CentraleSupélec

OPENCLASSROOMS

Parcours Data Scientist

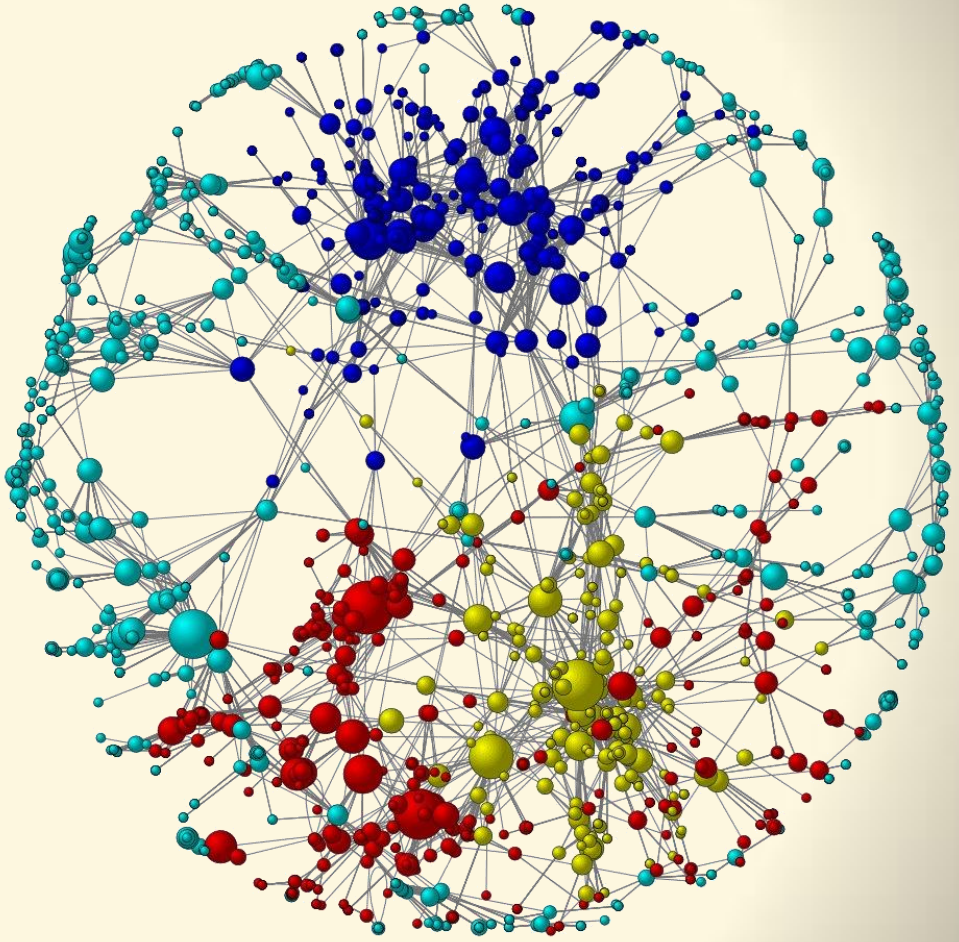
Projet 3 :

Développer un moteur de recommandations de films

IMDb

Sommaire

- Présentation et Objectifs
- Nettoyage
- Modélisation
  - K-means
  - DBSCAN
  - Clustering Hierarchique
  - Isomap
  - LLE
  - TSNE
  - PCA
  - Modèle simple
- Modèle final
  - Mise en place
  - API
- Pistes d'évolutions
- Conclusion



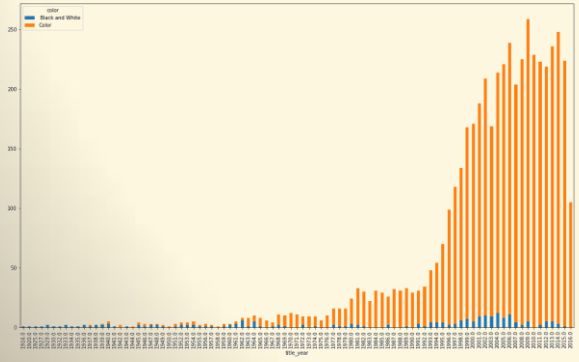
Presentation

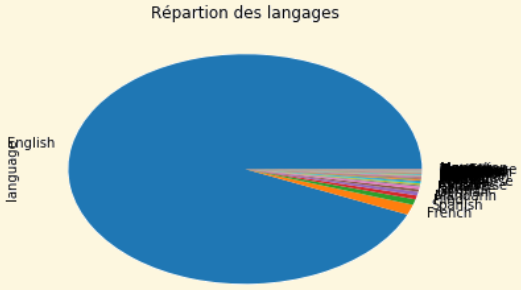
- Dataset
  - 5043 films
  - 28 features (genre, titre, année de sorties, …)
  - Issue de IMDb
  - 2698 pts manquants (1,91 %)

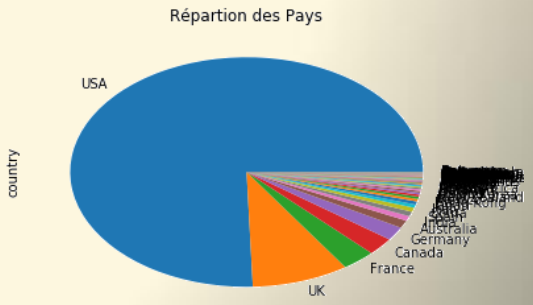
- Objectif :
  - Faire un moteur de recommandation de films par similarité
    - « Clustering » sans labels
  - Générer une simple API

Nettoyage

- Phase 1 : Données manquantes
  - Color : basé sur la date
  - Comptage (like/review/…): 0
  - Durée/vente/budget/year : moyenne
  - Langue/Country: Majorité (English/USA)
    - Validation par producteur
  - Noms : None







Nettoyage

- Phase 2 : Simplification
  - Genres :
    - Split et OHE
  - Rating :
    - Par âge (OHE+)
  - Acteurs:
    - Par occurrence

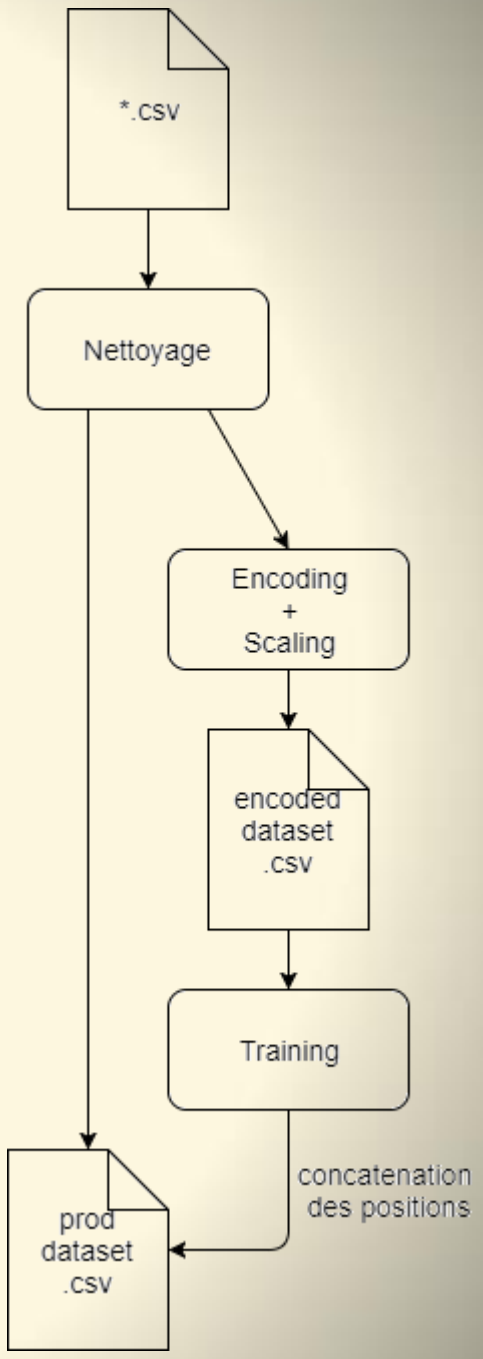
Nettoyage

- Phase 3.1 : Suppression features
  - Keywords
  - Ratio (pas estimable)
  - Nom du film
  - Lien IMDb
  - Après test (color, num_critic, num_user, budget, …)

- Phase 3.2 : Suppression Films
  - Directeur ou likes act. 3 manquant
  - Duplicata et reindexation

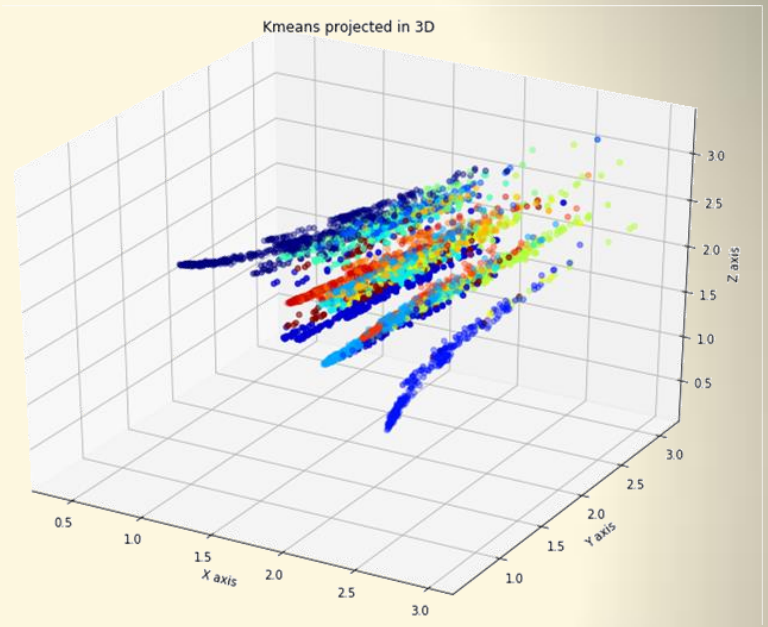
Nettoyage

- Phase 4 : Encodage/Clean
  - 1 version non-encodée allégée
    - Pour la production
  - 1 version encodée
    - Scaling (MinMax)
    - LabelEncoder pour textes



Modélisation

- K-means
  - Groupe « Clustering »
  - Nombre de clusters inconnu
    - A fournir
  - Retourne positions
    - Recommander possible

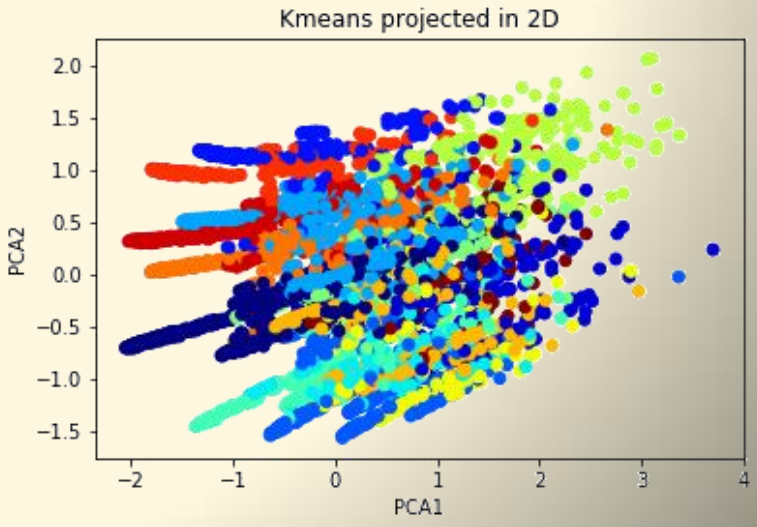


Entrée :

- Spider-Man 3

Sortie :

- Ironclad
- Prince of Persia: The Sands of Time
- Krrish
- The Three Musketeers
- The Musketeer



Modélisation

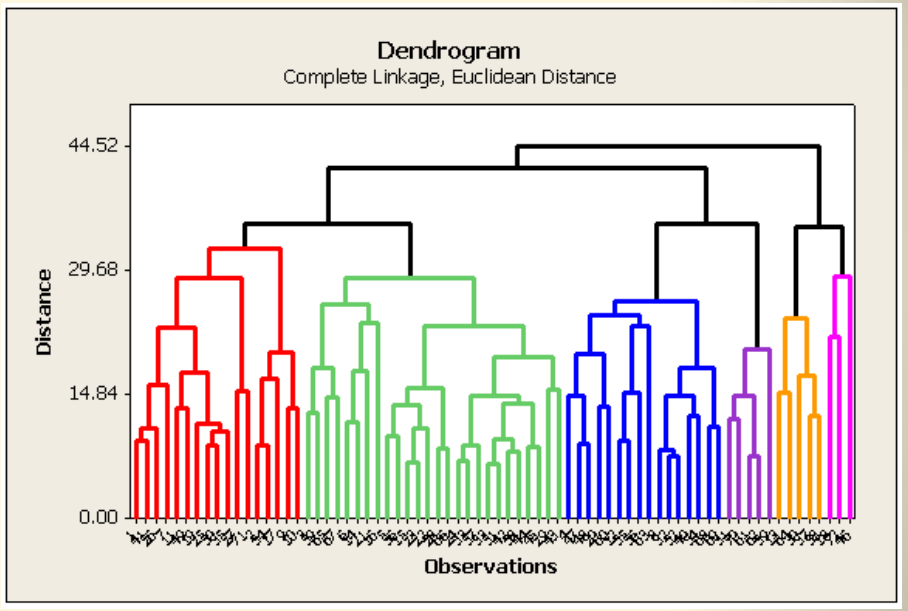
- DBSCAN
  - Groupe « Clustering »
  - Nombre de clusters inconnu
    - Pas nécessaire
  - Pas de positions
    - Recommander impossible
  - Contenu des clusters variable

Entrée :
- Cluster #6

Sortie :
- Terminator 3: Rise of the Machines
- The Matrix Reloaded
- Hulk
- Total Recall
- Terminator 2: Judgment Day
- Dredd
- Battle Los Angeles
- Æon Flux
- Universal Soldier: The Return
- The Black Hole
- Megaforce
- The Terminator
- Escape from New York
- Escape from the Planet of the Apes
- Battle for the Planet of the Apes
- Conquest of the Planet of the Apes

Modélisation

- AgglomerativeClustering
  - Groupe « Clustering »
  - Nombre de clusters inconnu
    - A fournir (optionnel)
  - Pas de positions
    - Recommander **possible**

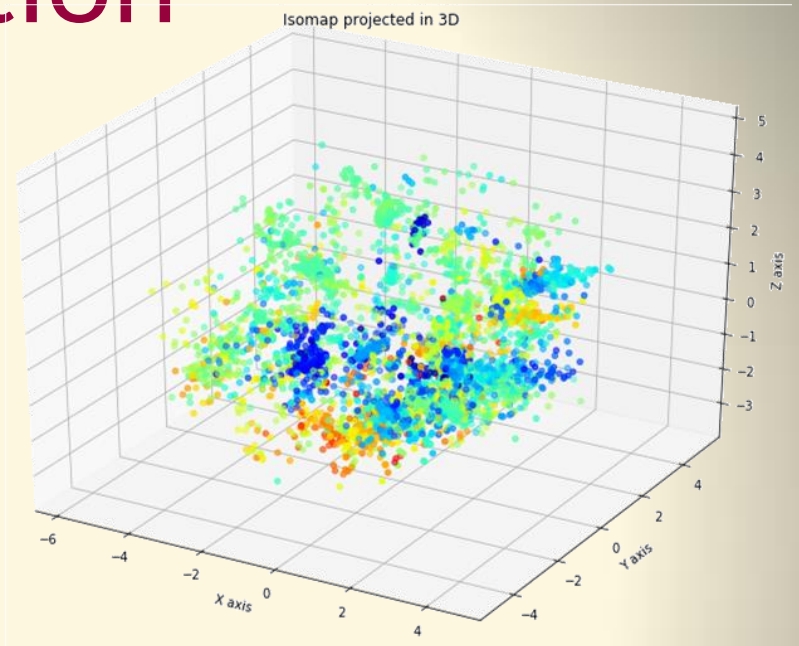


Entrée :
- Spider-Man 3

Sortie :
- A Knight's Tale
- **The Three Musketeers**
- **The Musketeer**
- Prince of Persia: The Sands of Time
- Spider-Man 2 et Spider-Man

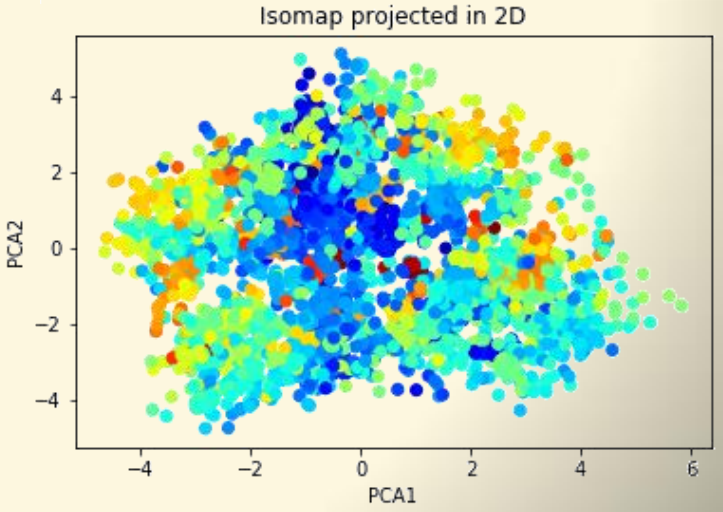
Modélisation

- Isomap
  - Groupe « manifolds »
  - Pas de clusters
  - Positions présente
    - Recommander possible



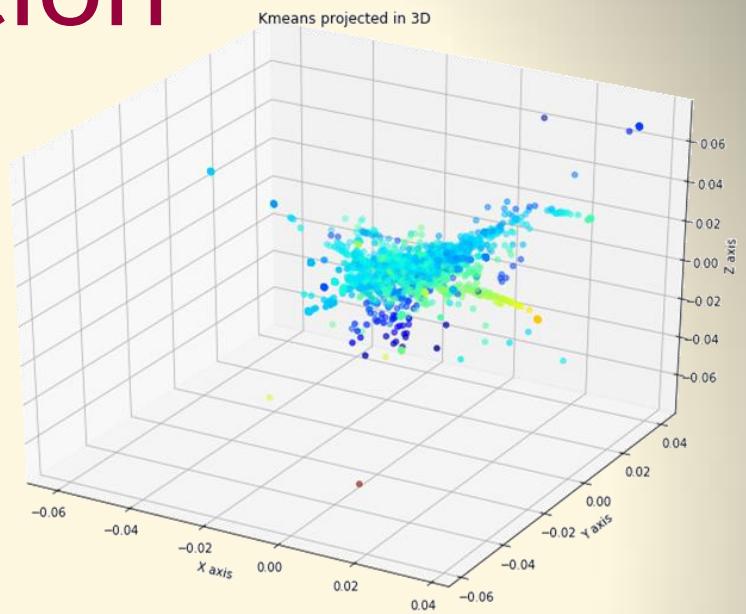
Entrée :
- Spider-Man 3

Sortie :
- Repo! The Genetic Opera
- A Knight's Tale
- The Three Musketeers
- Ironclad
- Bram Stoker's Dracula



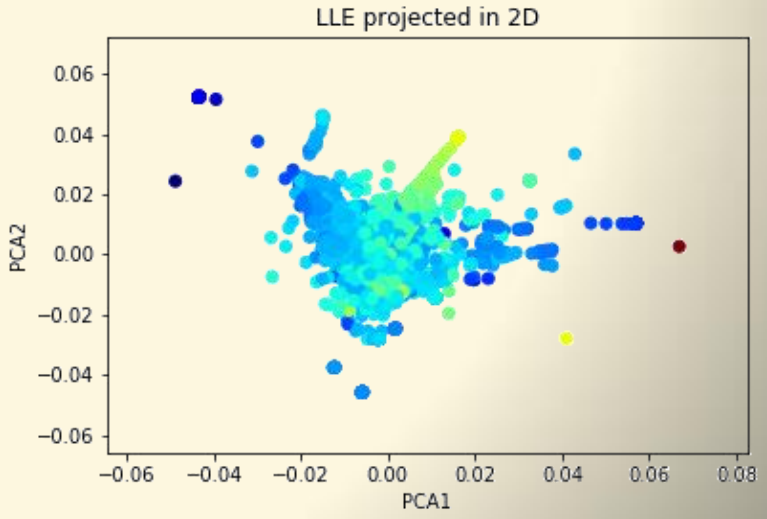
Modélisation

- Locally Linear Embedding
  - Groupe « manifolds »
  - Pas de clusters
  - Positions présente
    - Recommander possible
  - Sensible à l'initialisation



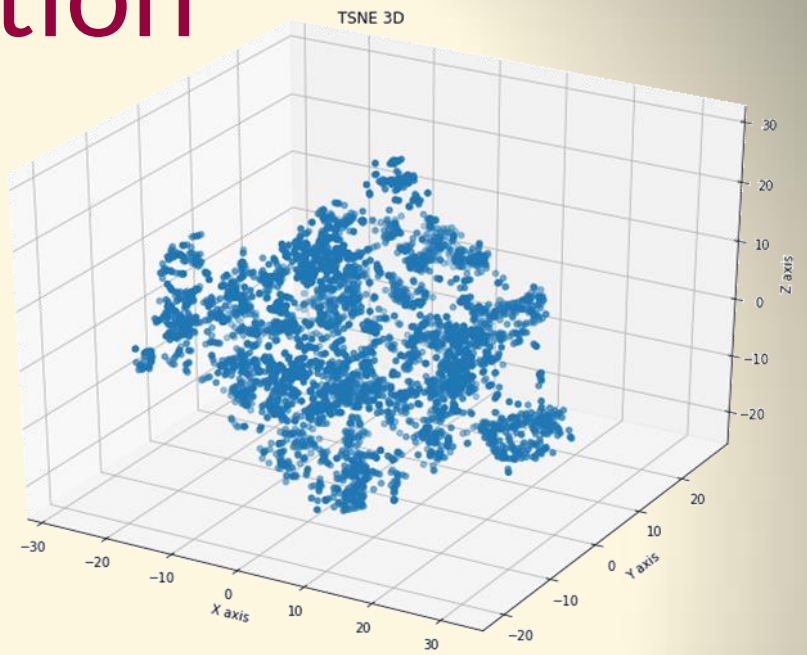
Entrée :
- Spider-Man 3

Sortie :
- A Knight's Tale
- The Three Musketeers
- The Shawshank Redemption
- Pulp Fiction
- It Happened One Night



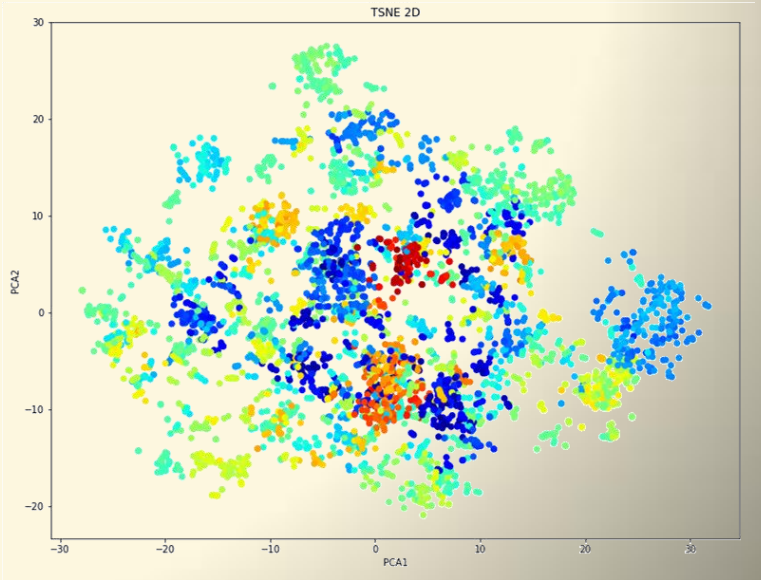
Modélisation

- TSNE
  - Groupe « manifolds »
  - Pas de clusters
  - Positions présente
    - Recommander possible



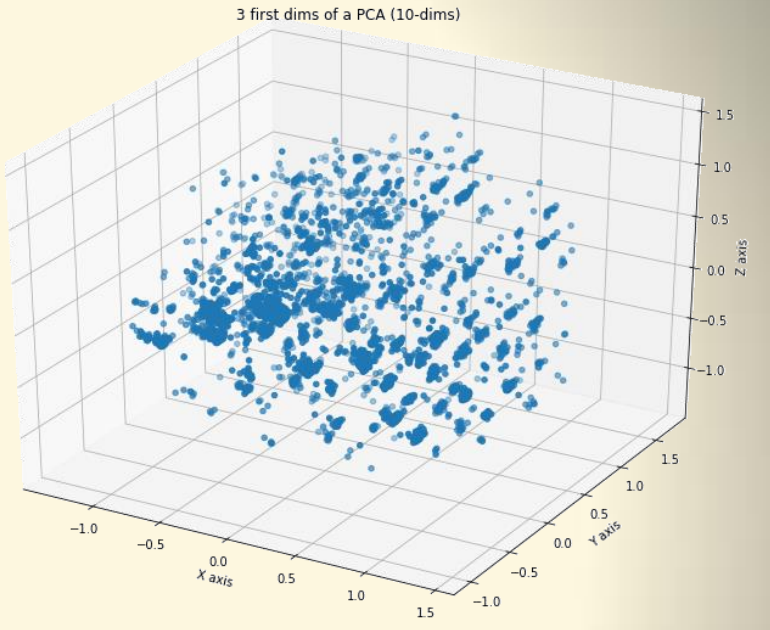
Entrée :
- Spider-Man 3

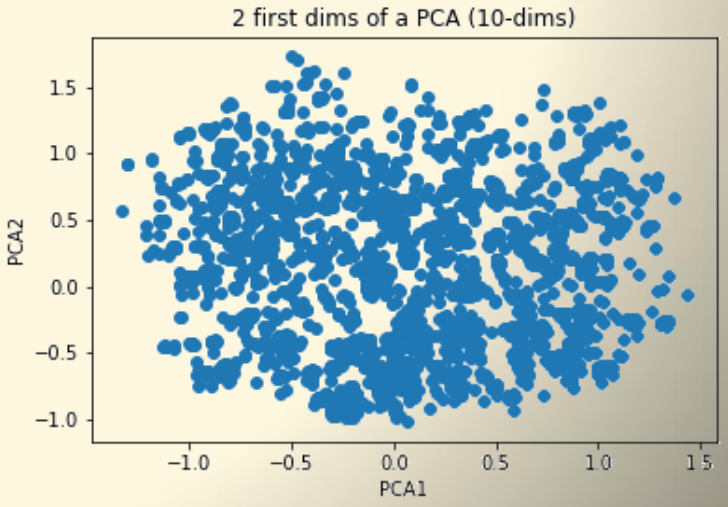
Sortie :
- Spider-Man
- Spider-Man 2
- The Three Musketeers
- The Musketeer
- A Knight's Tale



Modélisation

- PCA
  - Groupe « décomposition »
  - Pas de clusters
  - Positions présente
    - Recommander possible
  - Nécessite haute dimension
    - 10 dims => 70% de la variance





Entrée :

- Spider-Man 3

Sortie :

- The Three Musketeers
- The Musketeer
- The Charge of the Light Brigade
- Ironclad
- Red River

Modélisation

- Modèle simple
  - Basé sur la matrice encodée/scalée
  - Utilise la distance euclidienne
  - Avantage :
    - Online-Learning
    - Rapidité
    - Pas de perte d'information
  - Similaire au KNN

TSNE

Entrée :
- Spider-Man 3

Sortie :
- Spider-Man
- Spider-Man 2
- The Three Musketeers
- The Musketeer
- A Knight's Tale

Modèle Simple

Entrée :
- Spider-Man 3

Sortie :
- Spider-Man 2
- Spider-Man
- The Three Musketeers
- A Knight's Tale
- The Musketeer

Modèle final

- Mise en place
  - Modèle choisi :TSNE & Modèle Simple
  - Optimisation du TSNE basé sur kl_divergent
  - Agrégation des coordonnées au dataset non encodé
  - Récupération des n-pts les plus proche

Modèle final

- API
  - Utilisation du dataset non encodé
  - Sauvegarde de paramètres sur pickle
  - Creation petite API Flask (page d'acceuil et recommandation)
  - Utilisation du dataset pour recupérer les champs principaux basés sur la distance euclidienne

http://coni57.pythonanywhere.com/

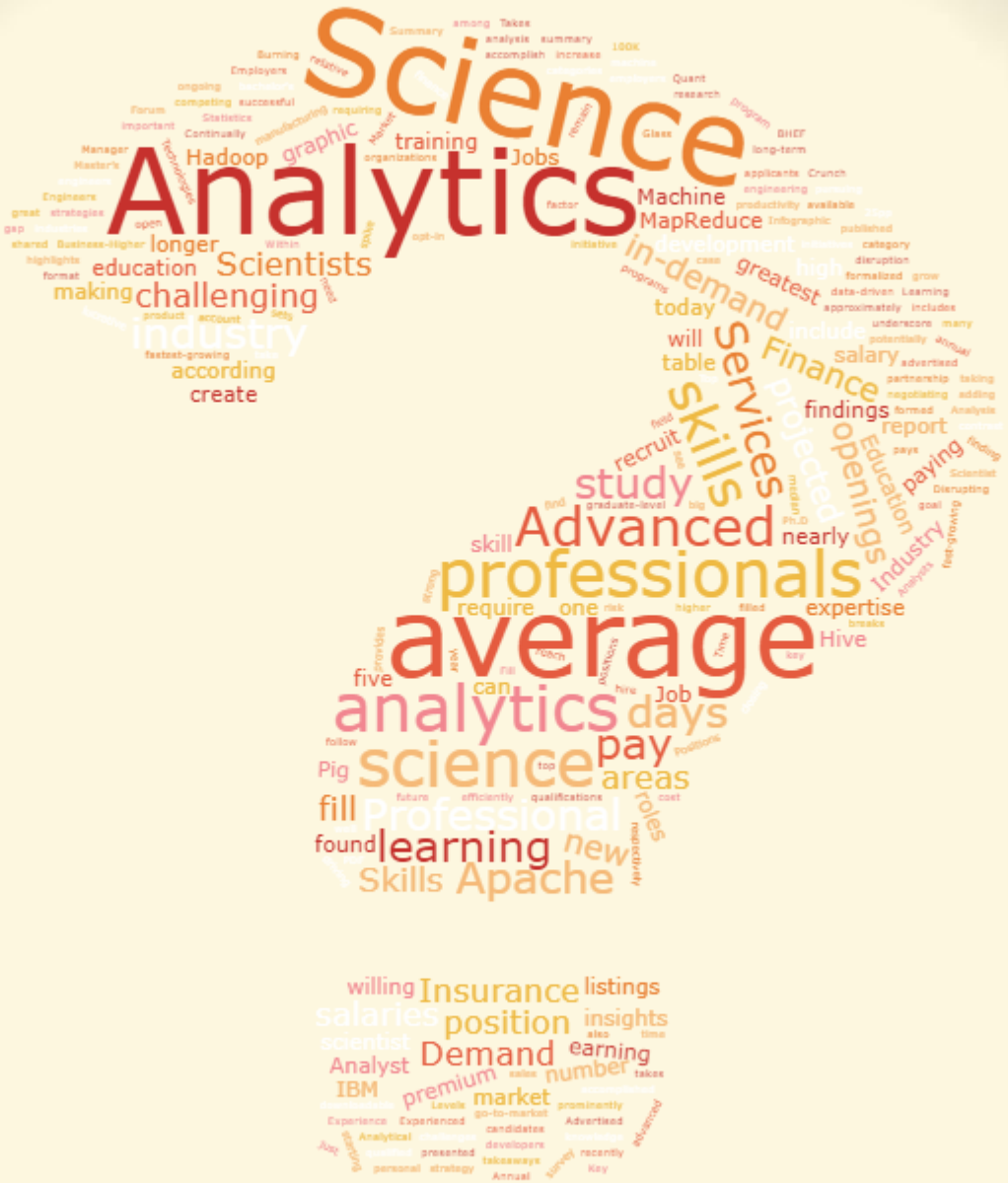
Pistes d'évolutions

- Coefficients
  - Permet de renforcer certaines features
    - Choix du Client
    - Choix de l'utilisateur
  - Evaluation semi-supervisée
    - Sur une centaine de films
  - Bagging ?

Conclusion

- Peu de critères initiaux
  - Seulement 28 features
- Modèles difficiles à évaluer
  - Pas de labels
  - Nombre de clusters inconnu
- Manifolds fonctionnent mieux que Cluster
  - Initialisation parfois importante
- Résultat satisfaisant surtout pour TSNE et modèle simple
  - Les autres ne prédisent pas les suites

- Le modèle Simple est le plus pertinent par sa simplicité


Science
Analytics
Hadoop
graphic
training
Jobs
Machine
MapReduce
longer
education
Scientists
making
challenging
industry
according
create
development
in-demand
greatest
high
today
include
Finance
Services
will
table
salary
findings
report
skills
recruit
study
Advanced
skill
nearly
professionals
require
one
expertise
Hive
average
five
can
Job
analytics
days
pay
science
Pig
areas
fill
Professional
found
learning
new
Skills
Apache
willing
Insurance
listings
salaries
position
insights
scientist
Demand
earning
Analyst
number
premium
IBM
market