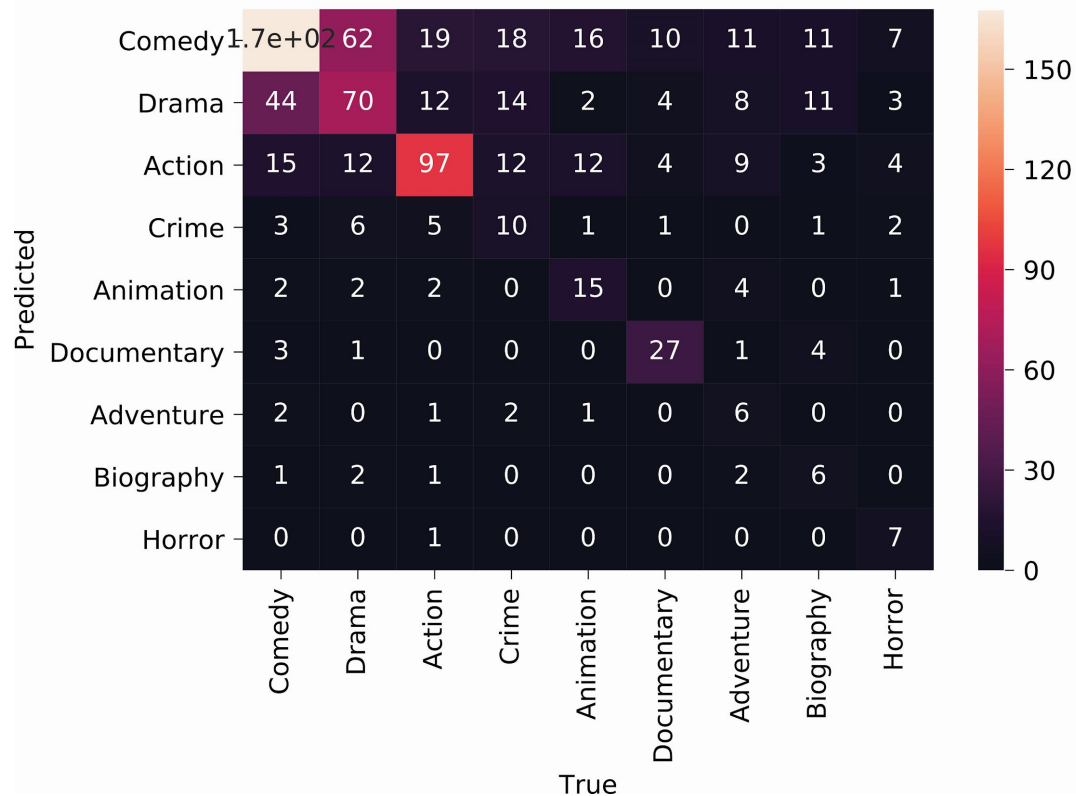
|             | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| Comedy      | 0.52      | 0.70   | 0.60     | 550     |
| Drama       | 0.54      | 0.09   | 0.15     | 162     |
| Action      | 0.52      | 0.19   | 0.28     | 190     |
| Crime       | 0.39      | 0.05   | 0.09     | 143     |
| Animation   | 0.49      | 0.81   | 0.61     | 951     |
| Documentary | 0.38      | 0.12   | 0.19     | 226     |
| Adventure   | 0.70      | 0.59   | 0.64     | 183     |
| Biography   | 0.38      | 0.27   | 0.32     | 619     |
| Horror      | 0.75      | 0.15   | 0.26     | 97      |

# Matrice de confusion (normalisée)



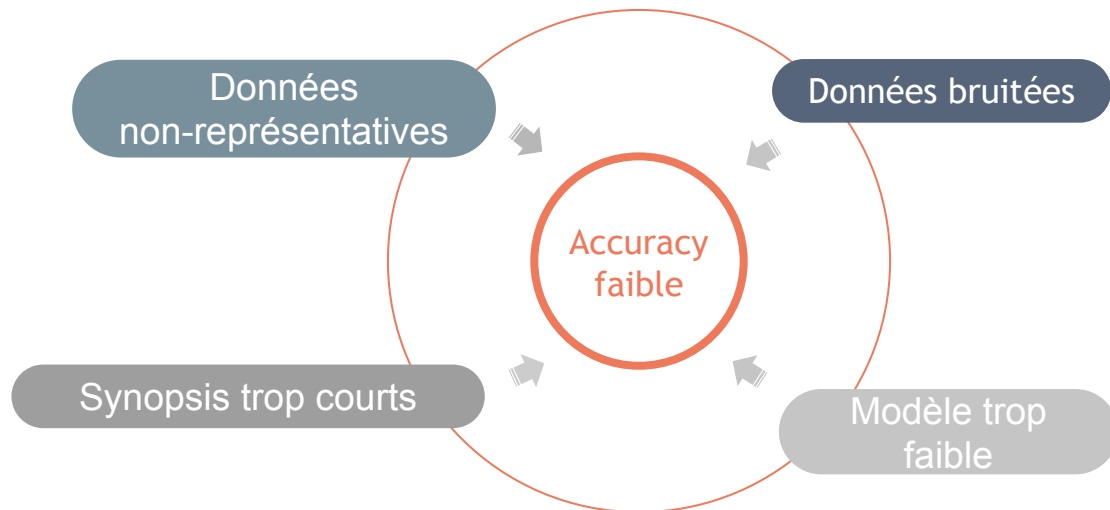
- Trop d'instances sont classées en Comedy par erreur
- Comedy et Action sont plutôt bien classés
- Drama confondu avec Comedy
- Animation confondu avec Action
- Les "petites catégories" ont des précisions mauvaises

# Matrice de confusion (brute)





# Hypothèses Accuracy Faible



# Quelle est la précision humaine ?

*Au bureau d'une société japonaise, lors d'une soirée, une maîtresse professionnelle est retrouvée morte, apparemment après des rapports sexuels brutaux. Web Smith, enquêteur de la police, est appelé pour enquêter mais, avant de s'y rendre, il reçoit un appel de quelqu'un qui lui ordonne de récupérer John Connor, ancien capitaine de police et expert des affaires japonaises. Quand ils arrivent là-bas, Web pense tout comprendre, mais Connor le prévient qu'il reste beaucoup à éclaircir.*

*Charlie, un poète, n'a jamais eu beaucoup de chance avec les femmes, avant de rencontrer Harriet, la fille de ses rêves ... ou de ses cauchemars. Charlie commence à soupçonner qu'Harriet est Mme X, une femme qui se marie puis tue ses époux.*

Trouver la catégorie des deux synopsis ci-dessus parmi:

**Comedy, Drama, Action, Crime, Animation, Documentary, Adventure, Biography, Horror**

# Quelle est la précision humaine ?

Au bureau d'une société japonaise, lors d'une soirée, une maîtresse professionnelle est retrouvée morte, apparemment après des rapports sexuels brutaux. Web Smith, enquêteur de la police, décide d'enquêter mais, à la suite d'un appel de Connor, il décide de récupérer John Connor, ancien capitaine de police et expert des affaires japonaises. Quand ils arrivent là-bas, Web pense tout comprendre, mais Connor le prévient qu'il reste beaucoup à éclaircir.

**ACTION**

Charlie, un poète, n'a jamais eu beaucoup de chance avec les femmes, avant de rencontrer Harriet, la fille de ses rêves ... ou de ses cauchemars. Charlie commence à soupçonner qu'Harriet est une menteuse qui se marie pour l'argent.

**COMEDY**

Trouver la catégorie des deux synopsis ci-dessus parmi:

**Comedy, Drama, Action, Crime, Animation, Documentary, Adventure, Biography, Horror**

# Plan

1

Equipe, Contexte, Objectifs, Livrables

2

Projet et Evaluation

3

Perspectives

4

Conclusion

# Perspectives

---

| Données  | Exploration   | Modèle   |
|--|---|--|
| <p>Récupérer plus de données</p> <p>Dupliquer les synopsis qui ont plusieurs genres</p> <p>Nouvelles features par clustering</p> | <p>Mesure de la proximité entre les synopsis vectorisés (calcul des angles)</p> | <p>Modèle plus “profond” (Réseaux de neurones récurrent, etc.)</p> |

# Plan

1

Equipe, Contexte, Objectifs, Livrables

2

Projet et Evaluation

3

Perspectives

4

Conclusion

# Conclusion

---

Natural Language Processing

Projet de Data Science complet

Exploration de méthodes et de modèles



**Merci de votre attention !**

**Avez-vous des questions ?**