Journal Report 1

9/1/19 - 9/7/19

Arnav Bansal
Computer Systems Research Lab
Period 5, Bansal

Daily Log

Monday September 2

Labor Day

Wednesday September 4

I searched for existing TED talk datasets and discovered a few: https://www.kaggle.com/rounakbanik/ted-talks, https://www.idiap.ch/dataset/ted, https://data.world/owentemple/ted-talks-complete-list, https://old.datahub.io/dataset/ted-talks. I also found a basic tutorial on some methods to analyze and process this data: https://www.kaggle.com/rounakbanik/ted-data-analysis. I settled on initially working with the Kaggle dataset and adapting the features for input to these APIs. This data set contains corresponding CSV files with 2,500 TED Talks, along with the video, transcripts, and metadata. I transferred this data into Excel after deleting redundant/necessary data entries for TED talks.

Friday September 6

Planning ahead for future steps and to keep in mind criteria for the feature sets we soon construct, I explored some of the best speech-to-text APIs, 5 of which are listed here: https://nordicapis.com/5-best-speech-to-text-apis/. Among these, I read information about what level of access I would have to the API for each of the following services: Google Speech-To-Text, Microsoft Cognitive Services, and IBM Watson. I also read up on the basics of Mellin transform and discrete cosine transform as we start to consider potential audio preprocessing algorithms for model inputs.

Timeline

Date	Goal	Met
August 23rd	N/A - School didn't start	N/A - School didn't start
August 30th	Finalize project idea, Review timeline, Make any modifications necessary, Focus on plan for upcoming weeks	Yes, Submitted Journal Report 0, Discussed plan with Mr. White
September 6th	Finish formatting dataset for initial processing and proof-of-principle model training	Yes, Have dataset with 2,461 entries of TED Talks corresponding to their transcripts and metadata
September 13th	Test various Speech/Audio to Text APIs on this dataset	
September 20th	Have initial results for baseline implementations of these APIs on this dataset	

Reflection

Finding all the resources and links was a good first step and helped us understand a bit of the practical lay of the land, in addition to the high-level, slightly theoretical ideas we already had about this project from initial literature review last year.

Manually processing through the CSV files we obtained from Kaggle with TED Talks, their transcripts, and corresponding metadata was not very exciting. After figuring out common patterns between excess/redundant entries, Arvind and I split up the work in excising them from the spreadsheet. In the end, we were able to establish a basic, sizeable dataset with 2,461 TED Talk entries that we hope will serve quite valuable for audio preprocessing and initial model configurations and testing. Please refer below to an excerpt of what our corresponding CSV files look like, ready to be worked on:

	A	B
1	transcript	url
2	Good mornir	https://ww
3	Thank you sc	https://ww
4	(Music: "The	https://ww
5	If you're here	https://ww
6	About 10 yea	https://ww
7	Thank you. I	https://ww
8	On Septemb	https://ww
9	I'm going to	https://ww
10	It's wonderfu	https://ww
11	I'm often ask	https://ww
12	I'm going to	https://ww
13	I can't help b	https://ww
14	I'm the lucki	https://ww

	A	B	C	D	E	F	G	H
1	comments	description	duration	event	film_date	languages	main_speake	name
2	4553	Sir Ken Robir	1164	TED2006	1140825600	60	Ken Robinsor	Ken Robinsor
3	265	With the san	977	TED2006	1140825600	43	Al Gore	Al Gore: Ave
4	124	New York Tir	1286	TED2006	1140739200	26	David Pogue	David Pogue:
5	200	In an emotio	1116	TED2006	1140912000	35	Majora Carte	Majora Carte
6	593	You've never	1190	TED2006	1140566400	48	Hans Rosling	Hans Rosling
7	672	Tony Robbins	1305	TED2006	1138838400	36	Tony Robbins	Tony Robbins
8	919	When two yc	992	TED2006	1140739200	31	Julia Sweene	Julia Sweene
9	46	Architect Jos	1198	TED2006	1140652800	19	Joshua Princ	Joshua Princ
10	852	Philosopher I	1485	TED2006	1138838400	32	Dan Dennett	Dan Dennett
11	900	Pastor Rick V	1262	TED2006	1140825600	31	Rick Warren	Rick Warren:
12	79	Accepting his	1414	TED2006	1140912000	27	Cameron Sin	Cameron Sin
13	55	Jehane Nouja	1538	TED2006	1140912000	20	Jehane Nouja	Jehane Nouja
14	71	Accepting th	1550	TED2006	1140652800	24	Larry Brillian	Larry Brillian