**Find Movie Similarity from Plot Summaries**

Use NLP and clustering on movie plot summaries from IMDb and Wikipedia to quantify movie similarity.

#### Project Description

Natural Language Processing (NLP) is an exciting field of study for data scientists where they develop algorithms that can make sense out of conversational language used by humans. In this Project, you will use NLP to find the degree of similarity between movies based on their plots available on IMDb and Wikipedia.

This Project lets you apply the skills from [Natural Language Processing Fundamentals in Python](https://www.datacamp.com/courses/natural-language-processing-fundamentals-in-python) and [Unsupervised Learning in Python](https://www.datacamp.com/courses/unsupervised-learning-in-python/). We recommend that you are familiar with the content in those courses before starting this Project.

The dataset contains the titles of the top 100 movies on [IMDb](https://www.imdb.com/) as well as each movie's plot summary from both IMDb and Wikipedia.

#### Project Tasks

* 1 Import and observe dataset
* 2 Combine Wikipedia and IMDb plot summaries
* 3 Tokenization
* 4 Stemming
* 5 Club together Tokenize & Stem
* 6 Create TfidfVectorizer
* 7 Fit transform TfidfVectorizer
* 8 Import KMeans and create clusters
* 9 Calculate similarity distance
* 10 Import Matplotlib, Linkage, and Dendrograms
* 11 Create merging and plot dendrogram
* 12 Which movies are most similar?