Jen Tat
2 October 2020

The Reading and Parsing of TED Talk Data

Overview of the Data

The datasets tabulated by Rounik Banik, a data scientist at Fractal Analytics, provide information of all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. For each of the talks, the TED main dataset contains the TED/TEDx event that the talk took place at, the official title of the talk, the full name of the speaker leading the talk, a description, the Unix timestamp of the publication date of the talk, and the duration, in seconds. The TED transcripts dataset includes the URL of the transcript of the talk, available on the official TED.com website, tabulated by Mauro Pelucchi, data scientist at Burning-Glass and WollyBI. Wei T, a student at the University of Edinburgh, tabulated a very similar data set to the TED main data set, except one that includes the URLs to the YouTube video of all of the TED Talks. Ahmad Fatani, a Kaggle user, also tabulated a similar data set, but one without URLs to the YouTube videos posted by the official TED channel, instead one that includes tags related to the topics of the TED Talks.

Purpose of the Data

The purpose of parsing the TED Talk data is to systematically organize the TED Talks and filter them on the official TED.com website using relevant fields, either specified by the user or chosen by the developers. This can be used to better transcribe the details of each of the talks for user comprehension and for better readability. The data can be categorized by its event and consequently, the year it took place, amongst even more specific categories, for the user to be informed of which of the TED Talks are older, and which are more relevant to recent times and affairs, and which correspond to their needs and interests. The data is used for the user to be able to filter through the sheer number of audio-video recordings of TED Talks, available on the official TED.com website. Currently, there are one hundred thirty-three audio-video recordings of TED Talks available on the official TED.com website, thus utilizing the data that is distinct for each talk or shares similarities with other talks, prevents each of the talks from becoming disregarded and having minimal user exposure.

Data Structure Description

The data structure used is a dictionary with keys that have values of another dictionary with keys that have values that are lists, and one of the items of this list is a list. A dictionary is returned where the TED/TEDx event is the key and for each of these keys, its value is another dictionary, where the key is the official title of the TED Talk that occurred at that TED/TEDx event, and its value is a list of information pertaining to that specific talk. The list contains the full name of the speaker of the talk, a description of that TED Talk; a list of tags related to the talk; the publication date of the audio-video recording of it, converted from its Unix timestamp to a human readable date, formatted as “year-month-day;” the duration of the talk that is casted from a string value to an integer value, and converted from seconds to minutes through integer division (no decimal); the URL for the transcript of the talk; and the URL of the YouTube video.

General Example

{TED/TEDx Event: {Official TED Talk Title: [Speaker Name, Description, [Tags], Publication Date, Duration, Transcript URL, YouTube Video URL], Official TED Talk Title: [...], ...}}

Data Based Example

{TED@BCG Berlin: {Happy maps: ['Daniele Quercia', 'Mapping apps help us find the fastest route to where we’re going. But what if we’d rather wander? Researcher Daniele Quercia demos “happy maps” that take into account not only the route you want to take, but how you want to feel along the way.', ['TED Fellows', 'art', 'activism'], '2015-01-06', 7, 'https://www.ted.com/talks/daniele_quercia_happy_maps', 'https://www.youtube.com/watch?v=AJg9SXIcPiM']

Banik, R 2017, *Banik posts: data about TED Talks*, electronic dataset, Kaggle, viewed 1 October 2020, <https://www.kaggle.com/rounakbanik/ted-talks>.

Wei, T 2017, *Wei posts: Ted Talks Transcript*, electronic dataset, Kaggle, viewed 5 October 2020, <https://www.kaggle.com/goweiting/ted-talks-transcript>.

Fantini, A 2020, *Fantini posts: Ted Talks dataset*, electronic dataset, Kaggle, viewed 5 October 2020, <https://www.kaggle.com/ahmadfatani/ted-talks-dataset>.

Pelucchi, M 2020, *Pelucchi posts: TEDx talk*, electronic dataset, Kaggle, viewed 5 October 2020, <https://www.kaggle.com/mauropelucchi/tedx-talk?select=tedx_dataset.csv>.