

Project3

Liqi Zhu & Xiaonan Hu

Part 1 Data Acquisition & Process

1. Data Source

For the data source, based on the research on playforms and related APIs, we decided to choose twitter as our main data source, we also managed to scrape Quora and use that dataframe as a supplement. Twitter's official API is great for small dataset, but far from enough for our positive anymore detectaion.

Official API limits the requesting rate and amount. We managed to conquer this issue by using an unofficial API based on Beautiful Soup Pack to scrap the search page of keyword: 'anymore'.

For Quora, we developed our own scrapper and that will be in Part 3.

Twitter data Link: https://drive.google.com/open?id=1V8ipRSBpPmnNGaBOZF98cxTIRh4rC1U_

Quora data link: <https://drive.google.com/open?id=1cJDZ6j6vK4M1tVvqGypM-DS3rd0WRCab>

2. Data process

For Twitthter data, We scrapped over 800,000 tweets that contains 'anymore' in total, over 60,000 for each month in 2017. We firstly dealt with normal punctuation, tokenized and translated punctuations to one format. Then with nltk.sent_tokenize package. After we dropped off the duplicated sentences, we finally have a dataframe consists of 808,490 sentences.

For Quora data, we scrapped over 6,727 sentences that contains'anymore'. Also tokenized and translated to be analyzed.

Code is in pyhton file attached.

Part 3 'Positive Anymore'Detection

With the concept of NPI and negative triggers, we detect the negative anymore based on the context. We weren't able to find an existing NPI corpus so we manually define as showed in codes. It turns out that we can detect most of the srtict positive anymore. Error can be those typo mistake and declarative questions.

```
1. df2['isQuestion']=b.str.contains("\?")
2. df2['isNegativeAnymore']=
```

```

(b.str.contains("before|prefer|without|surprised|lost|ion|\?
|than|rather|less|most|not|no|not|n\'t|nt|idk|idc|idek|idec|idfk|idfc|rela
ctant.*|deny|reject.*|refuse.*|decline.*|repulse.*|impossible|doubt|suspec
t.*|suspicious|hard|never|rarely|none|no|every|only|barely|scarcely|few|li
ttle|only|all|zero|0|if|whether|whatever|whenever|wherever|unless|lest|who
ever|than|rather|less|most|except|prevent.*|unlikely|improbable|dislike.*"
)|b.str.contains("keep&from")|b.str.contains("kept&from")|b.str.contains
("too&to")|b.str.contains("est&any")|b.str.contains("est&ever")|b.str.c
ontains("without|except")|b.str.contains("all&but"))
3.
4. df2['OvertNegatives']=
b.str.contains("ion|not|no|not|n\'t|nt|idk|idc|idek|idec|idfk|idfc")
5. df2['IncorporatedNegatives']=b.str.contains("lost|prevent.*|unlikely|imp
robable|dislike.*|reluctant|deny|reject.*|refuse.*|decline.*|repulse.*|imp
ossible|doubt|suspect.*|suspicious")
6. df2['NegativeFrequencyAdverbs']=b.str.contains("hard|never|rarely|none|n
o|every|only|barely|scarcely")
7. df2['Quantifiers&QuantifiedAdverbs']=b.str.contains("few|little|only|all
|zero|0")
8. df2['HypotheticalClauses']=b.str.contains("if|whether|whatever|whenever|
wherever|unless|lest|whoever")
9. df2['Comparatives, Superlatives, etc.']=
(b.str.contains('before|prefer|without|surprised|than|rather|more|less|mos
t')|b.str.contains("est&any")|b.str.contains("est&ever"))
10. df2["*Predicationsof'excess'withToo"]= b.str.contains("too&to")
11. df2["*NegativePrepositions"]=(b.str.contains("without|except")|b.str.co
ntains("all&but"))
12. df2['EvenUse']=b.str.contains("even")
13. df2['EveryUse']=b.str.contains("every")
14. df2['ModalVerb']=b.str.contains("shall|should|would|must|can|could|might|
will|won\'t|wont|might")
15. df2['EverUse']=b.str.contains("ever")

```

Part3 'Positive Anymore' Examples and Discussion

1. Example

The total number of sentences that we detected positive anymore is narrowed down to 12,990, about 1.6% of total. Here's some of the example of positive anymores:

Sentence Example
"but then again, what does "coon" mean anymore.."
**"care about me anymore and why the fuck did you block me."
"cod is shit anymore; infinity ward 3 strikes & out, & banning ppl w/ unfilled lobbies."
"@instagram hi there, i need a bit of help, i have forgotten my login and i the email address anymore that i have set it up with"
"so, the awards are reliable anymore.. "
"i'm supporting chelsea at epl but i will support everton too, anymore."
"im such a sad person anymore"
"@realdonaldtrump so anymore golden showers lately ."
"who smiles in pictures anymore pic.twitter.com/vugglyyzat "
"what is going on anymore."
"you should feel lucky anyone hits you up like that anymore"
"starbucks and an iphone 7 can i be anymore white"
"it's him protecting himself from anymore hate."
"shows how much i hang around /r/hockey anymore."
"previous said things doing apply anymore. "
"can afford anymore embezzelers split the port folios."
"and why did he, he just looked back at me that way anymore, because the past is overlay."
"im going to the gym with thea tomorrow and im scared because...she's a sag sun leo moon need i say anymoretake shit anymore"
"take shit anymore"
"i miss my after school naps, i get so tired anymore "
"in the name anymore."
"the world is watching @morning_joe anymore. "

Sentence Example
"i'm ok anymore coz of this "
.....etc.

2. Discussion

While we were enjoying our observation to some interesting sentences, as mentioned above, the percentage of 'positive anymore' is very low. And as we showed, some of the positive anymore can be hard to get the true meaning, we also found that if 'anymore' in questions can be considered as positive or neutral, then there's a myth how we gonna tell those anymores' classification. Based on our dataset, there are about 12,000 potential positive or semi-positive anymores in questions, not to mention there's always people tweeting without punctuations.

We believe our code helped us find a feasible method to observe those positive anymores, and based on addition modifications, the algorithm can be improved to better accuracy. We may investigate different characteristics based on categories of NPIs and key word usings, and that will be detailed in following parts.

Part 3 Quora

1.Quora Scraper

To enrich our data source, we decide to scrape Quora. And on Quora, people always write in a more nature way as there are in the particular context. But there is no official API for Quora, and most of the existing unofficial APIs were developed in Python 2, so we programmed a scraper by our own.

Our scraper is developed using the packages '**selenium**', '**time**' and '**bs4**'. First, the scraper open the searching result page on Quora (<https://www.quora.com/search?q=anymore>) using the '**webdriver**' from '**selenium**' package. Then, defining the function '**execute_times**' using functions '**driver.execute_script(selenium)**' and '**sleep(time)**' to scroll the page to load more results. When the scrolling finished, using the function '**BeautifulSoup**' to capture the questions' texts containing 'anymore', and the urls of answers respectively. After that, we scraped the urls we got, in the similar way, and obtained the answers' paragraphs containing 'anymore'. At last, using the '**sent_tokenize(nltk)**' to get all the sentences containning 'anymore', and saving as .txt files.

2.Quora Data Overview

With this dataset, we narrowed down to 392 potential positive anymores if questions are involed in discussion and 25 samples if questions are not considered.

It's not surprising that most of the sentences are questions, but quora data shows that percentage of positive anymore (0.3% without questions) can be much lower than that in twitter data, considering the fact that people use informal languages much more often on twitter, it may be an evidence.

Our evaluation

It turns out dealing with JSON file can be tough, we observed all kinds of informal usage of language, people may use English words with Chinese format punctuations, people may use 'idk', 'idc', 'idrc' to indicate some phrases, people may not use punctuation at all. And it took us much time for data acquisition and process.

It's a pity that our method is not able to grasp geographic data, and that will definitely be our next goal, we found the tool to grab the data but we aren't managed to involve that part in this report yet.

Hope our research and data helps anyway to this research.

Thanks.