**Mid-Semester Progress Report**

DSA5900 – Spring 2021

Pendleton Tompkins

03/22/2021

## Introduction

Social media has rapidly emerged as one of, if not the, most prominent mediums through which companies may increase their brand awareness. Twitter, a social media platform garnering over 330 million monthly active users, serves as an ideal environment for companies to engage with consumers (Statista). However, unlike traditional advertising mediums (TV, billboards, etc.), on Twitter, the size of the audience which ultimately sees a certain tweet (message) crafted by a company is related with how well the contents of the tweet are received by other users on Twitter. This is because tweets, initially, only appear in the home timeline, or “feed,” of users following the account which posted the tweet. However, Twitter users in the initial audience can “retweet” the post, causing it to appear in the feed of any Twitter users following them. This implies that any given tweet may potentially propagate through a massive network of people, much larger than the initial audience.

However, for this to occur, people must enjoy the contents of the tweet enough to retweet it themselves. This differs massively from traditional advertising mediums. For example, regardless of the time or effort that a company puts into creating a TV commercial, if they book a slot during the super bowl, they're guaranteed to reach a massive audience. Furthermore, the size of the audience has no relationship with the contents of the commercial. Regardless of whether the commercial is of exceptional or horrendous quality, everyone watching the super bowl would also watch the commercial. This is why social media differs so greatly from tradition, larger audiences cannot be bought, they must be earned.

Throughout the course of this semester, Dr. Yoon and I have been working towards our ultimate goal of creating computer programs able to predict the number of likes and retweets a message posted on Twitter will generate. Specifically, we’re interested in tailoring these programs such that they may be utilized by brand company Twitter accounts to predict likes and retweets prior to tweet publication, given the contents of the tweet and information regarding the company’s Twitter account.

This would allow companies to determine, beforehand, whether a certain message is worth posting on Twitter. Furthermore, a tool to predict likes and retweets is synonymous with having the means to predict the number of people who will have a positive response to your tweet. Making the tool, essentially, a free, quick focus group for advertisement and branding campaigns performed through Twitter. This should, in turn, increase the speed at which a company’s Twitter account grows, as posts reaching larger audiences are likely to yield more new followers than posts reaching smaller audiences are. Finally, the programs could be used to improve customer service responses and feedback. Responding to questions, comments, and concerns posted online about the company is an important component of every company’s official social media accounts. This could be used to help determine which formats of responses satisfy consumer questions or comments, and which formats leave consumers unsatisfied.

Dr. Yoon hopes to gain insights into human behavior in the online realm throughout the implementation of this project, and provide answers to the question of “what are the factors that separate a popular tweet from an underperforming tweet?” This project is directly related to those goals because identifying factors which separate tweet performance should, at least in theory, be the factors which are most useful in predicting likes and retweets.

## Objectives

The overarching project objective is to determine the processes and methods which can, most accurately and reliably, predict the final number of likes and retweets prior to a tweet’s publication (given tweet contents and information regarding the company’s Twitter account). To do this, we must first identify the specific factors which are most closely related with a tweet’s popularity. Our aim is to identify, or create, meaningful independent variables drawn from text analysis. These will be used in tandem with variables drawn from other sources (such as the number of people following the Twitter account) to best predict likes and retweets. At the conclusion of the project, we wish to publish a research paper detailing all our most important findings.

My two primary learning objectives are to: 1) gain experience with the complete process of addressing a data science question, and 2) become proficient in working with textual data. Of course, advancing my knowledge in forecasting and machine learning are learning objectives as well. However, personally, prior to beginning this project I felt that natural language processing (NLP) was perhaps a weak point in my data science repertoire. I no longer feel as if that’s the case, I’m already considerably more comfortable working with textual data than at the beginning of the semester.

## Data

### Ingestion

Fortunately, Dr. Yoon had already collected data prior to beginning the project and supplied it to me on the first week of the semester. Dr. Yoon was able to gather data directly from the source by utilizing ‘Orange3’ data mining software to scrape the Twitter API. Considering the data is of reasonable size (there are only a total of 32,247 tweets), and that we did not intend to continue gathering data throughout the course of the semester, data has simply been stored in Excel format on our PCs.

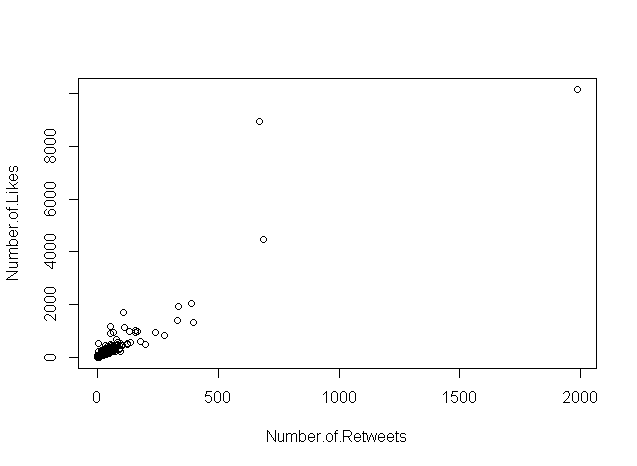
### Exploration

Data originates from ten separate brand company Twitter accounts: Amazon, BMW, Coca Cola, Disney, Google, McDonald’s, Microsoft, Mercedes Benz, Samsung, and Toyota. In total, there are 32,247 rows of data and 18 variables initially gathered. The variable names and definitions are given below:

|  |  |
| --- | --- |
| **Variable** | **Explanation** |
| Author | The Twitter ‘handle’ (username) of the company which composed the tweet |
| Content | The textual content of the tweet (includes emojis, links, etc.) |
| Date | The time-date stamp of when the tweet was posted |
| Language | The language in which the tweet was composed |
| Location | The approximate location of the author at the time when the tweet was posted |
| Number of Likes | The number of likes the tweet received |
| Number of Retweets | The number of retweets the tweet accumulated |
| In Reply To | If the tweet was in reply to another user, the user’s Twitter handle is available. Else, the field is left blank |
| Author Name | The name of the company which composed the tweet |
| Author Description | The ‘biography’ of the posting account |
| Author Statuses Count | The total number of tweets over the lifetime of the composing account, at the time of data collection |
| Author Favorites Count | The total number of tweets the composing account has ‘favorited’ (liked) over its lifetime, at the time of data collection |
| Author Friends Count | The number of accounts the author was following at the time of data collection |
| Author Followers Count | The number of accounts following the author at the time of data collection |
| Author Listed Count | The number of Twitter ‘lists’ the composing account has been placed in, at the time of data collection |
| Author Verified | Binary variable indicating whether the author is a verified Twitter account or not |
| Longitude | The longitude of the device from which the tweet was posted, at the time of posting |
| Latitude | The latitude of the device from which the tweet was posted, at the time of posting |

Initial exploration of the data has already yielded many insights. For example, there appear to be two distinct tweet categories for brand company Twitters. The first category may be referred to as “official” tweets. These are standalone tweets posted by the company’s Twitter account, they average a high number of likes and retweets. The second category of tweets may be referred to as “in reply to” tweets or may be thought of as “customer service” tweets. These occur when another Twitter user posts a question, comment, or concern related to the company, and the company Twitter account replies to address the issue. Although this is a very important component of a brand company’s Twitter account, these types of tweets average very few likes and retweets. Furthermore, some companies seem to engage in a healthy mixture of both tweet categories, while other companies seem to (almost) only post tweets belonging to one category or the other.

At the very beginning of the project, one component I was interested in examining was the relationship between likes and retweets. For all companies, I created scatter plots of retweets vs. likes, as in the following plot for Amazon data:



To save on time and space, I’ll simply post Amazon results, rather than post results for all 10 companies or results for a standardized dataset consisting of all companies. All of these plots seemed to have displayed a linear relationship between likes and retweets. However, that was truer for some companies than for others, as some seem to have only displayed a weak linear relationship. Furthermore, these plots made it simple to identify the presence of outliers in the data. As you can see above, Amazon’s best performing tweet garnered approximately 2,000 retweets, while Amazon’s second-best performing tweet only accumulated 700 or so retweets. A vast, vast majority of Amazon’s tweets garnered under 500 retweets.

Additionally, I created the following descriptive tables for each dataset. Keep in mind that I’m still only displaying results for Amazon data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Description** | **Number Observations** | **Average Likes** | **Average Retweets** | **Likes/RT** |
| All Amazon tweets | 3174 | 30.65 | 5.13 | 5.97 |
| Amazon tweets containing link | 1063 | 77.61 | 13.03 | 5.96 |
| Amazon tweets not containing links | 2111 | 7.01 | 1.16 | 6.04 |

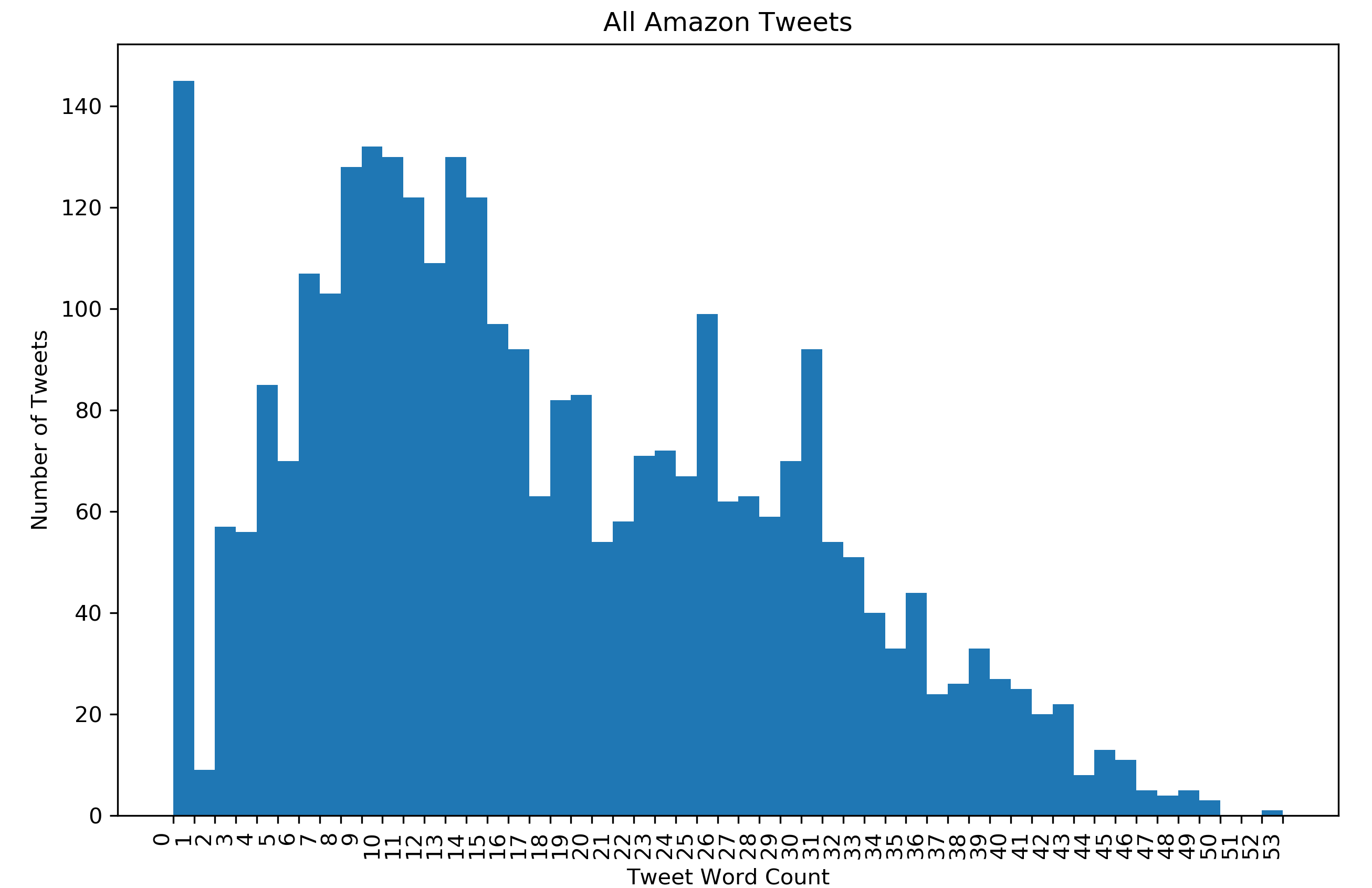
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Description** | **Number Observations** | **Average Likes** | **Average Retweets** | **Likes/RT** |
| All Amazon official tweets | 247 | 353.12 | 59.27 | 5.96 |
| All Amazon IRT tweets | 2927 | 3.44 | 0.57 | 6.04 |
| Official tweets with link | 221 | 359.76 | 60.74 | 5.92 |
| Official tweets without link | 26 | 296.65 | 46.77 | 6.34 |
| IRT tweets with link | 842 | 3.56 | 0.51 | 6.98 |
| IRT tweets without link | 2085 | 3.39 | 0.59 | 5.75 |

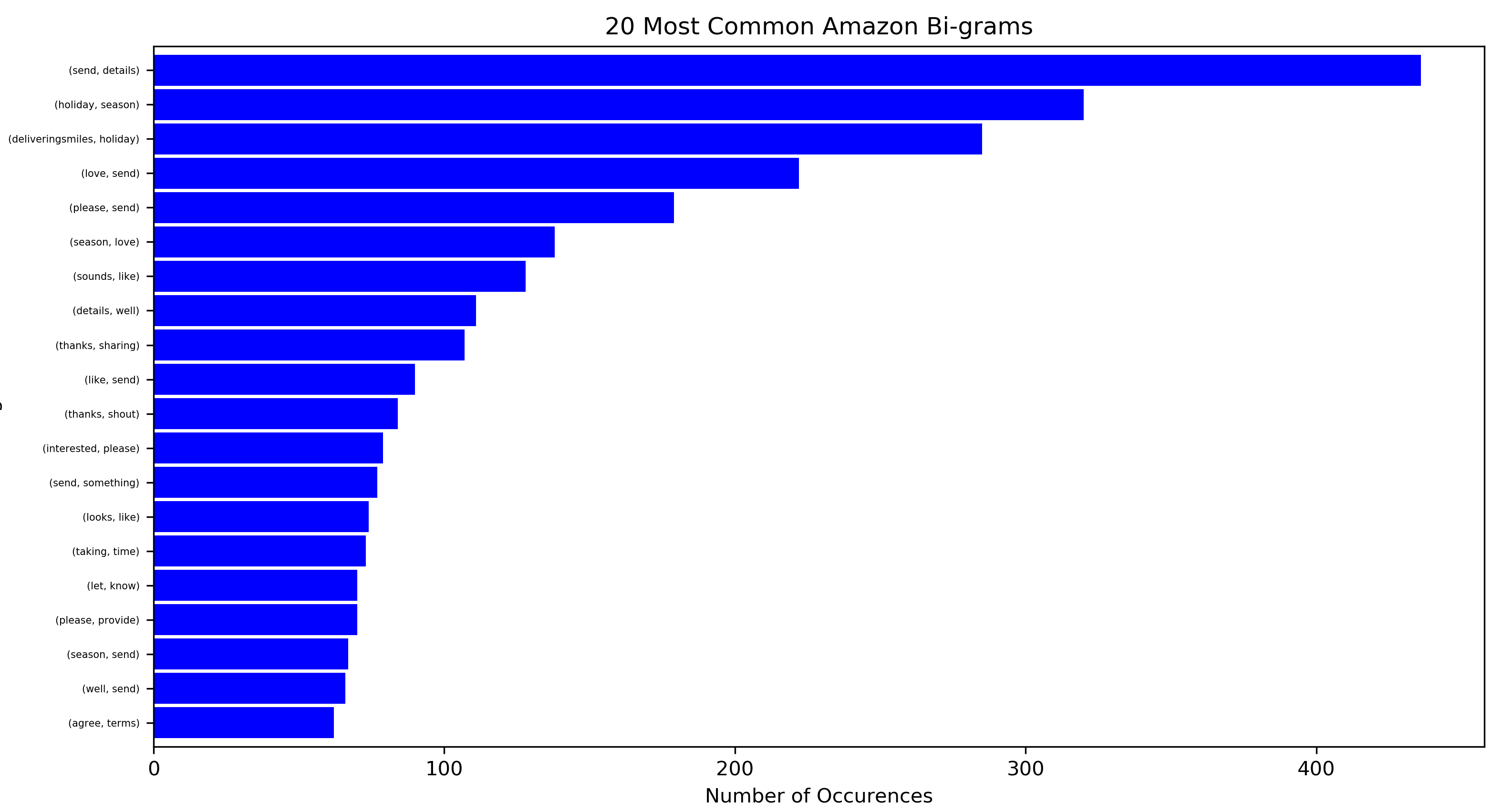
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Description** | **Number Observations** | **Average Likes** | **Average Retweets** | **Likes/RT** |
| Tweets with 0 # | 2465 | 27.88 | 4.58 | 6.09 |
| Tweets with 1 # | 697 | 37.73 | 6.61 | 5.71 |
| Tweets with 2 # | 11 | 191.82 | 36.91 | 5.20 |
| Tweets with 3 # | 1 | 148 | 13 | 11.38 |
| Official tweets w/ 0 # | 138 | 443.95 | 71.36 | 6.22 |
| Official tweets w/ 1 # | 98 | 241.84 | 44.62 | 5.42 |
| Official tweets w/ 2 # | 10 | 210.8 | 40.6 | 5.19 |
| Official tweets w/ 3 # | 1 | 148 | 13 | 11.38 |
| IRT tweets with 0 # | 2327 | 3.21 | 0.62 | 5.18 |
| IRT tweets with 1 # | 599 | 4.34 | 0.39 | 11.13 |
| IRT tweets with 2 # | 1 | 2 | 0 | NA |
| IRT tweets with 3 # | 0 | NA | NA | NA |

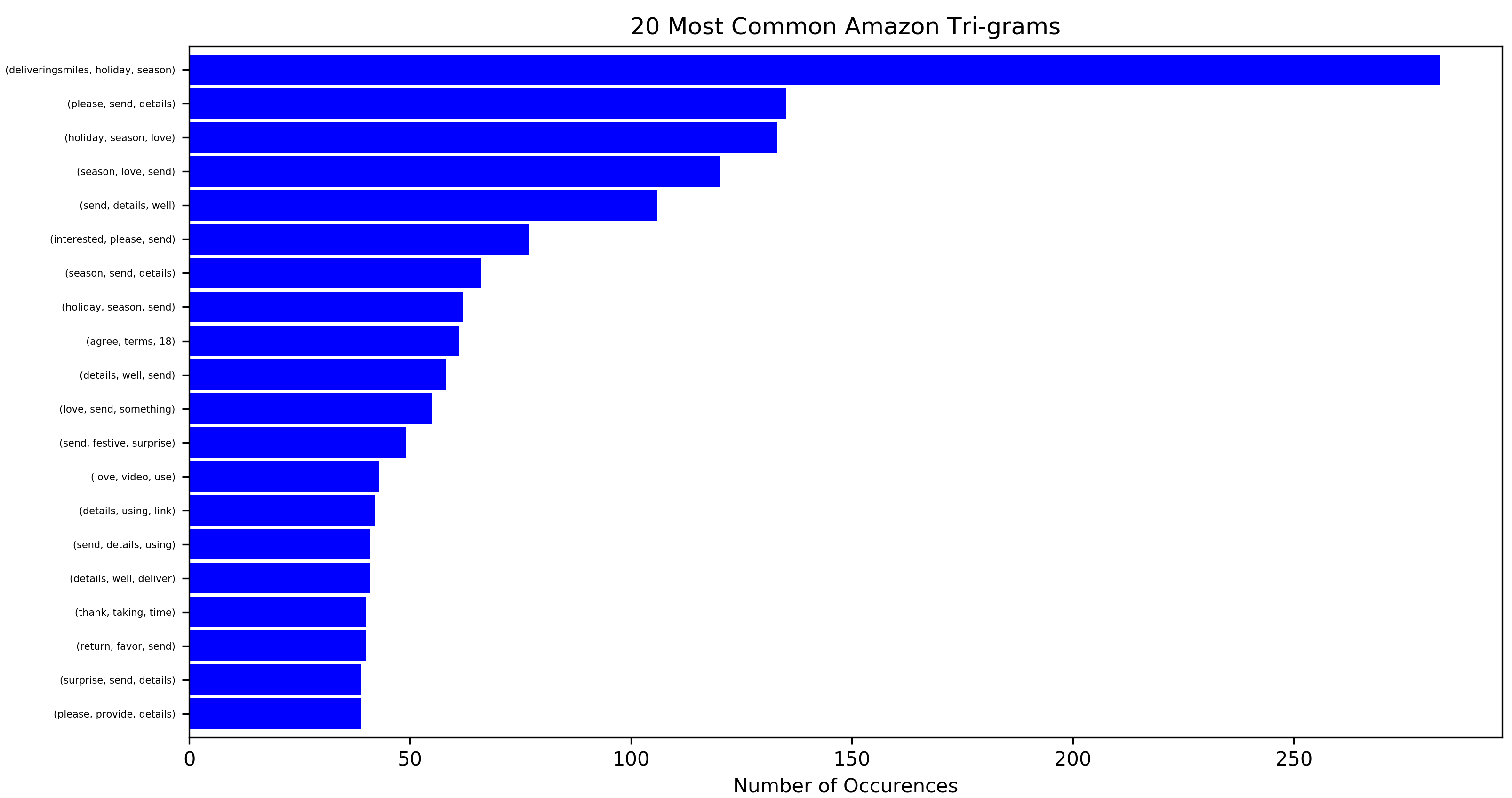
Admittedly, I should have also included standard deviations when recording these tables. However, at this stage it was not yet my intention to prove or disprove any theories and including those statistics in future tables should be straightforward. After having examined some basic statistics and getting a general feel of the data, I wanted to examine whether there seems to be a relationship between tweet length and performance. To do so, I generated the following figures and tables, for all companies:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tweet Category** | **No. Tweets** | **Avg. Word Count** | **Total Words** | **Unique Words** |
| All Tweets: | 3139 | 17.46 | 54797 | 4161 |
| Official Tweets: | 247 | 22.96 | 5671 | 1568 |
| Top 25% Official: | 62 | 23.68 | 1468 | 621 |
| Median 50% Official: | 123 | 22.90 | 2817 | 974 |
| Bottom 25% Official: | 62 | 22.13 | 1372 | 614 |
| IRT Tweets: | 2892 | 16.99 | 49126 | 3388 |
| Top 25% IRT: | 723 | 13.51 | 9767 | 1538 |
| Median 50% IRT: | 1446 | 16.98 | 24556 | 2292 |
| Bottom 25% IRT: | 723 | 20.50 | 14822 | 1535 |

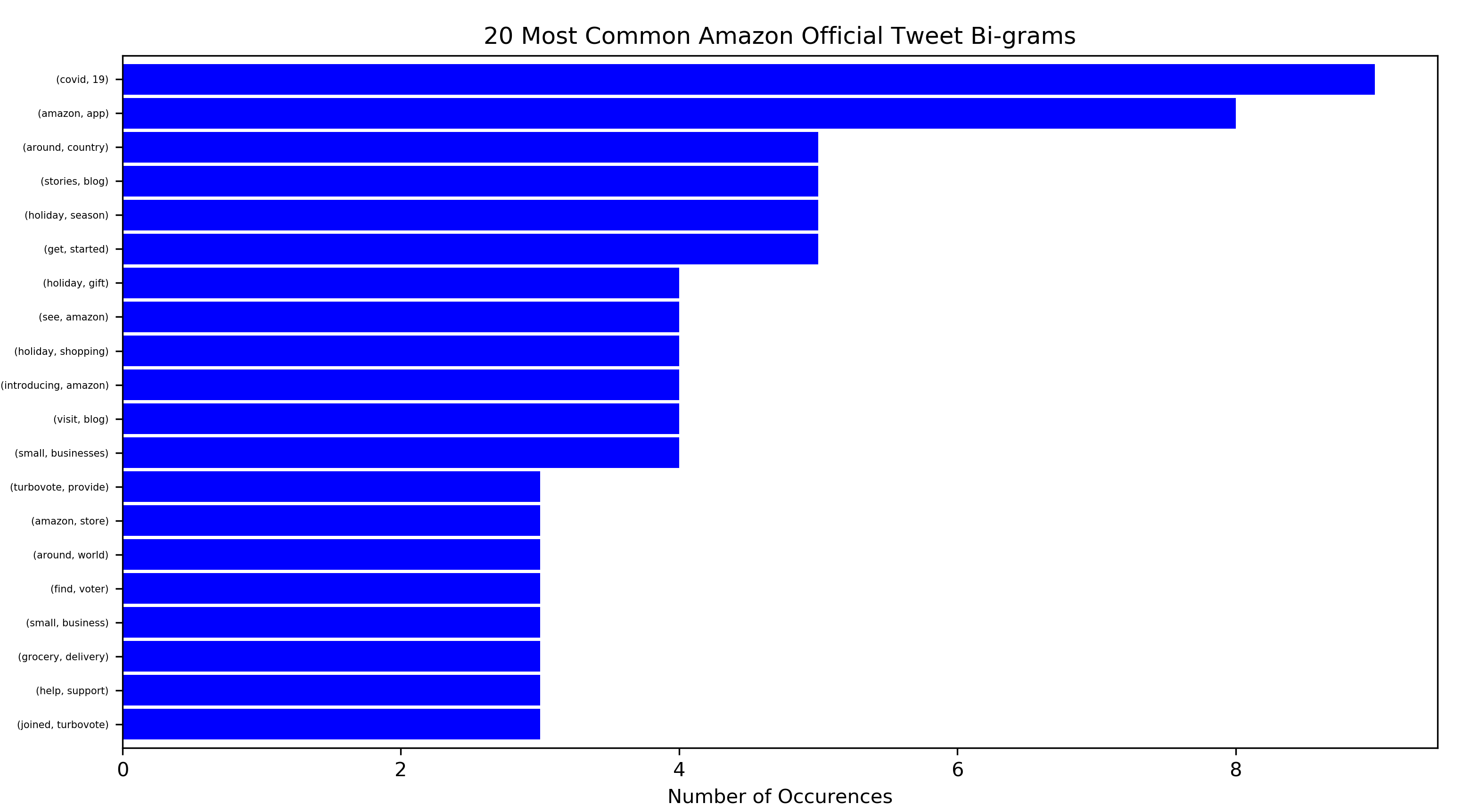
The ‘performance categories’ (top 25% official, etc.) are based on the number of likes the tweet generated. Thus, the ‘Bottom 25% Official’ category consists of Amazon’s least-liked official tweets.

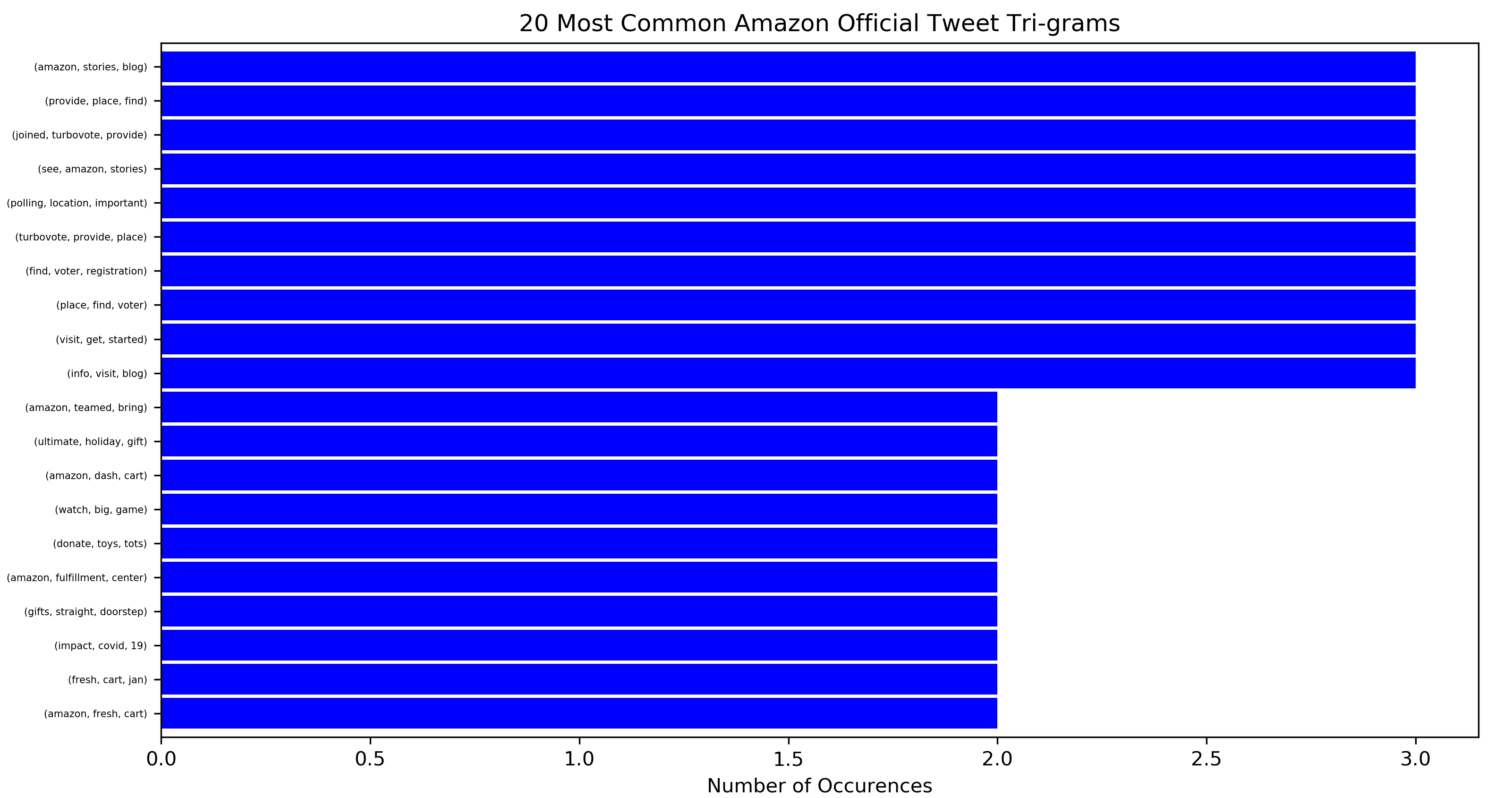


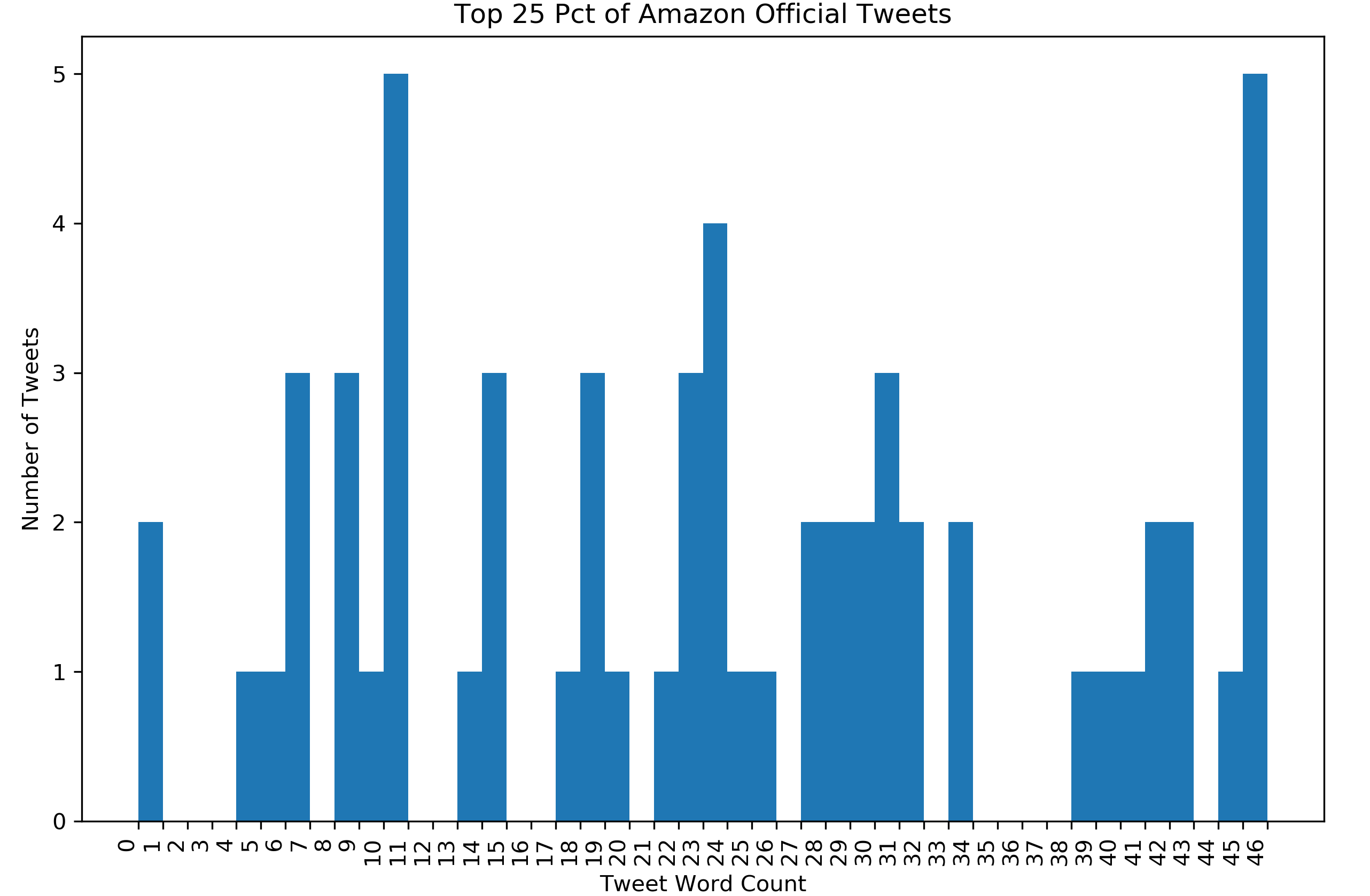


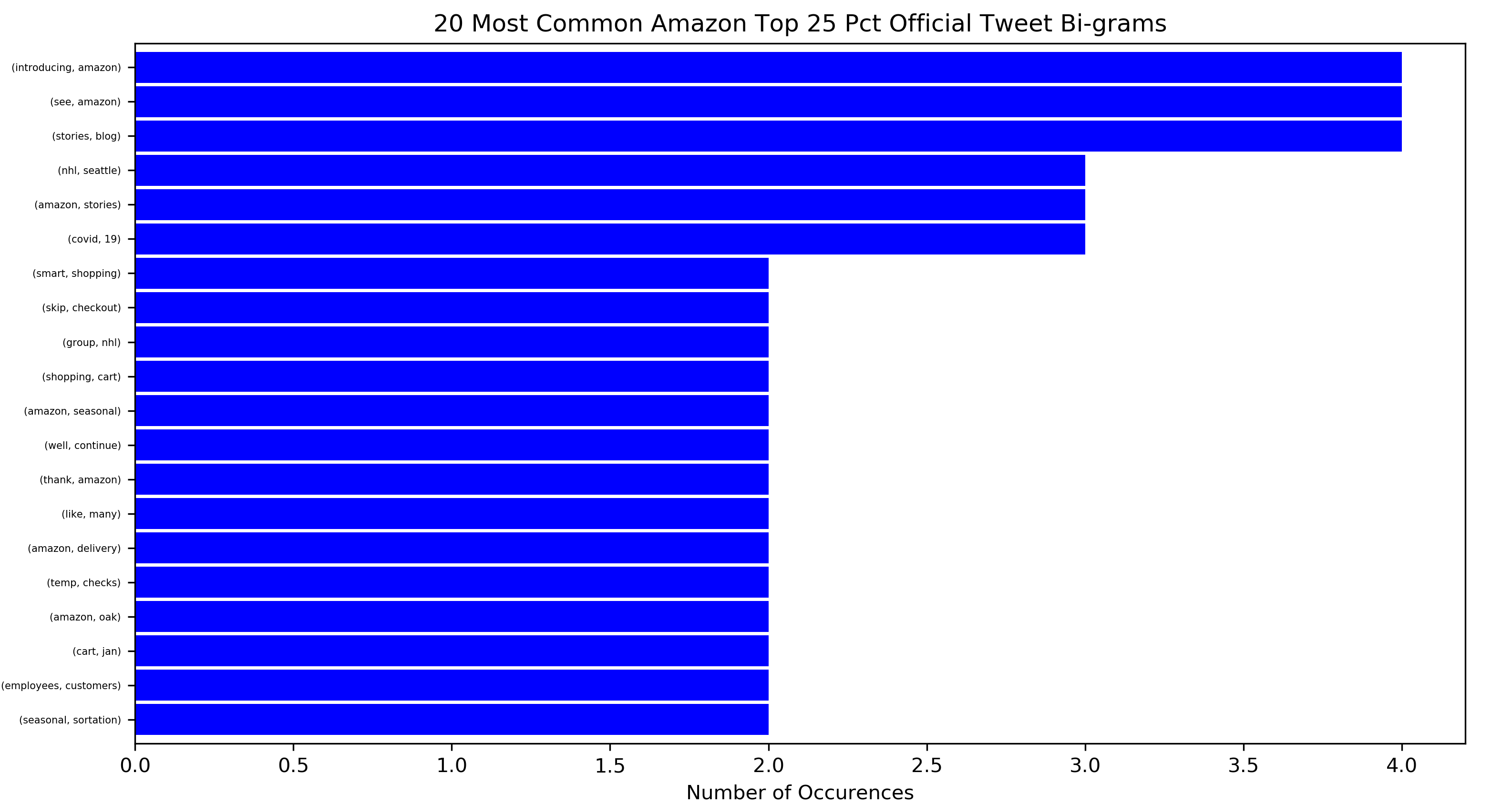


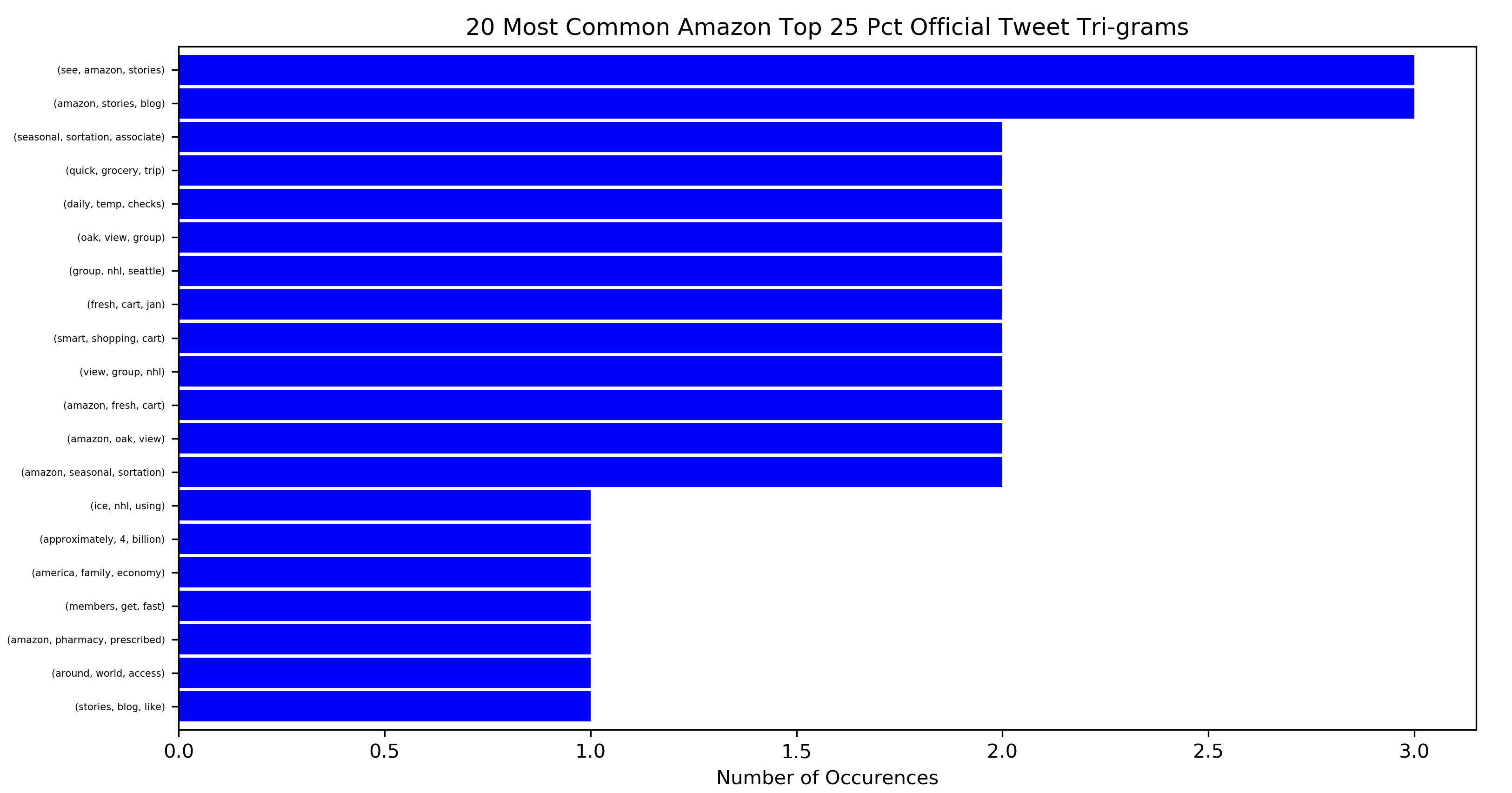


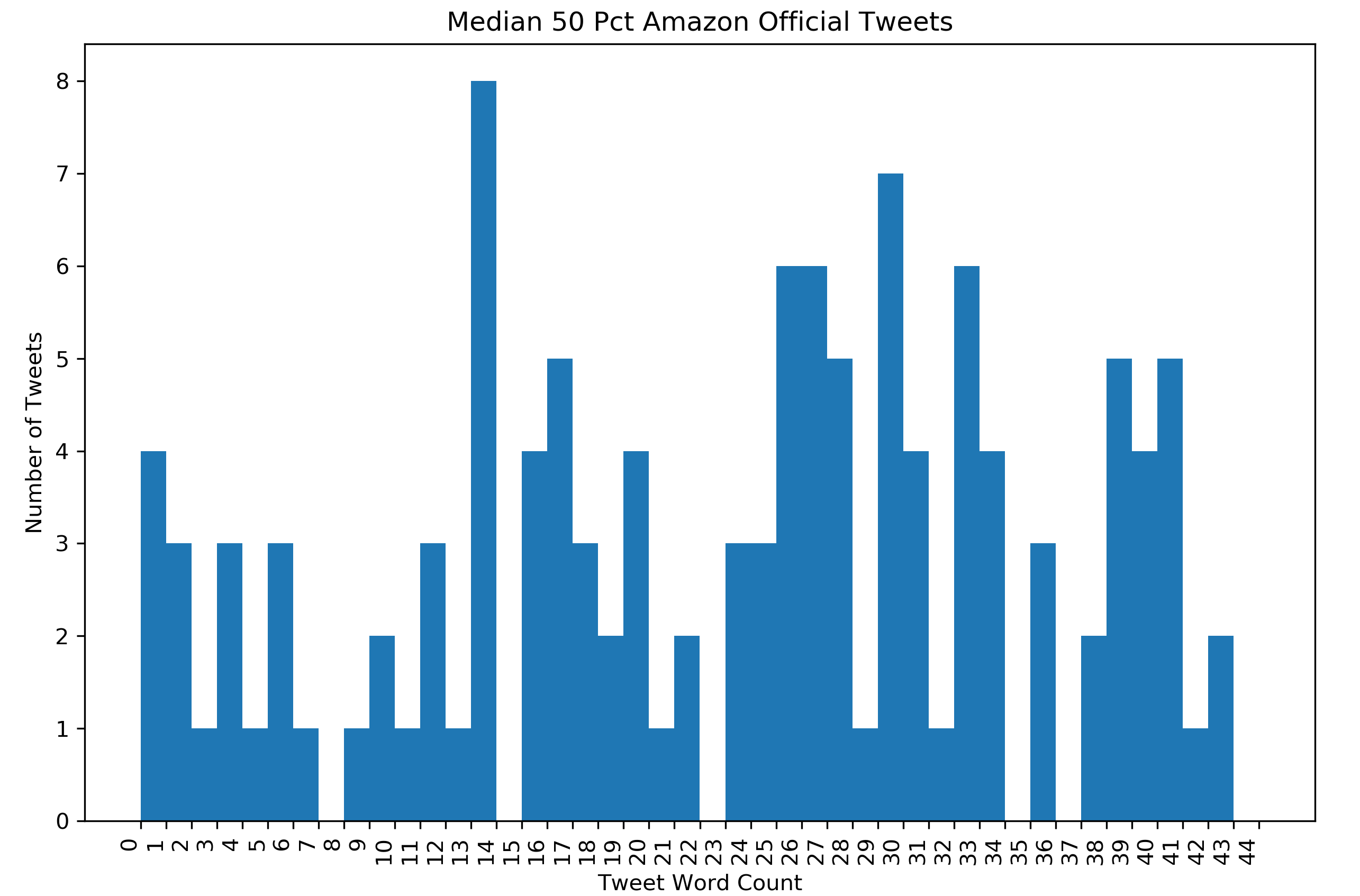


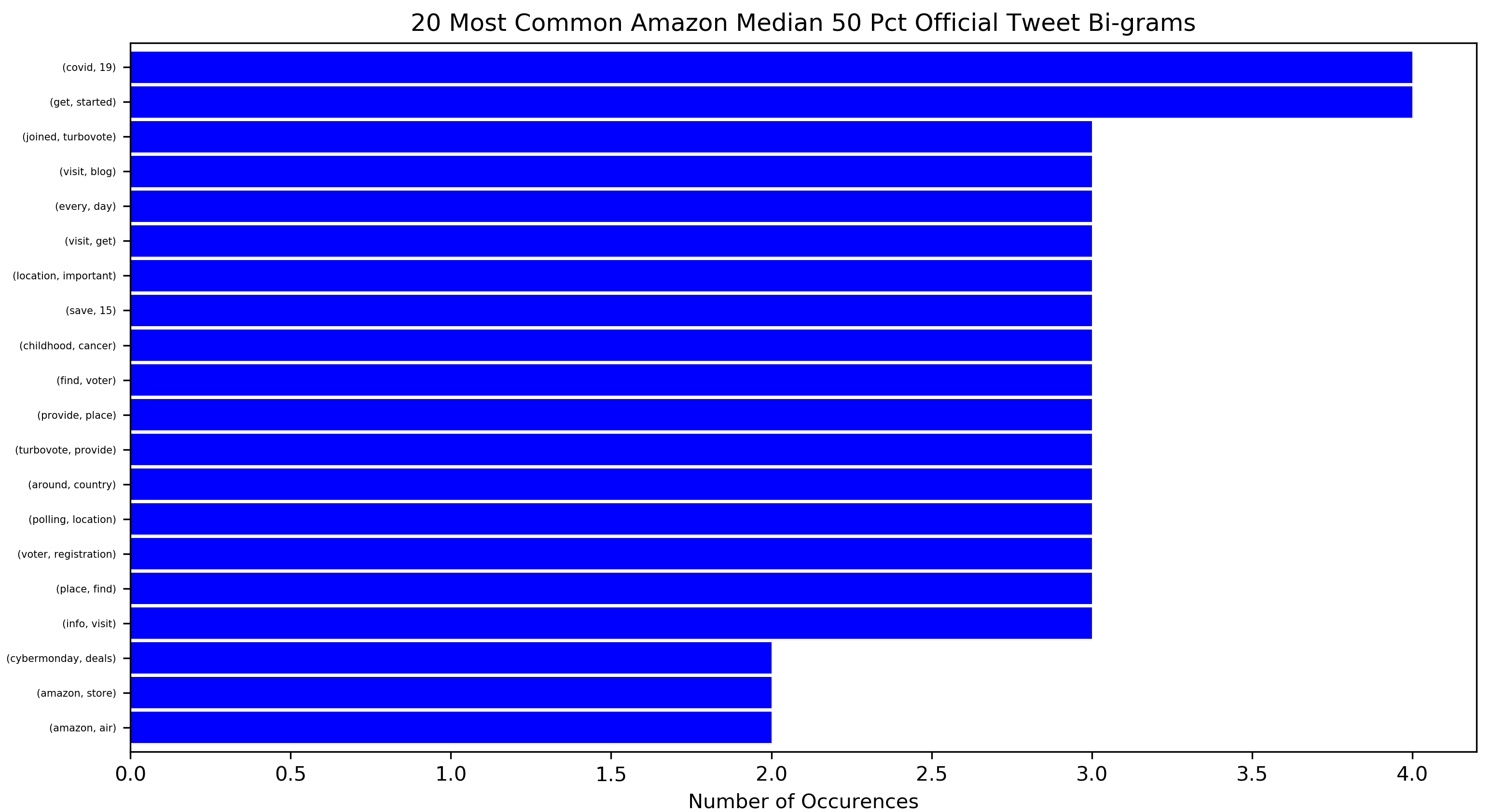


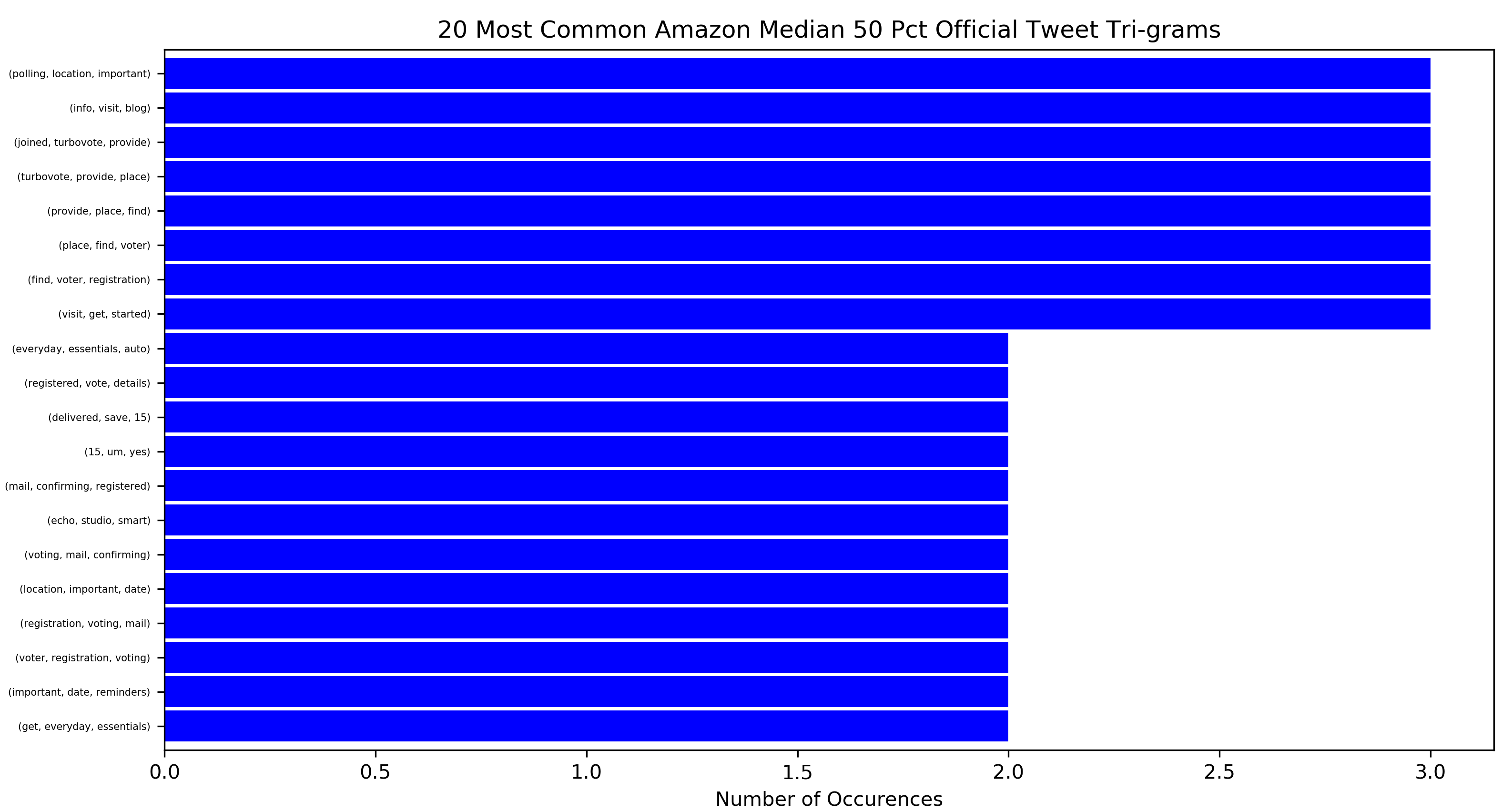


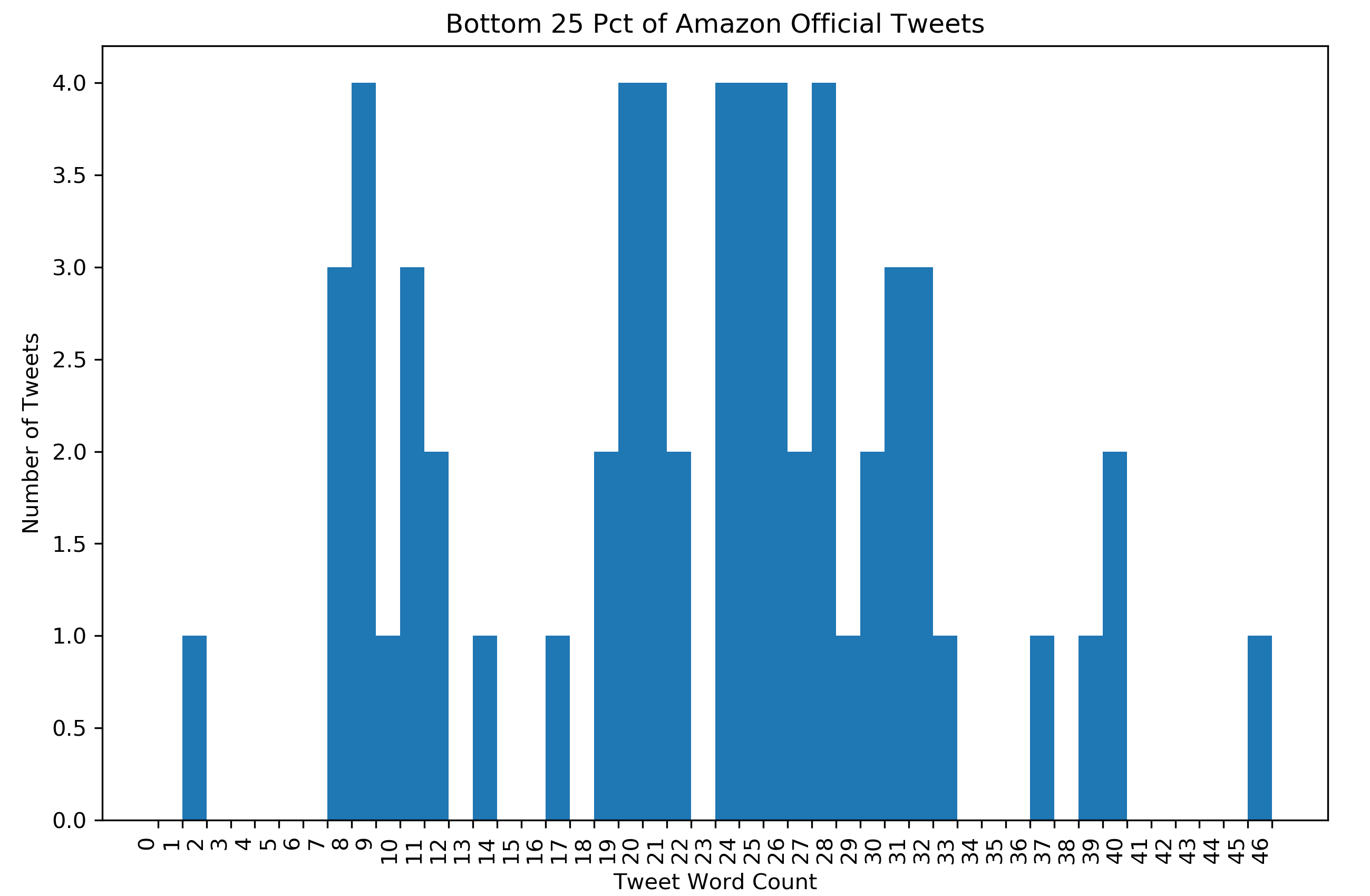


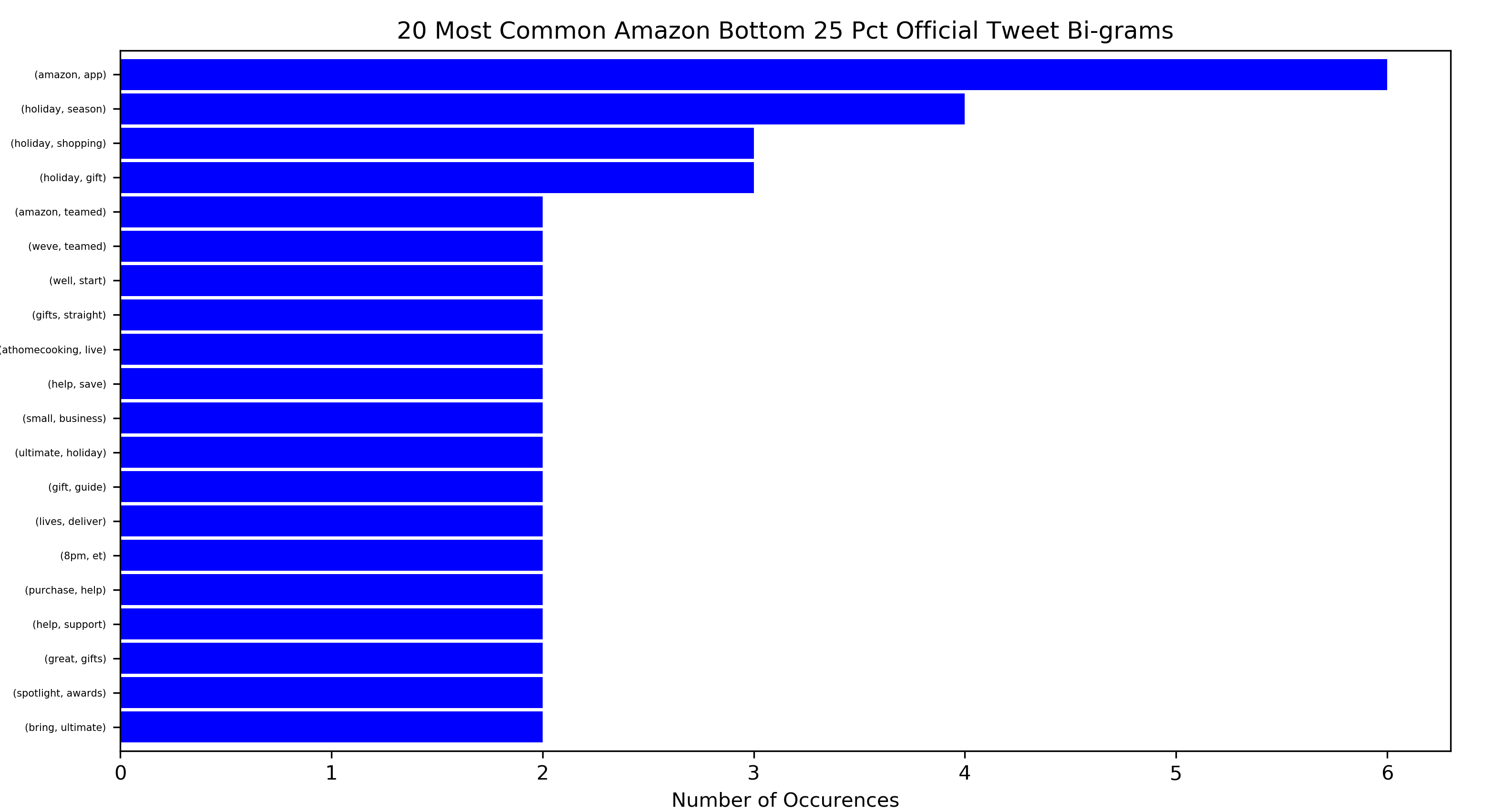


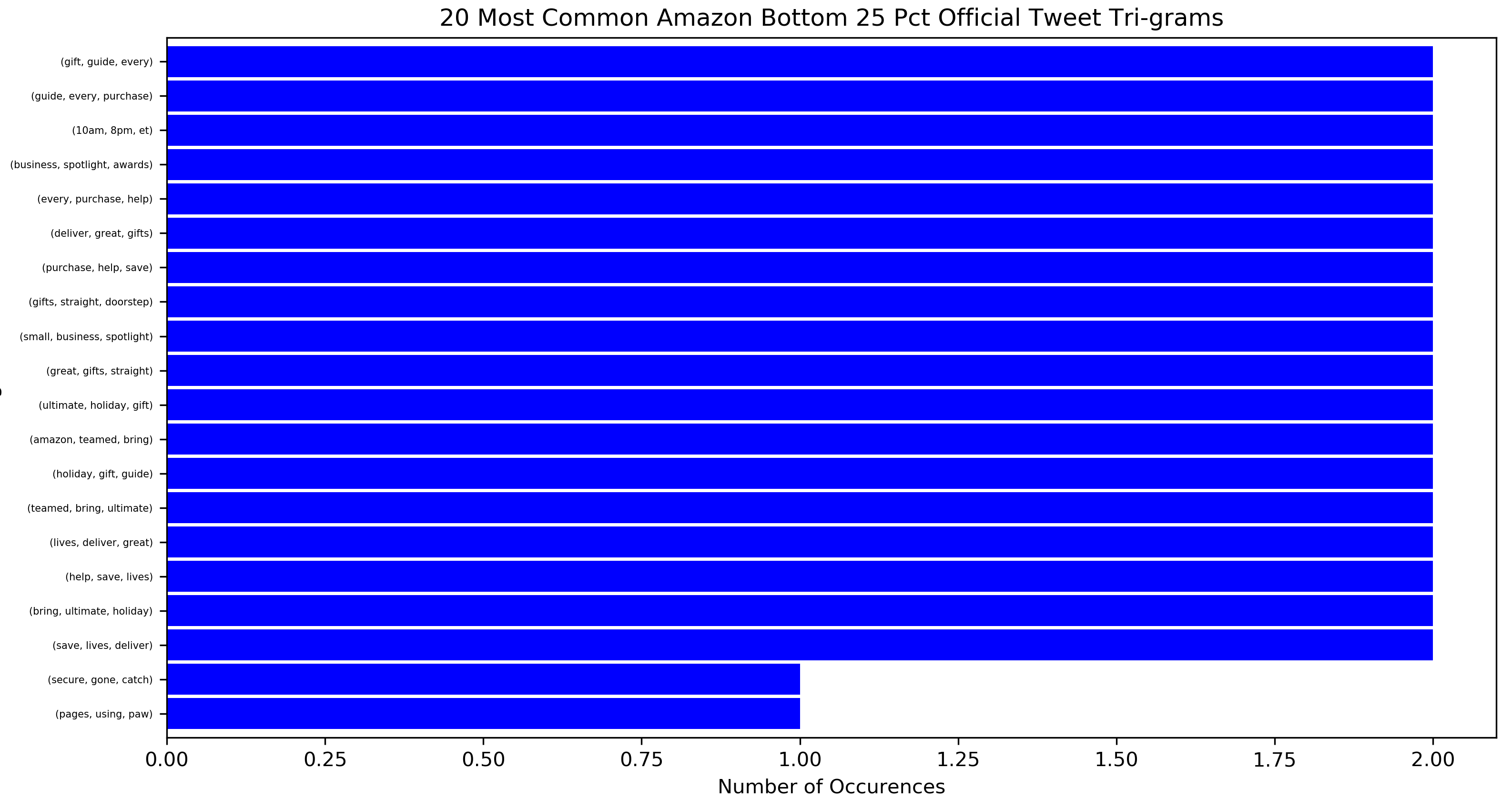


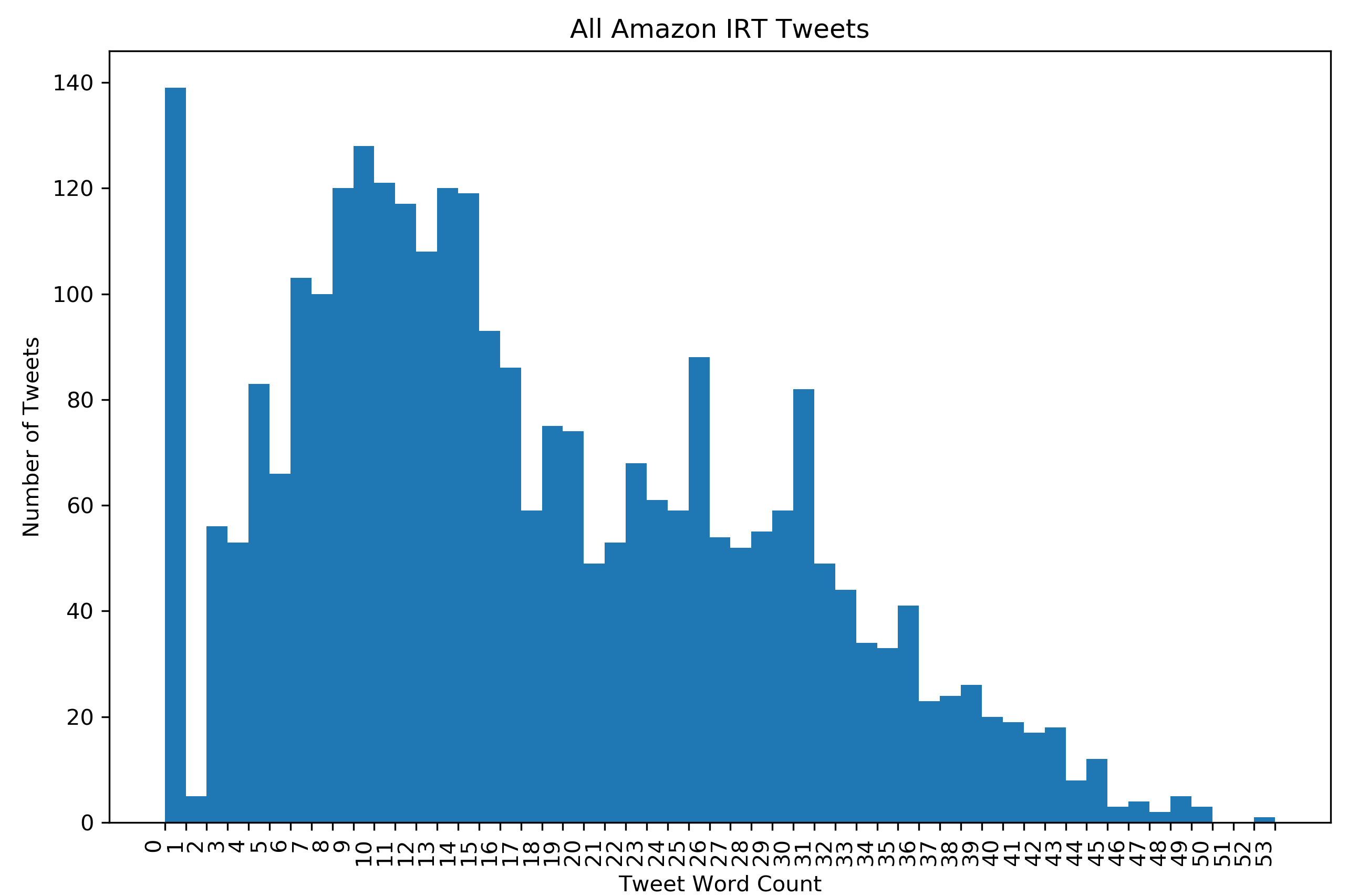


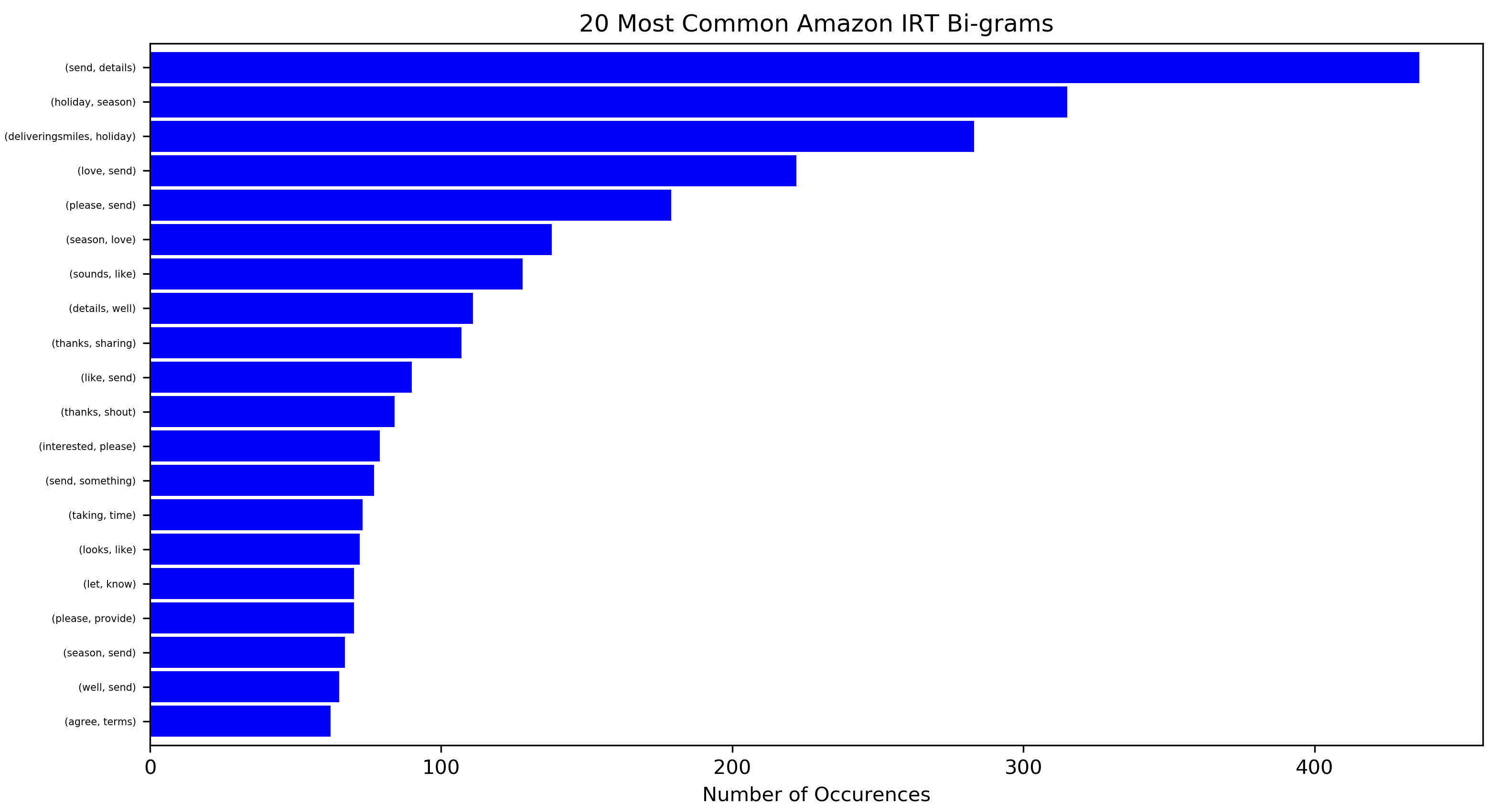


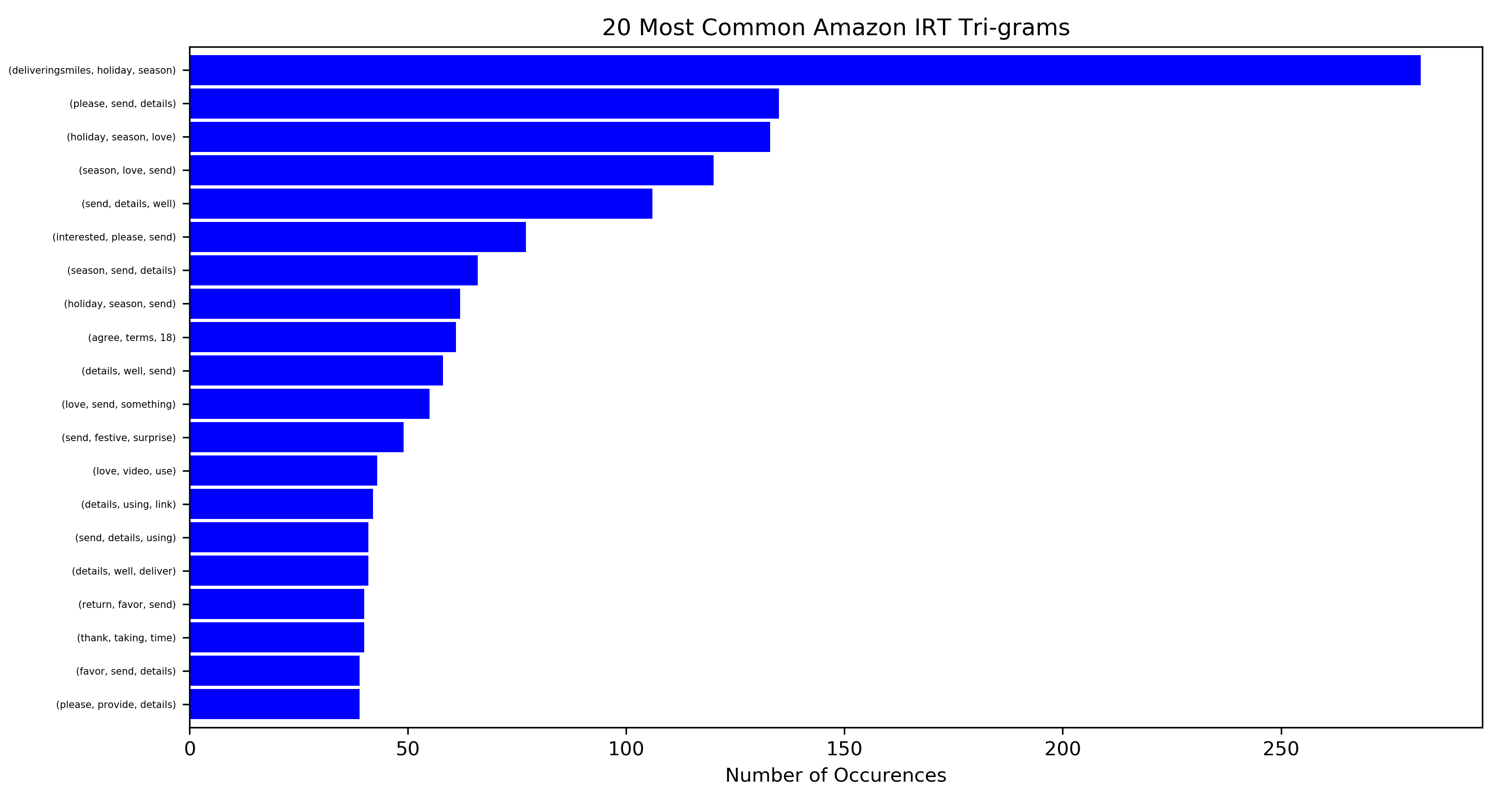


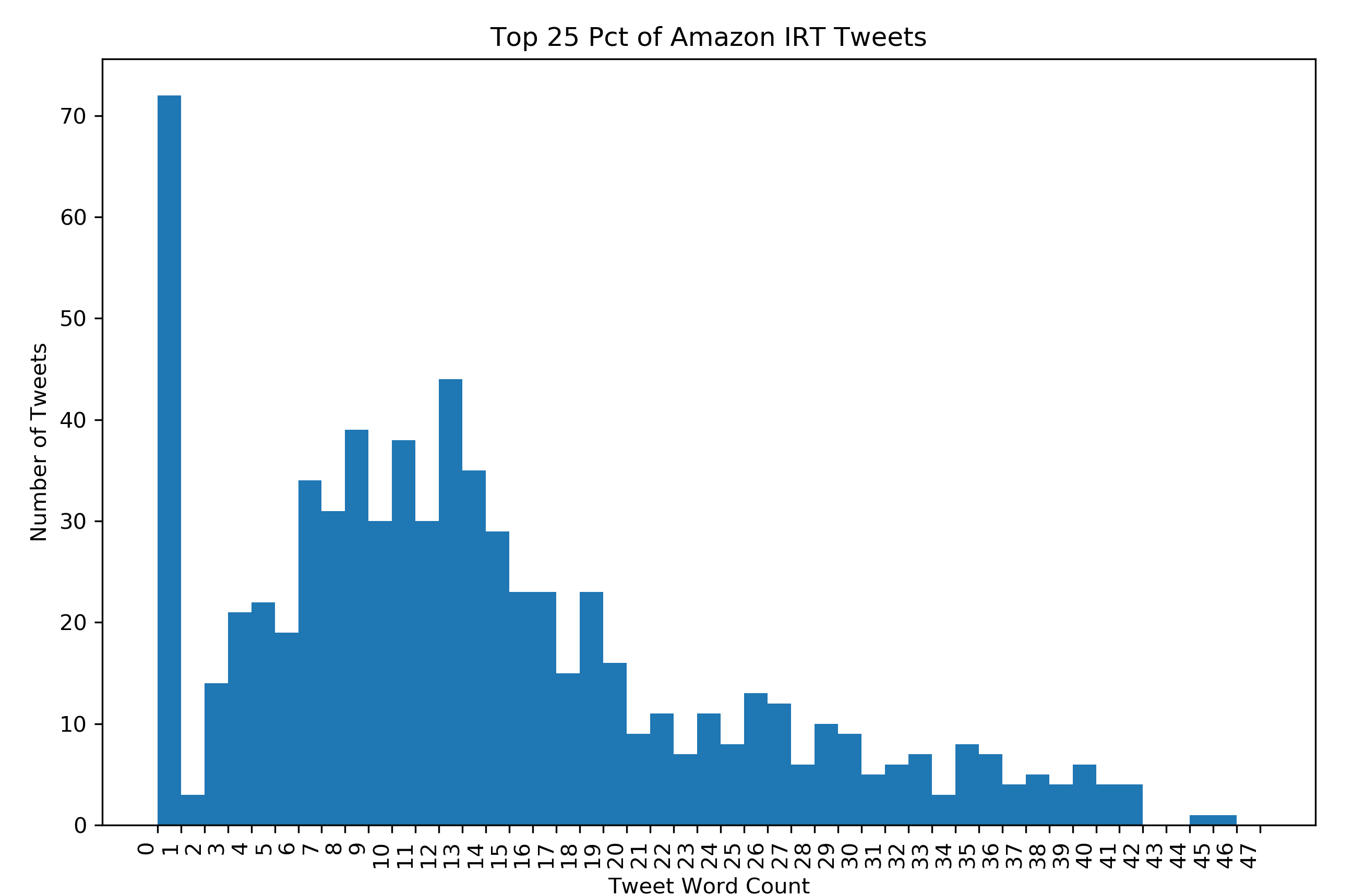


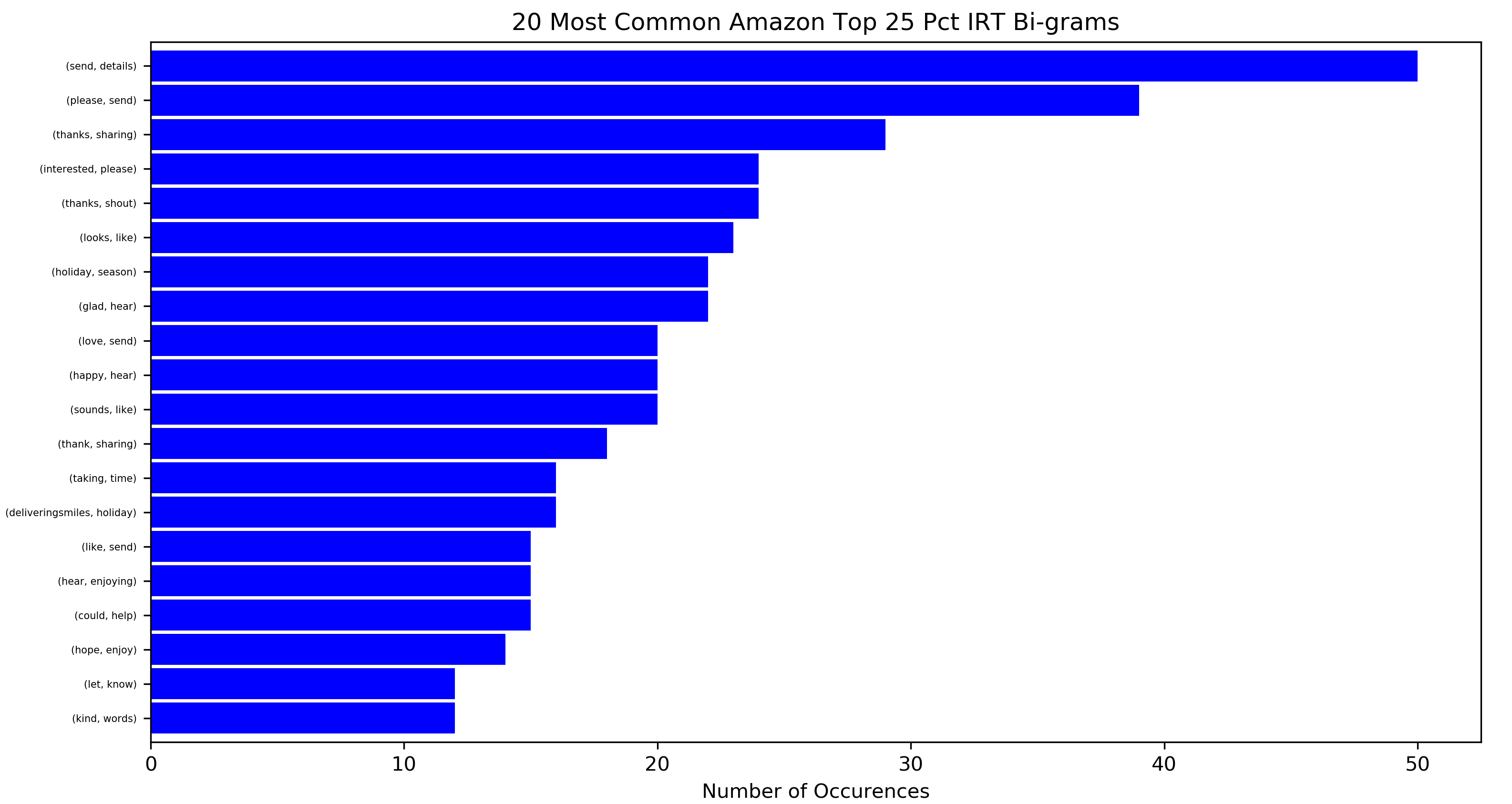


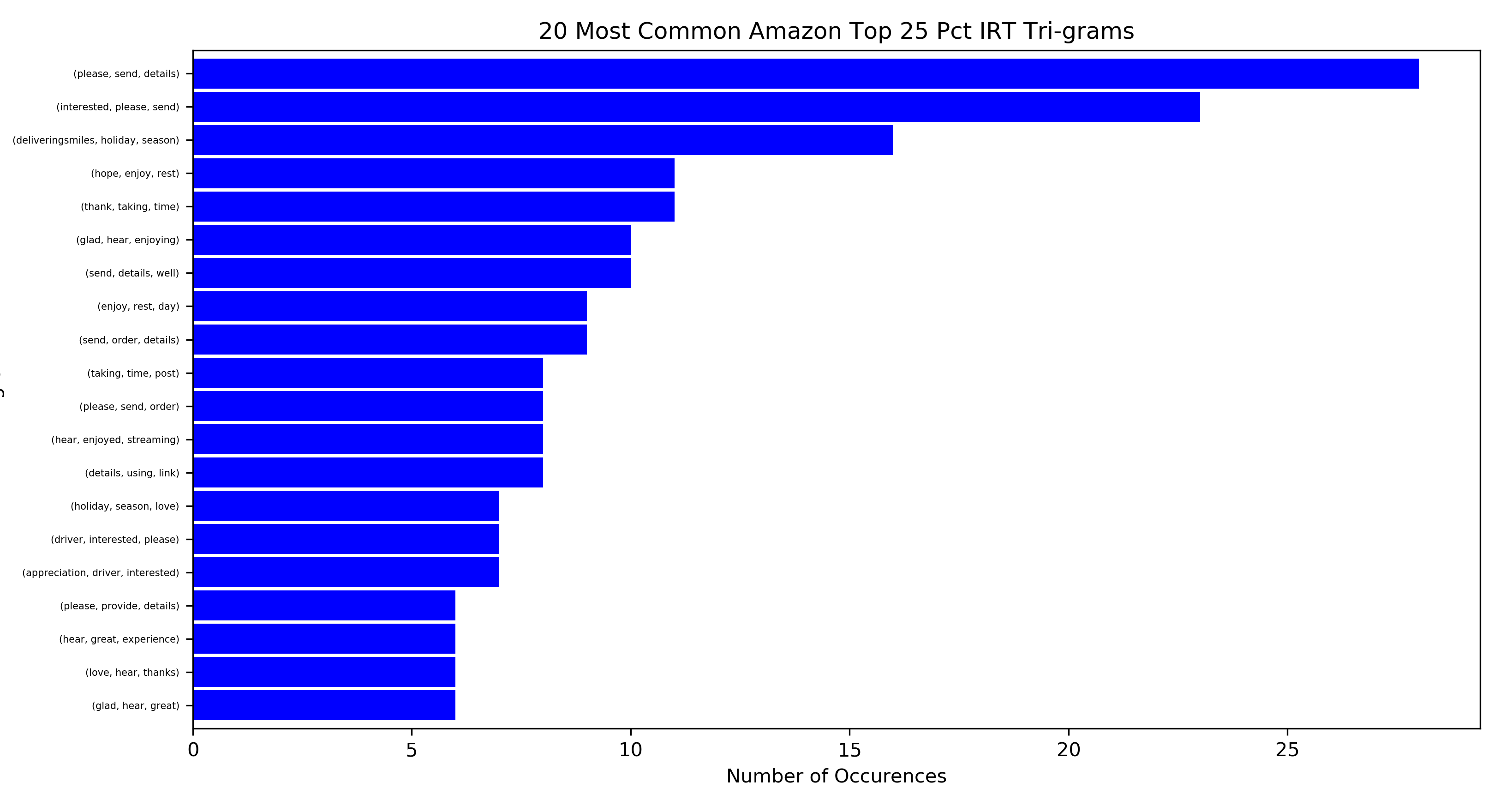


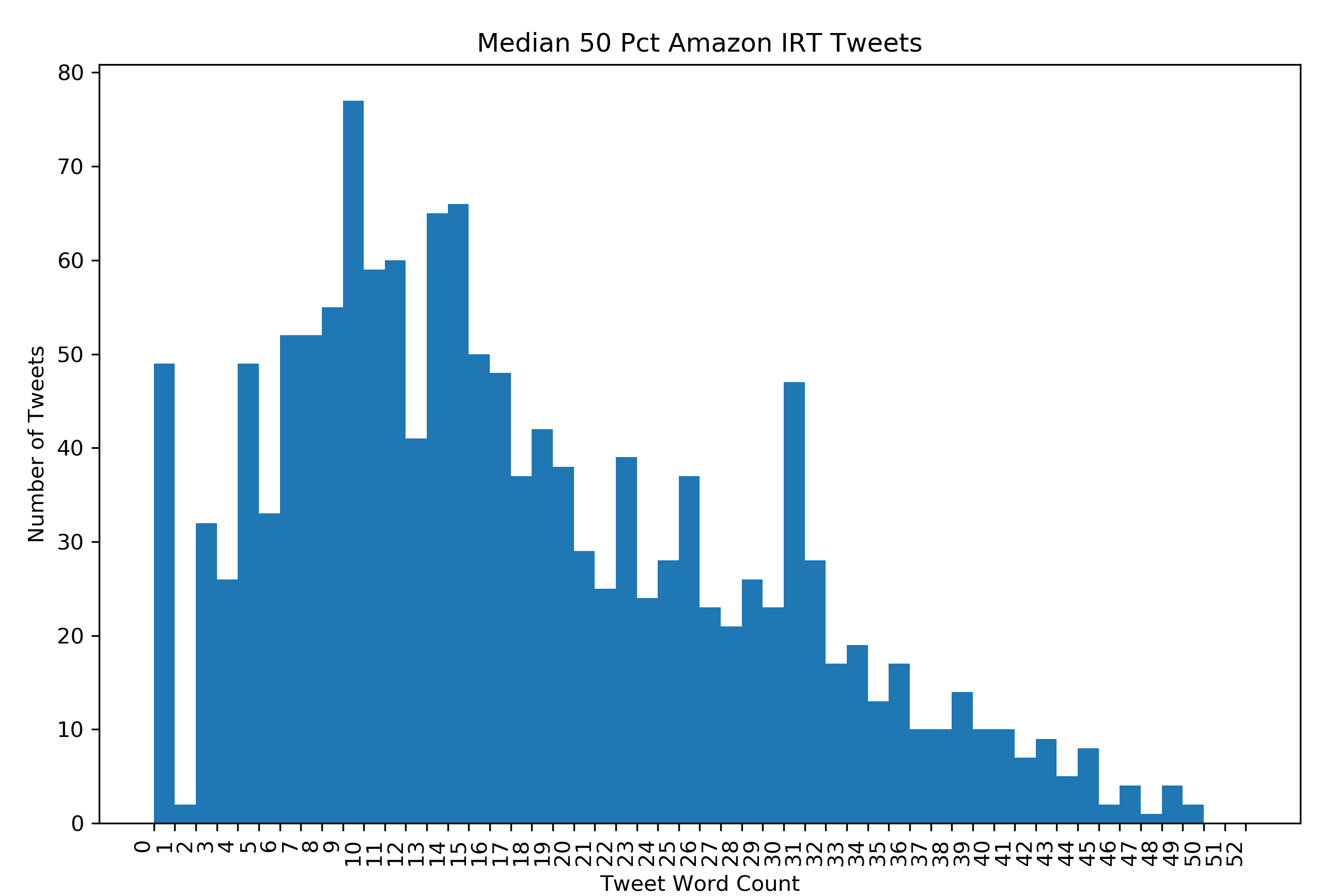


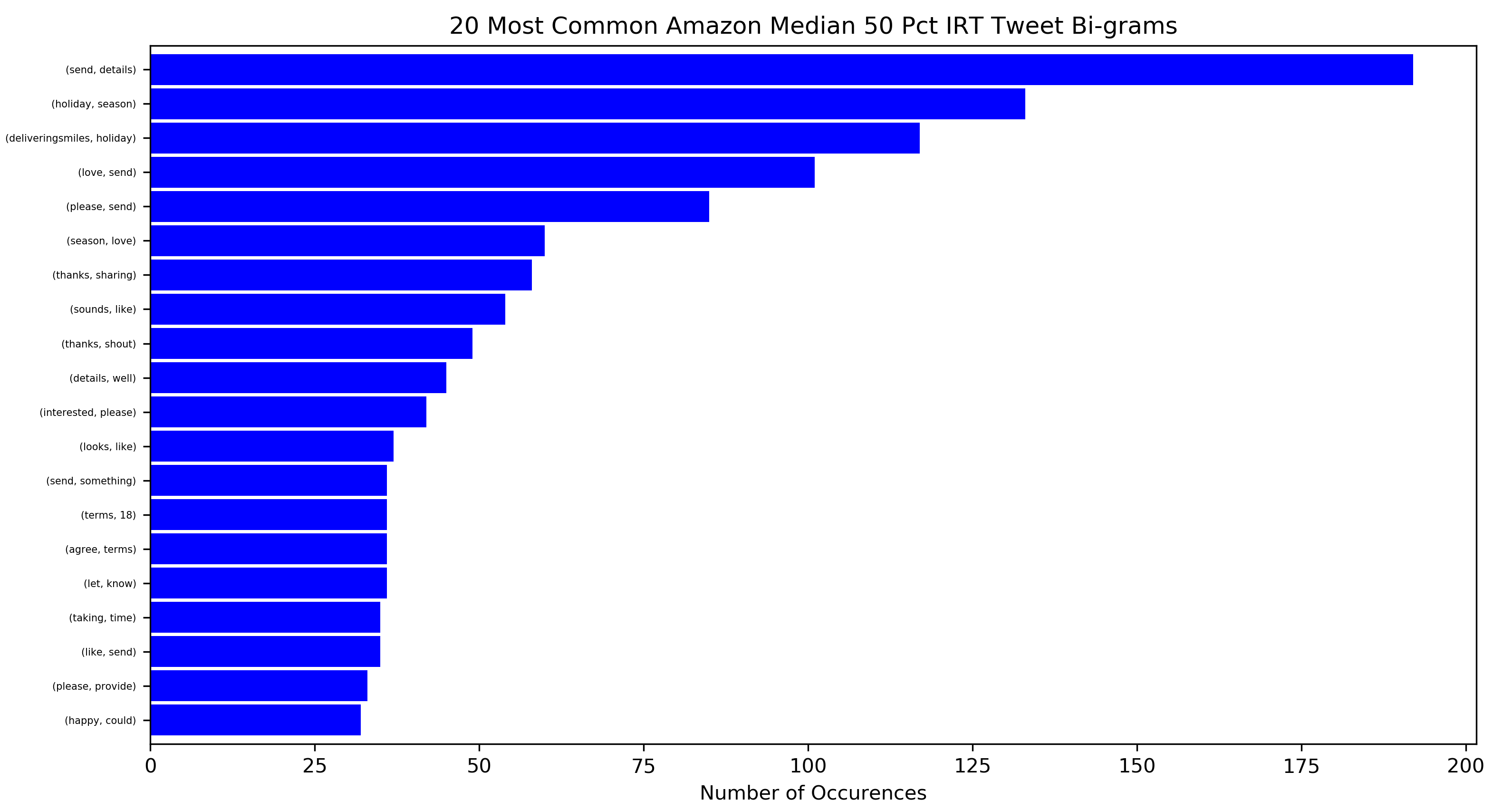


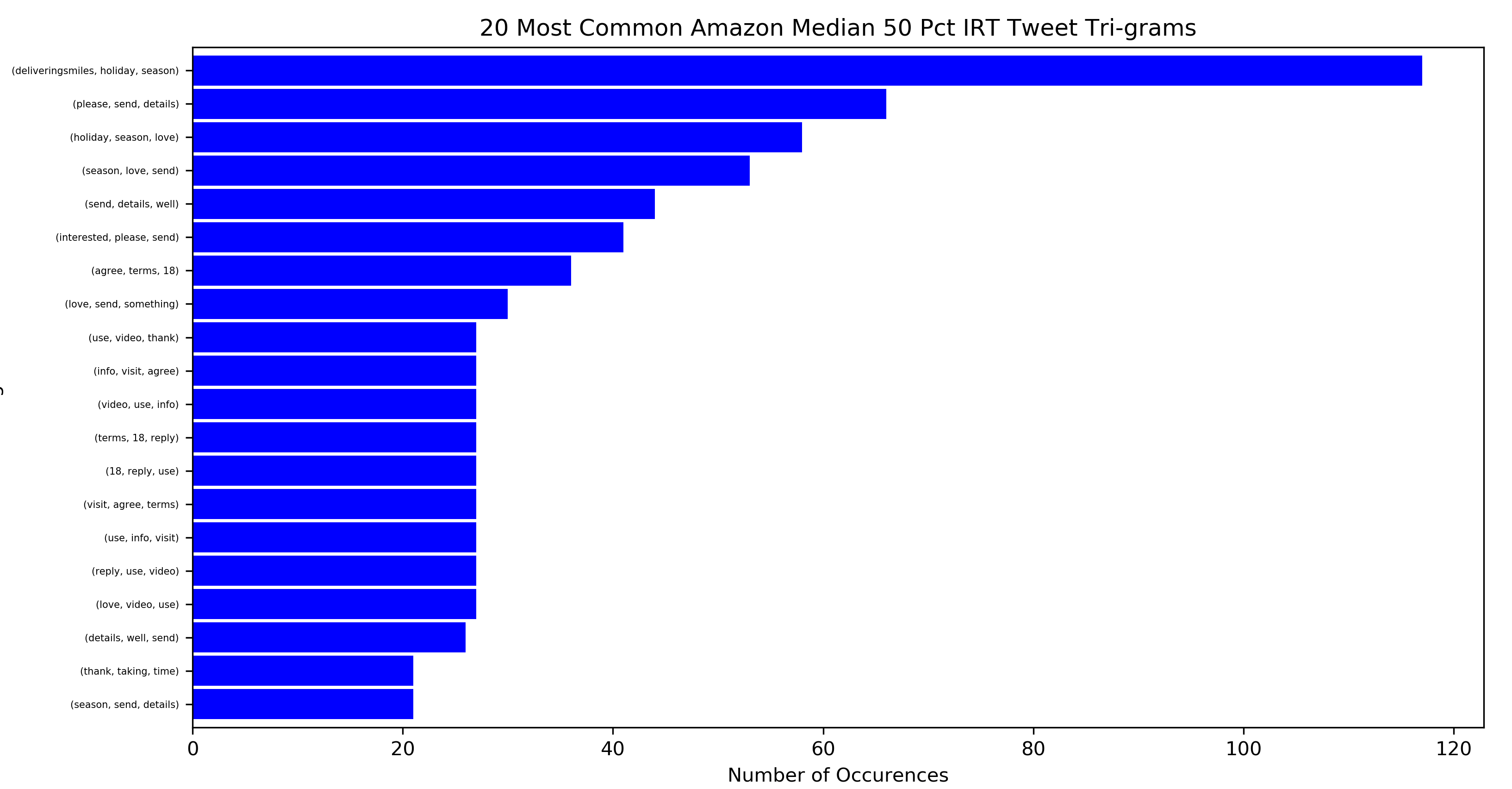


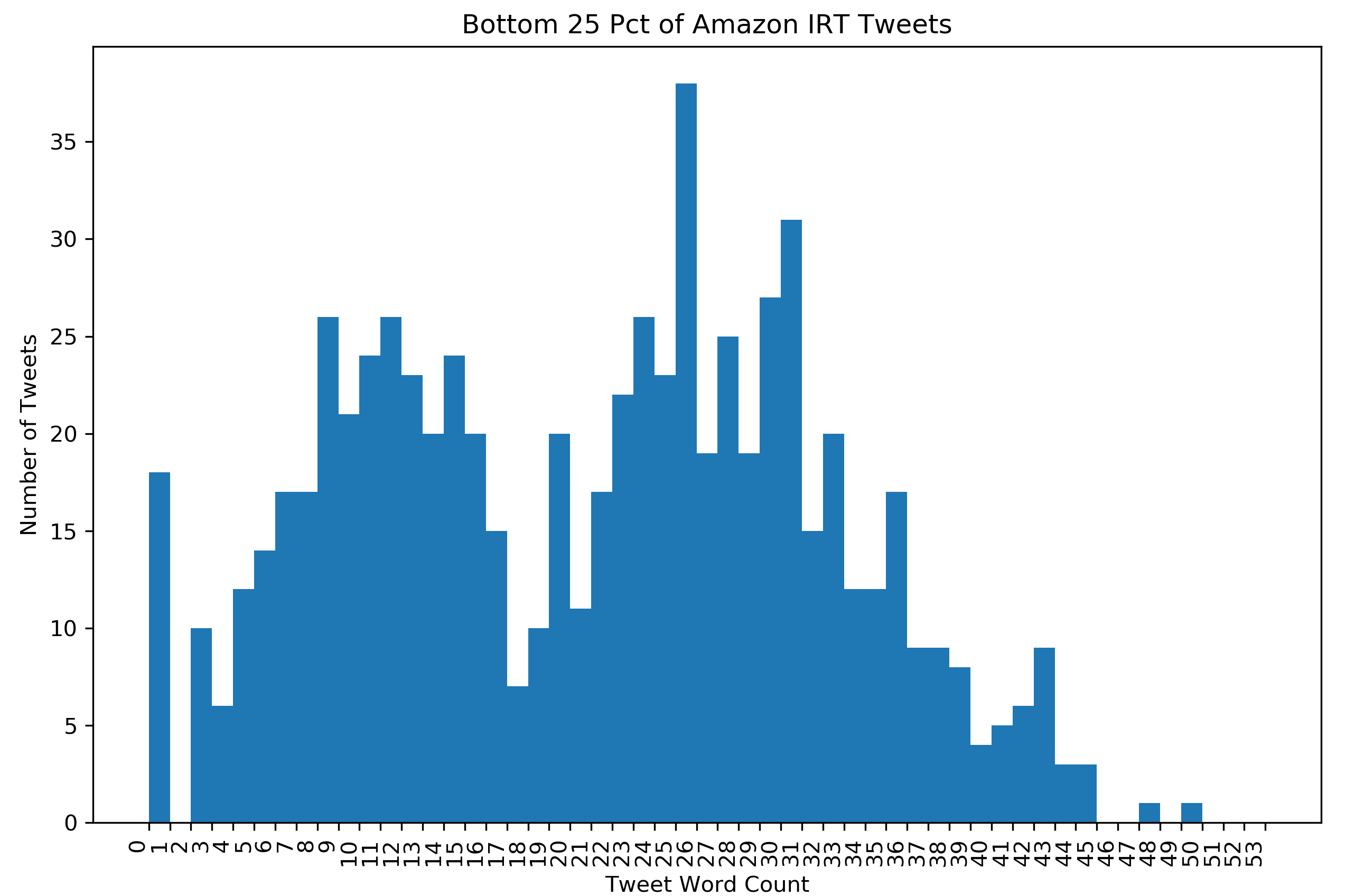


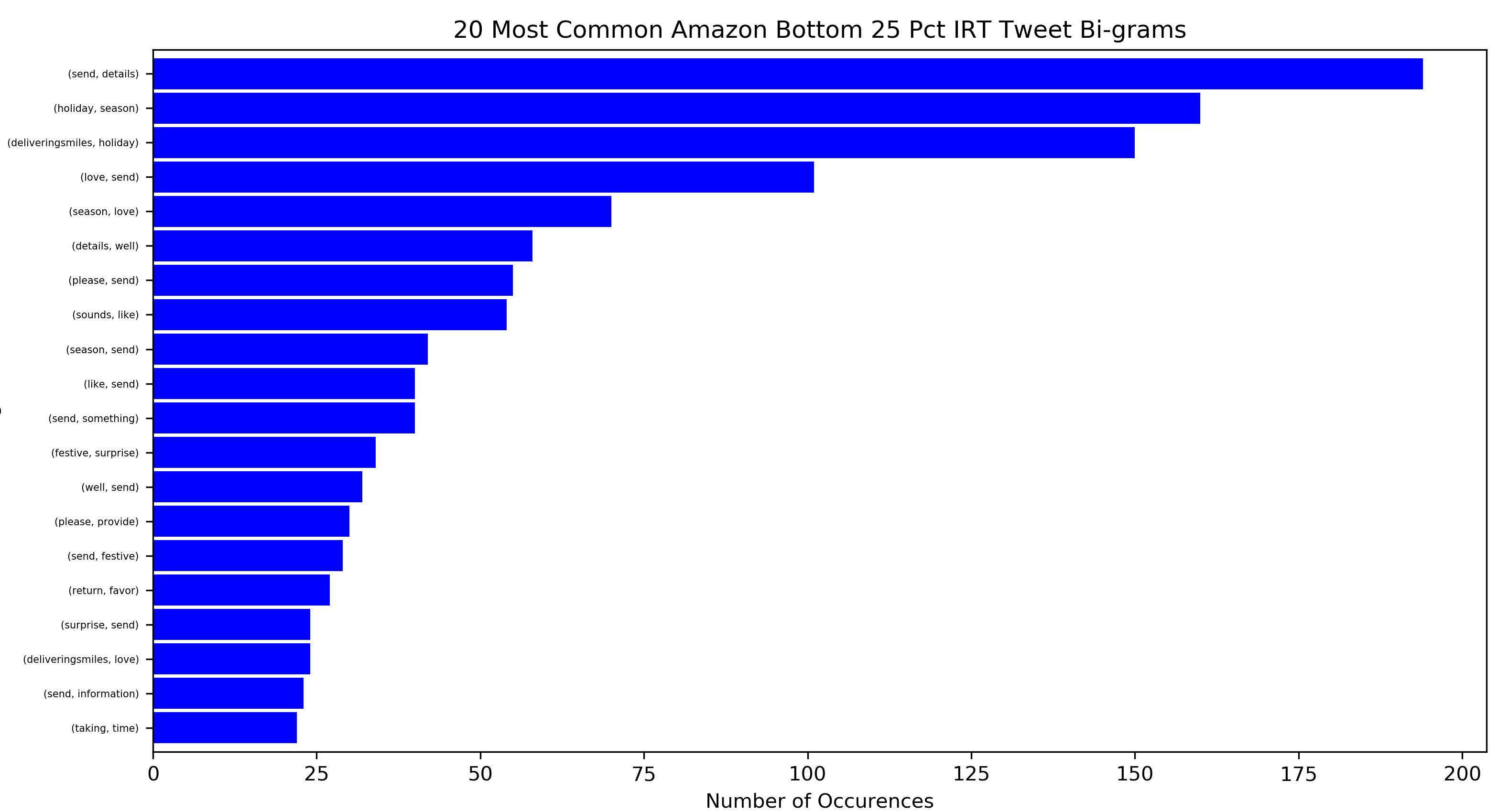


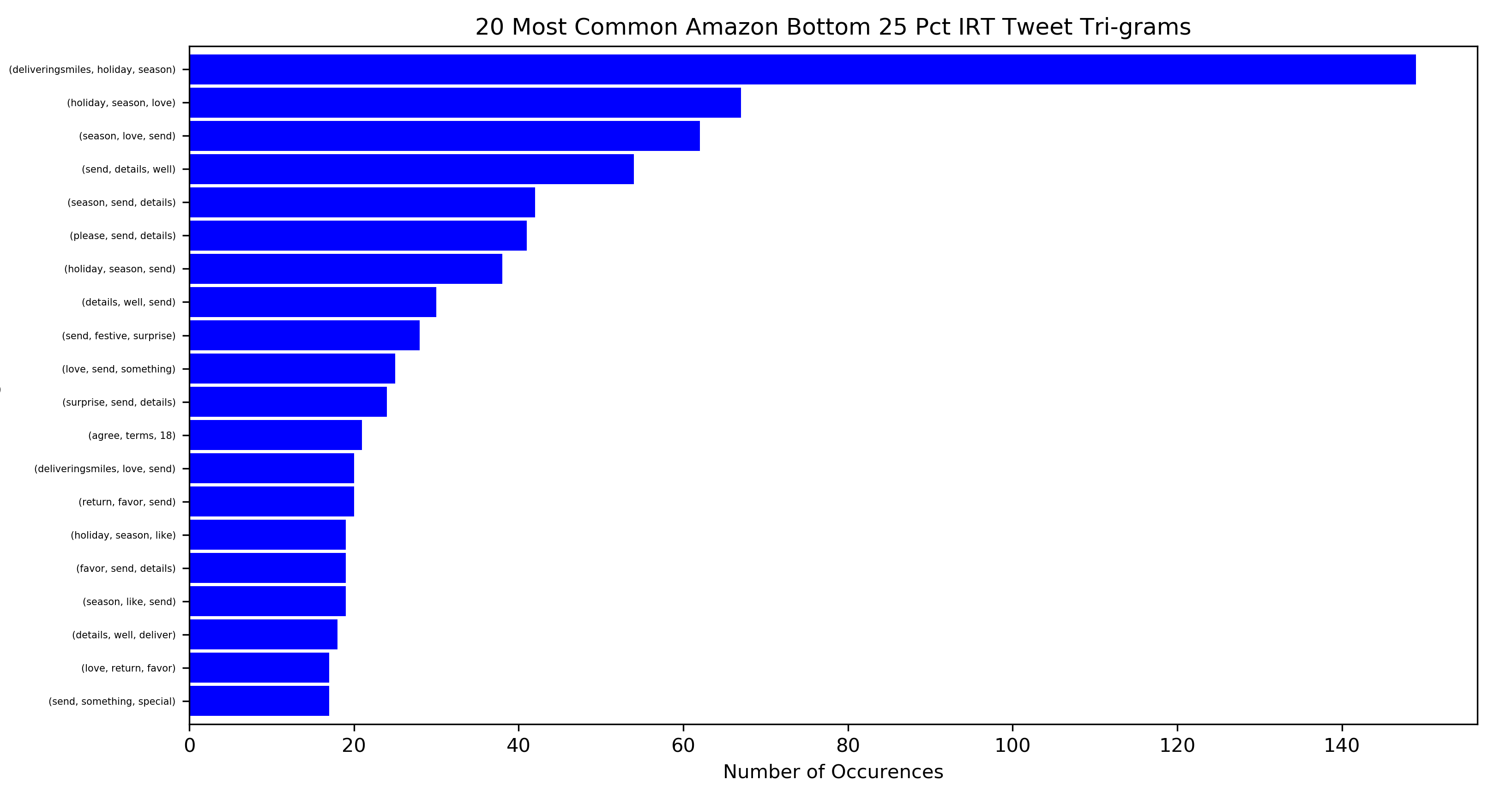












Additionally, at this stage I determined what words are common to each of the categories listed above. Furthermore, when a word was common to a particular category, but not common to any other categories, I considered the word to be ‘uniquely common’ to the category and made note of it:

**Amazon Top Words**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **All Tweets** | **Official Tweets** | **Top 25% Official** | **Median 50% Official** | **Bottom 25% Official** | **All IRT Tweets** | **Top 25% IRT** | **Median 50% IRT** | **Bottom 25% IRT** |
| send, 939 | amazon, 86 | amazon, 28 | amazon, 40 | amazon, 18 | send, 938 | love, 132 | send, 408 | send, 418 |
| love, 732 | new, 30 | delivery, 11 | new, 19 | holiday, 12 | love, 723 | send, 112 | love, 356 | details, 266 |
| details, 628 | see, 26 | customers, 8 | day, 17 | shop, 8 | details, 624 | happy, 104 | details, 282 | deliveringsmiles, 239 |
| like, 481 | delivery, 25 | thank, 8 | see, 15 | check, 8 | deliveringsmiles, 468 | thanks, 102 | please, 230 | love, 236 |
| deliveringsmiles, 481 | help, 23 | ok, 8 | help, 14 | help, 7 | like, 466 | like, 87 | thanks, 229 | holiday, 199 |
| please, 437 | holiday, 22 | family, 7 | get, 13 | favorite, 7 | please, 434 | hear, 84 | like, 224 | season, 175 |
| holiday, 420 | get, 22 | new, 7 | check, 11 | new, 6 | thanks, 405 | details, 76 | happy, 196 | like, 155 |
| thanks, 412 | check, 20 | like, 6 | delivery, 10 | gits, 6 | holiday, 398 | hope, 72 | deliveringsmiles, 194 | please, 137 |
| happy, 386 | today, 19 | today, 6 | items, 10 | app, 6 | happy, 377 | please, 67 | holiday, 168 | surprise, 92 |
| season, 357 | day, 19 | support, 6 | holiday, 10 | live, 6 | season, 347 | thank, 62 | hear, 163 | help, 77 |
| hear, 323 | people, 19 | world, 6 | customers, 9 | season, 5 | hear, 322 | sharing, 57 | thank, 151 | happy, 77 |
| help, 296 | customers, 18 | see, 6 | people, 9 | today, 5 | help, 273 | help, 52 | season, 147 | well, 77 |
| thank, 285 | support, 16 | learn, 5 | today, 9 | shopping, 5 | thank, 271 | great, 50 | help, 144 | hear, 75 |
| surprise, 238 | family, 15 | introducing, 5 | free, 8 | teamed, 5 | surprise, 233 | glad, 50 | great, 119 | thanks, 74 |
| hope, 232 | like, 15 | year, 5 | need, 8 | support, 4 | hope, 231 | day, 44 | hope, 118 | something, 65 |
| great, 231 | thank, 14 | employees, 5 | around, 8 | donate, 4 | great, 227 | enjoying, 44 | surprise, 106 | sounds, 65 |
| day, 192 | learn, 14 | proud, 5 | store, 8 | beauty, 4 | sharing, 186 | shout, 38 | sharing, 96 | thank, 58 |
| time, 188 | items, 13 | get, 5 | home, 8 | see, 4 | time, 179 | new, 37 | day, 93 | great, 58 |
| sharing, 187 | around, 13 | people, 5 | visit, 8 | well, 4 | well, 175 | looks, 36 | time, 91 | time, 58 |
| well, 187 | deliveringsmiles, 13 | community, 5 | deals, 7 | get, 4 | day, 173 | enjoy, 35 | shout, 85 | festive, 57 |

-The words uniquely common to each of the categories above are as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **All Tweets** | **Official Tweets** | **Top 25% Official** | **Median 50% Official** | **Bottom 25% Official** | **All IRT Tweets** | **Top 25% IRT** | **Median 50% IRT** | **Bottom 25% IRT** |
|  |  | ok | free | shop |  | glad |  | something |
|  |  | world | need | favorite |  | enjoying |  | sounds |
|  |  | introducing | store | gifts |  | looks |  | festive |
|  |  | year | home | app |  | enjoy |  |  |
|  |  | employees | visit | live |  |  |  |  |
|  |  | proud | deals | shopping |  |  |  |  |
|  |  | community |  | teamed |  |  |  |  |
|  |  |  |  | donate |  |  |  |  |
|  |  |  |  | beauty |  |  |  |  |

### Preparation

Quite a bit of work has gone into data preparation. Fortunately, there is no missing data, so there never was a need for imputation or anything of the sort. However, there were issues within the textual contents of tweets stemming from data collection. For example, ampersand signs were transformed from ‘&’ to ‘&amp;’ during data collection, the same is true for some other symbols as well. These were found and replaced with the appropriate symbols within the data files themselves, although it’s entirely possible to do so during textual pre-processing instead.

It’s vitally important that textual data is standardized or pre-processed in some fashion prior to being used in most any NLP model. However, considering the specifics are context dependent, I’ll refrain from speaking about how textual pre-processing was handled for each case (in this report).

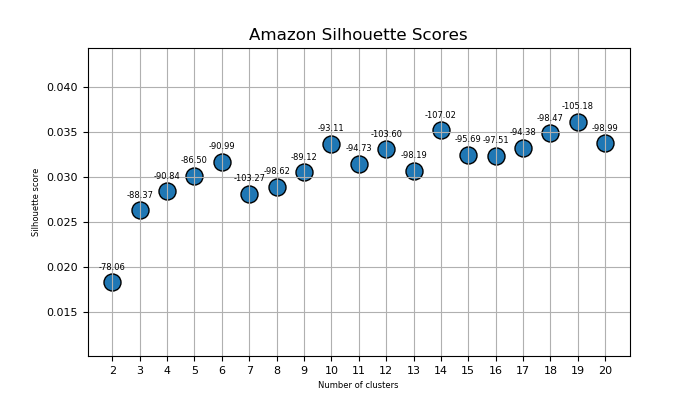
In preparation for analysis, I’ve created several new variables. The first two are very simple, they’re binary variables, named “OT” and “IRT”, indicating whether the tweet is an ‘official tweet’ (OT), or whether the tweet is an ‘in reply to’ tweet (IRT). Furthermore, I’ve created variables indicating the number of words in each tweet, as well as the number of hashtags in each tweet. Additionally, I’ve created a ‘sentiment’ variable, in which every tweet is assigned a sentiment score between -1 and 1. This allowed me to create the following type of table for every company:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Description** | **No. Observations** | **Avg. Sentiment** | **Avg. Likes** | **Avg. Retweets** |
| All Tweets: | 3136 | 0.7666 | 27.76 | 4.56 |
| All Negative Tweets: | 140 | -0.6812 | 87.66 | 15.16 |
| All Neutral Tweets: | 293 | 0.0 | 31.20 | 5.48 |
| All Positive Tweets: | 2703 | 0.9247 | 24.29 | 3.91 |
| Official Tweets: | 246 | 0.5190 | 313.27 | 51.42 |
| Negative Official: | 32 | -0.6949 | 369.81 | 64.94 |
| Neutral Official: | 39 | 0.0 | 198.54 | 34.44 |
| Positive Official: | 175 | 0.8566 | 328.50 | 52.74 |
| Top 25% Official: | 62 | 0.6044 | 775.42 | 120.56 |
| Median 50% Official: | 122 | 0.4938 | 186.39 | 33.25 |
| Bottom 25% Official: | 62 | 0.4930 | 98.34 | 17.84 |
| IRT Tweets: | 2890 | 0.7877 | 3.46 | 0.5709 |
| Negative IRT: | 108 | -0.6772 | 4.06 | 0.42 |
| Neutral IRT: | 254 | 0.0 | 5.51 | 1.03 |
| Positive IRT: | 2528 | 0.9295 | 3.23 | 0.53 |
| Top 25% IRT: | 723 | 0.8325 | 11.34 | 1.27 |
| Median 50% IRT: | 1444 | 0.7932 | 1.24 | 0.49 |
| Bottom 25% IRT: | 723 | 0.7306 | 0.00 | 0.03 |

In order to create sentiment scores, I used the “IBM-Watson Analyzer.” Seeing as that I lack labeled data, I had no choice but to choose an unsupervised sentiment analysis algorithm. I found that this had superior performance when compared to ‘TextBlob’ and ‘VADERSentiment analyzer’, two other unsupervised sentiment analysis algorithms. Furthermore, not only did IBM-Watson have better performance than these two across all evaluation metrics (accuracy, precision, and recall), it also outperformed an ensemble method consisting of a “majority wins” ruleset between the three algorithms (Anita).

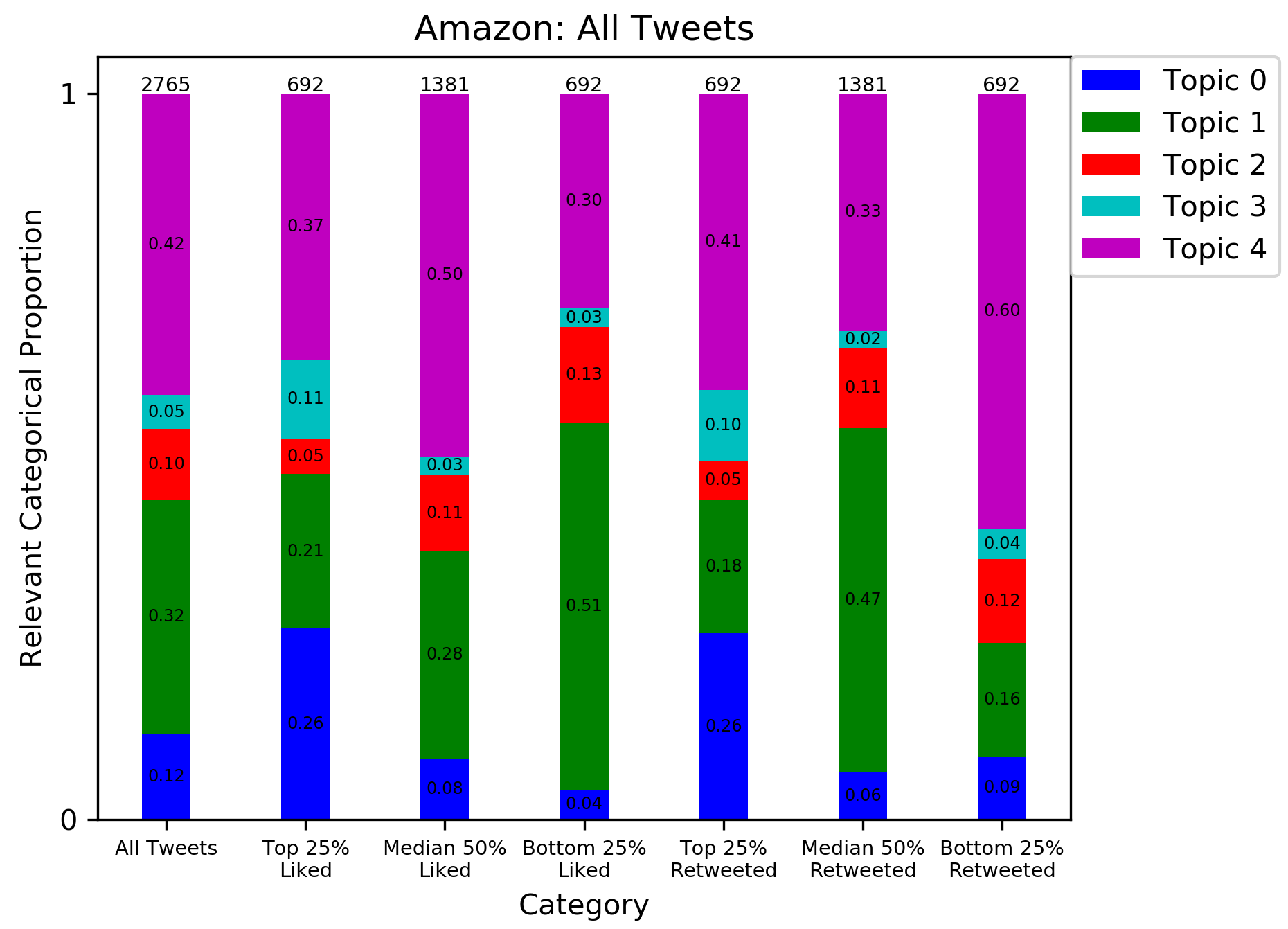
Additionally, I have created a ‘topic’ variable for each company’s data. Research into topic modeling with Twitter data consistently led me to the same model, the “Biterm Topic Model” (BTM). BTM was created specifically for use with short-texts and to overcome the difficulties that traditional topic modelling algorithms experience when faced with sparse data. Throughout all algorithms tested on Twitter data, BTM predicted topics “with the minimum intra distance and maximum inter distance” as well as, on average, “producing the most coherent topics” (Jonsson). Furthermore, a separate paper’s results “demonstrated that BTM not only can learn higher quality topics, but also can more accurately capture the topics of documents than previous methods” (Yan).

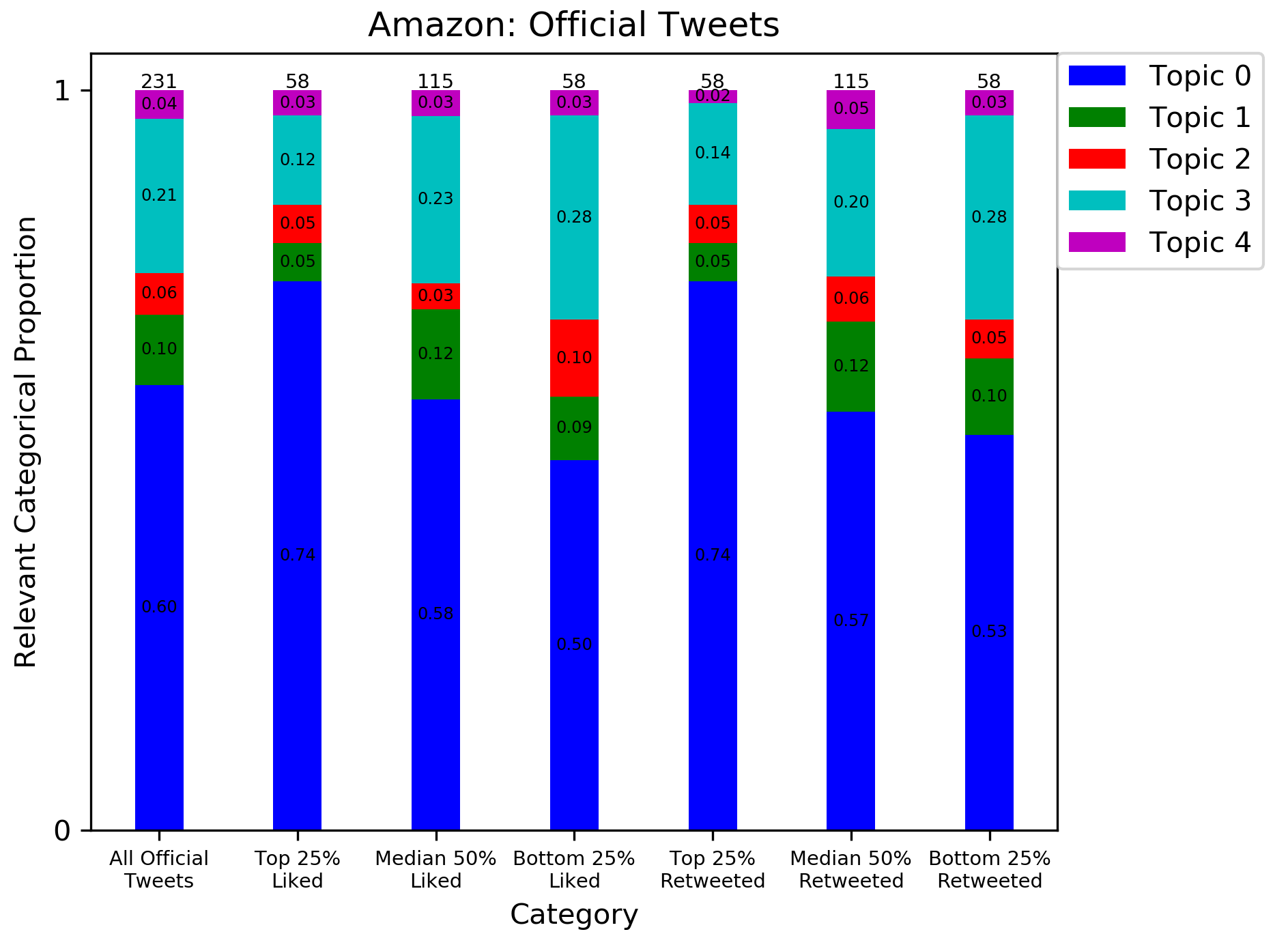
In order to determine the optimal number of topics for each company, I created a ‘quick’ version of BTM, which may be thought of simply as an estimate, and tested it out on 2-20 topics for each company. Ultimately, the number of topics producing the best average topic coherence score was selected for each dataset. While my version of BTM may be quicker than performing the full BTM process, it’s still rather slow. Interested in whether silhouette scores could be utilized to determine the optimal number of topics in a quicker fashion, I produced the following figure for each company:

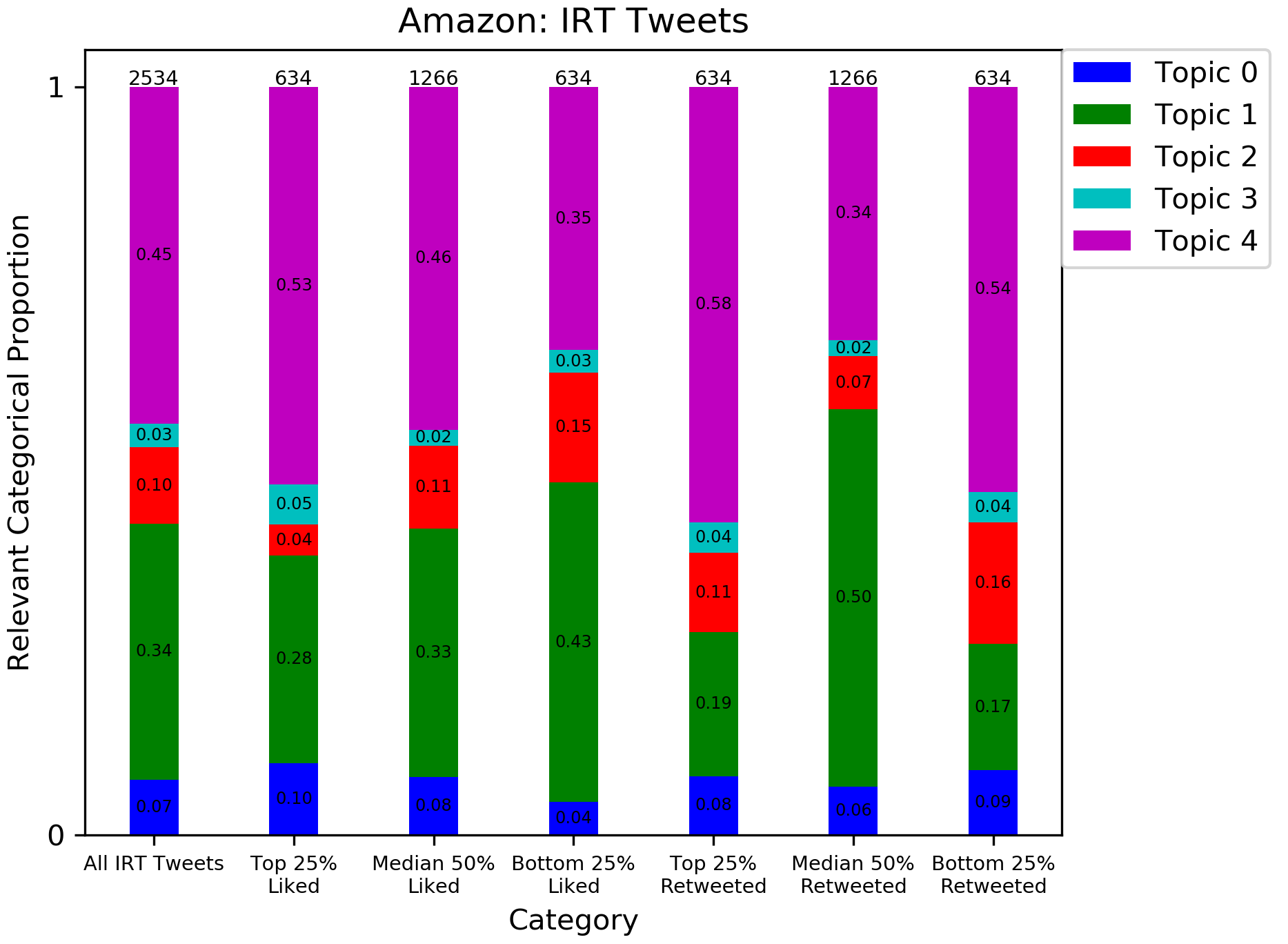


Silhouette scores are represented by the blue dots, the estimated average topic coherence has been overlayed above. In this case, it’s desirable for topic coherence scores to be less negative (Yan). One interesting observation resulting from this is that, for 8 of the 10 companies, the optimal estimated number of topics was within 1 position of a 15% or greater change in silhouette scores. I should also mention that, for companies in which k = 2 produced the least negative average topic coherence, I selected the number of topics producing the second least negative average topic coherence. This may have added a slight degree of subjectivity, but it’s not my belief that any of these companies only discuss two primary topics throughout their past 3,000 or so tweets.

Having created topic variables for each dataset, I was able to create the following figures and tables (the values at the top of bars represent the number of tweets belonging to that category):







|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Number Obs.** | **Data Proportion** | **Avg. Likes** | **Stdev: Likes** | **Avg. Retweets** | **Stdev. Retweets** |
| 0 | 326 | 0.1179 | 159.50 | 587.51 | 26.14 | 68.80 |
| 1 | 890 | 0.3219 | 8.79 | 77.30 | 1.44 | 13.12 |
| 2 | 271 | 0.0980 | 14.08 | 83.70 | 3.58 | 23.00 |
| 3 | 128 | 0.0463 | 93.08 | 205.93 | 14.0 | 22.76 |
| 4 | 1150 | 0.4159 | 4.45 | 28.21 | 0.76 | 2.75 |

**Amazon Official Tweets:**

Total Number of Official Tweets: 231

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Number Obs.** | **Data Proportion** | **Avg. Likes** | **Stdev: Likes** | **Avg. Retweets** | **Stdev. Retweets** |
| 0 | 139 | 0.6017 | 369.12 | 857.48 | 60.50 | 95.24 |
| 1 | 22 | 0.0952 | 290.14 | 406.41 | 48.09 | 70.22 |
| 2 | 13 | 0.0563 | 270.0 | 287.70 | 64.15 | 87.86 |
| 3 | 48 | 0.2078 | 208.0 | 178.75 | 34.29 | 22.10 |
| 4 | 9 | 0.0390 | 243.67 | 116.89 | 28.78 | 10.91 |

**Amazon IRT Tweets:**

Total Number of IRT Tweets: 2534

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Number Obs.** | **Data Proportion** | **Avg. Likes** | **Stdev: Likes** | **Avg. Retweets** | **Stdev. Retweets** |
| 0 | 187 | 0.0738 | 3.70 | 18.65 | 0.61 | 1.33 |
| 1 | 868 | 0.3425 | 1.66 | 8.16 | 0.25 | 0.59 |
| 2 | 258 | 0.1018 | 1.19 | 3.84 | 0.53 | 0.63 |
| 3 | 80 | 0.0316 | 24.13 | 190.59 | 1.83 | 11.93 |
| 4 | 1141 | 0.4503 | 2.56 | 15.84 | 0.54 | 0.74 |

## Methodology

### Techniques

It is our intention to test a wide variety of modelling techniques, as well as to test feature combinations within algorithms. This makes it exceedingly difficult to say that we have one specific “intended modelling technique.” However, models we intend to test include, but are not limited to: Random Forest Regression, K-Nearest Neighbors Regression, Support Vector Machines, neural networks, linear and other regression models, gradient boosted decision trees, and adaptive boosting methods.

Every algorithm will be tested on all datasets. Ideally, K-fold cross validation will be used. However, if time does not allow, data will simply be split into reproducible train/test splits, and comparisons will be made based off performance on testing data instead. Statistics such as RMSE, MAPE, AIC, and BIC will be computed for each run. In order to determine whether certain algorithms have statistically significant improvements in performance, a Wilcoxon test may be applied to determine whether algorithm A’s predictions for company X were superior to algorithm B’s predictions for the same company.

### Process Validation

Dr. Yoon and I both agree on generating predictions for companies individually, then comparing aggregate performances. This is because some variables, such as tweet topic, cannot be extrapolated to other datasets. For example, topic 0 for Amazon data has no relation to topic 0 for BMW data. I believe that my accuracy measures are sound for the occasion because these are the same general accuracy measures I’ve used for almost all predictive machine learning models. A Wilcoxon test is appropriate for comparing predictions stemming from different algorithms over the same dataset because that data is, by definition, paired. Furthermore, our target variables tend to not be Gaussian, making parametric tests ineligible. However, further research will still be put into validation.

## Results and Analysis

## Deliverables

## References

Anita, O. (2019, December 23). Evaluating Unsupervised Sentiment Analysis Tools Using Labeled Data. Retrieved March 22, 2021, from https://heartbeat.fritz.ai/evaluating-unsupervised-sentiment-analysis-tools-using-labeled-data-8d4bb1336cee

Jonsson, E., & Stolee, J. (n.d.). *An Evaluation of Topic Modelling Techniques for Twitter* [Scholarly project]. Retrieved from https://www.cs.toronto.edu/~jstolee/projects/topic.pdf

Statista. (2021, January 27). *Twitter: Number of monthly active users 2010-2019*. <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

Yan, Xiaohui & Guo, Jiafeng & Lan, Yanyan & Cheng, Xueqi. (2013). A biterm topic model for short texts. WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web. 1445-1456. 10.1145/2488388.2488514.