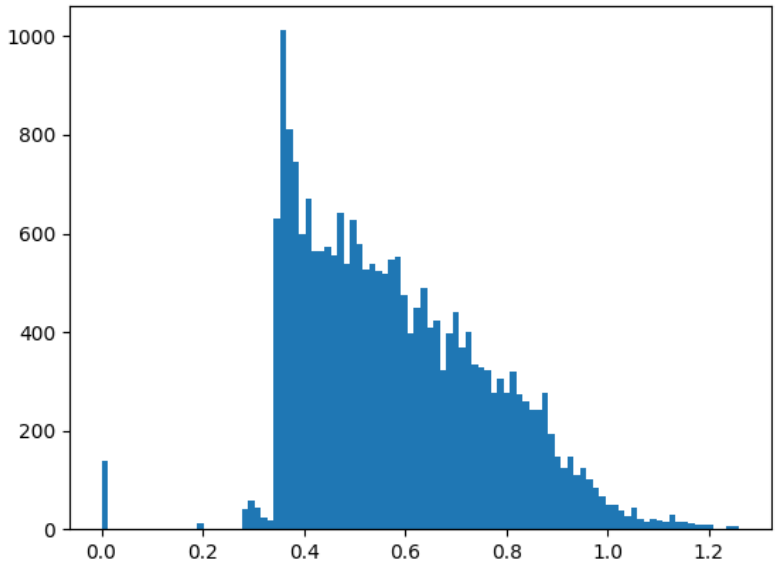
Matching graph

Ceci est le document qui va justifier les questions demandé dans le sujet. Le projet a été fait par Victor CHAU.

Dans ce projet, nous essayerons de voir si la popularité d'un film est dû à son année de sortie ou bien les acteurs qui y figurent. Le dataset est un CSV de film qui contient des données chiffrées ainsi que du texte. Il y a donc des données quantitatives ainsi que non quantitatives.

Les modules utilisés sont pandas pour traiter le CSV, pygraphviz afin d'afficher le graph et numpy pour les matrices. Lorsque nous lançons le projet, le graph est spécifié en dur dans un des dossiers du projet. Panda charge le CSV et nous affichons quelques données utiles pour la BDD. Une fois que les films ont été récupéré, nous calculons un indice entre chaque film afin de créer une matrice de dissimilaritée. Cet indice est basé par l'addition de: la différence de l'année de sortie divisée par l'écart maximum, la popularité divisée par 100 et les acteurs/actrices/réalisateurs communs. L'histogramme de ces donnés nous permet de remarquer que la matrice de dissimilarité semble suivre une loi demi-normale. Cela nous a permis de pouvoir configurer un seuil de 0.37 dans notre programme. Nous considérerons donc qu'un film est similaire à un autre si l'indice est inférieur à 0.37.



L'image de l'histogramme qui ressemble a une demi-normale.

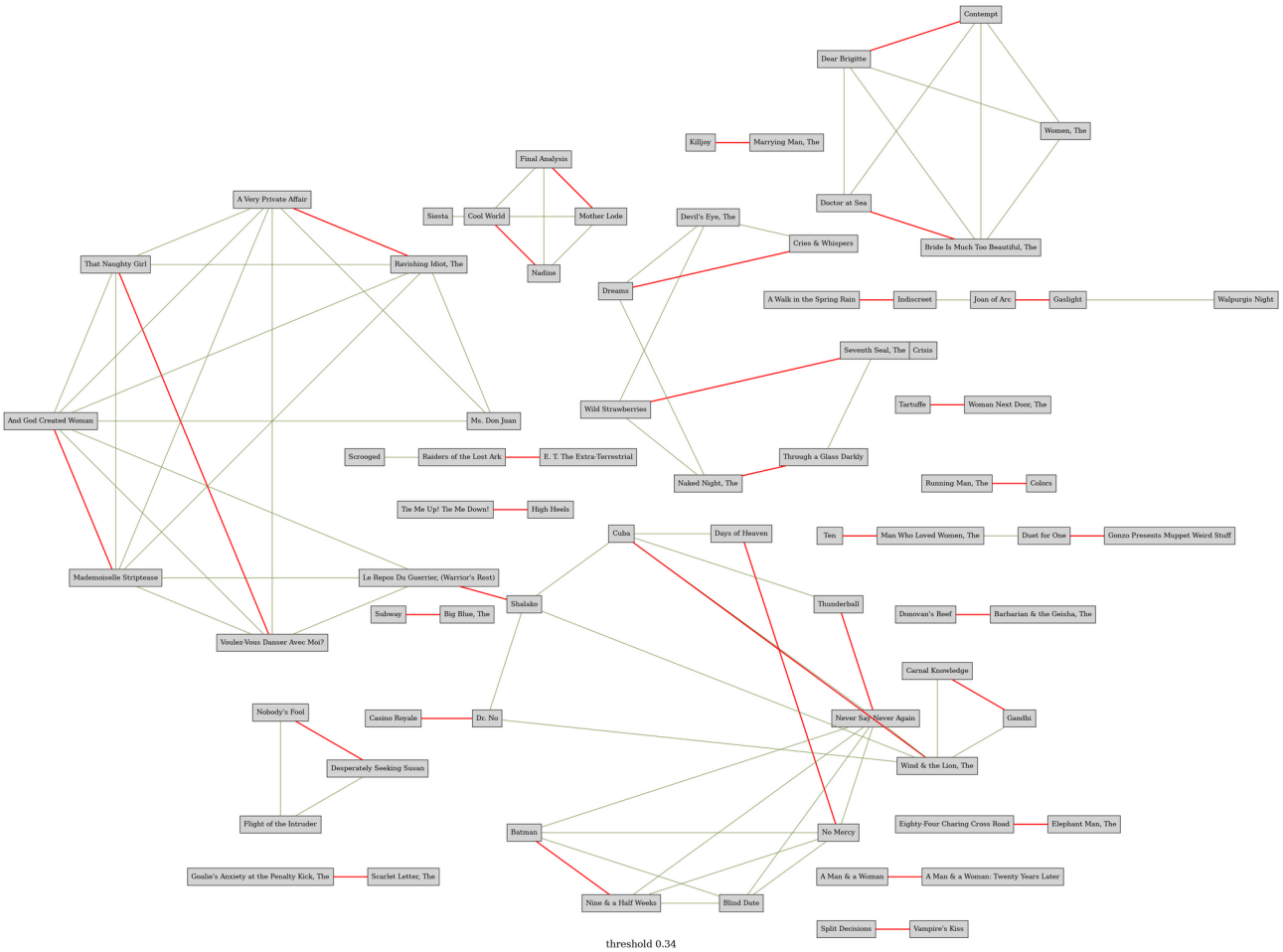
```
# we build a hybrid dissimilarity
dissimilarity = math.sqrt(
    ((movie_1_year - movie_2_year) / gapYear)**2
    + match_strings(movie_1_actor, movie_2_actor)**2
    + match_strings(movie_1_actress, movie_2_actress)**2
    + match_strings(movie_1_director, movie_2_director)**2
    + ((movie_1_popularity - movie_2_popularity) / 100)**2
```

L'image au-dessus montre comment nous calculons l'indice de dissimilarité. Nous faisons la racine de l'addition du carré de la distance d'année, de la distance de popularité et les poids retournés par la présence ou non de personnel commun.

```python
def match_strings(string1, string2):
    if string1 == string2:
        return 0
    return 0.2
```

Ici nous définissons la fonction match_string, si les personnes ne sont pas les mêmes, nous retournons un poids de 0.2.

Une fois que nous avons les informations récupérés et modelisé par pygraphviz, vous voulons avoir un dictionnaire qui permet de retourner tout les noeuds qui sont lié à un noeud donné. Avec cette variable nous allons parcourir 2 fois notre graph. La premiere fois c'est pour avoir tout les noeuds qui sont lié ensemble afin de les regrouper en un ensemble. Pour faire cela, nous parcourons en profondeur un noeud au hasard et récupérons tous ses n voisins. La deuxième fois c'est pour trouver le chemin le plus long du noeud sans repasser par le meme noeud. Cela est fait en parcourant le graph en largeur. Une fois que nous ayons récupéré le chemin maximum pour chaque sous graph, nous pouvons appliquer notre matching. Pour ce faire, nous parcourons 2 par deux le chemin maximum et associons ensemble les noeuds et enfin nous cherchons dans les feuilles pour associer le reste. Pour tracer le graph, nous ne traçons que les noeuds qui sont liés. Soit une moitié de films par rapport à la taille du CSV.



threshold 0.34

Ceci est l'image de la sortie de notre programme avec film2.csv. Nous pouvons remarquer qu'il y a un grand groupe et plusieurs petit groupes. Le gros groupe rassemble quasiment tout les films avec de très basse popularité ensemble. Les films du coté gauche comprenant "That Naughty Girl" ou "Ms Don Juan" on une moyenne de popularité de 25 et les films du coté droit comprenant "Cuba" ou "Days of Heaven" ont une moyenne de popularité de 10. La feuille avec "Gandhi", "Carnal Knowledge" et "Wind and the lion" ont une popularité très basse et ils ont la même actrice "Bergen, Candice". Néanmoins, il semble avoir une corrélation entre l'année de sortie et la populatité. Les films du coté droit ont été sorti dans les années 60 alors que le coté gauche sont sorti dans les annés

70. Une des explications de la hausse de popularité dans les films nuls peuvent-être du à la recrudescence de bons film non Hollywoodien.
D’autre ensemble dans le shémas ont rassemblé des acteurs commun tels de “Bardot, Brigitte” ou le même réalisateur tel que “Bergman, Ingmar”. L’ensemble incluent le réalisateur a une popularité basse, mais des films avec une bonne popularité produit par le même réalisateur sont non présents sur le graph! Pareil pour “Bardot, Brigitte”. Il semble donc avoir une légère influence de la popularité du film en fonction du casting. Néanmoins, est-ce que la popularisation du cinéma dans les années 60-70 influenceraient-ils d’une bonne manière les films peu connus?