Where Are Regionalisms on Reddit

Noelwiz

Introduction

Regionalisms are words that are associated with a region. In our reading of chapter two of Because Internet, a study was mentioned that found people use more regionalism when talking replying directly to someone (MCCULLOCH, 2019). More precisely, they looked at the intended audience, from one person being replied to, to the world with a hashtag. What they found was that people choose when to use regionalisms and "tweetspeak" based on the audience, shown by an increasing likelihood of using them the wider the intended audience.

My question is, is are people also more likely to use regionalisms when making a post on a subreddit, replying to a post, or mentioning a user, like twitter. Statistics were gathered from subreddits focused around various locations because location data isn't available from reddit. Reddit does not exactly have the same concepts of hashtags and @replies as twitter. Instead it's organized around subreddits which are communities people can post on focused around a single topic such as gaming, or Seattle. Each post has a comment section for discussion, which people can reply to each other creating threads. Additionally, people can /u/mention a user specifically.

I hypothesize that Comments mentioning users will have the most regionalisms, followed by comments, followed by posts mentioning users, followed by posts. My reasoning for my hypothesis is that posts have the widest audience while comments have the smallest, and if a user is mentioned, it's more likely to be a response or conversation.

Why is this important? I want to identify regionalisms automatically in the future, so knowing where to look in this data will be helpful when trying to find them. Speaking of the data, I used some of a large set of corpuses divided by subreddit, and processed with a library called ConvoKit. The data was collected by a website called Pushshift.io from the creation of that website until October 2018 (Convo Kit Doumentation, 2020).

The exact subreddits can be found listed at the end, as well as the list of regionalisms I used. I borrowed the word list from the paper "Audience-Modulated Variation in Online Social Media" by Umashanthi Pavalanathan, the article from Because Internet, and then used most of the same places to better compare my findings with theirs. Data was collected on the frequency of all regionalisms by subreddit as well as the percent of utterances that contained at least one regionalism divided by subreddit and by if the utterance was a comment, or post, which either did or didn't mention another user.

Research Question: On location centric subreddits, in what contexts do people use regionalisms most?

Hypothesis: Comments mentioning users will have the most regionalisms, followed by comments, followed by posts mentioning users, followed by posts.

Methodology

The raw data I used came from Pushshift.io and included metadata like what subreddit a post was from, the score (upvotes – downvotes), a user id and author, as well as a conversation id to organize by conversations. Unfortunately using the Convo Kit Library proved impossible for most of the data as loading up the corpuses took up too much memory. Instead I wrote a Python script that processed them into four text files for each subreddit based on weather an utterance was a post or a comment, and weather it mentioned a user or not.

While processing the data, I made all the text lowercase, and created a second text file with common words removed, using the list of stop words included with the Natural Language Tool Kit Python library. It took about an hour and a half to make the text files, but after that they were much faster to work with.

With that data, I then collected three statistics. First, I collected the percent of posts, posts that mention a user, comments, and comments that mention a user that contained any regionalism in the list at the end of the paper. I used the text file for each subreddit with the stop words filtered out to speed up the computational time, and wrote the results to a .scv file I could open in Microsoft Excel. The second statistic I collected was the term frequencies (what % of all words) for every word in my list of regionalisms for each subreddit. I did this because in Pavalanathan's article, they included a figure with the term frequencies they recorded for their list of regionalisms. Finally, I recorded stats about the stats I got. The number of unique words, the total number of words, and the number of posts regardless of if they had a regionalism.

Results

As I did data processing and started basing my work off of Pavalanathan's set of words I noticed some potential confounding factors I couldn't or didn't have time to control for.

Junk Data

Bots are an interesting part of reddit. People can write bots to post comments or posts automatically or do other things. Unfortunately I didn't have a way to filter out bots, so their automatic responses are in my data. Some subreddits use bots to remind users of the rules, mentioning them by name to alert the user. I think this reduced the number of comments with regionalisms, and posts. Additionally, I did not notice that deleted comments were left in as "[deleted]" until it was late, so that

increased the number of all types of posts. Finally, I completely forgot that urls may sometimes have "/u/" in them.

Regionalism selection

The data Pavalanathan used to choose their words was based on twiter data from 2009 to 2011 using some smart start stuff that I didn't have time for. So, the word selection is based on a different platform up to 7 years older than my data.

I also kept the smiley faces in my data, which aren't exactly what I'm looking for, but are still a marker of informal language

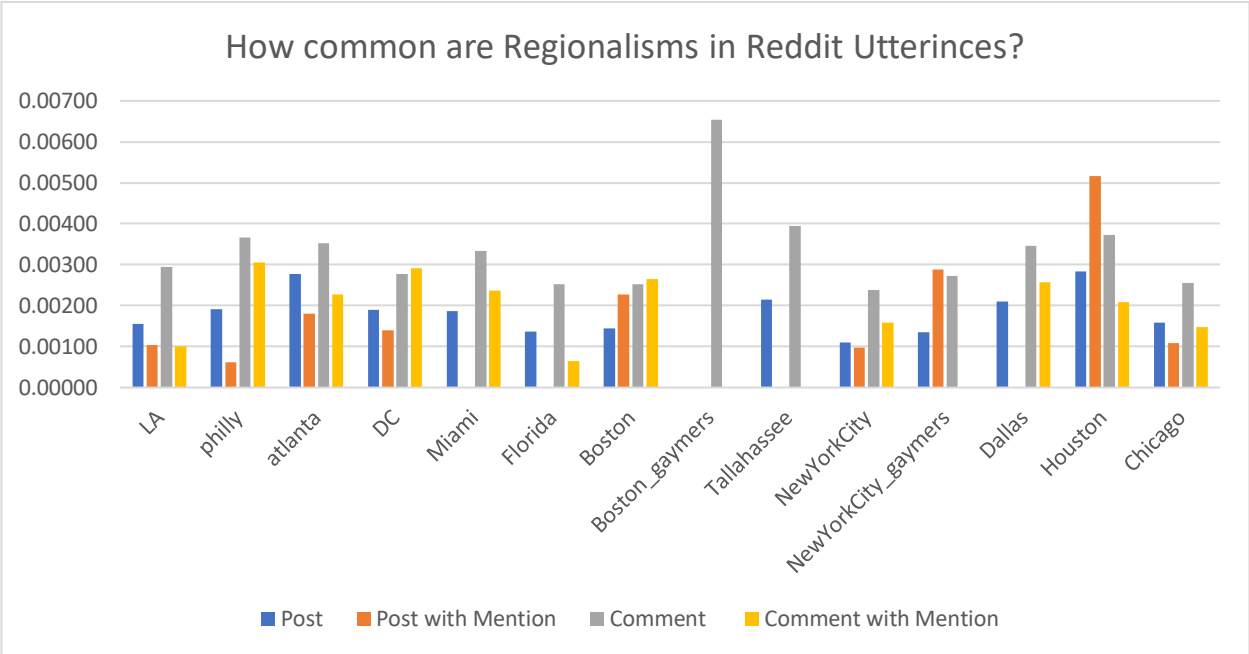
Analysis

In total, there were 504,348,448 words in the data across all subreddits. I defined words as anything with a space between it. So, "( )" would consist of the word "(" and ")". There were a total of 29,655,569 posts and comments overall.

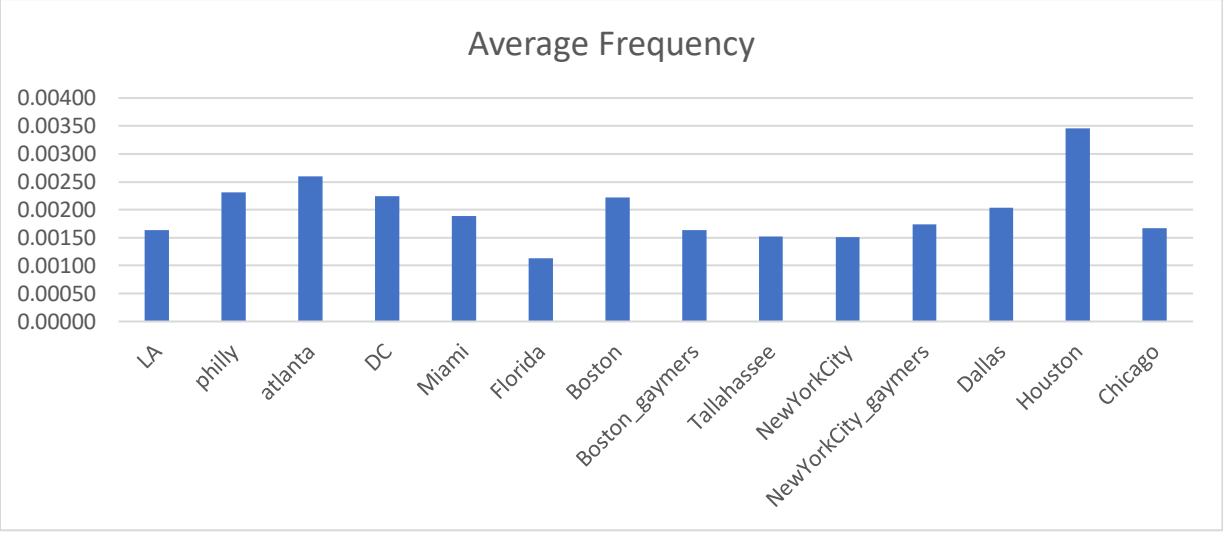
For every subreddit I analyzed, regionalisms were more common in comments than in posts, as seen bellow. However, mentioning a user was not necessarily correlated with the number of regionalisms used for posts or comments. Only Boston matched my predictions with this data.

Comparing posts and comments that do and don't mention another user, almost every single subreddit had less frequent regionalisms when another user was mentioned.

One interesting thing is that the two local gaming subreddits had more regionalisms than their local subreddits. This matches the idea that the more local the audience, the more likely a person is to use a regionalism.



Subreddit	Post Frequency Difference (mentions user – no mention)	Comment Frequency Difference (mentions user – no mention)
LA	-0.00052	-0.00195
philly	-0.00130	-0.00061
atlanta	-0.00097	-0.00125
DC	-0.00049	0.00014
Miami	-0.00187	-0.00096
Florida	-0.00136	-0.00187
Boston	0.00083	0.00013
Boston_gaymers	0.00000	-0.00654
Tallahassee	-0.00215	-0.00395
NewYorkCity	-0.00013	-0.00079
NewYorkCity_gaymers	0.00153	-0.00272
Dallas	-0.00210	-0.00089
Houston	0.00232	-0.00163
Chicago	-0.00049	-0.00108



Conclusion

In conclusion, it appears regionalisms are more common in posts than in comments, but it's impossible to say without more work weather or not mentioning a user has any correlation or effect. On reddit, like twitter, audience does change the frequency of the use of regionalisms and informal language. I'm disappointed I don't have enough time to go back and farther refine the data, but this would be good work for someone else to try.

One thing that needs exploring is the effect of bots on this. This was probably one of the biggest oversights, along with deleted comments when doing this analysis. During testing, I found that the deleted comments were the most common word for the LA subreddit at least, and probably more, and there was a large number of them. Given these oversights im surprised how common they were still. Part of that is likely the inclusion of some smiley faces from Pavalanathan's word list, but still. The differences are large enough that I still think that mentioning a user decreases the number of regionalisms used, but I would not be too surprised if the opposite were true.

Other future work that could be done is comparing this to more general audiences. I had a few larger subreddits downloaded like /r/gaming but it took so long to get the text files to analyze for this set of data that I couldn't in the time I spent on this. I also think it would be worth more closely replicating Pavalanathan's process for this data. I tried to record frequencies, but that isn't entirely functioning, but would be useful to compare to the older twitter data to see how much of an effect platform and time have had on my results for better comparison. Also, looking at city, county, state and country subreddits. Given the results so far, comments that don't mention a user use the most regionalisms, but with a wider audience, people may use more or less regionalisms in posts compared to comments. People might use more to distinguish which area they're from, or less to blend with everyone.

Finally, I plan on making the code a little more user friendly for my own use. So far, I want to get input from users, download the corpus and unzip it in the code rather than myself, and save the processed text files in one place. Saving the processed text files would make it possible to iterate over all of them at once. Also, making a small script that converts a set of corpus to the processed text file than analyzes them seems like a good improvement. The .csv files for the analysis results, and the source code will be available on github, linked in the See Also section. Some more data is also included below.

Regionalism List

- lml
- deadass
- od
- odee
- werd
- cud
- nuttin
- nicee
- sed
- lata
- buggin
- wrd
- noe
- w|
- layin
- okk
- lols
- lolrt
- crazyy
- sour
- wid
- fasho
- ahah
- cuh
- koo
- cuhz
- fkn
- ahahah
- ;o
- mfs
- goofy
- nbs
- lbvs
- bogus
- 2ma
- lbs
- mf
- ikr
- lmmfao
- hoop
- crackin
- ion
- nun
- oomf
- tf
- (;
- finna
- dang
- fa
- (:
- <<
- >>
- <--
- .!
- trippin
- y'all
- mayne
- fwm
- jammin
- shid
- jamming
- tripping
- azz
- bck
- ma'am
- bae
- whoop
- ole
- sho
- fck
- lowkey
- lawd
- fa
- trippin
- ard
- jawn
- cdfu
- bul
- wya
- 1omf
- jawns
- ctfu
- ctfuu
- hbu
- rd
- foh
- sike
- hype
- nut
- bull
- lt
- lrt
- llss
- bait
- fakin
- stamp
- ji
- brova
- siced
- hu
- wholetime
- guh
- bol
- jit
- bih
- vibe
- preciate
- fye
- frfr
- slick
- shid
- fr
- ain
- ikr
- followback
- flex
- gotcha
- legit
- deff
- gunna
- yall

Subreddits Used

- r/LosAngeles
- r/Philadelphia
- r/Atlanta
- r/washingtondc
- r/Miami
- r/florida
- r/boston
- r/BostonGaymers
- r/Tallahassee
- r/nyc
- r/nycgaymers
- r/Dallas
- r/Houston
- r/chicago

Unused Regional Subreddits

I had these downloaded but not enough time to use them

- r/Portland
- r/PortlandGaming
- r/Bellevue
- r/BellevueWA
- r/Bellingham
- r/Seattle
- r/Washington
- r/WashingtonStateMaps
-

Raw Data

Frequency of Regionalism

Subreddit	Post	Post with Mention	Comment	Comment with Mention	Average
LA	0.00156	0.00104	0.00295	0.00100	0.00164
philly	0.00191	0.00061	0.00367	0.00306	0.00231
atlanta	0.00278	0.00181	0.00352	0.00227	0.00260
DC	0.00189	0.00140	0.00277	0.00291	0.00224
Miami	0.00187	0.00000	0.00333	0.00237	0.00189
Florida	0.00136	0.00000	0.00252	0.00065	0.00113
Boston	0.00144	0.00227	0.00252	0.00265	0.00222
Boston_gaymers	0.00000	0.00000	0.00654	0.00000	0.00164
Tallahassee	0.00215	0.00000	0.00395	0.00000	0.00153
NewYorkCity	0.00110	0.00097	0.00238	0.00159	0.00151
NewYorkCity_gaymers	0.00135	0.00288	0.00272	0.00000	0.00174
Dallas	0.00210	0.00000	0.00346	0.00257	0.00203
Houston	0.00284	0.00516	0.00372	0.00209	0.00345
Chicago	0.00158	0.00109	0.00255	0.00147	0.00167

Number of Words Overall

Subreddit	N - num words	Num Utterences
LA	62414741	3743256
philly	40913311	2472926
atlanta	57069042	3282875
DC	39048277	2177489
Miami	9049440	551743
Florida	7778808	466582
Boston	61907433	3507582
Tallahassee	2519834	156242
Boston_gaymers	9391	674
NewYorkCity	61174507	3670782
NewYorkCity_gaymers	310104	25142
Dallas	26102077	1531530
Houston	59616506	3612348
Chicago	76434977	4456398
Total	**504,348,448**	**29,655,569**

Number of Utterances (a post or a comment)

	Post		Comment	
Subreddit	No user mention	user mention	no user mention	user mention
LA	305708	3842	3391500	42206
philly	175063	1632	2279878	16353
atlanta	252288	3312	3000790	26485
DC	200729	2863	1961865	12032
Miami	69134	287	479374	2948
Florida	51603	322	411558	3099
Boston	258488	2644	3227173	19277
Tallahassee	62663	352	92655	572
Boston_gaymers	368	0	306	0
NewYorkCity	258861	1026	3392006	18889
NewYorkCity_gaymers	6689	347	17653	453
Dallas	165311	4043	1353600	8576
Houston	302170	4458	3285148	20572
Chicago	320178	10968	4087828	37424
Total	2,429,253	36096	26981334	208886
	29,655,569			

Number of words

Subreddit	Post		Comment	
	No user mention	user mention	no user mention	user mention
LA	3725225	44446	58172814	472256
philly	2200972	18846	38481090	212403
atlanta	3606082	42300	53074784	345876
DC	2924564	42906	35856778	224029
Miami	834006	3293	8171661	40480
Florida	543646	4334	7194572	36256
Boston	3521565	32990	58019536	333342
Tallahassee	815888	5526	1687062	11358
Boston_gaymers	4583	0	4808	0
NewYorkCity	2697568	15545	58199420	261974
NewYorkCity_gaymers	83562	4248	218480	3814
Dallas	2064626	33557	23871056	132838
Houston	3971989	42260	55332877	269380
Chicago	4298390	97619	71625127	413841

Number of Unique Words

Subreddit	Post		Comment	
	No user mention	user mention	no user mention	user mention
LA	156150	10588	1018312	49499
philly	116168	5308	787867	29954
atlanta	154617	9932	957173	44538
DC	135353	9551	728993	31008
Miami	59955	1078	262732	9450
Florida	41682	1551	233449	8769
Boston	153136	8613	1000540	41175
Tallahassee	60093	1700	91039	3677
Boston_gaymers	1596	0	1500	0
NewYorkCity	130194	4235	1013524	36507
NewYorkCity_gaymers	14133	1619	26566	1591
Dallas	111686	7924	541254	22065
Houston	170502	9669	996063	37374
Chicago	166189	21444	1159092	48615

Works Cited

Convo Kit Doumentation. (2020). Retrieved from https://convokit.cornell.edu/documentation/subreddit.html

Harvard University Press. (n.d.). *Dictionary of American Regional English | DARE*. Retrieved from Dictionary of American Regional English: https://www.daredictionary.com/

Hickey, W. (2013, June 5). *22 Maps That Show How Americans Speak English Totally Differently From One Another*. Retrieved from Buisiness Insider: https://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6

Labov, W. (n.d.). Retrieved from TELSUR PROJECT: https://www.ling.upenn.edu/phono_atlas/home.html

MCCULLOCH, G. (2019). *Because Internet.*

See Also

- Data set used
  - https://zissou.infosci.cornell.edu/convokit/datasets/subreddit-corpus/corpus-zipped/
- Code Repository
  - https://github.com/Noelwiz/RegionalismsOnReddit