AN ANALYSIS OF THE TED_MAIN DATASET

BRENDAN WHITAKER

Abstract. We analyze the dataset `TED_main` regarding TED Talks sourced from Kaggle (kaggle.com). We explore the data from the TED videos with the objective of determining if there are interesting observations or insights to uncover. We treat the basic questions of which are the most popular talks in the dataset, and the possibility of relationships between popularity and and several other variables. We construct a popularity measure from the frequency of tags in the dataset, and analyze the summary statistics, the least and most popular talks in the dataset according to our measure, and the correlation values and scatterplots of popularity with a number of variables.

Contents

1. Introduction

We give some preliminary information about the dataset. `TED_main` contains metadata for 2550 talks given from 2006 to Fall of 2017. There are 18 column headers with information concerning the speaker, topic of the talk, date of filming, etc. The data was imported and analyzed using `python 2.7` in the Jupyter Notebook web application. The `NumPy`, `Matplotlib`, and `pandas` python libraries were imported in order to aid in data exploration, cleaning, and predictive modeling. the data was imported in a `pandas` dataframe, a structure similar to an Excel workbook.

To get a simple summary of the data in our dataframe, we use the `describe()` function to obtain some simple summary statistics:

Date: SP18.

	comments	**duration**	**film_date**	**languages**	**num_speaker**	**published_date**	**views**
count	2550	2550	2.55E+03	2550	2550	2.55E+03	2.55E+03
mean	191.562353	826.5102	1.32E+09	27.326275	1.028235	1.34E+09	1.70E+06
std	282.315223	374.0091	1.20E+08	9.563452	0.207705	9.46E+07	2.50E+06
min	2	135	7.46E+07	0	1	1.15E+09	5.04E+04
25%	63	577	1.26E+09	23	1	1.27E+09	7.56E+05
50%	118	848	1.33E+09	28	1	1.34E+09	1.12E+06
75%	221.75	1046.75	1.41E+09	33	1	1.42E+09	1.70E+06
max	6404	5256	1.50E+09	72	5	1.51E+09	4.72E+07

Table 1. `df.describe()`

Just from this basic summary we already have some meaningful information available about this series of talks. For example we now know the average number of comments per video around a couple hundred, the average duration is about 14 minutes, and most talks are translated into at least 20 languages. We can also see that the most there has ever been in a single TED talk is five, and the least-watched talk still has over 50,000 views.

An immediate revelation from this table is that `python` has failed to recognize the unix timestamps given for the `film_date` and `published_date` columns, so we will have to clean the data such that the dates are interpreted correctly. Something else to note is that we only have these numerical statistics for the numerical data. The other 11 or so columns of information were text, and thus the `describe()` function omitted them from the table. We also note that since there are 2550 rows, and the count for each of the 7 columns in Table 1 is also 2550, however, this does not necessarily mean there are no missing data points in these categories. For example, in the languages category, we see that the minimum number of languages is 0, but this is impossible, since sorting the dataset using the command `df.sort_values(by=['languages'])` yields a talk called *6 ways to save the internet* by Roger McNamee which has a value of zero for the number of languages in which the talk is available. A quick Google search yields a video of the talk which is clearly in English. So we there's a bit of work to be done to clean up the data and make sure it all makes sense.

2. Skew and cleaning the data

We can also get a rough idea of the skew of the numerical variables from Table 1 by comparing the median and mean of each of the columns. Let M denote the median, s the standard deviation, and $\overline{x}$ the mean. We define a normalized measure γ of skew:

$$\gamma = \frac{\overline{x} - M}{s}.$$

And we will arbitrarily say that a variable with skew $|\gamma| < 0.1$ is (approximately) symmetric. If $\gamma \geq 0.1$ we say the mean is significantly greater than the median and call this right-skewed data. If $\gamma \leq -0.1$ we say the mean is significantly less than the median and call this left-skewed data. Then for the following ratio variables we have:

variable	γ	*skew*
comments	0.261	right-skewed
duration	-0.057	symmetric
languages	-0.070	symmetric
num_speaker	0.135	right-skewed
views	0.232	right-skewed

Table 2. Skew values.

So we now have some rudimentary information about the distributions of these metadata columns.

We now move on to the task of cleaning the data such that it outputs readable dates instead of unix timestamps in scientific notation, which are essentially useless to the casual reader. We make use of the `datetime` library, which gives us a convenient function to convert unix timestamps into a date variable which we can then output in any format of our choosing. We then create a new `pandas DataFrame` variable called `dfdate` to store our newly converted date values along with the rest of the dataset. Defining a new function `dateConversion` to make the dates readable, we have the following code snippet:

```
def dateConversion(x):
    return datetime.datetime.fromtimestamp(
    int(x)
    ).strftime('%Y-%m-%d %H:%M:%S')


dfdate['film_date'] = df.apply(lambda row:
                      dateConversion(row['film_date']),axis=1)
```

which yields entries in the `dfdate['film_date']` column which look like this:

```
2006-02-24 19:00:00
```

And applying the same process to the `published_date` column yields successfully reformatted date columns.

3. Construction of a popularity measure

We now move on to answering the question of which are the most popular talks. It may be tempting to derive a measure of popularity solely from the number of views, but this may be an ill-conceived method for the following reason: we do not know what the system which collected this data counts as a view. There is no metadata regarding the views column available on `kaggle.com`, and we can't be sure on which platform the videos were viewed. A quick look at TED's website reveals that they have their own player for videos distinct from the youtube player. If all the views were counted on youtube we could use general knowledge about what youtube counts as a view to determine whether or not these quantities would be good measures of popularity. For example, if the website on which the talk was hosted counts a click on the video link as a view, then this measure wouldn't be giving us an idea of how popular the talk is so much as it would be giving us an idea of how popular/enticing the link or thumbnail of the video appears. On the other hand, if a view was defined as the number of people who watch through the entire (or a good portion of) the video, then perhaps it would be a good measure of popularity, since we would have some idea of viewer retention.

For this reason, we will instead use the net number of ”positive” ratings/tags of the video as a measure of popularity. Using the following code snippet, we make a list of each of the distinct tags along with their overall frequency across the entire dataset:

```
dftags = {}
for i in range(0,2549):
    length = len(literal_eval(dfdate['ratings'].iloc[i]))
    for j in range(0,length - 1):
        currentName =
         literal_eval(dfdate['ratings'].iloc[i])[j]['name']
        if currentName not in dftags:
            print currentName
            dftags[currentName] = literal_eval
            (dfdate['ratings'].iloc[i])[j]['count']
        else:
            dftags[currentName] += literal_eval
            (dfdate['ratings'].iloc[i])[j]['count']
```

We note here that there was no way of knowing whether there would be a relatively small number of distinct tags to make it a good measure of popularity before actually writing the code to generate the list. However, after generating the list of tags, we have 14 different tags, and sorting them into positive, negative, and neutral categories yields:

tag	*positivity*
Funny	$+$
Beautiful	$+$
Ingenious	$+$
Courageous	$+$
Longwinded	$-$
Confusing	$-$
Informative	$+$
Fascinating	$+$
Unconvincing	$-$
Persuasive	$+$
Jawdropping	$+$
Ok	0
Obnoxious	$-$
Inspiring	$+$

Table 3. Tag categories.

We denote the aggregate rating of a talk as r_t and let Λ be set of 14 distinct tags given above. Let p_i be the positivity $(+1, -1, 0)$ of tag i. Let c_i be the count of tag i for the specified talk. Then we have:

$$r_t = \sum_{i \in \Lambda} p_i \cdot c_i.$$

We implement this formula and generate rating values as integers for each talk, and import these into a copy of the `DataFrame` object. The code for this was lengthy and for brevity we omit it. We then sort these in ascending and descending order to see which talks are most popular, and also which are the least popular.

4. Popularity of talks and correlations

So which was the most popular talk? This was one titled *My stroke of insight* by a neuroanatomist named Jill Bolte Taylor, which had a net positivity rating of 67241. The top 10 highest rated talks are:

title	*year*	*rating*
My stroke of insight	2008	67241
Do schools kill creativity?	2006	65476
Your body language may shape who you are	2012	62986
The power of vulnerability	2010	58293
How great leaders inspire action	2009	51596
How to live before you die	2005	36234
The happy secret to better work	2011	31471
Underwater astonishments	2007	29008
The danger of a single story	2009	28777
The power of introverts	2012	28705

Table 4. 10 most popular TED talks.

It is interesting to note that the while the median date for the set of all talks was sometime in 2012, since it's unix timestamp was around 1.33×10^9, the average date of the most popular talks is sometime in late 2008. And the most recent talk out of these was in 2012, which is the same year of the median of the whole dataset. This suggests that it may several years for a talk to become significantly popular. Another more qualitative thing to note about these 10 talks is that most of them are observed to be focused on self-empowerment/self improvement. In this category you could safely place all but the *My stroke of insight*, *Do schools kill creativity?*, *the danger of a single story*, and *Underwater astonishments*. One possible explanation for this is that talks are popular and highly rated when the viewers feel as though the information they gained from the talk is directly applicable to their own lives.

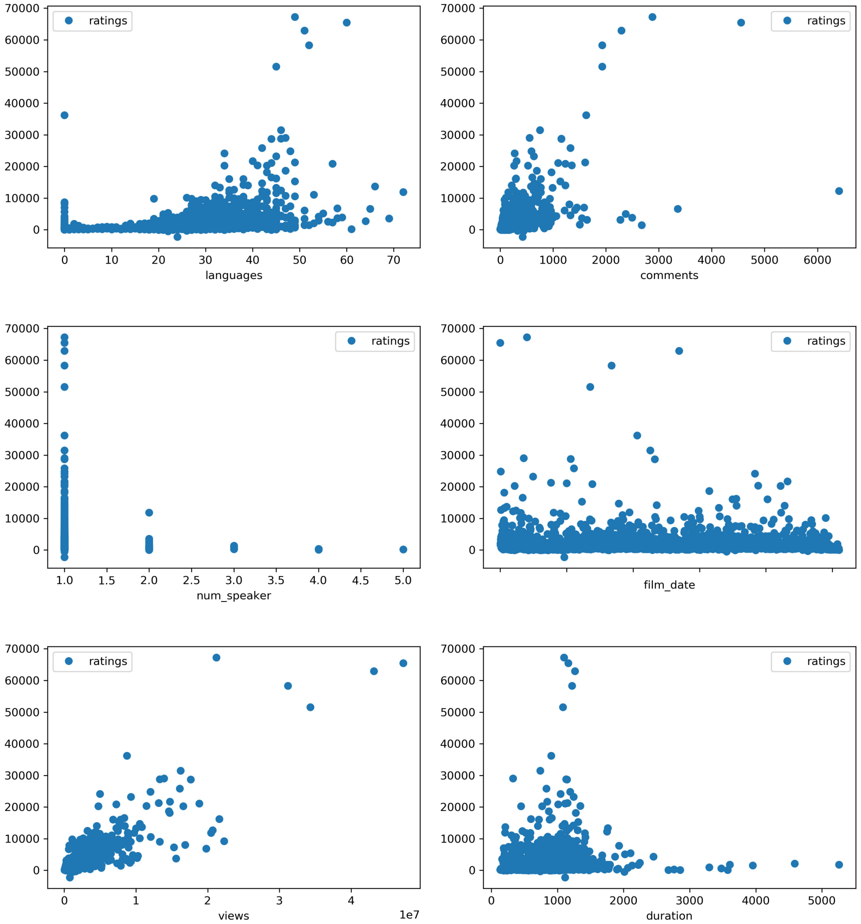
The least popular out all the TED talks was *17 words of architechural inspiration* by Daniel Libeskind with a rating of -2241. There doesn't appear to be an immediately clear reason why so many people disliked this talk, except perhaps that maybe people find architecture exceptionally boring. Other notable unpopular talks were *The NSA responds to Edward Snowden's TED talk*, *Enough with the fear of fat*, and *Is religion good or bad?*. These were all within the top 10 most unpopular TED talks, and had ratings below 0. The reasons for these talks' unpopularity is more easily imagined being in line with political, social, and cultural strife. This suggests that one thing people are not looking for in lecture-style talk is controversy.

We now move on to the task of discovering if there are any relationships or meaningful correlations between popularity as measure with the rating metric constructed in the previous section, and other data from the set. We give the correlation coefficients for ratings with a number of other variables:

	comments	**duration**	**languages**	**num_speaker**	**views**
ratings	0.609881	0.097217	0.333162	-0.038618	0.853597

Table 5. Correlations with ratings.

And we also give the corresponding scatterplots to these pairs, as well as one for ratings vs. film date:



We immediately see that there is no correlation between popularity (ratings) and the variables `duration`, `num_speaker`, and `film_date`. These values are so close to zero that their signs are essentially meaningless. As expected, we see a significant positive correlation between our new measure of popularity and views, which suggests that for the most part, videos that looked like they would enjoyable and informative from the title and thumbnail alone (influencing the number of views) tended to have content that was in line with those positive expectations. There was similar, but slightly weaker relationship between popularity and comments, which is also expected, as more popular talks would incite more discussion. It is interesting to note that there is a slight (0.33) correlation between languages and popularity, which suggests that only the most popular talks get translated into a large number of languages. From the scatterplot we can see that 5 out of the 6 most popular talks were all translated into more than 40 languages. It is also interesting to note that none of the talks with more than 1 speaker had a popularity rating higher than 15,000. Thus suggests that a single speaker is a more effective lecture-style.

We note here that we omitted any discussion of the occupation of the speakers or the description of the the talks. This is because a cursory look at the number of distinct entries for these two columns on the kaggle website reveals that there are very few duplicate entries in these columns: most of datapoints are unique and thus it would be difficult to do an analysis of their distribution without some nontrivial processing of their semantics, which would likely require some sort of NLP model.