TED TALKS DATA ANALYSIS

Natural Language Processing Project Proposal

Hayat Aldhahri & Juri AlSayigh 11/7/21 NLP

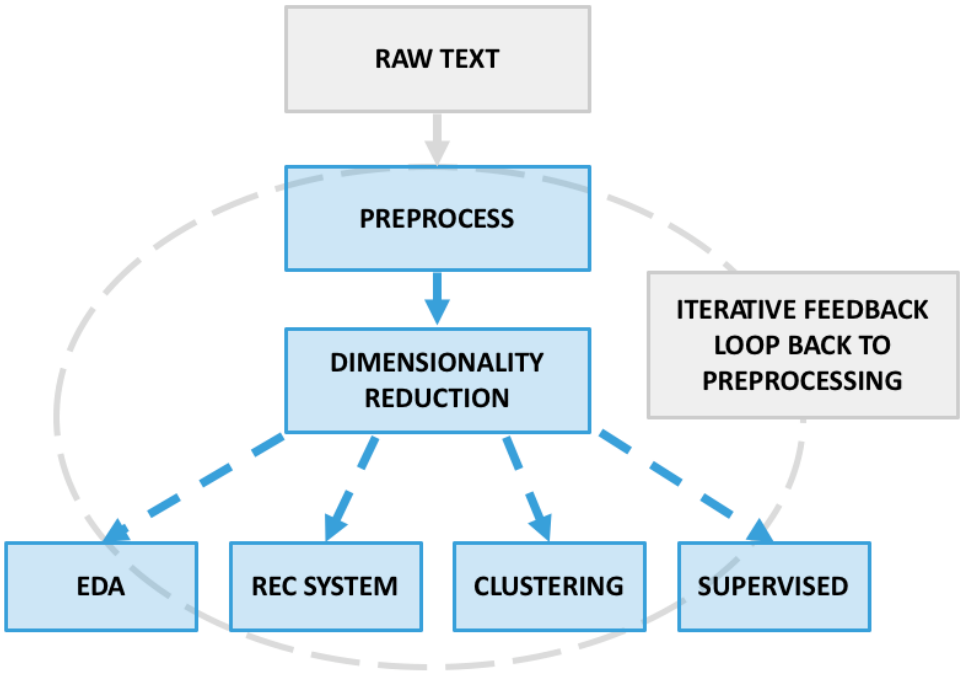
أكاديمية سدايا
SDAIA Academy

Overview

TED talks have been inspiring people since 1987 when it was founded by Richard Saulman as non-profit organization. Gathering speakers and experts from different fields such as technology, history, art, entertainment, and designs to share their ideas and thoughts for the public. Operating under the theme “ideas worth spreading”, TED talks videos and audio data are available online, uploaded in TED official website ever since it was first filmed in 1994.

In this project, we will be analyzing the dataset using machine learning algorithms to find the most inspiring topics and talks among the 2,550 talks.

As for the model development, we will be following the presented workflow for data analysis.



Data Description

There are two dataset used for this project.

- ted_main.csv
- transcripts.csv

The dataset contains audio-video information of TED talk uploaded in their official website. The recording are from 1994 to 2017. The dataset has been scraped from the official website and it is available as csv files in Kaggle.com for data analysis. The below table highlights are the features from the file “ted_main.csv” 17 features and 2,550 rows.

FEATURE	DESCRIPTION
COMMENTS	The number of comments for the topic
DESCRIPTION	Brief description on the topic presented
DURATION	The duration of the TED talk in seconds
EVENT	The event where the TED talk took place
FILM_DATE	The Unix timestamp of the filming The seconds count started on January 1st, 1970
LANGUAGES	The number of languages in which the TED talk is available
MAIN_SPEAKER	The main speaker of the TED talk The first named speaker of the talk
NAME	The official name of the TED talk (title and speaker)
NUM_SPEAKER	The number of speakers in the talk
PUBLISHED_DATE	The Unix timestamp for The published date of the TED talk
RATINGS	A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)
RELATED_TALKS	A list of dictionaries of recommended talks to watch next
SPEAKER_OCCUPATION	The occupation of the main speaker
TAGS	The themes associated with the talk
TITLE	The title of the talk
URL	The URL of the talk
VIEWS	The number of views on the talk

The transcripts.csv file has the following feature:

- Transcripts: The official English transcript of the talk.
- URL: URL of the talk.

Tools

- Pandas and NumPy packages to data manipulation.
- Seaborn and Matplotlib libraries for data visualization.
- Nltk and spaCy for text preprocessing.
- sklearn's CountVectorizer.
- Sklearn clustering and preprocessing.
- re: Regular expression.
- Gensim for LDA Topic modeling.
- Jupyter notebook that hosts the code.
- Prezi for presentation.

Conclusion

This project will showcase a list of topics by Natural Processing Language (NLP) on raw text scraped from TED talks official website found on kaggle . The analysis will be helpful for those interested in findings the most viewed topics and inspirational talks.