# Analyse de co-occurence: le loup de wall street

GHADDAR Mohamad Ali, DENFIR Khaled, BELAIBA Yannis,
ABDALLAH Ali,  BANKUMUKUNZI Juste

# INTRODUCTION

- choix du film: le loup de wall street
- **Problématique**: analyse des relations entre les différents personnages
- analyse par scène: comptage du nombre d'apparition par personage dans chaque scene

# PRÉSENTATION DES DONNÉES

- Script data
- 289 Scenes
- personnages
- Text to table

# PRÉSENTATION DES DONNÉES

| | scene | jordan | donnie | naomi | patrick | max | brad | mark | manny | jean-jacques | emma | teresa | leah |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1a | 1a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1b | 1b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1c-1d | 1c-1d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1e | 1e | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1f | 1f | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3-3b | 3-3b | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 4 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 6 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7-7c | 7-7c | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 8 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | 30 | 8 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 10 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 11 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 12 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 13 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 14 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14a | 14a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 19 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 20 | 20 | 12 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 |
| 21 | 21 | 12 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 0 |
| 22 | 22 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | 23 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

DE CE FAIT, NOUS AVONS CHOISIT DEUX MÉTHODES AFIN D'ANALYSER LES DONNÉES NÉTOYÉES:

# MÉTHODE 1: REGLES D'ASSOCIATION

# MÉTHODE 2: TF-IDF

# ASSOCIATION RULES

# MÉTHODE 1

Définition:

- Les règles d'association sont une technique d'exploration de données utilisée pour identifier des relations intéressantes entre des variables binaires dans une base de données transactionnelle.
- Elles sont basées sur la mesure de deux quantités : la fréquence d'apparition de chaque élément individuel (l'itemset) dans la base de données, et la probabilité conditionnelle d'apparition d'un itemset donné, étant donné l'apparition d'un autre itemset.

# MÉTHODE 1

Les mesures utilisées:

- le support: la fréquence d'apparation d'un élément dans un document ou texte
- le confidence: apparition d'un élement B en fonction d'un élément A. P(B|A)
- le lift :comparaison de la fréquence observée d'une paire d'items à leur fréquence attendue si les deux items étaient indépendants.

# EXEMPLES

# TF-IDF

# MÉTHODE 2

Classer les documents en fonction de leur pertinence par rapport à une requête donnée.

- Evaluer l'importance d'un mot
- Rechercher des informations
- La classification de documents
- La recommandation de contenu
- L'analyse de sentiment

# TF-IDF

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

- La fréquence de terme (TF) : le nombre de fois qu'un terme apparaît dans un document. Plus un terme apparaît fréquemment dans un document, plus il est important pour ce document.

- L'inverse de la fréquence de document (IDF) : le nombre de documents dans le corpus qui contiennent le terme. Plus un terme est présent dans un grand nombre de documents, moins il est important pour un document en particulier.

# TF

$$\mathrm{tf}(t,d) = \frac{n_{t,d}}{n_d}$$

- n t,d : nombre d'occurence t dans le texte d
- n d : nombre de mot dans le texte d

# IDF

**Sklearn:**

$$\text{IDF}(t) = \log \frac{1+n}{1+df(t)} + 1$$

**Standard:**

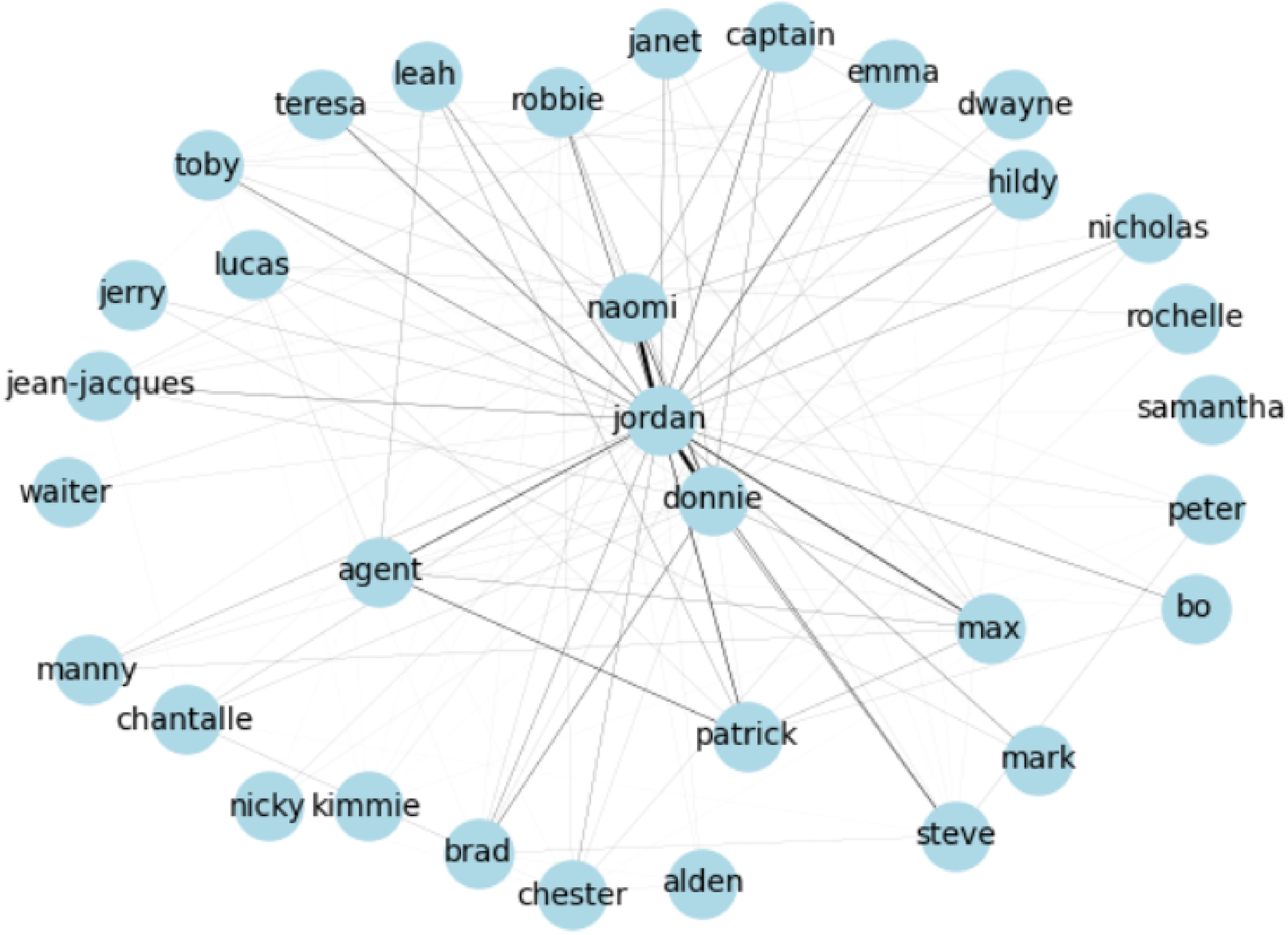
$$\text{IDF}(t) = \log \frac{n}{df(t)}$$

- N : représente le nombre total de documents dans le corpus

- DF(t) : représente le nombre de documents dans le corpus qui contiennent le terme t

# EXEMPLES

# Le Réseau


Réseau de Co-occurence
janet
captain
leah
emma
teresa
robbie
dwayne
toby
hildy
nicholas
lucas
jerry
naomi
rochelle
jean-jacques
jordan
samantha
waiter
donnie
peter
agent
bo
max
manny
chantalle
patrick
mark
nicky
kimmie
steve
brad
chester
alden