

**MOHAMEDATNI  
AYA**

**L2 MIASHS  
22106289**

# **Science des données : Projet individuel**

## **PLAN :**

### **1 - Apprentissage non supervisé**

#### **1.1 - Jeu de données**

\_Description de votre jeu de données

\_Description du processus que vous avez suivi pour constituer votre jeu de données.

#### **1.2 - Nettoyage et prétraitement des données**

\_Description de votre démarche pour nettoyer et/ou prétraiter les données.

#### **1.3 - Clustering**

\_Description de la chaîne mise en place et analyse des résultats obtenus.

### **2 - Apprentissage supervisé**

#### **2.1 - Jeux de données**

\_Description de votre jeu de données étiquetées

\_Description du processus pour constituer votre jeu de données étiquetées

\_Description de votre jeu de données à prédire

\_Description du processus que vous avez suivi pour constituer votre jeu de données à prédire

#### **2.2 - Nettoyage et prétraitement des données**

\_Description de votre démarche pour nettoyer et/ou prétraiter les données.

#### **2.3 - Modèles d'apprentissage**

\_Description de la chaîne mise en place et analyse des performances.

#### **2.4 - Prédictions**

\_Description et analyse des résultats prédits.

## 1 - Apprentissage non supervisé

### 1.1 - Jeu de données

La première étape est de trouver **un jeu de données** avec un maximum d'**informations intéressantes** pour pouvoir les exploiter par la suite.

Les variables qui m'intéressais au début étaient le titre, un ou plusieurs genres/catégories et un synopsis pour chaque film.

Après plusieurs recherches j'ai trouvé, sur le site web **kaggle**, le fichier **tmdb\_5000\_movies.csv** qui contient des renseignements interessant sur une grande quantité de film.

**Source :** [https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb\\_5000\\_movies.csv](https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv)

On nous propose un fichier csv avec les variables **budget, genres, homepages, id, keywords, original\_language, title, overview (synopsis), popularity, production\_company, production\_counties, release\_date, vote\_average, vote\_count, tagline** pour 5000 films.

Ce jeu de données contient les valeurs que je recherchais et plus encore.

**Néanmoins**, il avait besoin d'être nettoyé car en l'ouvrant avec un tableur j'ai pu constater qu'il y avait beaucoup de valeurs manquantes. De plus toutes les variables ne m'intéressaient pas forcement.

### 1.2 - Nettoyage et prétraitement des données

Apres avoir trouver le jeu de données qui m'intéressait, il était temps de **nettoyer mes données pour pouvoir les exploiter**.

J'ai décidé de **supprimer** les colonnes des variables budgets, homepages, id, keywords, original\_language et production\_counties.

Ensuite j'ai crée un nouveau tableau nommées **mohamedatni\_aya\_data1.csv** où j'ai copier une partie du fichier **tmdb\_5000\_movies.csv** .

Puis supprimer tous les films dont les données n'étaient pas complète.

Par conséquent, le fichier **mohamedatni\_aya\_data1.csv** est le jeu de donnés que je vais utiliser **pour répondre à la première partie de mon rapport individuel**.

## 1.3 - Clustering

Avant de commencer ma chaîne de traitement, il a fallut que je me renseigne sur **les objectifs et les limites** d'un clustering.

Le clustering (appelé aussi classification non supervisée,) a pour objectif de **séparer un ensemble d'observations en groupes homogènes**. Il est attendu d'un clustering de générer des groupes où les observations appartenant au même groupe sont plus similaires que les observations appartenant à des groupes différents.

Son objectif initial est de décrire les données à partir des groupes qui les constituent. On cherche ainsi à **faire émerger des groupes qui ont une signification**.

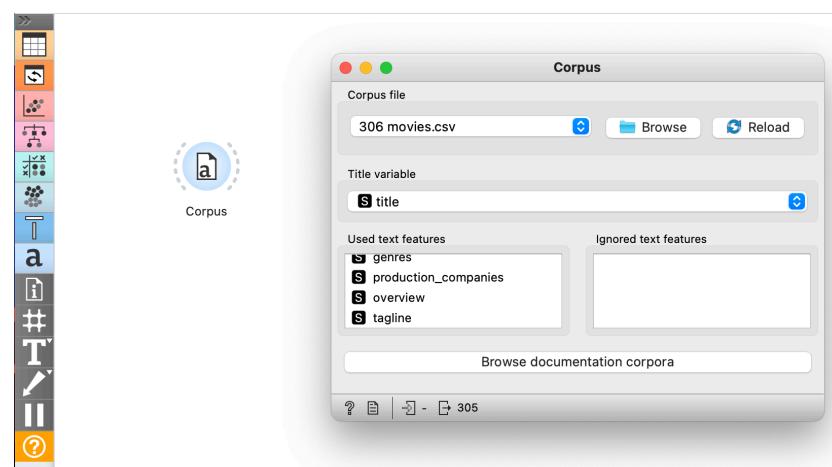
Ainsi, j'ai mis en place **une chaîne de traitement en vue de produire une classification basés sur un algorithme de clustering hiérarchique**.

Pour cela j'ai ouvert un nouveau fichier **Orange**, où j'ai ajouter l'outil **Corpus**. Cet outils permet de charger un corpus de textes. J'ouvre le corpus pour chargez le fichier **mohamedatni\_aya\_data1.csv**

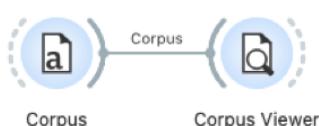
Je rappelle donc que ce fichier contient un tableau où chaque ligne correspondant à un film.

Pour chacun on connaît :

- **title**
- **tagline**
- **overview**
- **genres**
- **popularity**
- **remlease\_date**
- **vote\_average**
- **vote\_count**
- **production\_companies**



Afin de voir le contenu du corpus, j'ajoute l'outil **Corpus Viewer** en sortie de **Corpus** :



## Double clique sur **Corpus View** :

Corpus Viewer

Info  
Tokens: n/a  
Types: n/a  
Matching documents: 305/305  
Matches: n/a

Search features

- popularity
- release\_date
- vote\_average
- vote\_count
- genres
- overview
- production\_companies

Display features

- popularity
- release\_date
- vote\_average
- vote\_count
- genres
- overview

Show Tokens & Tags

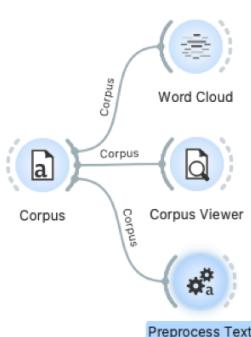
Auto send is on

RegEx Filter:

Rank	Title
1	Avatar
2	Pirates of the Caribbean: At World's End
3	Spectre
4	The Dark Knight Rises
5	John Carter
6	Spider-Man 3
7	Tangled
8	Avengers: Age of Ultron
9	Harry Potter and the Half-Blood Prince
10	Batman v Superman: Dawn of Justice
11	Quantum of Solace
12	Pirates of the Caribbean: Dead Man's Chest
13	The Lone Ranger
14	Man of Steel
15	The Avengers
16	Pirates of the Caribbean: On Stranger Tides
17	Men in Black 3

popularity: 150.438  
release\_date: 2009-12-10  
vote\_average: 7.2  
vote\_count: 11800  
genres: [{"id": 28, "name": "Action"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]  
overview: In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes torn between following orders and protecting an alien civilization.  
production\_companies: [{"name": "Ingenious Film Partners", "id": 289}, {"name": "Twentieth Century Fox Film Corporation", "id": 306}, {"name": "Dune Entertainment", "id": 444}, {"name": "Lightstorm Entertainment", "id": 574}]  
tagline: Enter the World of Pandora.  
title: Avatar

Pour visualiser les mots les plus souvent utilisés dans ces textes, j'ajoute l'outil **Word Cloud** (dans le menu Text Mining) en sortie de **Corpus**.



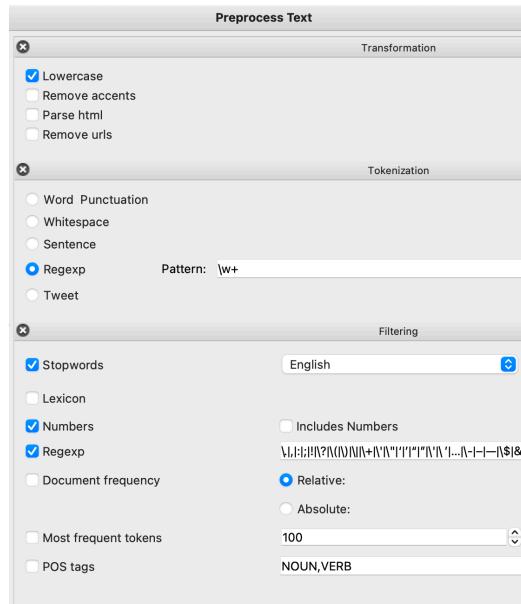
Les mots les plus fréquents ne sont pas forcément les plus intéressants pour caractériser le corpus. Il va falloir « nettoyer » le corpus .

C'est l'étape de prétraitement vue en cours.

J'ajoute donc l'outil **Preprocess Text** (menu Text Mining) en sortie de **Corpus**.

Dans le cadre **Transformation**, par défaut *Lowercase* transforme tout le texte en minuscule.

Ensuite il supprime les signes de ponctuation. Le cadre **Filtering**, il supprime les mots d'arrêt lorsqu'on coche *StopWords*, il faut ensuite sélectionner la langue des mots d'arrêts souhaité.



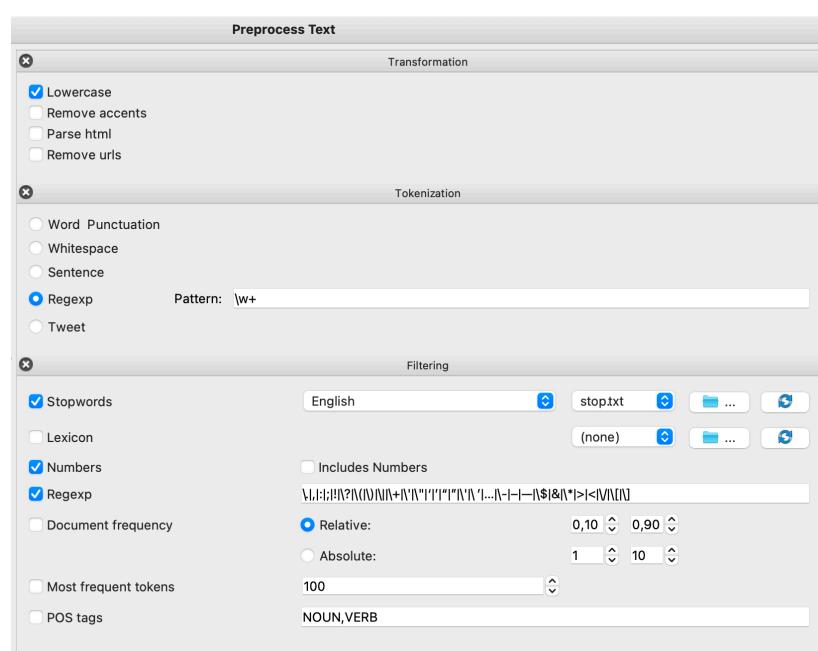
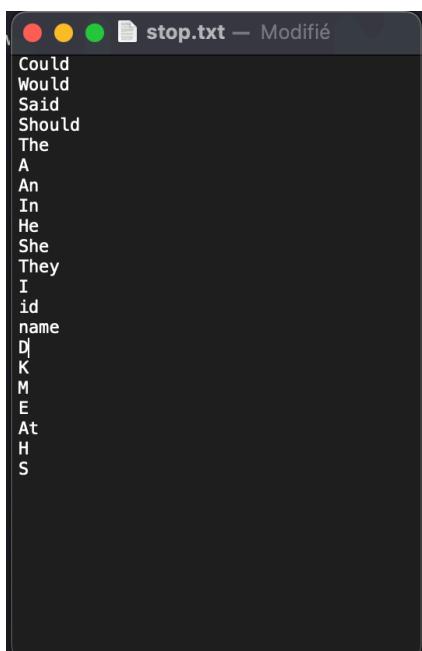
J'ajoute un **nuage de mots** en sortie de **Preprocess Text** :



J'ai remarqué que dans le fichier de départ certaines variables sont écrit de cette manière : **[{"id": 28, "name": "Action"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]**

Au début j'ai pensé que cela pouvait me poser problème.

Mais il a simplement fallut que je crée un fichier texte et ajoute les mots que je voulais supprimer. Ensuite j'ai chargé le fichier dans **Process Text** sur le dossier à droite de stopwords :



Le nuage de point est maintenant plus intéressant :



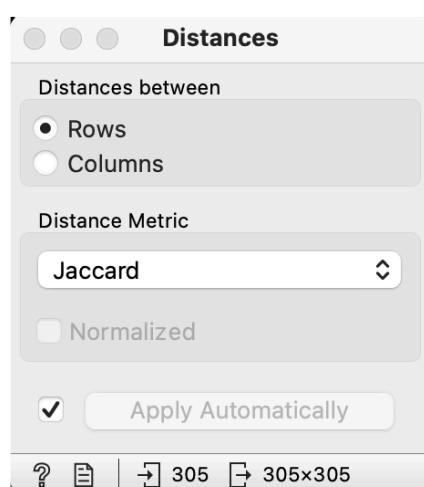
En sortie de **Preprocess Text**, j'ajoute l'outil **Bad of Words** (menu Text Manning)

En effet on a vu qu'un algorithme de clustering prend en entrée une matrice de distances qui peut être calculée grâce à des sacs de mots.

Je souhaite donc **calculer la matrice de distances entre les documents**. Pour cela, j'ajoute l'outil **Distances** (menu Unsupervised) en sortie de **sac de mots**.

Pour voir le résultats de Distances, j'ajoute **Distance Matrix** (menu Unsupervised) en sortie de distances.

Double clique sur l'outil **Distances** :

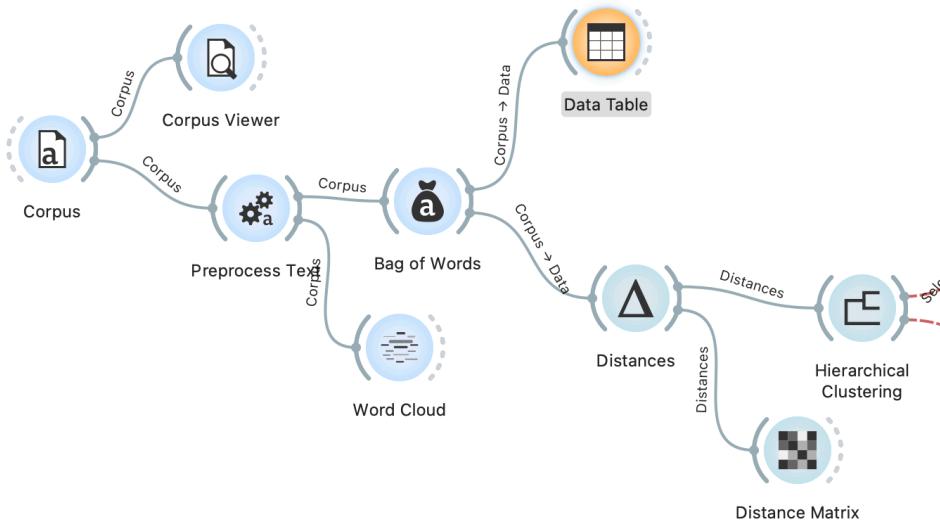


Dans les paramètre, j'ai décidé que la mesure utilisée serait **Jaccard**, car c'est le paramètre qui me renvoyer la matrice de distances la plus intéressante.

Double clique sur **Distance Matrix** :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		0,896	0,929	0,918	0,871	0,905	0,946	0,902	0,921	0,898	0,922	0,910	0,945	0,855	0,892
2	0,896		0,884	0,932	0,843	0,878	0,901	0,905	0,871	0,900	0,910	0,611	0,833	0,843	0,865
3	0,929	0,894		0,920	0,907	0,895	0,926	0,919	0,908	0,925	0,853	0,882	0,900	0,907	0,896
4	0,918	0,932	0,920		0,924	0,928	0,917	0,932	0,918	0,875	0,927	0,931	0,945	0,862	0,926
5	0,871	0,843	0,907	0,924		0,914	0,895	0,885	0,929	0,905	0,929	0,857	0,892	0,868	0,859
6	0,905	0,878	0,895	0,926	0,914		0,919	0,912	0,915	0,918	0,905	0,845	0,906	0,886	0,904
7	0,946	0,901	0,926	0,917	0,895	0,919		0,926	0,932	0,930	0,942	0,900	0,911	0,929	0,920
8	0,902	0,908	0,918	0,932	0,885	0,912	0,926		0,938	0,935	0,937	0,917	0,944	0,898	0,824
9	0,921	0,871	0,908	0,914	0,929	0,918	0,932	0,938		0,893	0,932	0,869	0,930	0,864	0,917
10	0,898	0,900	0,925	0,875	0,905	0,918	0,930	0,935	0,893		0,920	0,911	0,920	0,805	0,919
11	0,922	0,910	0,853	0,927	0,929	0,905	0,942	0,937	0,932	0,920		0,909	0,908	0,929	0,931
12	0,910	0,611	0,882	0,931	0,857	0,845	0,900	0,917	0,869	0,911	0,909		0,831	0,857	0,893
13	0,945	0,833	0,900	0,945	0,892	0,906	0,911	0,944	0,930	0,920	0,908	0,831		0,916	0,918
14	0,855	0,843	0,907	0,862	0,868	0,886	0,929	0,884	0,864	0,805	0,929	0,857	0,916		0,873
15	0,892	0,865	0,896	0,926	0,859	0,904	0,920	0,824	0,917	0,919	0,931	0,893	0,918		0,873
16	0,928	0,792	0,894	0,943	0,886	0,901	0,917	0,931	0,899	0,906	0,903	0,757	0,883	0,899	0,802
17	0,915	0,905	0,929	0,939	0,921	0,943	0,909	0,935	0,905	0,944	0,916	0,914	0,899	0,913	
18	0,937	0,892	0,928	0,910	0,920	0,900	0,933	0,928	0,872	0,904	0,943	0,903	0,913	0,851	0,933
19	0,932	0,891	0,924	0,934	0,917	0,798	0,938	0,942	0,918	0,920	0,929	0,901	0,937	0,864	0,919

Je peux maintenant ajouter l'outil **Hierarchical Clustering** (menu Unsupervised ) en sortie de **Distances**.



Double Clique sur l'outil **Hierarchical Clustering** :

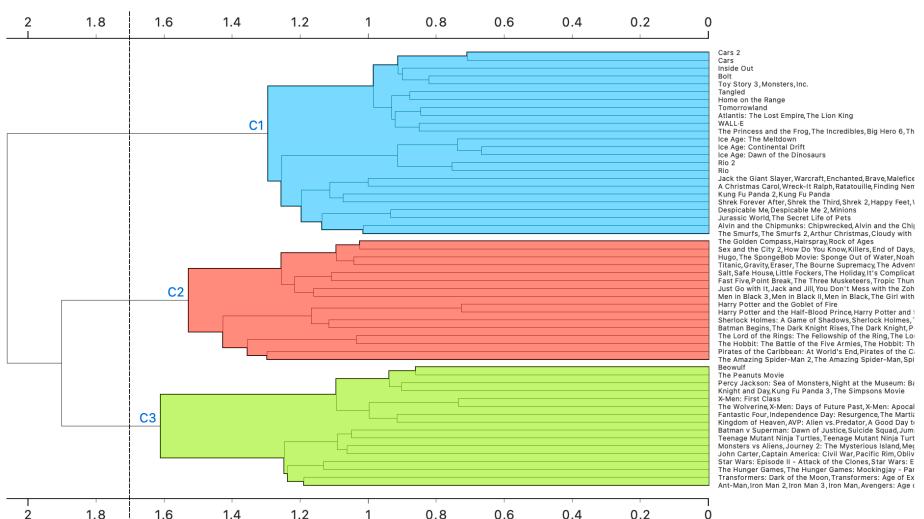
Dans le cadre *Linkage*, je sélectionne Ward.

Pour les annotations, je sélectionne le Titre.

Voici le **dendrogramme** que j'obtient lorsque je place la ligne de coupe de façon à obtenir **les 3 clusters principaux** :

Pour pouvoir interpréter ce classement, j'ajoute en sortie de **Hierarchical Clustering** l'outil **Corpus Viewer**.

Il prendra en compte seulement les films dans les cluster sélectionnées et non l'ensemble du corpus.



J'ajoute aussi l'outil **Word Cloud** en sortie de **Hierarchical Clustering** pour pouvoir comparer les nuages de mots de chaque cluster.

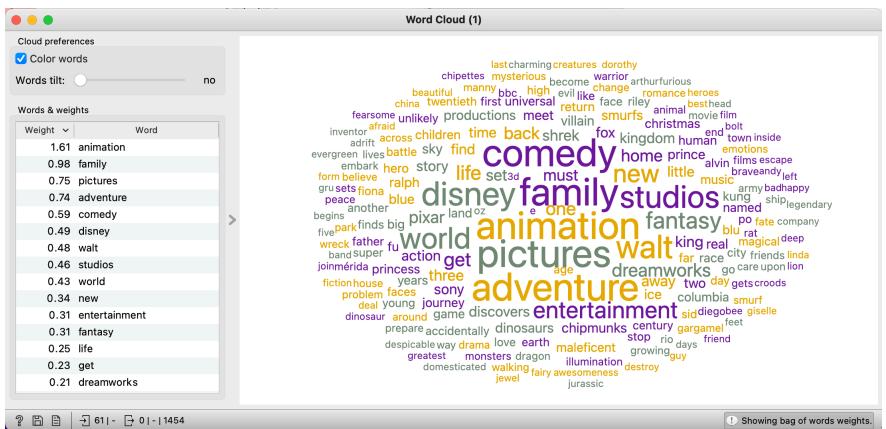
### **Observation individuel de mes clusters :**

Le nuage de mots de ce cluster met en évidence les mots **animation, family, adventure, comedy, disney...**

Grâce au nom des films on peut déjà observer que le premier cluster contient majoritairement des films animés comme Cars, Shrek, Minions.

Il rassemble aussi beaucoup de Walt Disney Animation Studios comme The Lion King, Finding Nemo, Ice Age.

Néanmoins, il y a quelques films qui ne sont pas animé comme Maleficent et Alvin and the Chipmunks mais qui sont des films familiales.



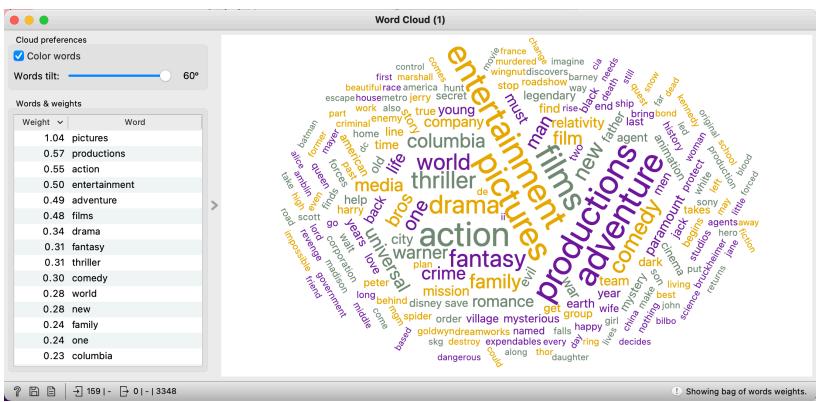
Cars 2  
Cars  
Inside Out  
Bolt  
Toy Story 3  
Monsters, Inc.  
Tangled  
Home on the Range  
Tomorrowland  
Atlantis: The Lost Empire  
The Lion King  
WALL-E  
The Princess and the Frog  
The Incredibles  
Big Hero 6, The Good Dinos  
Ice Age: The Meltdown  
Ice Age: Continental Drift  
Ice Age: Dawn of the Dinos  
Rio 2  
Rio  
Jack the Giant Slayer  
Warcraft  
Enchanted  
Brave, Maleficent  
A Christmas Carol  
Wreck-It Ralph, Ratatouille  
Finding Nemo  
How to Train Your Dragon,  
Kung Fu Panda 2  
Kung Fu Panda  
Shrek Forever After  
Shrek the Third, Shrek 2  
Happy Feet, Walking With D  
Madagascar: Escape 2 Afri  
Despicable Me  
Despicable Me 2  
Minions  
Jurassic World  
The Secret Life of Pets  
Alvin and the Chipmunks: C  
Alvin and the Chipmunks: T  
Alvin and the Chipmunks: T  
The Smurfs  
The Smurfs 2  
Arthur Christmas  
Cloudy with a Chance of M

Les films du **deuxième cluster** parlent principalement d'**action, adventure, films, drama, fantasy, thriller, comedy, family, romance, cinema ...**

Lorsqu'on observe les données, on constate que ce cluster rassemble des films comme *Titanic* mais aussi comme *Twilight* ou même *Spider-Man*.

Ma première impression est que ce cluster rassemble des films d'actions avec des histoires romantiques, avec ou sans science fiction, des thrillers ect..

On peut remarquer aussi qu'il y a plusieurs franchises : Harry Potter, The Lord of the Ring, The hobbits

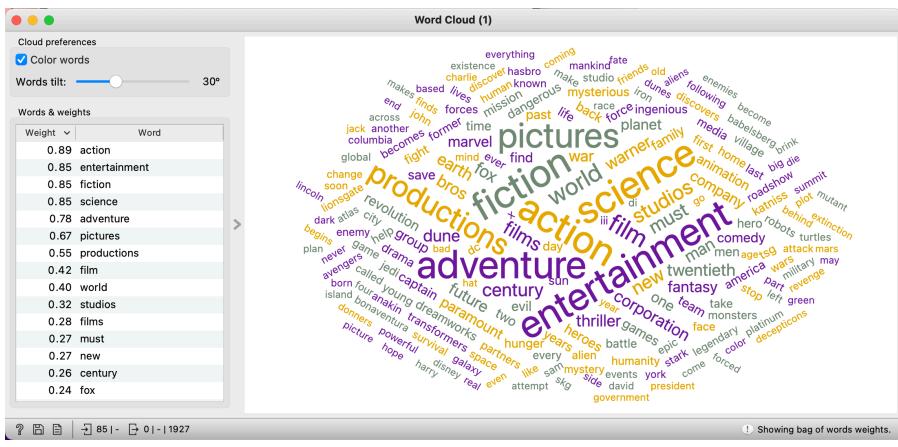


The Amazing Spider-Man 2  
The Amazing Spider-Man  
Spider-Man 3  
Spider-Man 2  
Mission: Impossible - Ghost Protocol  
Mission: Impossible - Rogue Nation  
Mission: Impossible III,Mission: Impossible  
The Expendables 3  
The Expendables 2,The Expendables  
Spectre,Quantum of Solace  
Skyfall,XXX  
Just Go with It  
Jack and Jill  
You Don't Mess with the Zohan  
Funny People,Grown Ups 2,Get Him to the  
Men in Black 3  
Men in Black II,Men in Black  
The Girl with the Dragon Tattoo  
RoboCop,The Other Guys,Funny People  
Salt,Safe House,Little Fockers  
The Wolfman,The Kingdom,Battle of  
Fast Five,Point Break,The Thing  
The Tourist,The Rundown,Thunderbirds  
The Monuments Men,Charlie Wilson's  
Hugo,The SpongeBob Movie:  
The Golden Compass,Hairspray  
Sex and the City 2,How Do You  
Harry Potter and the Goblet of Fire  
Harry Potter and the Half-Blood Prince  
Harry Potter and the Order of the Phoenix  
Sherlock Holmes: A Game of  
Sherlock Holmes  
The Hangover Part II,Fool's Gold  
Ocean's Eleven,Ocean's Thirteen  
Batman Begins  
The Dark Knight Rises,The Dark Knight  
Titanic,Gravity,Eraser,The Bourne Ultimatum  
Poseidon,Collateral Damage,The Siege  
Pirates of the Caribbean: At World's End  
Pirates of the Caribbean: Dead Man's Chest  
Pirates of the Caribbean: On Stranger Tides  
Pirates of the Caribbean: The Curse of the  
Snow White and the Huntsman  
The Huntsman: Winter's War  
Alice in Wonderland,Alice Through the  
Looking Glass  
The Lone Ranger,The Sorcerer's Apprentice  
The Lord of the Rings: The Fellowship of  
The Lord of the Rings: The Return of the  
Lord of the Rings: The Two Towers  
The Hobbit: The Desolation of Smaug  
The Hobbit: An Unexpected Journey

Dans le troisième cluster, on trouve parmi les mots les plus fréquentés : **action, science fiction, adventure, studios, film...**  
J'ai aussi noté qu'il y avait les mots **future, Marvel, hero, mission.**

On voit que ce cluster rassemble plusieurs Marvel comme les X-Men, Captain America, Iron Man. Mais aussi des films comme Star Wars ou Hunder Games.

On peut en conclure que cette catégorie possède de nombreux films de super héros, mais aussi des films d'action avec de la violence. On retrouve beaucoup de films de fictions, et de bagarre dans ce cluster.



Beowulf
The Peanuts Movie
Percy Jackson: Sea of Monsters
Night at the Museum: Battle of the Smithsonian
Eragon
Knight and Day
Kung Fu Panda 3
The Simpsons Movie
X-Men: First Class
The Wolverine
X-Men: Days of Future Past
X-Men: Apocalypse
Fantastic Four
Independence Day: Resurgence
The Martian
Kingdom of Heaven
AVP: Alien vs. Predator
A Good Day to Die Hard
Avatar, Rise of the Planet of the Apes
Batman v Superman: Dawn of Justice
Suicide Squad
Jumper
I Am Legend, Green Lantern
Teenage Mutant Ninja Turtles
Teenage Mutant Ninja Turtles: Out of the Shadows
Allegiant, Divergent
Ender's Game, Now You See Me
Mars Attacks!, Soldier Monkeys vs Aliens, Journey to the Center of the Earth
Contact, Sphere, Interstellar
The Island, The Time Machine
John Carter, Captain America: The Winter Soldier
Pacific Rim, Oblivion
Edge of Tomorrow, The Last Witch Hunter
Cloud Atlas, TRON: Legacy
Star Wars: Episode II - Attack of the Clones
Star Wars: Episode III - Revenge of the Sith
Star Wars: Episode I - The Phantom Menace
The Hunger Games
The Hunger Games: Mockingjay - Part 1
The Hunger Games: Catching Fire
Transformers: Dark of the Moon
Transformers: Age of Extinction
Transformers: Revenge of the Fallen
Ant-Man
Iron Man 2, Iron Man 3, Iron Man
Avengers: Age of Ultron, Guardians of the Galaxy
The Avengers, Captain America: Civil War

L'analyse de cluster n'est pas une tâche automatique, mais un processus de découverte de connaissances qui implique des essais et des échecs. Il a été souvent nécessaire de **modifier les paramètres de prétraitement des données et le nombre de cluster jusqu'à ce que le résultat final soit satisfaisant.**

En analysant l'ensemble de mes clusters, j'ai pu admettre quelques conclusions:

- Tout mes clusters ont une grande apparition du mot **adventure**.

- Le **cluster 1** rassemble tous les films d'animations Disney, mais pas tout les films animés;

Exemple : Bob l'éponge classé dans le cluster 2, je précise quand même que c'est un film bob l'éponge inédit où les personnages sorte de l'animation pour rentrer dans le monde réelle)

- Les **deux derniers clusters** sont souvent composé de films **d'action**.
- Le **cluster 3** rassemble les films Marvel. Pour arrivé à cette conclusion j'ai du me renseigner car on pense souvent que Spider Man (cluster 2) est un Marvel. Mais c'est Sony qui produit et possède les droits cinématographique des sagas Spider-Man depuis 1999.
- En regardant les titres, on remarque aussi que le **3e clusters** possèdent des films plus violent que le **1er et le 2e cluster**.

Le classement de certains films, comme The Simpsons Movie ou Kunfu Panda 3 dans le cluster 3 m'ont amenées à me poser plus de questions sur la façon dont les films ont été repartie entre les clusters.

- Le classement de Kunfu Panda 3 dans le cluster 3 alors que Kunfu Panda et Kunfu Panda 2 sont classé dans le cluster 1, peut être justifier par le changement de productions lors de la réalisation de Kunfu Panda 3.
- Pour the Simpsons Movie, je pense que c'est justifier par son synopsis : « After Homer accidentally pollutes the town's water supply, Springfield is encased in a gigantic dome by the EPA and the Simpsons are declared fugitives. » Ainsi on pourrait croire que the Simpsons Movie est un film apocalyptique, ce qui correspond plutôt bien aux films du cluster 3.

**Pour conclure notre analyse**, je remarque que chaque catégorie correspond à un public différent.

Malgré plusieurs limites, je peux affirmer que mon clustering hiérarchique propose trois clusters qui regroupent les films en trois catégories :

- **Le premier cluster** rassemble les films pour enfants et des films familiales. On y trouve fréquemment des compagnies de production comme DreamWorks Animation et Walt Disney Animation Studios.
- Dans **le deuxième cluster**, on retrouve des films cultes, souvent avec une histoire touchante, des personnages attachants et de l'action. On retrouve beaucoup de films avec un monde fantastique.
- **Le dernier cluster** qu'en a lui a une dimension apocalyptique, avec des supers héros qui sauvent le monde ect.. Le **cluster 3** rassemblent les films les plus violent adapter à un public averties. On retrouve beaucoup de productions Marvel Studios et TSG Entertainment.

## 2 - Apprentissage supervisé

### 2.1 - Jeux de données

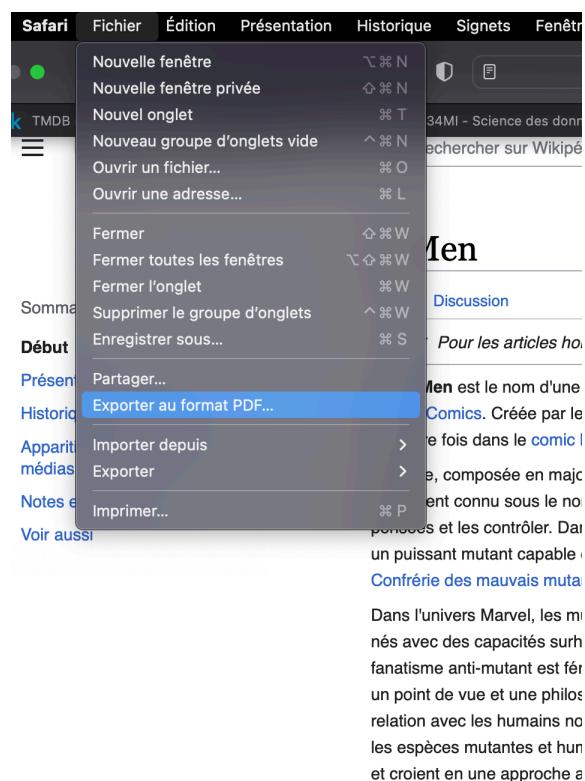
Lors d'analyse de données, la première étape consiste à **récupérer un jeu de données et de le formater**.

Pour la deuxième partie de mon projet, j'ai décidé de créer mon propre jeu de données.

J'ai alors créé un dossier nommé **Film Wik** avec cinq fichiers nommés Comédie, Romance, Science fiction, Documentaire et Horreur.

Nom	Date de modification	Taille	Type
> Comedie	27 nov. 2022 à 09:06	--	Dossier
> Documentaire	27 nov. 2022 à 09:35	--	Dossier
> Horreur	27 nov. 2022 à 09:12	--	Dossier
> Romance	hier à 23:02	↑ 153 ko	Dossier
> Science fiction	27 nov. 2022 à 09:07	↑ 291 ko	Dossier

Ils contiennent un ensemble de texte issus de wikipedia classée par thème. Chaque texte porte sur un film et a été généré directement de la page wikipedia correspondante en cliquant sur Télécharger comme PDF.



## 2.2 - Nettoyage et prétraitement des données

J'ai ouvert un nouveau fichier **Orange**.

Pour charger un dossier contenant des fichiers, j'utilise l'outil **Import Corpus** qui se trouve dans le menu Text Mining. A la suite j'ajoute l'outil **Corpus Viewer** pour pouvoir visualiser mes données. On peut remarquer qu'Orange a automatiquement créé une variable **category** contenant le nom du dossier dans lequel se trouvait le fichier.

**C'est la variable que nos modèles vont apprendre à prédire.**

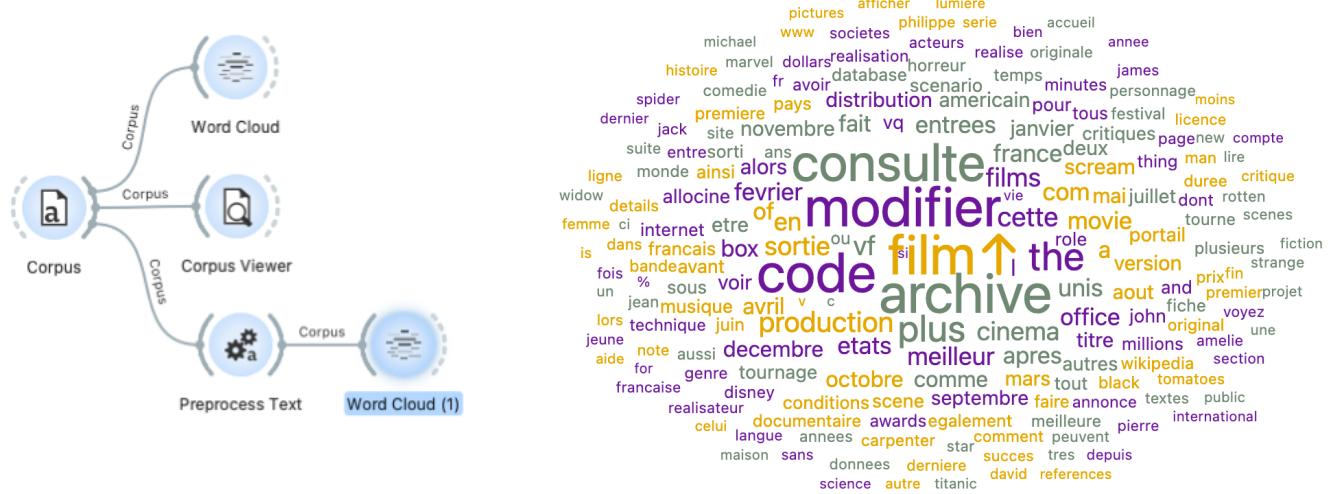


A la suite, je vais prétraiter mes données en ajoutant l'outil **Preprocess Text**. J'y indique que je souhaite que les mots d'arrêt soit en français et je rajoute mon fichier « mot interdit.txt »

Pour supprimer les nombres, il faut filtrer les **Numbers**. Je sélectionne **Regexp** pour supprimer les signes de ponctuation et **Remove URT** pour supprimer les mots issus des URL des sites Web.

The screenshot shows the 'Preprocess Text' window in Orange. On the left, there's a sidebar with 'Preprocessors' and 'Preview' sections. Under 'Preprocessors', 'Stopwords' is selected. The main area has two tabs: 'Filtering' and 'Transformation'. In the 'Filtering' tab, 'Stopwords' is set to 'French' and 'Mot interdit' is set to '(none)'. There are checkboxes for 'Numbers' (unchecked), 'Regexp' (checked), and 'Document frequency' (unchecked). 'Regexp' has a field containing a regular expression pattern. In the 'Transformation' tab, there are checkboxes for 'Lowercase', 'Remove accents', 'Parse html', and 'Remove urls'. On the right, the 'Output' pane shows the resulting text from the preprocessing steps, which includes several French words and some punctuation.

Ensuite, j'ajoute l'outil **Word Cloud** pour visualiser les mots les plus fréquent issus du **Preprocess Text**



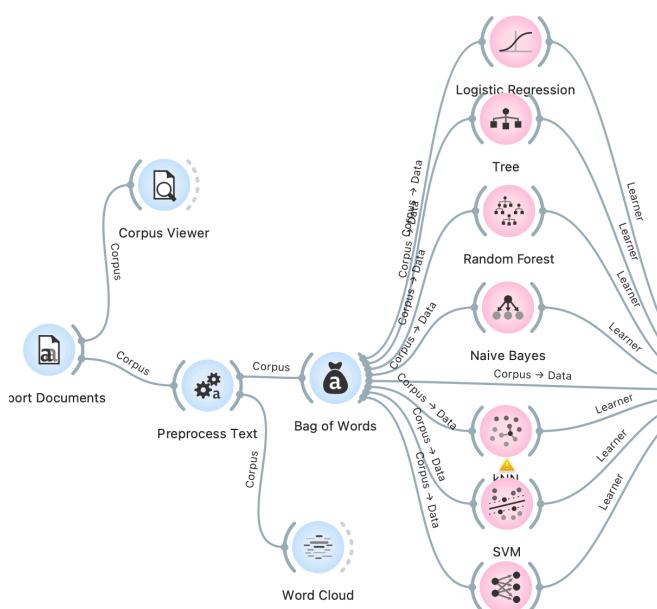
## 2.3 - Modèles d'apprentissage

Maintenant que mes données sont propres, je peux mettre en place un algorithme de classification supervisée.

Je commence par ajouter l'outil **Bag of Words** (menu text Mining) en sortie de **Proprocess Text**.

Je met en place une chaîne de traitement permettant d'entraîner un modèle **Logistic Régression** à reconnaître les catégories.

J'ai ensuite comparée **Logistic Regression** avec les autres modèles de classification évoqués en cours. J'ai ajouté les modèles disponibles dans le menu Model : **kNN, Naive Bayes, SVM, Tree, Random Forest, Neural Network**



En sortie des modèles de classifications , j'ajoute l'outil **Test and Score**. Cet outil doit aussi prendre en entrée les données issues du **Corpus de texte**.

Il permet de générer une **validation croisée du modèle**.

Pour chaque modèle entraîné, l'outil **Test and Score**, permet d'observer son **accuracy** (CA), sa **f-mesure** (F1), sa **précision** (Precision) et son **rappel**(Recall).

Lorsqu'on ouvre la **matrice de confusion** la première colonne indique que le modèle a une précision..

Le modèle le plus performant est **Logistic Régression** car son accuracy (0,756), sa f-mesure (0,756), sa précision (0,763) et son rappel (0,756) sont les plus élevées.

Je note aussi que le modèle **Tree** et **Neural Network** ont des résultats proche du modèle **Logistic Régression**.

The screenshot shows the Weka interface with the 'Test and Score' configuration panel on the left and the 'Evaluation results' table on the right.

**Test and Score Configuration:**

- Number of folds: 5
- Stratified (checked)
- Cross validation by feature (unchecked)
- Random sampling (unchecked)
- Repeat train/test: 5
- Training set size: 66 %
- Stratified (checked)
- Leave one out (unchecked)
- Test on train data (unchecked)
- Test on test data (unchecked)

**Evaluation Results Table:**

Model	AUC	CA	F1	Precision	Recall
KNN	0.732	0.439	0.431	0.475	0.439
Tree	0.818	0.695	0.684	0.685	0.695
SVM	0.555	0.256	0.104	0.066	0.256
Random Forest	0.828	0.598	0.585	0.621	0.598
Neural Network	0.836	0.634	0.602	0.747	0.634
Naive Bayes	0.652	0.171	0.099	0.421	0.171
Logistic Regression	0.901	0.756	0.756	0.763	0.756

En sortie de **Test and Score**, j'ajoute l'outil **Confusion Matrix** (menu Evaluate)

The screenshot shows the Weka interface with the 'Confusion Matrix' result table on the right.

**Learners:**

- Logistic Regression
- kNN
- Tree
- Random Forest
- Naive Bayes
- SVM
- Neural Network

**Confusion Matrix Table:**

		Predicted					$\Sigma$	
Actual	Comedie	Comedie	Horreur	Science fiction	Romance	Documentaire		
		Comedie	16	0	4	1	0	21
		Horreur	1	12	3	1	0	17
		Science fiction	1	1	11	2	0	15
		Romance	4	1	0	12	1	18
		Documentaire	0	0	0	0	11	11
$\Sigma$	22	14	18	16	12	82		

Pour le modèle **Logistic Regression**, la première, 3e et 4e colonnes nous indiquent que le modèle n'a pas une très bonne **précision** pour les catégories : **comédie, science fiction et romance**. Mais le nombre de films bien prédit est toujours supérieur au nombre de films mal prédit. Les autres colonnes nous indiquent que la précision pour ces catégories est meilleure.

On peut observer que 62 films sont bien classés.

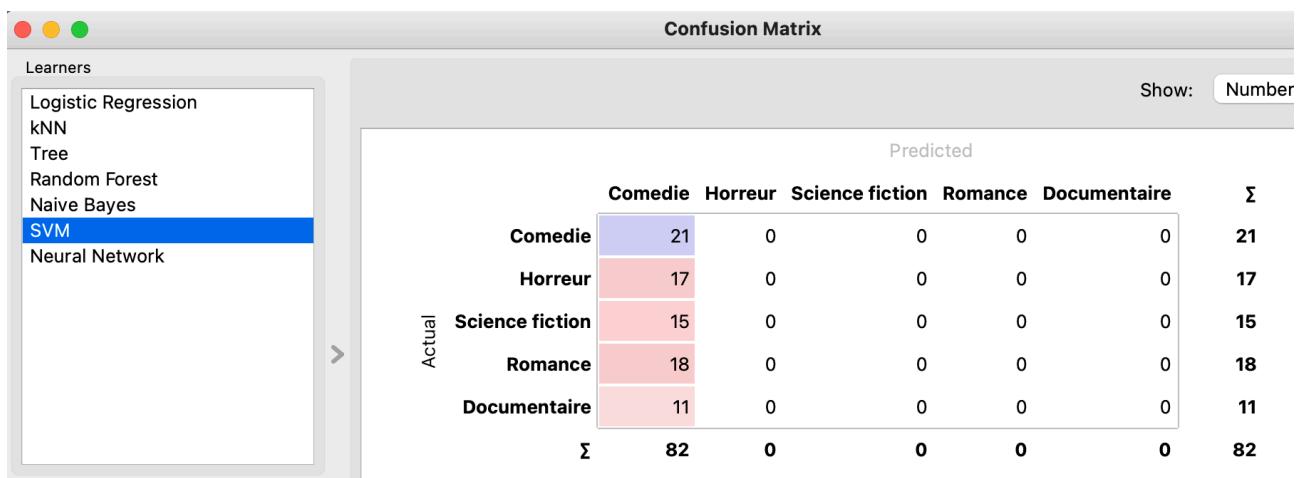
Soit 11 films de science fiction sur 15, 11 documentaires sur 11, 12 films d'horreur sur 17, 12 films romantique sur 18 et 16 comédies ont été bien classée sur 21.

Néanmoins 3 films d'horreur et 4 comédies ont été classées comme des films de science fiction, 1 film romantique a été prédit comme un documentaire; un film de romance et un film de science fiction ont été prédit comme film d'horreur; 2 films de science fiction, un film d'horreur et une comédie ont été prédit comme des films romantiques; et finalement un film de science fiction, un film d'horreur et 4 romances ont été prédits comme des comédies.

Malgré le fait que **Logistic Regression** soit le meilleur modèle, il n'est pas infaillible, il comporte beaucoup d'erreur.

Dans **Test and Score** on peut voir que par conséquence, le rappel est très bon pour documentaire (1,0) et moyenne pour horreur (0,706), comédie (0,762), science fiction (0,733) et romance (0,667) pour le modèle **Logistic Regression**.

On remarque que pour ce jeu de données, le modèle **SVM** est très mauvais.



## 2.4 - Prédictions

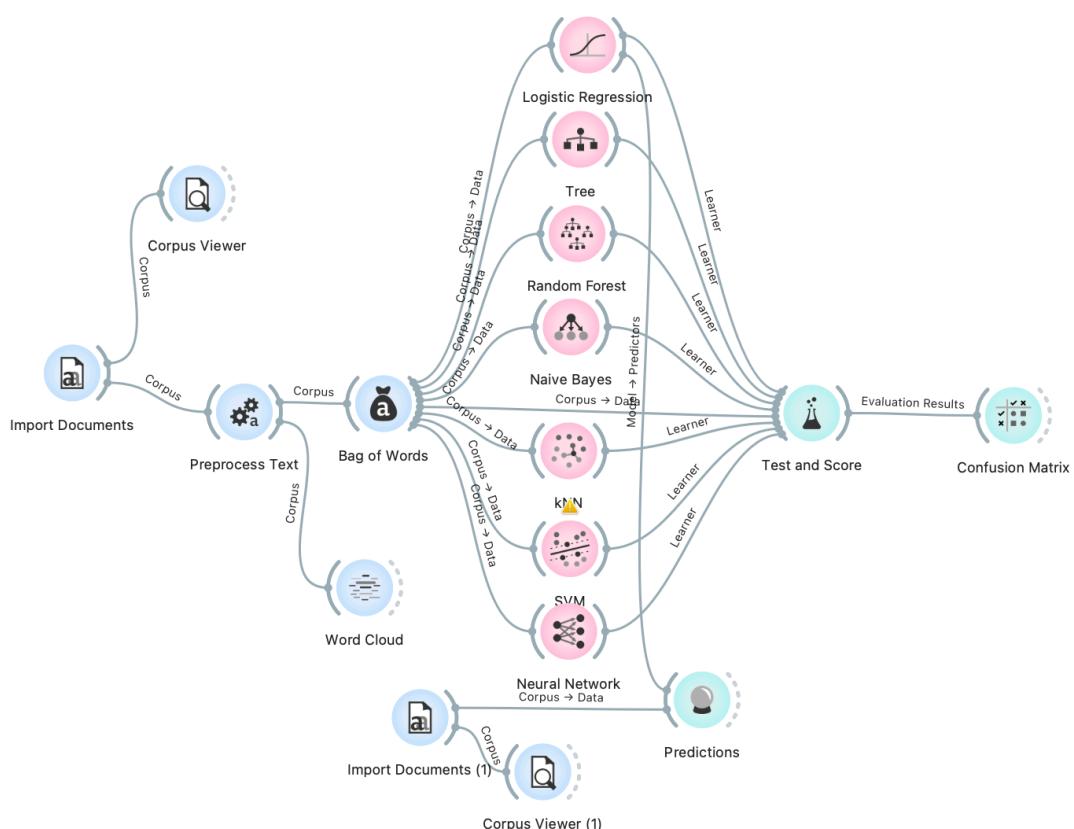
Maintenant que j'ai comparé mes modèles et identifié celui qui produit les meilleures résultats, on va pouvoir faire des prédictions.

Il va falloir tester le modèle sur des films non classés.

J'ai donc créé **un seul et unique** dossier avec des films non classés obtenue à l'aide du même procédé que celui des récoltes des données supervisé avec Wikipedia.

Par la suite j'ai chargé ce fichier dans **Orange** pour prédire les catégories des films à l'aide de l'outil **Import Documents**. En sortie j'ajoute l'outil **Corpus View**. On peut voir que le fichier contient 8 films. Par contre, il ne contient pas de variable classification permettant de déterminer si une ligne concerne un film d'horreur, de romance, de science fiction, comédie ou documentaire.  
**C'est cette variable que vous allez prédire grâce aux modèles entraînés précédemment.**

Donc en sortie **Import Documents**, j'ajoute un **Prédictions** (menu Evaluate). De plus, j'ajoute une connexion de façon à ce que **Prédictions** prenne en entrée les modèle le plus performant.



## Réelle catégorie de mes films :

**Double clique sur Prédictions :**

Titre	Genre
Bumblebee	Science fiction
Chucky	Horreur
Fire of Love	Documentaire
L'armée des douze singes	Science fiction
Five	Comedie
Lucy	Science fiction
Finch	Horreur
Remember me	Romantique

Show probabilities for		(None)	▼
	Logistic Regression		name
1	Science fiction	Bumblebee (film) — Wikipédia	
2	Comedie	Chucky (série de films) — Wikipédia	
3	Horreur	Finch (film) — Wikipédia	
4	Documentaire	Fire of Love (film) — Wikipédia	
5	Comedie	Five (film,2016) — Wikipédia	
6	Science fiction	L'Armée des douze singes — Wikipédia	
7	Comedie	Lucy (film,2014) — Wikipédia	
8	Comedie	Remember Me (film,2010) — Wikipédia	

**On peut observer que 5 films sur 8 ont été bien classé.  
Ce qui est plutôt cohérent avec le score et la matrice de confusion.**

Pour finir, j'ajoute une connexion de façon à ce que **Prédictions** prenne en entrée tous les modèles mis en place.

## Double clique sur **Prédictions** :

Predictions							Restore Original Order	
	Logistic Regression	Neural Network	SVM	kNN	Naive Bayes	Random Forest	Tree	name
1	Science fiction	Science fic...	Com...	Science fic...	Document...	Science fic...	Romance	Bumblebee (film) — Wikipédia
2	Comedie	Science fic...	Com...	Horreur	Document...	Horreur	Horreur	Chucky (série de films) — Wikipédia
3	Horreur	Science fic...	Com...	Horreur	Document...	Horreur	Horreur	Finch (film) — Wikipédia
4	Documentaire	Science fic...	Com...	Romance	Document...	Horreur	Documenta...	Fire of Love (film) — Wikipédia
5	Comedie	Science fic...	Com...	Horreur	Document...	Comedie	Comedie	Five (film,2016) — Wikipédia
6	Science fiction	Science fic...	Com...	Comedie	Document...	Comedie	Science fic...	L'Armée des douze singes — Wikipédia
7	Comedie	Science fic...	Com...	Comedie	Document...	Science fic...	Science fic...	Lucy (film,2014) — Wikipédia
8	Comedie	Science fic...	Com...	Romance	Document...	Comedie	Horreur	Remember Me (film,2010) — Wikipédia

On peut clairement voir que les modèles **Neural Network**, **SMV** et **Naive Bayes** ne sont pas fonctionnelle.

Contrairement aux modèles **kNN** qui a bien classé 4 films sur 8 et pour **Random Forest** 5 films sur 8 ont été bien classé.