Human language technologies // COMP40020

Thomas Igoe // 17372013

Assignment 1

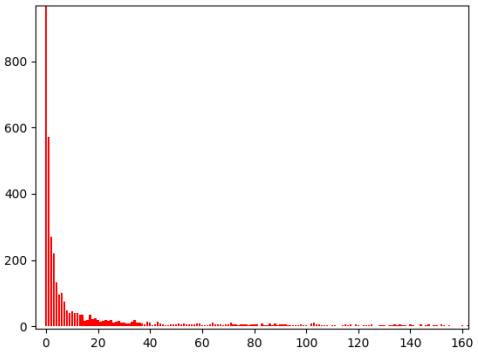
Research Question: Can we spot trends among high scoring comments vs low scoring comments using tools like lexical diversity or common n-grams?

Like any social media, reddit has metrics to determine the success or engagement of a post or comments. For reddit, this is primarily the points system. Users can both upvote and downvote posts and comments, and this is used to determine popular posts to show to other users. Naturally people would like their posts to succeed, therefore I will investigate if there are any common themes between a comment's score and it's language through use of lexical diversity analysis and the common n-grams which occur in popular posts.

To use as a baseline for this, I also looked at the overall lexical diversity of different subreddits, so I was better able to compare the lexical diversity among different subreddits, as well as within the subreddit. This can give me a better idea of the average ranges for different subreddits.

Method

Data collection is a modified version of the supplied method. The code still opens a connection to reddit using the PRAW library, however the code has been changed to not only write the data to a main text file (a corpus of all comments) but to also write subset corpora each containing only comments which scored over a certain threshold. These subset corpora can then be analysed to give information about comments at those thresholds. The comments were divided into progressively larger ranges (for example, comments with scores 1-10, then 10-100, 100-1000 etc as well as a bracket for all negatively scoring comments). The aim of this was to somewhat normalize the distribution of comments. As can be seen in the graph[1], the quantity of comments at different linear thresholds follows a hyperbolic curve. Scaling the data intervals logarithmically, this can be somewhat addressed. (However more work could be done to address this problem by finding a more appropriate scale instead of assuming 10^x).



I also used only top level comments (ie: replies to the original post) as replies to comments tend to score less than top level comments. Therefore I excluded them to remove this effect on the dataset.

[1] This image graphs the number of comments with a certain number of points.

I also do not include comments that have been removed ie: where the body is “[removed]” or “[deleted]”, as these contain no useful data for the purposes of this experiment.

Because the data has already been subdivided into files containing comments of each threshold, calculating lexical diversity, and common n-grams for each becomes relatively easy.

When calculating the most common n-grams, I pass each text through some preprocessing steps. These steps include converting all tokens to lowercase, and removing all punctuation. In this case however, I deliberately chose *not* to use stemming or to remove stop words. If I did so, many common phrases could have been missed (eg “shoot your shot”, or “it is what it is”). This did unfortunately mean many of the most frequent n-grams were not particularly interesting (eg “one of the best”). However this can be manually overcome by searching the n-grams by hand. (Although an improvement that could be made to this code would be to remove n-grams that consisted of *only* stop words). I apply this calculation to each of the different score corpora using NLTK.
The methods I have written simply calculate the lexical diversity for each text file. (i.e. number of unique tokens divided by number of overall tokens in a corpora). Like the previous calculation, this calculation is done for all point thresholds, as well as the overall corpus of the subreddit in question. However, the lexical diversity calculation does use additional preprocessing steps which I intentionally neglected in the common n-gram part, namely stemming and removing stop words. Because the focus is on individual words, and the overall diversity of the corpus, we do not need to worry about words that don’t add much meaning to the text, or words that mean the same, or similar things. Finally one additional step I took was to ensure that a fair comparison was being made across the different sub-corpora, which was to only take the same number of tokens from each corpora. This is because it can be much easier for shorter texts to be lexically diverse (the most extreme example being a 1 word text with a lexical diversity of 100%). The reason for this is because despite my attempt to normalize the different score tiers with logarithmically scaled brackets, the 1-10 bracket still ended up being the largest one. So to combat this, I took the size of the smallest corpora size (the +10,000 bracket) and used only that many tokens from each of the other corpora, ultimately ensuring that each bracket was using an equal number of tokens.

Results:
Out of the main subreddits I investigated “todayilearned", "funny", "Science", "Gaming", and "leagueoflegends" there was a decent amount of variance between their lexical diversity values. Out of those five, r/leagueoflegends (a popular MOBA video game) actually had the highest lexical diversity of those five. Meanwhile r/todayilearned actually had the least.

While there was significant variance between different subreddits, different point tier corpora within each subreddit was less varied. However, in most subreddits a small, yet distinct trend was occurring, where lower scoring comments would consistently have slightly lower diversity scores. Generally it was not a large difference, (usually 1-2 percent but can be as large as 5-6). However it consistently occurred across multiple subreddits.

r/todayilearned

Lexical diversity of all comments:	0.1810831223066329
Lexical diversity of posts w score >10,000:	0.20445155271224844
Lexical diversity of posts w score >1,000:	0.1988982184903816
Lexical diversity of posts w score >100:	0.1919232307077169
Lexical diversity of posts w score >10:	0.1938335776800391
Lexical diversity of posts w score >1:	0.18050557554755875

r/funny

Lexical diversity of all comments:	0.2367074210975263
Lexical diversity of posts w score >10,000:	0.27921524026158656
Lexical diversity of posts w score >1,000:	0.27779357406880867
Lexical diversity of posts w score >100:	0.26201307932897355
Lexical diversity of posts w score >10:	0.26158657947114017
Lexical diversity of posts w score >1:	0.2477964174011942

r/science

Lexical diversity of all comments:	0.2362179487179487
Lexical diversity of posts w score >10,000:	0.2721153846153846
Lexical diversity of posts w score >1,000:	0.2955128205128205
Lexical diversity of posts w score >100:	0.29599358974358975
Lexical diversity of posts w score >10:	0.28669871794871793
Lexical diversity of posts w score >1:	0.2693910256410256

r/gaming

Lexical diversity of all comments:	0.2067319169252805
Lexical diversity of posts w score >10,000:	0.2554308904273096
Lexical diversity of posts w score >1,000:	0.2557889711148245
Lexical diversity of posts w score >100:	0.24600143232275007
Lexical diversity of posts w score >10:	0.2431367868226307
Lexical diversity of posts w score >1:	0.22511339221771307

r/leagueoflegends

Lexical diversity of all comments:	0.36019736842105265
Lexical diversity of posts w score >10,000:	0.40789473684210525
Lexical diversity of posts w score >1,000:	0.4194078947368421
Lexical diversity of posts w score >100:	0.421875
Lexical diversity of posts w score >10:	0.42680921052631576
Lexical diversity of posts w score >1:	0.3560855263157895

As for the common n-gram experiment, certain phrases would more commonly appear on higher scoring comments, rather than lower scoring ones. For example the 4-gram “thanks for the gold” was common among multiple subreddits, many popular memes, and jokes can also often appear such as “rule of fight club” or “brick in the wall”.

Discussion

Overall this experiment was interesting, but did not produce results that were as clear as I would have liked. Subdividing the dataset into five subsets meant that even when processing 300 different posts, some corpora were relatively small. Perhaps with more computation time, more definitive results could have been achieved.

The inter-subreddit lexical diversity experiment yielded interesting results. Subreddits which approach a variety of different topics, such as r/todayilearned and r/science actually have lower lexical diversity than a subreddit focused on a specific topic such as r/leagueoflegends. After some investigation, I would attribute this to the wide variety of names used in the r/leagueoflegends subreddit. There are over 140 different characters (or "legends"), as well as a wide variety of names, streamers, teams and strategies being discussed. These can all be attributed to the subreddit's higher lex diversity Meanwhile, r/todayilearned covers a wide variety of topics and interests, however without a large number of enthusiasts, most discussion occurs at a surface level, using language the average person would know about a topic. This means that many people will be using the same language to talk about a topic.

The lexical diversity experiment shows that the diversity of lower scoring comments is consistently lower than higher scoring ones. I would attribute this to the large number of comments that are short, and don't add much to the discussion (eg "nice", "this is amazing", "haha"). Comments that are short, and almost entirely positive do not tend to generate much engagement, and are swiftly ignored for longer comments expressing a more complex opinion. There is an enormous amount of comments like this, with one or two word comment bodies, heavily influencing this brackets' corpus.

The common n-gram experiment also shows some useful data. Common n-grams at high levels can be jokes or references like "rule of fight club" (a reference to the eponymous movie). However a common n-gram across multiple subreddits was "thanks for the gold". This is a tradition on reddit as to thank people after receiving "reddit gold" , a paid award that users can give to one another. This however does not imply that the phrase "thanks for the gold" results in high score (what I was looking for) but instead shows that high scoring comments often receive gold awards, and therefore often thank people for awards.

I also looked at negatively scoring comments for patterns, however many of them were removed or deleted, which usually happens when platform moderators remove something that breaks reddit's rules, usually as a result of harassment or bullying. Because many of the rules are about harassment and respectful conduct, it makes sense that vitriolic comments will score poorly before being removed. Many of the remaining comments would have negative sentiment as well as a proportionally higher number of curse words and slurs. This makes sense, as people tend to downvote users who are being openly hostile towards other users either through racisim or sexisim, or other harassing speech such as threats of violence or bullying.

From these points I can draw some conclusions about popular comments on reddit. Many popular comments will simply make a reference to a popular meme or piece of media (such

as the “brick in the wall example”). However this is not a guarantee of success as there are also many examples of similar jokes in lower scoring brackets.
If a user is going to express an opinion on reddit, and wishes to get a good score they cannot be too positive nor too negative. If a comment is purely positive, it will rarely get enough points to leave the 1 to 10 bracket, as people don’t tend to engage with those comments. Purely negative comments are also received poorly, and often end up with negative scores. Commenters must strike a good balance between the two.

(Word count 1700)

Note Folder Python project was just the environment for using pycharm for much of the coding and the .py file in there should be the same code as in the jupyter notebook.