Projet de BDA

Description du dataset

Nous allons nous baser sur le dataset publique et gratuit de `IMDb` qui est un site notant et donnant des avis sur des films et des séries.

Le dataset est disponible ici dans des fichiers tsv.

Ces datasets fournissent des informations concernant les titres, les membres de l'équipe réalisant les films et séries, concernant les épisodes en cas de séries, concernant les acteurs et concernant les votes attribués par le site.

Ces dataset rassemblent quelques plusieurs millions d'entrées, ce qui classe logiquement ces données dans la catégorie `Big Data`.

Idées de projet

Vu la diversité de choix de projet nous avons souhaité énumérer ci-dessous nos premières idées.

- Appliquer des techniques d'analyse et de statistique sur l'entièreté du dataset. Idée vague, mais se concrétisera après avoir suffisamment pris les données en main.
- Simuler un problème de machine learning non-supervisé en regroupant les films par catégorie (les catégories sont ici données, mais nous pourrions nous en passer afin de le faire de façon automatique). Nous nous baserions sur la description du film en utilisant des technologies NLP (NE, POS, Chunks) puis en les vectorisant afin de les projeter dans un espace multidimensionnel qui nous permettrait de comparer les distances entre les différentes descriptions. Nous utiliserions les outils NLP de Stanford.
- Même idée que le dernier point, sauf que nous ferions de l'apprentissage supervisé en utilisant dans ce cas-là les catégories comme classes pour entrainer un modèle.
- Créer un modèle de régression sur les notes données par le site en se basant sur les acteurs, réalisateurs, catégories, etc... afin de prédire la note d'un film.