FINAL PROJECT REPORT

Raya Young

This project aims to analyze a dataset describing over 4,000 TED Talks, sourced from Kaggle.com, and updated as of May 2020. After analyzing this data, the findings will then be compared against Tweets scraped from Twitter under the hashtags #TED, #tedtalk, and #TEDx

Part 1 - Structured Data

TED data from Kaggle was formatted into a .csv file and read into Python. Data was read in from the file row by row, and assigned to column to be converted and processed later as a panda dataframe. Data Types such as duration and date were converted appropriately, as well as removal of NAs and blank spaces were stripped.

An initial look into the pandas dataframe shows some fields displayed as dictionaries. For example, Occupations, Topics, and Related Talks all display as a list / dictionary, describing more than one entry. Most of what we will be focusing on for this project are basic descriptions of each talk like Speaker, Title, Transcript, and Views.

Talk ID	Title	Speaker 1	Occupations	About Speakers	Views	Date Recorded	Date Published	Comments	Duration	Topics	Related Talks	Description
0001	Averting the climate crisis	Al Gore	{0: ['climate advocate']}	{0: 'Nobel Laureate Al Gore focused the world'...	3523392	2006-02-25	2006-06-27	207020	00:16:17	['alternative energy', 'cars', 'climate change...	{243: 'New thinking on the climate crisis', 54...	With the same humor and humanity he exuded in ...
0002	Simple designs to save a life	Amy Smith	{0: ['inventor', 'engineer']}	{0: 'Amy Smith designs cheap, practical fixes ...	1724438	2006-02-24	2006-08-15	100010	00:15:06	['MacArthur grant', 'alternative energy', 'des...	{1561: 'Energy from floating algae pods', 1072...	Fumes from indoor cooking fires kill more than...
0003	How to rebuild a broken state	Ashraf Ghani	{0: ['president-elect of afghanistan']}	{0: 'Ashraf Ghani, Afghanistan's new president...	981920	2005-07-12	2006-10-18	7050	00:18:45	['business', 'corruption', 'culture', 'economi...	{127: 'Want to help Africa? Do business here',...	Ashraf Ghani's passionate and powerful 10-minu...
0004	The real future of space exploration	Burt Rutan	{0: ['aircraft engineer']}	{0: "In 2004, legendary spacecraft designer Bu...	2427994	2006-02-24	2006-10-25	109060	00:19:37	['NASA', 'aircraft', 'business', 'design', 'en...	{141: "Inside the world's deepest caves", 264:...	In this passionate talk, legendary spacecraft ...
0005	Great cars are great art	Chris Bangle	{0: ['car designer']}	{0: 'Car design is a ubiquitous but often over...	978483	2002-02-02	2007-04-05	8010	00:20:04	['business', 'cars', 'design', 'industrial des...	{4: 'The real future of space exploration', 26...	American designer Chris Bangle explains his ph...

FINAL PROJECT REPORT

Raya Young

We begin with EDA on several attributes of the .csv dataset:
With over 4,000 rows of data, are there any speakers who gave more than one talk?

Speaker 1	Title
Alex Gendler	34
Iseult Gillespie	19
Emma Bryce	12
Daniel Finkel	11
Hans Rosling	10

In fact, some speakers had over 15 talks! In delving further, we see from Duration, that rather than the large formal talks, sometimes shorter lessons are published in the 5 minute range. Many others gave just one talk.

By contrast, the talk with the longest duration came in at just over an hour, on the topic of climate change and environmental activism.

Talk ID	Title	Speaker 1	Occupations	About Speakers	Views	Date Recorded	Date Published	Comments	Duration	Topics	Related Talks	Description	Transcript
54715	How we can turn the tide on climate	Chris Anderson	{0: ['head of ted'], 1: ['climate advocate']}	{0: 'After a long career in journalism and pub...	1493370	2019-12-04	2019-12-12	4040	01:05:22	['climate change', 'environment', 'activism', ...	{32560: 'The disarming case to act right now o...	Witness the unveiling of Countdown, a major gl...	[Citizens of the world] [We face a global cris...

Is there a pattern with recording date? At a glance, we can see that several talks were recorded sometimes days apart, though this could be all over the world. This dataset contains TED Talks as well as TEDx Talks, which are usually smaller events at the local level. These seem to follow a format similar to a convention, with a packed schedule over several days, with audiences choosing which talks to attend. For example, almost 100 talks were recorded on a single day, according to the data. Others still, seem to be standalone talks, and so were perhaps a special event that invited a single speaker. Corresponding recording location and event name may be helpful for future analysis.

Date Recorded	Title
2017-04-24	97
2018-04-10	95
2019-04-15	90
2007-03-03	54
2017-08-27	53
...	...
2014-03-16	1
2014-03-24	1
2014-03-25	1
2014-03-27	1
2020-04-30	1

FINAL PROJECT REPORT

Raya Young

Next we begin to explore the popularity and success of the talks themselves. The average number of views is 2,148,006, and this talk by Sir Ken Robinson titled, “Does Education Kill Creativity?” had the most number of views 6,505,1954!

Talk ID	Title	Speaker 1	Occupations	About Speakers	Views	Date Recorded	Date Published	Comments	Duration	Topics	Related Talks	Description	Transcript
0066	Do schools kill creativity?	Sir Ken Robinson	{0: ['author', 'educator']}	{0: "Creativity expert Sir Ken Robinson challe...	65051954	2006-02-25	2006-06-27	40903010	00:19:24	['children', 'creativity', 'culture', 'dance',...	{865: 'Bring on the learning revolution!', 173...	Sir Ken Robinson makes an entertaining and pro...	Good morning. How are you? (Audience) Good. It...

The average number of comments per talk was 239073.9, and the talk with the most number of comments was titled “Militant Atheism” by Richard Dawkins with 60,404,090 comments!

Talk ID	Title	Speaker 1	Occupations	About Speakers	Views	Date Recorded	Date Published	Comments	Duration	Topics	Related Talks	Description	Transcript
0113	Militant atheism	Richard Dawkins	{0: ['evolutionary biologist']}	{0: 'Oxford professor Richard Dawkins has help...	5788514	2002-02-02	2007-04-16	60404090	00:29:10	['God', 'atheism', 'culture', 'religion', 'sci...	{86: 'Letting go of God', 94: "Let's teach rel...	Richard Dawkins urges all atheists to openly s...	That splendid music, the coming-in music, "The...

The relative popularity of these two talks begs an analysis of sentiment in the comments and transcripts for further insight on the impressions emotions these two talks represent.

Conclusion and Next Steps:

More pattern inspection on this dataset could include: what is the most common occupation of the speakers? What is the average sentiment and word frequency in the transcripts? What about comparing the transcripts of the most commented and most viewed talks? Many of the speakers tend to be leaders and academics, so it would follow that there would be some notable patterning to their speech.

Data on the recording location and event name for each talk would be helpful; it could even be combined with demographic information to provide insight into the interest and values of the local population.

Sentiment analysis on the comments and for each talk could prove insightful as well. It is not hard to imagine vibrant debate in both of the most popular talks on education and evolution, respectively.

Another opportunity for analysis would be to try to predict the number of comments based on the subject matter and number of views of a talk.

FINAL PROJECT REPORT

Raya Young

Part 2 - Unstructured Data

In Part 2 of this assignment, we explore the processing of unstructured data and its subsequent analysis. Because my final project is based on a dataset around TED Talks, I chose to use a Twitter API to scrape tweets with the hashtags #TED, #tedtalk, and #TEDx. Twitter hashtags are not case-sensitive, so this captured lots of data. In my initial scraping of 500 tweets each, for a total of 1500, I realized that this left too few tweets available in English for later analysis. For a second scraping (08-29), I called for the retrieval of 4000 tweets for each hashtag, which gathered a total of 4601 available tweets.

Each tweet has several attributes, but the fields of particular interest are:
'_id' (index); 'created_at' (time), 'text' (Tweet), 'entities' (hashtags, etc.); 'user' (name, etc.); 'retweet_count', 'favorite_count', and 'lang' (language).
'created_at' was converted to datetime using pd.to_datetime(), and counts were converted to integer data type.

First, we explore the frequency of each attribute:
'created-at' was binned by date, and binned by hour, to investigate for any emerging patterns. This collection of scraped Tweets spanned 2020-08-21 through 2020-08-29. Of these, the date with the most Tweets 08-22, with 784 Tweets total.

created_date	created_at
2020-08-22	784
2020-08-26	604
2020-08-23	588
2020-08-25	507
2020-08-24	492
2020-08-28	442
2020-08-21	426
2020-08-27	425
2020-08-29	333

	created_date	text
1021	2020-08-22	RT @TEDxCDMX: Hoy más que nunca, recordemos es...
1022	2020-08-22	그럼 어째서 우리는 결코 사무엘 피어퐁 랭리에 대해서 들어본 적이 없는가? \nTh...
1023	2020-08-22	RT @fahadalahmdi: هل تعرف ماهو الفرق بين السوا...
1024	2020-08-22	@SirKenRobinson murió ayer 21 agosto 2020, gra...
1025	2020-08-22	RT @thewetbaguett3: Like for part 7! #ted #mov...
1026	2020-08-22	그들은 없었다. 우리가 성공의 레서피라 여기는 것들을 \nthey had none ...
1027	2020-08-22	Like for part 7! #ted #movie #moviescene #4u #...
1028	2020-08-22	ダニエル・H・コーエン: よい\議論をするために\nhttps://t.co/llqGnURP...
1029	2020-08-22	RT @TEDxCDMX: Hoy más que nunca, recordemos es...
1030	2020-08-22	If you've never watched @SirKenRobinson's orig...

What is being tweeted on 08-22 that is so popular? As it turns out, one of the most popular speakers, Sir Ken Robinson, passed away on 08-21. He gave a very popular talk entitled "Do Schools Kill Creativity?" From the text output on this day, we can see the messages about the speaker shared by many users in many languages. We know from our earlier analysis, that this was the most popular talk by the number of views, as of May 2020.

FINAL PROJECT REPORT

Raya Young

As an aside, we can also see some tweets in reference to a #ted movie. The initial scraping for this project took place earlier in the week, before I realized I had too few tweets. In the interim, the third BIll and Ted movie was released (08-28), and people were excited!

Next, we further parse this datetime value to Tweets per hour. This output is not very specific, because this scraping does not specify the time zones of individual users. Regardless, it seems that the most common time to tweet for these specific hashtags is around noon.

Finally, we combine these two subdivisions and see the tweets per hour, per day. The most tweets on 08-22 were at 1PM.

		created_at
created_date	**created_hour**	
2020-08-21	**5**	5
	6	24
	7	43
	8	25
	9	30
...	**...**	...
2020-08-29	**15**	21
	16	12
	17	18
	18	23
	19	3

		created_at
created_date	**created_hour**	
2020-08-22	**13**	90
	20	79
	18	56
	15	54
	14	46
	12	45
	19	44
	17	44
	22	43
	16	42

Utilizing some scripts provided by Dr. Landowski, 25 languages were represented in the dataset, 2645 unique users, and the top 20 hashtags follow:

```
Top 20 Frequency Hashtags
TedTalk 925
TED 704
Leadership 576
Management 382
TEDx 283
tedx 265
Tedx 210
ted 160
TEDTalk 107
教育 91
Ted 77
動画 74
tedxtcet 70
فهد_الأحمدي 67
تيد 67
صباح_الخير 67
talk 58
IFTHH 56
tedtalk 55
technology 52
```

FINAL PROJECT REPORT

Raya Young

Of the 4601 tweets scraped that contained the TED hashtags, 3418 had the lang attribute ‘en’, indicating English. These were saved to a separate dataset named dfen for further analysis.

The average number of favorites in this collection is 2, and the maximum is 818.

	_id	created_at	text	entities	user	retweet_count	favorite_count	lang	created_date	created_hour
2920	5f4aa86f678bd9aa94f14169	2020-08-22 12:45:33	Good morning. If you haven't watched my #TedTa...	{'hashtags': [{'text': 'TedTalk', 'indices': [...	{'id': 17375057, 'id_str': '17375057', 'name':...	286	818	en	2020-08-22	12

The average number of retweets in this collection is 28.8 while the maximum is 1235.

	_id	created_at	text	entities	user	retweet_count	favorite_count	lang	created_date	created_hour
4600	5f4aa89bd6a56ce85825ae65	2020-08-21 05:44:10	RT @SandyAhlawat89: When civilian countrymen a...	{'hashtags': [], 'symbols': [], 'user_mentions...	{'id': 2880746036, 'id_str': '2880746036', 'na...	1235	0	en	2020-08-21	5

Placing top Retweets into their own table, we can start to see a pattern among the attributes. For instance, some of the RTs have slight variations on the text, perhaps because they are a RT of an RT, and so are counted differently.

	_id	created_at	text	entities	user	retweet_count	favorite_count	lang	created_date	create
4600	5f4aa89bd6a56ce85825ae65	2020-08-21 05:44:10	RT @SandyAhlawat89: When civilian countrymen a...	{'hashtags': [], 'symbols': [], 'user_mentions...	{'id': 2880746036, 'id_str': '2880746036', 'na...	1235	0	en	2020-08-21	
3298	5f4aa89bd6a56ce85825a94f	2020-08-28 21:57:27	RT @AttorneyGriggs: “From Civil Rights to Soci...	{'hashtags': [{'text': 'justicefighter', 'indi...	{'id': 32931825, 'id_str': '32931825', 'name':...	492	0	en	2020-08-28	
2821	5f4aa86f678bd9aa94f14106	2020-08-22 14:25:13	RT @hmcghee: Good morning. If you haven't watc...	{'hashtags': [{'text': 'TedTalk', 'indices': [...	{'id': 16586846, 'id_str': '16586846', 'name':...	286	0	en	2020-08-22	
2804	5f4aa86f678bd9aa94f140f5	2020-08-22 15:14:06	RT @hmcghee: Good morning. If you haven't watc...	{'hashtags': [{'text': 'TedTalk', 'indices': [...	{'id': 1476103248, 'id_str': '1476103248', 'na...	286	0	en	2020-08-22	
2805	5f4aa86f678bd9aa94f140f6	2020-08-22 15:12:45	RT @hmcghee: Good morning. If you haven't watc...	{'hashtags': [{'text': 'TedTalk', 'indices': [...	{'id': 1454712913, 'id_str': '1454712913', 'na...	286	0	en	2020-08-22	

FINAL PROJECT REPORT

Raya Young

Text was then grouped to see how many unique tweets were being shared. Again, we can see a particular tweet by an @hmcgee rising to the top. The first two entries have similar text.

text	retweet_count
RT @hmcghee: Good morning. If you haven't watched my #TedTalk, please do — and let me know what you think.	50050
RT @hmcghee: Good morning. If you haven't watched my #TedTalk, please do — and let me know what you think. https://t.co/1sOAaRf0Ng	15730
RT @maysoonzayid: TALK TO THE #DISABLED PERSON NOT THE NON DISABLED PARENTS. Thanks for coming to my #TedTalk	6810
RT @DuncanJWardle: Which is your favorite place to #ThinkDifferent? Mine is at my @TEDx talk on the Theory of #Creativity at @Tedx_AUK: htt…	2288
RT @EmergMedDr: If you get bit by an insect , it stings and hurts but you don't need to seek help straight away in an A&E.\n\nBuy some antihi…	1634
…	…
RT @LollyDaskal: A caring attitude is one of the great leadership qualities of a great leader.\n~@LollyDaskal https://t.co/ppfSHzIgez #Leade…	8
RT @LollyDaskal: You don't INSPIRE OTHERS by speaking about how amazing you are.\n\nYou INSPIRE OTHERS by showing them how amazing THEY ARE.…	8
RT @TEDxTCET: Even in these difficult times, a sister's love always shines. Here's to a quarantine Raksha Bandhan! #tedxtcet #tedxtcet2020…	8
A caring attitude is one of the great leadership qualities of a great leader.\n~@LollyDaskal https://t.co/ppfSHzIgez… https://t.co/jFj35CfF76	8
Get inspired for some family time! 😍 👇 What's your favourite #TEDtalk?\n#ece #ecd\nhttps://t.co/ObshPEhraS	8

The tweet by @hmcgee is indeed topical. Speaker Heather C. McGee gave a TED Talk titled, “Racism has a Cost for Everyone.”

Prior to the release of the new Bill and Ted movie, I also anticipated some tweets that were not exactly relevant like the meme, ‘Thank you for coming to my TED Talk’. You can see this above with a Retweet count of 6810.

Text tokenization on all Tweets brought up a notable frequency with LollyDaskal, although not in the other metrics like comments and views previously discussed. Her talk is entitled, “We Cannot Lead Others Without First Leading From Within”, and we can see Leadership also appears in the inventory of frequent terms. Once stopwords are removed, clearer analysis can emerge. It is also possible that because LollyDaskal’s username is one word, other names mentions may not appear because they were parsed as two words.

```
[('#', 9433),
 (':', 6521),
 ('@', 4530),
 ('https', 3466),
 ('RT', 2665),
 ('.', 2662),
 ('LollyDaskal', 1894),
 (',', 1886),
 ('you', 1604),
 ('to', 1531),
 ('a', 1243),
 ('the', 1238),
 ('and', 978),
 ('is', 956),
 ('of', 939),
 ('TedTalk', 908),
 ('//t.co/ppfSHzIgez', 824),
 ('’', 800),
 ('~', 800),
 ('TED', 787),
 ('in', 635),
 ('what', 593),
 ('my', 578),
 ('for', 552),
 ('!', 549),
 ('do', 534),
 ('Leadership', 515),
 ('I', 456),
 ('de', 441),
 ('t', 431)]
```

FINAL PROJECT REPORT

Raya Young

Conclusion and Next Steps

In conclusion, the tweets did show a pattern in the days and times they were tweeted, and seemed to follow the news somewhat. This tweet collection is incredibly small in scope, but potential next steps would be to analyze the attributes of popular tweets and predict the popularity of new tweets.

Next steps would be to filter the tweets by a particular attribute, most intuitively English tweets, and see what is being said about TED Talks generally. How many memes are there? How closely does the subject matter follow the news cycle? Beyond the passing of Sir Ken Robinson, are people more interested in education because they have to be more involved due to homeschool and quarantine? Do the relative topics of the talk being discussed seem to follow the news cycle, as in the civil rights events happening across the country? Or in the case of coping, are people more drawn to “happy” talks to take a break from all of the stress of daily life? What about sentiment analysis on the popular vs. unpopular talks? In terms of prediction, can we anticipate the ebbs and flows of a certain talk (either new or aged), that will rise in popularity due to relevance, and predict the amount of time it will take to be discussed on Twitter? Will publish date and time affect the potential popularity of a talk?