Using Collocates to Examine Anonymous Expression of Self and Perception of Others

Maxwell Wenzel
LING 3430

December 16, 2018

Throughout your everyday life there are countless factors that contribute to how others perceive you. Anything from your tone of voice to what you are wearing will either consciously or subconsciously change the way people you interact with think of you. Of course, many of these factors can be directly controlled by you. What you wear, say, and do are largely determined by your direct actions and in that way you can manage to a degree how you express yourself and how others perceive you. While these few that you can control do contribute a large amount to how your overall outwards identity is expressed, there are still the many you can not. The two groups of characteristics can be seen to be reflective of two concepts of yourself. The parts of your outwards expression that you have control of can be thought of as a representation of yourself determined by how you see yourself and how you want others to think of you. The second part which is what you can not control can be furthermore divided into categories, there's that such as inherent attributes such as your voice, skin color, eye color, and so on which are simply immutable features. However, there are also things such as behavior and actions which are subconscious and in a way can convey information to who you really are.

A key component of how others perceive you is the fact that you often are not interacting with a completely new set of people from day to day. During your habitual and regular activities throughout the day, it is reasonable to assume that most of the people you interact with are people you have interacted with before. From your coworkers, friends, family, or even just the employee that helped you at the store you're a regular of, it is more likely than not that the two of you have history. Even a passing glance is enough to plant the seed for the mental model of how an individual thinks of another. What starts off as a first impression or simple observation snowballs, as there are more and more interactions between two people, into a full fledged representation of another person. This sort of representation is from where we draw our opinions of others and build familiarity.

We tend to say thing such as “That was out of character for them” when someone does something that appears to be dissonant with our idea of them. The previously mentioned factors that control your outwards appearance and personality is what is observed by others and then filtered by their preexisting beliefs and experiences to form their own idea of you. This model directly dictates how that person interacts with you. These factors all combine to create a complex system of bias and uncontrollable factors that influence interaction.

It may seem that it is impossible to have human interaction without the influence caused by all of these factors. It is true that when observing a conventional conversation between two people that it is not possible to not have the content and manner of their conversation be influenced by the unconscious projections of their selves and biased perceptions of each other. It is likely possible that there is no way to completely eliminate these factors when considering conversation due to the nature of how the human mind processes it. The closest you can get to achieving such a state of conversation unbiased by previous interactions and uncontrollable expressions of yourself is through a completely anonymous text based correspondence. Given the people involved speak the same language, having this anonymity completely eliminates past bias with this person besides from what has been given in the conversation. This allows for each individual in the correspondence the ability to carefully and fully choose what they “say” or type. In turn this eliminates much of the unintentional information given away about yourself in a conversation. Of course, factors such as vocabulary, and in some cases dialect, will be unintentionally conveyed through this correspondence which could give hint to the other person’s characteristics. These are some of the things that would be very difficult to remove without the use of some sort of equalizing intermediary.

The purpose of my research was to observe how others are perceived and how people express themselves in these sorts of anonymous correspondences. My first problem with examining this was finding a source of data which would suitably meet these guidelines. The primary problem is that there is no way that I could find to gather conversational data of this sort. By this I mean data that was produced in the setting described above with pure anonymity between parties. The conditions required for this sort of correspondence require a great deal of effort to orchestrate and record outside of digital correspondence, for this reason I limited myself to finding a source of data to some sort of web based communication. While the pure anonymity that I described above would be best for my research goal, sources of these types of conversations on the Internet are fairly rare, and the ones that do exist do not have a useful amount of data. I decided eventually to gather data from www.reddit.com which has an environment in which individuals converse in a sort of soft anonymity. Reddit is structured into small sub-communities called subreddits, these each have a name which often denotes the theme of that community and then leaders or moderators of that

community regulate that only relevant content is contributed. Subreddits are further structured into posts which can contain a link to a web page, image file, or contain a body of text. Each post then has a comment section where conversations can take place among users. Each user has an account which can be viewed to show their previous posts and comments. However, a user has no identifying information which links them to their "real" identity unless such information is given in a post or comment by the user which is rare, discouraged, and frowned upon. This structure provides a soft anonymity which shares many characteristics of the pure anonymity desired, except with a Reddit user it is possible to view their previous activity or to be recognized by their user name, however this is not often done. Users more often than not do not recognize or build familiarity with each other with the exception of subreddits with a small number of users. Due to the rarity of these sort of breaches in anonymity, Reddit comments seem a suitable source of data in which most of the correspondence is done in an anonymous way. By being able to hand choose which subreddits to gather data from I was able to avoid small communities where there were more likely to be interactions with familiarity.

In order to collect Reddit comment data I created a python script (source code below) that would go through the two hundred most popular posts on the given subreddits and save the comments to a text file using the python library PRAW (Python Reddit API Wrapper) if they contained certain keywords. Since I wanted to observe interactions where either a user was expressing themselves are directly expressing their opinion of another user I used the keywords I, you, I'm, and you're. I chose five different subreddits to gather comments from; casualconversation, changemyview, happy, legaladvice, and iama. I chose these subreddits as they are all relatively large communities ranging from happy with 300k users to iama with eighteen million. I chose casualconversation as a source for comments as it, as the name suggests, is focused on the general topic of just conversing with no general rules except to just talk to people and be civil. I figured this would give a good baseline of people conversing about a variety of topics. Next I chose changemyview as it would provide data where people are more directly speaking to an individual and addressing them about something they have described about themselves. The concept of changemyview is that the user posting states a view that they have on something, usually somewhat controversial, decisive, or against norms, and the users that comment provide arguments that go against the poster's view. I also chose happy which has a more simple form of the concept being that you post something that makes you happy or something that happened to you that made you happy. I chose this subreddit as it provides conversations similar to casualconversation but with a bit more focus and direction towards the original poster. For my fourth choice I chose the subreddit legaladvice which has the concept of the commenters discussing the poster's situation and offering legal advice. I included this subreddit as I figured it would add more conversation data in a more professional

yet still largely anonymous situation. The reasons for why I chose iAma is a little more complex. The concept of iama is that the poster is someone of interest, such as someone with unique life experiences, profession, or a celebrity who then asks the commenters to ”ask me anything”. The poster will then respond to the questions from other users. For example the most popular post of all time for this subreddit was made by Barack Obama. As you might expect, this does, in some circumstances, remove part of the element of anonymity. However, I still believe this data to be valuable as the anonymity is only broken in one direction. Additionally, this is the subreddit I chose with the most users meaning it was able to provide a lot of data.

Once I had saved the comments from the top 200 posts from each of these five subreddits, I loaded the comments from the text document into a Jupyter notebook. From there I was able to process the comments in several ways to get the information I wanted. First I separated the comments from each other, from that I was able to see that I had gathered a total of $136,006$ comments across the five subreddits. I then further split up each comment into individual words and removed punctuation from the words. I then was able to count the total number of words that I had gathered in my data which ended up being $6,043,108$. From here I had to decide how I wanted to process the data to extract something meaningful. From the data provided I figured that the best way to extract meaningful information pertaining to how the keywords that I described above were used in the data was to find the twenty most common collocates for the four word positions after each of my keywords. Doing this involved going through each comment and finding where each keyword was and then increasing a counter keeping track of how many times each word appeared adjacent to one of my keywords. After doing this, I was able to query this counter to find the most common collocates of each keyword in my data. In order to make this data more easily observable and easy to work from in the process of understanding trends I wrote a python script that would take this counter, the total number of words in each position of adjacency to my keywords, and the appropriate keyword as input. It would then print out the LaTeX code to produce a nice well formatted table that is a little easier on the eyes than a python print statement. These tables can be found below at the end of the document and I will be referring to them throughout my analysis. I will now go through each table and discuss my analysis of their contents.

In order to make it easier to consult this table and compare it to my conclusions stated here I will briefly describe its structure. The general structure applies to each table as they are identical besides from their contents. There are four primary columns separated by double lines in the table each split into two sub-columns. The first-sub column indicates in the first row which position this column corresponds to following the convention displayed below:

keyword 1st 2nd 3rd 4th

Then, further down this first sub column, the collocates of this position are listed from most to least frequent. As you may have noticed, there is one sub column on the far left side of the table which indicates the ranking of that word in accordance to frequency for easy reference. The second sub column indicates the percentage of the total words that appear in this position that this collocate makes up. Other than this, the very last row of the table indicates the total number of words that appear in each position. Before moving on to the individual analyses of the collocate tables I will make a few general notes on how I will be performing my analysis. Firstly I will largely be ignoring the collocates that carry very little semantic meaning the except when observing the possible lexical bundles that can be observed from the tables. I should also mention here that when performing my analysis that I will be mostly ignoring the fourth position for all tables as they seem to contain little to no significant trends worth consideration. On a cursory glance you can quickly see that the fourth column in each table has near identical contents aside from a few slight changes in ranking.

The first table is showing the most common collocates up to four positions after the word I as it appeared in the comment data. The first collocate I will look at in this table is the most common in the first position as it is the one you could say is most closely linked to the word I. In this case it is "have" which is not too surprising of a result but nonetheless worthy of analysis (is producing hard evidence for what may be taken for granted or seen as mundane not one of the most important roles of science?). The appearance of this particular collocate as the most common is likely due to it having a wide variety of ways in which it can be used. One of the most important, if not the most important as supported by this finding, things we have when it comes to contributing to a conversation is our past experiences. So it is not surprising to find a word that is incredibly useful when referring to our past experiences as one of the most common words to appear with "I". The next most common word in the first position is "am" which goes to show the importance of expressing our current state of being, emotions, and thoughts to others. This can be thought to be paired with the next most common in the first position "was" which conveys similar semantic meaning but instead in the past. It can be observed that a simple lexical bundle can be constructed with "am" and the 18th most common word in the second column "wondering", the 16th in the third "if", and the fourth most common in the fourth column "you". From this you get the lexical bundle 'I'm wondering if you" which has some very interesting implications. This phrase demonstrates an individual's interest into what another individual in the conversation has to offer. This seems to suggest that there is a prevalent sense of engagement and respect for the others within these conversations. This is then followed by another pair worth mentioning, the fourth and fifth most common collocates in the first position are "think" and "know" respectively. These could possibly suggest that when communicating without the aid of things such as expression, tone, and body lan-

guage that it is more difficult to discern when someone is stating something as fact or opinion. It could also be seen as “think” appearing higher up than “know” as an indication of users hesitancy to claim something as fact due to the inherent fact of this form of communication necessitating that each participant has the ability to fact check literally at their fingertips. The hesitancy stemming from the fear of being called out and discredited. A collocate in the second position that can be used to create several lexical bundles is the seventh most common “like”. This combined with some of the collocates from the first column produces the bundles “I would like”, “I don’t like”, and “I feel like”. All of these interestingly are furthermore being used to express the individual’s desires, opinions, and feelings. This indicates that these features are among the most important that a person considers in what they want to express themselves. Another point of interest in this table are the words that appear in multiple positions. For example, “I” itself appears in the second, third, and fourth column ranked at 11th 6th and fifth respectively. “Think” also appears in both the first and second column ranking at fourth and fourteenth. Appearing in the first column goes to show how common it is to directly state your thoughts on a matter in these conversations and the second column can be used to create many more fine grained three word lexical bundles which perform the same function but with several different modifications. Those are just some of the more interesting observations that can be made from this table as a full analysis would require a paper of its own.

The first thing you may notice moving on to the collocates of “you” table is the numerous similarities in contents with the previous table. This alone reveals some insights into the treatment of the self and others. Primarily that the same focus on the expression of one’s own thoughts, feelings, and condition transfers over to a similar focus and respect for the same of other users in the conversation. It is in a way comforting to see this maintained care and respect for each-other even when removing many humanizing elements of a conversation. One major difference between the appearance of these words in the data is that there are approximately twice as many instances of “you” than “I” in the data. These seems to indicate a focus on the other person when in a conversation which in consequence suggests that despite a lack of familiarity that users are able to establish a rapport that is healthy for the conversation by focusing on the other person rather than themselves. The first word that appears in the first column that distinguishes it from the “I” table is the third entry “are” which is used to make statements about an users opinion of their partner in the conversation. From this you can see a strong tendency to engage with the other participants of a conversation by then adding their direct opinion of a person’s thoughts or action. This is certainly a difference between normal conversing and this anonymous conversing in that people are usually much more hesitant to share their opinion on another person’s idea, especially if they don’t know them well. A significant trend that can be seen among these collocates is the offering of advice or help towards the other person. This can be seen in ranks 10 “can”, 11 “could”, and 16 “would”.

This points towards a general willingness to assist others and provide constructive input to another person's problem. I feel as though this differs from normal conversations between individuals who are unfamiliar with each other where simply providing sympathies would be more common. This prevalence could also be due to the bias created by gathering data from legaladvice and changemyview where the giving of advice and constructive criticism is highly common. Other from these differences the trends seem to closely mirror those seen in the previous table right down to the high prevalence of "you" where there was "I" in the last.

Moving on to the table for "I'm" the first thing you may notice which shouldn't come as a surprise is that the overall length of any given word in this table is much longer than in the previous ones. This is due to the increased amount of verbs and adjectives that appear to the nature of "I'm". A standout feature of this table is that the word "a" which is the most common in the first position makes up a whopping 12.6% of all words that directly follow "I'm" which is more than double of the top collocates in the previous tables. This seems to be caused, at least in part, by a few evident lexical bundles that are tightly grouped towards the top of each column. These are "I'm a huge fan of" and "I'm a big fan of", which are without a doubt caused by the inclusion of iama. It is very common to see people prefacing a question for a celebrity with something along the lines of the above lexical bundles which likely skewed the data. More interestingly there are the high prevalence of adjectives which are optionally paired with the adverbs "really" in rank 11 of the first position and "very" in rank 13 of the first position. Among these there is "happy" in the sixteenth position of the first column and the ninth position of the second column, "glad" at rank 7 of the first column, and "sorry" at rank 15 of the first column to name a few. This display a similar prevalence as seen in the first table of a high rate and variety of sharing how one is feeling. A possible reason for this high rate of appearance could be that it is necessary due to +the lack of other ways to convey and display emotion in a normal face to face conversation. You also have the collocates "curious" in rank 8 of the first position and rank 17 of the second position as and "wondering" at rank 19 of the first position. These expose the trend of either inputting into the conversation something that they are thinking of and asking others for what they are thinking which builds into the valuing of the thoughts of others brought up previously. While the contents vary greatly from the previous tables you once again see the same trends throughout.

Much in the way that a mirroring was seen between the collocates of "I" and "You", there is that same relationship seen here between "I'm" and "You're". Something very important to note is that the non-contraction version of "I'm" and "You're" are some of the most common collocates present in the tables for "I" and "You". Once again here you see the very common first position collocate of "a" and its pair "an" in fifth place of the first position. These can be seen to pair with

other collocates to form the lexical bundles “You’re a hero” which utilizes the rank 19 collocate of the second position and “You’re an inspiration” using the rank 13 collocate of the second position. The first of these can be interpreted to mean directly declaring that someone is a hero for their actions, or more likely in the more common context to be a hyperbolic way of thanking another person for their response to some sort of request or question. In contrast the second is most likely arising from use in both happy and iama. From my research into each of these communities to see if they would be suitable for this analysis I noticed that many of the posts in happy have a premise of someone recently rising above adversity and becoming happy. It is very likely that the commonality of this trope led to the appearance of this lexical bundle as people are likely to comment that the person overcoming pain to reach happiness has inspired them to do the same. The last trend apparent among these collocates is the direct declaration of positivity towards another participant in the conversation. This can be seen in the first position collocates at rank 9 “awesome”, 16 “amazing”, 19 “happy”, and 20 “looking”. The last collocate listed here comes into its sense of praise when paired into a lexical bundle with the collocate at rank 12 in the second position to form “You’re looking great”. An odd one out in this is the pairing of “you’re happy” which seems to come off as a bit awkward as it is if someone is declaring another person happy. For this to make sense it is necessary to speculate that the preceding words forming a lexical bundle using this collocate would be something similar to “you look like you’re happy” which would make sense in the context of the happy subreddit. The overall theme observed in the collocates of “you’re” is an overwhelming amount of positivity directed towards others.

Having taken each table of collocates into consideration, they each add a piece to a puzzle showing several larger trends observed in soft anonymous conversation. The first of these I attribute to the lack of functions that are heavily used in normal face to face conversation. These being tone, body language, gaze, gesture, and facial expression. All of these are very useful in a conversation to convey one’s opinions, thoughts, and feelings in a subtle and nuanced manner. By introducing this sterile environment of near anonymity you removing these channels of conveying information and you are limited down to almost exclusively your words. It is for this reason that you see in many of these tables many words that are used to express and articulate one’s own feelings. Not only do you find these words but also words that are used to ask the other participants in the conversation to express themselves when it is left unclear. This ties into the next overarching observation you can glean from these tables which is the tendency to draw or direct focus to others in the conversation rather than the self which in turn further drives the conversation. I believe this stems from conversations in this digital space are often involving people who wish to and enjoy conversing. While this seems speculative, I believe it is a very reasonable assumption to make. Throughout your everyday life it is not an uncommon occurrence to find yourself in a

conversation you do not wish to be a part of. When this happens it is not socially acceptable, in most cases, to halt the conversation and move on to something else. However, this problem does not exist when conversing in an anonymous situation, there is no negative consequence to simply not having a conversation, conversations in this medium require the explicit interest and will of both parties. For this reason all parties involved are incentivized to be productive in the conversation in the majority of cases, thus resulting in this trend of engagement in the conversation and its constituents which can be observed throughout the collocate tables. This same phenomenon seems to additionally contribute to the last trend that can be seen. This is the trend of the tendency towards positivity which in part could be due to people who wish to converse with others in general wish to do so in a productive and positive way. However, another likely cause of this is the specific places I gathered the data from. All of the subreddits I collected data from have a rule enforced by the moderators of something along the lines of "Be civil, constructive, and kind". When someone is negative in a non-constructive way or just plain mean their comment will often be deleted by a moderator resulting in its omission from the data I collected. While conversing in an anonymous environment does strip individuals of ways of expressing themselves, it also strips away many social conventions in rules. The outcome of this results in a high priority of self expression and mutual understanding, in consequence this allows people to converse in a way that would normally be impossible. The format of typing out your response forces each participant to think through what they contribute, further contributing towards the avoidance of negativity and other non-constructive inputs. This enhanced freedom and ability of expression in an anonymous setting appears to increase the interest, respect, and understanding despite its dehumanizing properties. While this form of communication initially appears to be lacking in many ways, it results in conversations that trend towards positivity and a push for egalitarian representation of feeling and thought.

1 Collocate Tables

Collocates of “I”								
Rank	1st	%	2nd	%	3rd	%	4th	%
1	have	6.756	to	6.301	to	6.075	to	3.754
2	am	5.923	a	5.378	a	3.446	a	3.136
3	was	5.735	you	4.436	the	3.235	the	3.108
4	think	4.144	the	2.948	you	2.828	you	2.901
5	know	3.553	that	2.787	that	1.934	i	2.388
6	don’t	3.181	it	2.294	i	1.809	and	2.308
7	would	3.136	like	1.958	it	1.557	of	2.104
8	just	2.7	in	1.683	and	1.396	that	1.691
9	can	2.67	have	1.649	my	1.245	in	1.519
10	love	1.982	this	1.613	in	1.238	for	1.471
11	feel	1.818	i	1.544	of	1.08	my	1.302
12	want	1.81	your	1.517	for	1.007	it	1.3
13	hope	1.707	my	1.36	your	1.006	is	1.245
14	do	1.641	think	1.247	is	1.0	your	1.075
15	had	1.506	want	1.152	this	0.996	this	0.934
16	could	1.214	know	1.068	if	0.965	with	0.908
17	can’t	1.188	not	1.034	what	0.898	be	0.863
18	really	1.117	wondering	0.925	with	0.776	on	0.802
19	get	1.1	be	0.848	have	0.756	but	0.801
20	will	0.936	for	0.728	on	0.674	what	0.751
Total	124663		123989		122827		121088	

Collocates of “You”								
Rank	1st	%	2nd	%	3rd	%	4th	%
1	think	6.508	to	7.078	the	5.455	the	3.908
2	have	5.786	the	4.714	to	4.029	to	3.623
3	are	4.285	a	3.687	a	3.092	a	2.605
4	for	2.981	your	2.794	you	2.296	you	2.578
5	ever	2.662	that	2.333	your	1.805	and	2.181
6	do	2.108	about	2.217	in	1.533	of	2.119
7	feel	2.042	any	1.849	for	1.497	in	1.947
8	and	2.027	you	1.819	and	1.426	your	1.679
9	were	2.009	it	1.747	that	1.259	for	1.474
10	can	1.811	in	1.618	of	1.258	that	1.204
11	could	1.592	have	1.422	this	1.256	on	1.078
12	to	1.421	for	1.348	it	1.181	i	1.005
13	want	1.116	of	1.23	be	1.071	with	0.99
14	get	1.099	on	1.223	do	0.919	is	0.862
15	like	1.091	doing	0.941	with	0.856	it	0.86
16	would	1.049	this	0.929	is	0.855	be	0.846
17	had	1.044	be	0.9	on	0.829	or	0.808
18	see	1.029	and	0.881	i	0.764	do	0.781
19	know	1.025	with	0.845	about	0.663	have	0.752
20	believe	0.85	is	0.82	what	0.657	what	0.75
Total	213466		208755		202892		194967	

Collocates of “I’m”								
Rank	1st	%	2nd	%	3rd	%	4th	%
1	a	12.621	to	7.775	to	4.692	you	4.048
2	not	9.481	a	3.599	fan	3.954	and	3.772
3	sure	5.512	you	2.555	you	3.486	of	3.597
4	so	3.206	huge	2.464	the	2.872	the	3.34
5	just	2.534	for	2.416	for	2.558	i	2.876
6	going	2.474	the	2.361	and	2.367	to	2.688
7	glad	2.36	sure	2.258	i	2.324	a	2.419
8	curious	2.065	of	2.137	of	2.072	your	1.516
9	in	1.921	happy	2.027	a	1.826	in	1.504
10	really	1.717	and	2.009	in	1.746	this	1.397
11	currently	1.615	in	1.754	that	1.353	my	1.372
12	an	1.525	big	1.517	if	1.107	that	1.303
13	very	1.477	that	1.39	your	1.033	it	1.184
14	from	1.429	about	1.36	my	1.015	but	1.072
15	sorry	1.297	i	1.238	about	0.984	for	1.028
16	happy	1.249	this	1.038	but	0.972	what	0.99
17	pretty	1.081	curious	1.008	late	0.941	have	0.859
18	trying	1.045	if	0.983	this	0.904	is	0.821
19	wondering	0.985	it	0.959	what	0.855	with	0.802
20	still	0.961	on	0.959	with	0.762	fan	0.727
Total	16655		16475		16263		15958	

Collocates of “You’re”								
Rank	1st	%	2nd	%	3rd	%	4th	%
1	a	10.264	to	7.163	the	3.889	you	3.283
2	doing	7.203	a	4.696	to	3.641	the	3.243
3	not	4.944	the	3.068	and	2.845	and	3.079
4	the	3.462	and	2.479	of	2.245	a	2.629
5	an	2.757	of	2.417	a	2.114	to	2.616
6	going	2.684	in	2.029	you	1.866	of	2.52
7	still	2.648	for	1.903	i	1.749	i	2.071
8	in	2.417	with	1.54	for	1.475	in	1.253
9	awesome	1.956	on	1.453	in	1.174	that	1.226
10	my	1.445	about	1.302	my	1.161	your	1.172
11	one	1.385	an	1.139	person	1.109	for	1.117
12	on	1.093	great	1.027	that	1.109	it	1.035
13	just	1.081	inspiration	1.014	inspiration	1.096	but	1.008
14	so	1.045	this	1.014	your	0.953	what	0.981
15	trying	0.899	good	1.002	it	0.913	my	0.831
16	amazing	0.826	that	0.927	do	0.887	do	0.817
17	right	0.826	is	0.927	but	0.874	have	0.79
18	talking	0.777	i	0.927	is	0.848	with	0.763
19	happy	0.692	hero	0.927	man	0.822	is	0.722
20	looking	0.692	your	0.826	with	0.783	so	0.708
Total	8233		7986		7663		7340	

2 Source Code

```
import pandas as pd
import numpy as np
import itertools as it
from collections import Counter
import copy

import praw
r = praw.Reddit( user_agent='Comment Extraction'
                ,client_id='', client_secret="")
r.read_only = True
sub = r.subreddit("iama") #cc,change my view, happy, legal advice
fil = open("comment1.txt", 'a', encoding='utf-8')
key_words = ["i ", "you ", "i'm ", "you're", "im ", "youre "]
submissions = 200
for submis in sub.top('all'):
    comforest = submis.comments
    comforest.replace_more(limit=None)
    for com in comforest:
        bod = com.body.lower()
        for key in key_words:
            if key in bod:
                fil.write(bod + '\n::::::::::::\n')
                break
    submissions -= 1
    if submissions == 0:
        break
fil.close()

the_dat = []
with open("comment1.txt",'r', encoding='utf-8') as f:
    ii = 0
    for key,group in it.groupby(f,lambda line: line.startswith('::::::::::::')):
        if not key:
            the_dat.append([])
```

```
            for line in group:
                the_dat[ii].append(line)
            ii += 1
f.close()

for ii in range(len(the_dat)):
    the_dat[ii] = ' '.join(the_dat[ii]).split()

for ii in range(len(the_dat)):
    for yy in range(len(the_dat[ii])):
        the_dat[ii][yy] = the_dat[ii][yy].strip('?').strip(',').strip('!').strip('.')

total_words = 0
for comment in the_dat:
    total_words += len(comment)

st_words = ['i', 'you', "i'm", "you're"]

the_corp = {wor : [Counter() for ii in range(5)] for wor in st_words}

for jj in range(len(the_dat)):
    for ii in range(len(the_dat[jj])):
        word = the_dat[jj][ii]
        if word in st_words:
            if word == "im":
                word = "i'm"
            elif word == "youre":
                word = "you're"
            coll = the_dat[jj][ii+1:ii+6]
            for kk in range(len(coll)):
                the_corp[word][kk][coll[kk]] += 1

for word in the_corp:
    for ii in range(5):
        tot = 0
        ent = the_corp[word][ii]
```

```
        for key in ent:
            tot += ent[key]
        ent['TOTAL'] = tot

freq_corp = copy.deepcopy(the_corp)
for k in freq_corp:
    for ii in range(5):
        ent = freq_corp[k][ii]
        tot = ent['TOTAL']
        for key in ent:
            ent[key] /= tot
            ent[key] = round(ent[key] * 100,3)

for ii in range(5):
    if ii != 0:
        print('&', end='')
    else:
        print("Total" + '&', end='')
    current = the_corp[entry][ii]['TOTAL']
    print("\multicolumn{2}{|c||}{" + str(current) +"}", end='' )
print("\\\\")
print("\\hline")

for jj in range(1,21):
    for ii in  range(4):
        if ii != 0:
            print('&', end='')
        else:
            print(str(jj) + '&', end='')
        current = the_corp[entry][ii].most_common(21)[jj]
        curfreq = freq_corp[entry][ii].most_common(21)[jj]
        print(current[0] +'&' + str(curfreq[1]),end='')
    print("\\\\")
    print("\\hline")
```

3 References

https://praw.readthedocs.io/en/latest/index.html
https://www.reddit.com/r/changemyview/
https://www.reddit.com/r/IAmA/
https://www.reddit.com/r/happy/
https://www.reddit.com/r/legaladvice/
https://www.reddit.com/r/CasualConversation/