أكاديمية سدايا
SDAIA Academy

TED talks using Topic Modelling

Banan Alhethlool banan.alhethlool@gmail.com

Ghadah Alharbi Ghadah.msh@gmail.com

Abstract:

Our aim is to use topic modeling to understand what is the topics that makes the most persuasive talks and to predict which Ted-Talks belongs to which topics.

TED is devoted to spreading powerful ideas in just about any topic. These datasets contain over 4,000 TED talks including transcripts in many languages.

Design:

Data was downloaded from Kaggle as a CSV file into pandas. Unnecessary columns were dropped and transcripts were cleaned using NLTK for modeling we used TFIDF and NMF. 4 topics were clearly identified and named (['Geology', 'stories', 'Economy', 'Computer Science']

Data:

The dataset is a TED Talks dataset found on Kaggle that has over 4000 talks, almost all of them in English. It has a column that has the transcription of each talk. Additional features of the dataset include: "views, speaker, (speaker) occupations, recorded_date, published_date, event, available_languages, duration, (number of) comments, topics" which could be used for aggregating info/ modeling later on.

Algorithms:

Data Cleaning: nltk used to perform preprocessing which included: Removing unnecessary character

Modeling: TFIDF used for vectorization and (NMF, LSA, LDA for performing topic modeling

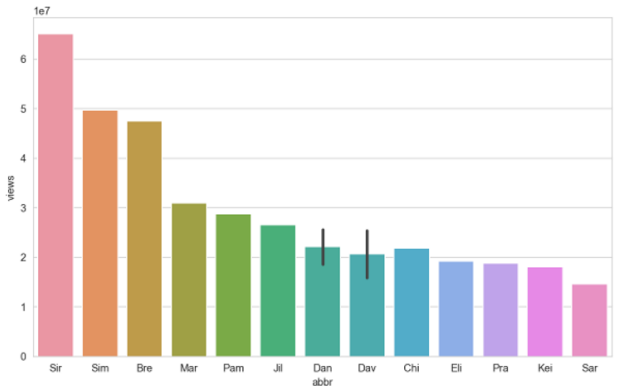
Tools:

The Tool That we need is Python and Jupyter notebook to execute the code**.** Pandas packages to manipulate data

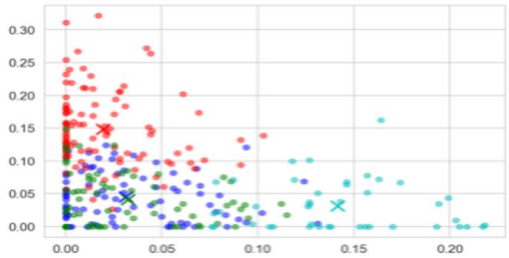
Feature extraction (TFIDF) Decomposition (NMF, LDA, LSA) and visualisation library (such as Seaborn and Matplotlib), Sklearn Library for model_selection

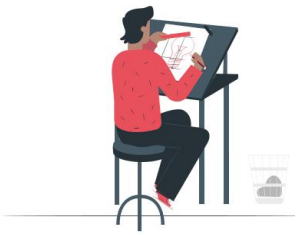
Communication:

EDA



Clustering





In Results, the most frequently topic at TED Talks is stories.

Results

Geology	water, planet, earth, ocean, citi
stories	said, music, love, stori, feel
Economy	countri, africa, percent, dollar, govern
Computer Scince	brain, comput, design, technolog, kind