1 Report on Adoption Research

1.1 Webscraping and Challenges Faced

I have used urllib3 in Python to scrape the data. urllib3 uses http to get the content of data from the url provided as input. Once the content is extracted from url, I parsed it using the newline characters first to get the lines neatly. Then I parsed it using the HTML tags intelligently in order to get only the important content out. In this case, I have targeted the content to be the comments in the forum threads. Once I could pull out the comments I used Case-Folding to make it usable for our analysis.

I have faced the following challenges during web-scraping:

1. Exploration of entire forum threads and multiple nested sub threads using scripts. For now I did not try to automate the way to explore all these nested threads as of now. I need to think of a way to do this in future.

2. I have used only Case Folding as the Data Preprocessing technique as of now. Based on increase in dataset or text, I can try to implement other techniques like punctuation removals or stemming/lematization etc.

1.2 Take a sample of data, i.e. a discussion thread, and tell us how many birth mothers in the thread decided to give up their child for adoption. Describe the steps you took to do this and any considerations that went into this

Since the dataset is pretty small at the moment, I could not use any Machine Learning or Deep Learning based techniques like we discussed in the first round of interview. Rather, what is suitable for such small datasets and the given problem is text analysis using regular expressions. Possibly, after automated scraping of entire website, if we find that we have sufficient data for model training, I can take an AI based approach. I have also searched for online datasets, but I could not find any related to our problem.

I have used the parsed data from the threads i.e. the comments and tried to find various regular expressions. $pattern = "(gave|give|place|placed|put|(lether(go)*))([\cdot])((((up)*)foradoption(.))|$
$to(.*?)(families|family))|$
$speakingasa((b|birth)(|-*)(mom|mother))|$
$iam((a)*(b|birth)(|-*)(mom|mother))|$
$tofind(wonderful)*(adoptive)*(parents|parent)formy$
$(child|children|kid|kids|baby|babies|newborn|new-born|twin|twins|daughter|daughters|son|sons).$
$(?!w))"$

Basically I have understood the patterns in the data and figured out simple regular expressions can represent the data we want to extract and took advantage of the same.

I have run the experiment of 7 positive sample and 7 negative sample data. By positive sample, I mean any thread which has a comment where a birth mother has spoken about giving out her child. Negative sample refers to no such occurrence of birth mother speaking about giving her child for adoption.

The results are as followed(**format= [url: number of mothers given their child for adoption**]:

- **Positive Sample:**

1. https://adoption.com/forums/thread/390261/open-adoption-never-again/: 3,
2. https://adoption.com/forums/thread/485730/giving-my-1-year-old-up-for-adoption/: 2,
3. https://adoption.com/forums/thread/146988/input-from-any-adoptees-i-am-birthmom/: 1,
4. https://adoption.com/forums/thread/41623/can-i-give-my-baby-up-in-america-if-i-am-in-t 1,
5. https://adoption.com/forums/thread/392740/want-to-give-my-twins-for-adoption-but-do-no 1,
6. https://adoption.com/forums/thread/10692/can-she-give-my-grandchild-up-for-adoption/: 6,
7. https://adoption.com/forums/thread/393036/adoption-by-choice-tampa-fl/: 2

- **Negative Sample**

1. https://adoption.com/forums/thread/487316/newly-adopted-teenager-faking-chest-pains/: 0,
2. https://adoption.com/forums/thread/390223/why-is-this-group-only-in-the-quot-post-quot 0,
3. https://adoption.com/forums/thread/376889/donor-embryo-adoptive-parents/: 0,
4. https://adoption.com/forums/thread/376955/is-there-a-service-that-does-the-scrapebook/ 0,
5. https://adoption.com/forums/thread/296220/sewing-baby-blankets-clothes/: 0,
6. https://adoption.com/forums/thread/482304/sending-to-birth-parents/: 0,
7. https://adoption.com/forums/thread/365054/start-to-finish-how-long/:0.

To the best of my analysis, the above provided regex has successfully detected the problem statement in positive cases and also did not detect any instance for negative data.

1.3 Give us a broad view of the process you would use to achieve the research objectives we described in our meeting and the timeline and time it would take.

Overview:

1. I plan to take initial time in exploring the data and figuring out ways to automate the webscraping process. 2. Then I would take some time on data analysis after I have the data in hand . 3. If the data is not large in volume, I will stick to traditional techniques like RegEx or others. However if we have sufficient data, I can manually tag them for positive, negative and irrelevant samples and train a model using Machine Learning or Deep Learning models depending on the data.

Timeline:

1. Automate webscraping and exploring data- 2-3 weeks.
2. Data Analysis - 1-2 weeks
3. Performing experiments-

- a. Preprocessing-1-2 weeks
- b. Actual implementation 2-3 weeks
- c. Results, analysis - 1 week

But these are all approximate timeline, If the data is smaller certain timelines may dicrease or if it is too big, I might need some timeline like analysis little more time

1.4 Script Link

`https://colab.research.google.com/drive/1hsFTD3Ddi9PaS-v5wzRc36A305eRyktr?usp=sharing`