



# Introduction

- Analyse du contenu proposé sur Netflix France
- Pourquoi choisir ce thème ?
  - Souhait de travailler sur une base de données filmiques ;
  - Netflix = le premier service de streaming au monde ;
  - Souhait de créer un système de recommandation.
- Comment recommander du contenu à un utilisateur sachant son historique à l'aide de réseaux ?

# Sommaire



**I - Les données**

**II - Les graphes**

**III - Les systèmes de recommandation**

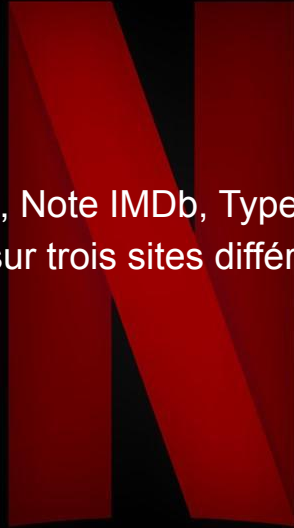
# I - Les données

- Données pour le graphe
- Données pour la recommandation



# I - Les données

- Données pour le graphe :
  - Titre, Identifiant, Genres, Casting, Note IMDb, Type du contenu
  - Récupérées par “web scraping” sur trois sites différents :
    - <https://fr.flixable.com>
    - <https://www.netflix.com/fr>
    - <https://www.imdb.com/>











# I - Les données

- Extraction des données sur [flixable.com](https://fr.flixable.com) :
  - Récupération du nombre de pages total ;

https://fr.flixable.com/?min-rating=0&min-year=1920&max-year=2020&order=title&page=1

FLIXABLE Parcourir Populaire Prochainement Bientôt déprogrammé Programmes originaux Ma liste Se connecter

|  |   |   |   |
|--|---|---|---|
| <b>48 vœux de Noël</b><br>2017 <b>IMDb</b> 4.0/10<br><br><b>5 Minarets à New York</b><br>2010 <b>IMDb</b> 5.8/10<br> | <b>4L</b><br>2019 <b>IMDb</b> 5.9/10<br><br><b>6 ans</b><br>2015 <b>IMDb</b> 5.6/10<br> | <b>4th Man Out</b><br>2015 <b>IMDb</b> 6.7/10<br><br><b>6 Balloons</b><br>2018 <b>IMDb</b> 5.8/10<br> | <b>5 Fingers</b><br>2018<br><br><b>6 Days</b><br>2017 <b>IMDb</b> 6.2/10<br><br><b>90 minutes au paradis</b><br>2015 <b>IMDb</b> 4.8/10 |
|--|---|---|---|

Affichage des titres de 1 à 40 sur 2,515

« 1 2 3 4 5 63 »

© 2020 Flixable  
À propos de nous | Nous contacter | Règles de confidentialité | Autres pays

# I - Les données

- Extraction des données sur [flixable.com](#) :
  - Récupération du nombre de pages total ;
  - Récupération des identifiants en parcourant toutes les pages de listage ;

¡Ay, mi madre!  
2019 IMDb 3.9/10

#HORROR  
2015 IMDb 3.0/10

#REALITYHIGH  
2017 IMDb 5.2/10

#RUCKER50  
2016 IMDb 5.4/10

#celle69

Inspecteur Console Débugueur Réseau Éditeur de style Performances Mémoire Stockage Accessibilité Adblock Plus

Rechercher dans le HTML

```
<a class="title" href="/title/80997613/">¡Ay, mi madre!</a>  
</strong>  
</div>  
<div class="rating-container">  
<div class="d-inline-block">2019</div>  
<div class="d-inline-block ml-2 plugin">
```

# I - Les données

- Extraction des données sur [flixable.com](https://fr.flixable.com/title/80027042/) :
  - Récupération du nombre de pages total ;
  - Récupération des identifiants en parcourant toutes les pages de listage ;
  - Récupération des données en parcourant toutes les pages descriptives des films/séries.



## Flash

2018 13+ 4 saisons

Dans cette comédie d'action de super-héros, un expert de la police sort du coma avec des facultés surhumaines et part en guerre contre tous ceux qui menacent sa ville.

Genres : Action, Policier, SF et fantastique

Réalisateur : Glen Winter

Acteurs : Grant Gustin, Candice Patton, Danielle Panabaker, Carlos Valdes, Tom Cavanagh, Jesse L. Martin, Neil Sandilands, Britne Oldford, Danielle Nicolet, Keiynan Lonsdale, Kim Engelbrecht

IMDb 7.8/10

Cliquez pour noter : ★★★★★

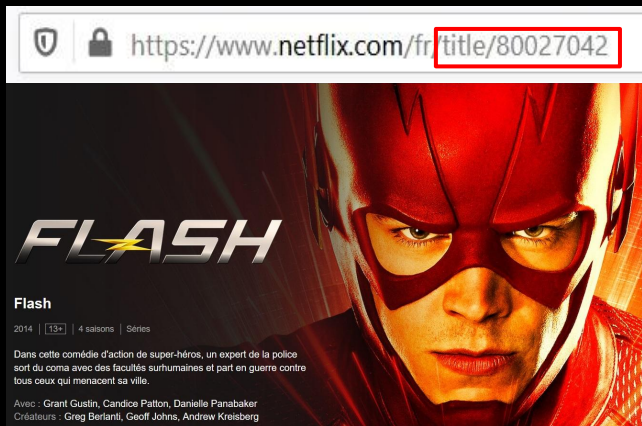
Date d'ajout de la nouvelle saison : 1/3/2019

REGARDER SUR NETFLIX + MA LISTE



# I - Les données

- Extraction des données sur [netflix.com/fr](https://www.netflix.com/fr/title/80027042) :
  - Récupération des données en parcourant toutes les pages descriptives des films/séries.



## Plus d'infos

Lecture hors ligne  
Téléchargeable

Sous-titres  
anglais, français

Distribution  
Grant Gustin  
Tom Cavanagh  
Hartley Sawyer

### Genres

Séries SF,  
Séries inspirées de bandes dessinées,  
Séries US, Séries d'action et d'aventure

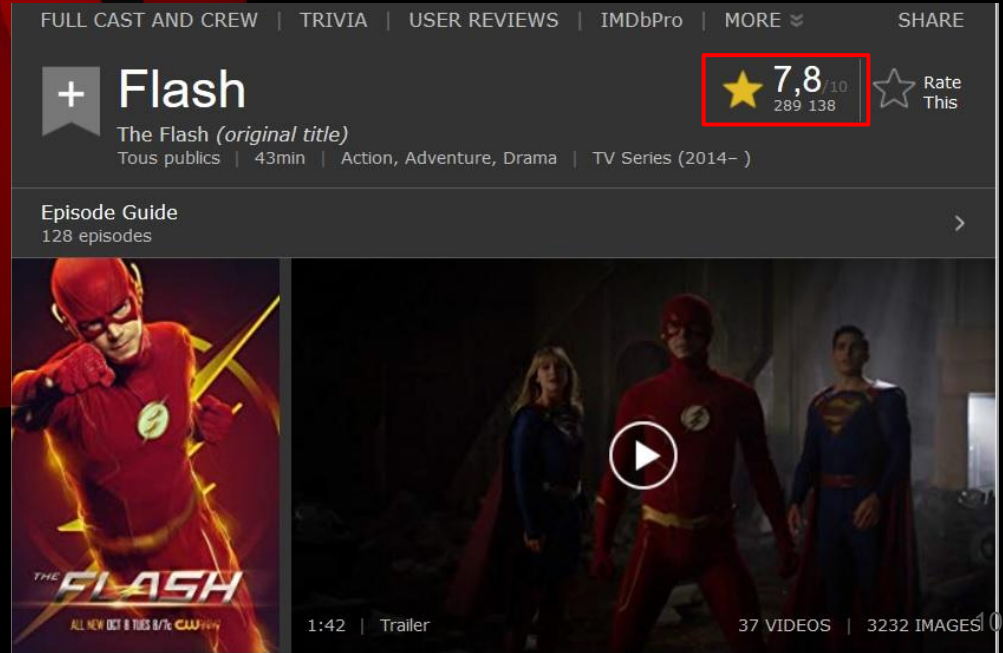
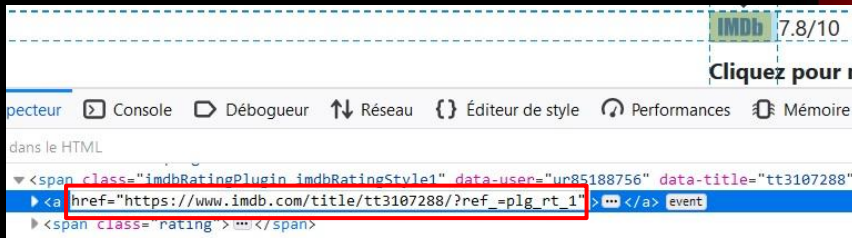
Ce programme est...  
Montée d'adrénaline, Palpitant

Candice Patton  
Jesse L. Martin  
Jessica Parker Kennedy

Danielle Panabaker  
Keiynan Lonsdale

# I - Les données

- Extraction des données sur imdb.com :
  - Récupération des données en parcourant toutes les pages descriptives des films/séries.



# I - Les données

- Tableau de données :

```
BDD_Netflix.head()
```

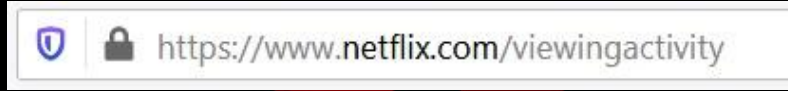
|   | Titre          | Genres Netflix                                  | Casting                                       | Note IMDb | Date d'ajout | Lien Netflix     | Type |
|---|----------------|---|---|-----------|--------------|------------------|------|
| 0 | ¡Ay, mi madre! | Comédies Films espagnols                        | Estefanía de los Santos Secun de la Rosa T... | 3.9       | 2019-07-19   | /title/80997613/ | film |
| 1 | #Horror        | Films d'horreur Films indépendants              | Chloë Sevigny Timothy Hutton Natasha Lyonn... | 3.0       | 2016-04-30   | /title/80074904/ | film |
| 2 | #realityhigh   | Comédies Films pour ados                        | Nesta Cooper Kate Walsh John Michael Higgi... | 5.2       | 2017-09-08   | /title/80125979/ | film |
| 3 | #Rucker50      | Docus société et culture Films documentaires... | NaN   | 5.3       | 2016-11-30   | /title/80147908/ | film |
| 4 | #Selfie        | Comédies Films roumains Drames Films pou...     | Flavia Hojda Crina Semciuc Olimpia Melinte... | 6.1       | 2019-06-01   | /title/81092768/ | film |

# I - Les données

- Données pour la recommandation
  - Titre
  - Récupérées sur Netflix
  - Nettoyées à l'aide d'expressions régulières



# I - Les données



|    | A   | B          |
|----|---|------------|
| 1  | Title   | Date       |
| 2  | En bref: Saison 1: La crise mondiale de l'eau     | 22/01/2020 |
| 3  | The End of the F***ing World: Saison 2: Épisode 8 | 19/01/2020 |
| 4  | The End of the F***ing World: Saison 2: Épisode 7 | 19/01/2020 |
| 5  | The End of the F***ing World: Saison 2: Épisode 6 | 19/01/2020 |
| 6  | The End of the F***ing World: Saison 2: Épisode 5 | 19/01/2020 |
| 7  | The End of the F***ing World: Saison 2: Épisode 4 | 19/01/2020 |
| 8  | The End of the F***ing World: Saison 2: Épisode 3 | 19/01/2020 |
| 9  | The End of the F***ing World: Saison 2: Épisode 2 | 19/01/2020 |
| 10 | The End of the F***ing World: Saison 2: Épisode 1 | 19/01/2020 |

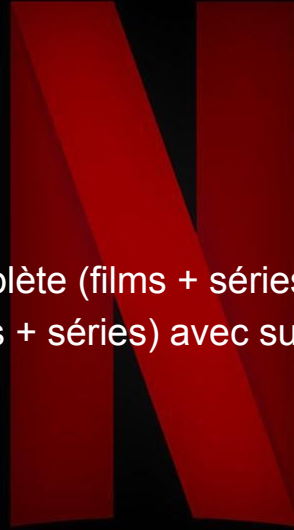
CleanHistory\_user1.txt

- 6 Underground
- After Life
- Ahmed Sylla - Avec un grand A
- American Psycho
- American Vandal
- Ano Hi Mita Hana no Namae o Bokutachi wa Mada Shiranai
- Archer
- Atlantide : L'empire perdu
- Atypical
- Avatar
- B: The Beginning
- BTOOOM!
- Bill Burr: I'm Sorry You Feel That Way
- Bill Burr: Paper Tiger
- Bill Burr: Walk Your Way Out
- Black Bullet
- Black Butler
- Black Butler : Book of the Atlantic
- Black Mirror

## II - Les graphes

- Création du graphe :
  - **Noeud** :
    - film/série (id\_netflix)
  - **Lien** : deux films/séries ont un lien si ils ont, soit :
    - au moins un genre en commun
    - au moins un acteur en commun
  - **Poids du noeud** :
    - note IMDb
  - **Poids d'un lien** : correspond à la somme :
    - du nombre de genres en communs entre deux films/séries
    - du nombre d'acteurs en communs entre deux films/séries

## II - Les graphes



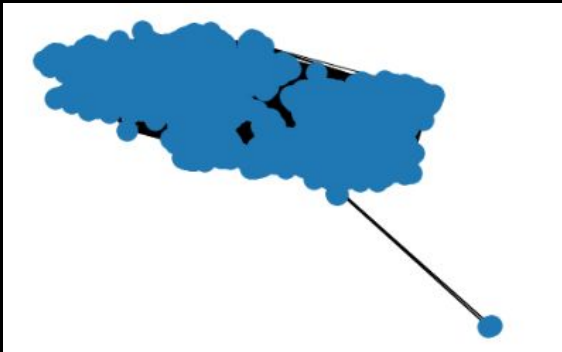
- Deux graphes :
  - Sur notre base de données complète (films + séries)
  - Sur notre base de données (films + séries) avec suppression des 3 genres les plus fréquents

## II - Les graphes

- Visualisation des graphes :

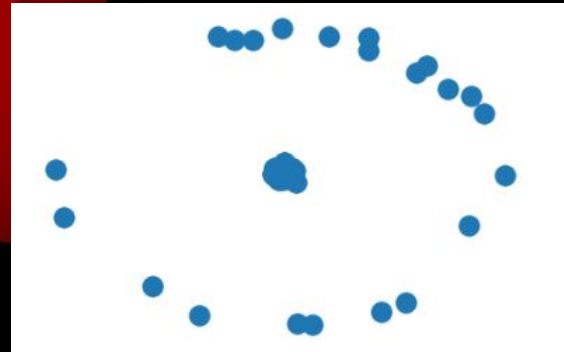
Avec les 3 genres les plus fréquents (Séries + Films) :

Graphe hyper-connecté



Sans les 3 genres les plus fréquents (Séries + Films) :

Films/Séries en dehors du cercle ayant auparavant  
comme genres : Drame, Comédie ou Films documentaires





## II - Les graphes

- Définition du graphe :

**Base de données complète (Films + Séries) :**

| number_nodes | number_edge | degree min | degree max | degree mean | density graph |
|--------------|-------------|------------|------------|-------------|---------------|
| 3996.0       | 1383195.0   | 4.0        | 1714.0     | 692.29      | 0.17          |

**Base de données complète (Films + Séries) sans les 3 genres les plus fréquents :**

| number_nodes | number_edge | degree min | degree max | degree mean | density graph |
|--------------|-------------|------------|------------|-------------|---------------|
| 3974.0       | 829014.0    | 4.0        | 1100.0     | 417.22      | 0.11          |

- Diminution du degré maximal avec suppression des 3 genres les plus fréquents dans les bases de données
- Densité du graphe faible : beaucoup de noeuds ne sont pas reliés entre eux
- Diminution de la densité avec suppression des 3 genres les plus fréquents : suppression de liens

## II - Les graphes

- Définition du graphe :
  - Les films ayant le degré minimal et maximale :
    - sur la base complète (film + series) :

| movie      |   |
|------------|---|
| degree min | Une cheminée à la maison : Version Bois de bou... |
| degree max | The Dirt  |

- sur la base complète (film + séries) avec suppression des 3 genres les plus fréquents :

| movie      |             |
|------------|-------------|
| degree min | Maktub      |
| degree max | Bleu Saphir |

## II - Les graphes

### - Centralité de degré :

Avec les 3 genres les plus fréquents (Séries + Films) :

|   | title  | degree centrality   |
|---|--|---------------------|
| 0 | The Dirt   | 0.4290362953692115  |
| 1 | Saint-Jacques... La Mecque                       | 0.4210262828535669  |
| 2 | My Deer Hunter Dad                               | 0.4157697121401752  |
| 3 | El Alien y Yo                                    | 0.4135168961201502  |
| 4 | Dave Chappelle: Equanimity & The Bird Revelation | 0.4127659574468085  |
| 5 | La Guerre selon Charlie Wilson                   | 0.4117647058823529  |
| 6 | Julie & Julia                                    | 0.4112640801001251  |
| 7 | Docteur Patch                                    | 0.4110137672090112  |
| 8 | Colkatay Columbus                                | 0.40826032540675844 |
| 9 | Dave Chappelle                                   | 0.3997496871088861  |

Sans les 3 genres les plus fréquents (Séries + Films) :

|   | title                                     | degree centrality   |
|---|---|---------------------|
| 0 | Bleu Saphir                               | 0.27534418022528157 |
| 1 | Rouge rubis                               | 0.27534418022528157 |
| 2 | Tiffany Haddish Presents: They Ready      | 0.2658322903629537  |
| 3 | Irish Gangster                            | 0.2618272841051314  |
| 4 | HUMORISTES du monde                       | 0.25957446808510637 |
| 5 | Historical Roasts                         | 0.2585732165206508  |
| 6 | Bumping Mics with Jeff Ross & Dave Attell | 0.2575719649561952  |
| 7 | Daniel Sloss: Live Shows                  | 0.2575719649561952  |
| 8 | Ellen DeGeneres: Relatable                | 0.2575719649561952  |
| 9 | Stand Up and Away! with Brian Regan       | 0.2575719649561952  |

Films contenant les genres : Drame et Comédie

- Suppression des 3 genres : modifie la liste des films ayant la meilleure centralité de degré

## II - Les graphes

### - Centralité de proximité :

Avec les 3 genres les plus fréquents (Séries + Films) :

|   | title  | closeness centrality |
|---|--|----------------------|
| 0 | The Dirt   | 0.6320202499604493   |
| 1 | Dave Chappelle: Equanimity & The Bird Revelation | 0.6300268096514745   |
| 2 | El Alien y Yo                                    | 0.6258812470625098   |
| 3 | Dave Chappelle                                   | 0.6249022368215236   |
| 4 | Kabhi Haan Kabhi Naa                             | 0.6182296502630764   |
| 5 | Holy Campl                                       | 0.6164172195648819   |
| 6 | The Clapper                                      | 0.6136712749615976   |
| 7 | Shakespeare in Love                              | 0.6127300613496932   |
| 8 | Alex Strangelove                                 | 0.611791730474732    |
| 9 | The Laundromat : L'affaire des Panama Papers     | 0.6106695200244574   |

Sans les 3 genres les plus fréquents (Séries + Films) :

|   | title                    | closeness centrality |
|---|--------------------------|----------------------|
| 0 | Bleu Saphir              | 0.5634799132115372   |
| 1 | Rouge rubis              | 0.5634799132115372   |
| 2 | Irish Gangster           | 0.5589363631969584   |
| 3 | Les Misérables           | 0.5586992578392673   |
| 4 | Glee                     | 0.5527589747396893   |
| 5 | Avril et le Monde truqué | 0.5507556664955812   |
| 6 | Le cœur a ses raisons    | 0.5494536436433457   |
| 7 | Crazy Ex-Girlfriend      | 0.5489955747449353   |
| 8 | Neo Yokio                | 0.548919304173284    |
| 9 | Chair de Poule - le film | 0.5478537370270796   |

Films contenant les genres : Drame et Comédie

- Suppression des 3 genres : modifie la liste des films ayant la meilleure centralité de proximité

## II - Les graphes

### - Centralité d'intermédierité :

Avec les 3 genres les plus fréquents (Séries + Films) :

|   | title   | betweenness centrality |
|---|---|------------------------|
| 0 | Dave Chappelle: Equanimity & The Bird Revelation  | 0.026157159918781033   |
| 1 | Dave Chappelle                                    | 0.02196147470146899    |
| 2 | James Acaster: Repertoire                         | 0.019186778746769377   |
| 3 | Jenny Slate: Stage Fright                         | 0.009951570016259736   |
| 4 | Black Mirror: Bandersnatch                        | 0.006406353760270023   |
| 5 | Sur le gril : Quand se nourrir est un défi        | 0.004640184360373949   |
| 6 | Notes on Blindness                                | 0.004430841656475235   |
| 7 | Capitaine Superslip et le conte terrifiant du ... | 0.004006386640633281   |
| 8 | Lost & Found : Le studio                          | 0.003971545985069681   |
| 9 | La véritable Miyagi                               | 0.003942845113996617   |

Sans les 3 genres les plus fréquents (Séries + Films) :

|   | title                                      | betweenness centrality |
|---|--|------------------------|
| 0 | Black Mirror: Bandersnatch                 | 0.008468370983950877   |
| 1 | The Creative Brain                         | 0.005260341953212939   |
| 2 | Tig  | 0.00525269445643461    |
| 3 | Poursuis tes rêves                         | 0.005156712539119102   |
| 4 | Müslüm                                     | 0.005113684419084162   |
| 5 | La véritable Miyagi                        | 0.005029508553221354   |
| 6 | Sur le gril : Quand se nourrir est un défi | 0.004739450501657141   |
| 7 | Russell Peters vs. the World               | 0.0046092635018642825  |
| 8 | Lost & Found : Le studio                   | 0.004544813653110298   |
| 9 | Naruto the Last : Le film                  | 0.004535741959527441   |

Films contenant les genres : Drame et Comédie

- Suppression des 3 genres : modifie la liste des films ayant la meilleure centralité d'intermédierité

## II - Les graphes

### - Centralité de vecteur propre :

Avec les 3 genres les plus fréquents (Séries + Films) :

|   | title                          | eigen vector centrality |
|---|--------------------------------|-------------------------|
| 0 | Saint-Jacques... La Mecque     | 0.03631153761617287     |
| 1 | My Deer Hunter Dad             | 0.03613677014994721     |
| 2 | The Dirt                       | 0.03602386763667886     |
| 3 | The Clapper                    | 0.03600237511064578     |
| 4 | Colkatay Columbus              | 0.03590479985732871     |
| 5 | La Guerre selon Charlie Wilson | 0.03587123023442041     |
| 6 | Julie & Julia                  | 0.0358440825456381      |
| 7 | Docteur Patch                  | 0.03584245144722309     |
| 8 | 2 States                       | 0.035842391964049106    |
| 9 | The Zoya Factor                | 0.035842391964049106    |

Sans les 3 genres les plus fréquents (Séries + Films) :

|   | title  | eigen vector centrality |
|---|--|-------------------------|
| 0 | Girlboss                                       | 0.04637151897687547     |
| 1 | Special  | 0.04637151897687547     |
| 2 | The Politician                                 | 0.04596808711833725     |
| 3 | Haters Back Off                                | 0.04591783813949129     |
| 4 | Neo Yokio                                      | 0.04587184513296487     |
| 5 | Tiffany Haddish Presents: They Ready           | 0.04560424233793661     |
| 6 | Les curieuses créations de Christine McConnell | 0.045263817113500665    |
| 7 | IZombie  | 0.0452612816805138      |
| 8 | Petits coups de pouce avec Carol Burnett       | 0.0451038562820157      |
| 9 | Un Noël chez Mamilia                           | 0.04509976982922144     |

Films contenant les genres : Drame et Comédie

- Suppression des 3 genres : modifie la liste des films ayant la meilleure centralité d'intermédierité

# III - Les systèmes de recommandation

- 4 systèmes de recommandation :
  - content-based recommendations
- Jaccard et Adamic/Adar : mesures de similarité pour de la prédiction de lien
  - Disponible sur le module networkx
  - Mesures basiques

$$J(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

$$A(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log(|\Gamma(w)|)}$$

$\Gamma(u)$  : l'ensemble des voisins de u.

# III - Les systèmes de recommandation

- Concernant l'historique étudié

## Genres

|                              |         |
|------------------------------|---------|
| Comédies                     | 31,33 % |
| Séries japonaises            | 30,12 % |
| Séries d'animation japonaise | 30,12 % |
| Séries US                    | 25,30 % |
| Séries comiques              | 24,10 % |
| Séries inspirées de mangas   | 18,07 % |
| Anime SF et fantastique      | 16,87 % |
| Séries primées               | 15,66 % |
| Stand-up                     | 14,46 % |
| Anime d'action               | 12,05 % |
| Séries                       | 12,05 % |
| Thrillers TV                 | 12,05 % |
| Séries british               | 9,64 %  |
| Anime dramatique             | 9,64 %  |
| Séries pour ados             | 9,64 %  |



## Casting

|                  |        |
|------------------|--------|
| Yuki Kaji        | 7,95 % |
| Maaya Sakamoto   | 5,68 % |
| Yukari Tamura    | 5,68 % |
| Jun Fukuyama     | 4,55 % |
| Kana Hanazawa    | 4,55 % |
| Ai Kayano        | 4,55 % |
| Takahiro Sakurai | 4,55 % |
| Kazuya Nakai     | 4,55 % |
| Aoi Yuki         | 4,55 % |
| Bill Burr        | 3,41 % |
| Daisuke Ono      | 3,41 % |
| Junichi Suwabe   | 3,41 % |
| Dave Chappelle   | 3,41 % |
| Mariya Ise       | 3,41 % |
| Hiroki Yasumoto  | 3,41 % |



# III - Les systèmes de recommandation

- Pour le graphe avec les 3 genres fréquents

| Jaccard et avec poids des liens           |         | Adamic/Adar et avec poids des liens              |           |
|---|---------|--|-----------|
| Saiki Kusuo no Ψ Nan : Le retour          | 83,878  | La Délicatesse                                   | 7841,8651 |
| Levius                                    | 80,9072 | Paris  | 7528,2332 |
| Magi: Adventure of Sinbad                 | 80,6887 | Chroniques sexuelles d'une famille d'aujourd'hui | 7507,6824 |
| 7SEEDS                                    | 78,8554 | Avril et le Monde truqué                         | 7358,9727 |
| Witchcraft Works                          | 76,3877 | Saint-Jacques... La Mecque                       | 7236,1345 |
| Jaccard et sans poids des liens           |         | Adamic/Adar et sans poids des liens              |           |
| Aziz Ansari Live at Madison Square Garden | 23,2049 | Saint-Jacques... La Mecque                       | 4599,7183 |
| Aziz Ansari: Buried Alive                 | 23,2049 | My Deer Hunter Dad                               | 4581,9886 |
| Aziz Ansari: RIGHT NOW                    | 23,2049 | Colkatay Columbus                                | 4479,7199 |
| Adam Devine: Best Time of Our Lives       | 23,2048 | Avril et le Monde truqué                         | 4469,9205 |
| ADAM SANDLER 100% FRESH                   | 23,2048 | Band of Robbers                                  | 4469,5217 |

# III - Les systèmes de recommandation

- Pour le graphe sans les 3 genres fréquents

| Jaccard et avec poids des liens  |         | Adamic/Adar et avec poids des liens       |           |
|----------------------------------|---------|---|-----------|
| Saiki Kusuo no Ψ Nan : Le retour | 83,878  | HUMORISTES du monde                       | 6387,3762 |
| Levi's                           | 80,9072 | Historical Roasts                         | 6386,6381 |
| Magi: Adventure of Sinbad        | 80,6887 | Bumping Mics with Jeff Ross & Dave Attell | 6385,7446 |
| 7SEEDS                           | 78,8554 | The Degenerates                           | 6385,7446 |
| Witchcraft Works                 | 76,3877 | The Honeymoon Stand Up Special            | 6385,7446 |
| Jaccard et sans poids des liens  |         | Adamic/Adar et sans poids des liens       |           |
| Insatiable                       | 20,8145 | Tiffany Haddish Presents: They Ready      | 3507,2939 |
| The Politician                   | 20,5997 | HUMORISTES du monde                       | 3457,471  |
| Haters Back Off                  | 20,5925 | Historical Roasts                         | 3455,305  |
| Santa Clarita Diet               | 20,5875 | Bumping Mics with Jeff Ross & Dave Attell | 3452,1117 |
| Astronomy Club: The Sketch Show  | 20,4925 | Daniel Sloss: Live Shows                  | 3452,1117 |

# Conclusion

- Systèmes de recommandation :
  - semblent correct
  - pas moyen de vérifier leur précision
  - Allez plus loin : quelques améliorations possibles :
    - prendre en compte le poids des noeuds (notes IMDb)
    - changer le poids des liaisons (importance plus forte pour le genre que le casting, tags)
    - système de recommandation en séparant films et séries
  - Ce qu'il manque :
    - notes du contenu regardé par l'utilisateur
    - retour utilisateur sur la recommandation
- Codes : [https://github.com/valentinecarteronGitHub/Projet-Magist-re\\_Graphes](https://github.com/valentinecarteronGitHub/Projet-Magist-re_Graphes)

A large, stylized red letter 'N' with a 3D effect, positioned in the center of the slide. It has a slight shadow and a gradient, giving it a sense of depth.

Merci pour votre attention !  
Avez vous des questions ?