
What are People Tweeting? Identifying “Trending” Topics in the Medical Field Using Twitter Data

SooHyun Kim

Client: ZS Associates (Wenhao Xia, Albert Whangbo)

November 11, 2022

ABSTRACT

This is the first of three reports for ZS Associates that aims to identify trending topics in the medical field using Twitter data. The objective of this report is to create a dictionary of possible topics that may be trending. For this research, we focus on Twitter text data related to prostate cancer. We use uni-, bi-, and tri-grams as definition of phrases. To identify such phrases, we implement the *bag of n-grams* feature extraction method. After adjustments, we identify the top 30 most commonly mentioned topics for each method. The topics we identify include phrases related to urology, metastatic castration-resistant prostate cancer (mCRPC), prostate cancer-specific mortality (PCSM), and various treatments and medication for prostate cancer. One key finding is with the strong discussion of a recent paper published which looks into the disparity between black and white men with prostate cancer. We also create wordclouds to visualize the results. This method can be replicated for other medical fields of interest.

1 INTRODUCTION

One of the most cost-effective ways to understand the demand of a target group is to listen in on their conversations. With the development of platforms like Twitter, it is easier for companies to capture popular topics that are discussed by the target groups. For this project, ZS Associates requested us to identify possible "trending" topics for seminars using Twitter data. This project is divided into three parts. First, this report looks into text pre-processing and topic identification. The second report, by Alexander Murph, analyzes whether these topics are "trending." The final report, by Taebin Kim, performs topic detection using Latent Dirichlet Allocation (LDA).

In this project, we use Twitter data that was collected by ZS Associates. We limit our focus on prostate cancer data. This report contains the first two steps in topic identification. Using methods discussed in Section 3.2, we pre-process the text data. After, we create a dictionary of possible topics that could be "trending." In order to identify the possible topics, we extract features using the *bag of n-grams* method. In this report, we define topics as phrases in the form of uni-, bi-, and tri-grams. Detailed explanations of this will be given in Section 3.3. We visualize the results using wordclouds. The deliverables for this project include the dictionary of possible "trending" topics given in Tables 2.1, 2.2, and 2.3 and the **R** code accessible via the GitHub repository¹.

The paper is organized as follows. Section 2 summarizes the findings of this report. Section 3 briefly discusses the data used and the methodology. Finally, we conclude with a short discussion about the replication in other medical areas and future work.

2 SUMMARY OF FINDINGS

Based on the *bag of n-grams* method, we identify the 30 most frequently mentioned phrases for uni-, bi-, and tri-grams. A uni-gram is a single word, a bi-gram is two words, and a tri-gram is three words that appear together. For each n-gram, we provide a table for the top 30 most frequently mentioned terms, a bar plot to compare the frequency of the top 20 terms, and a wordcloud to visualize the results.

¹https://github.com/sirmurphalot/STOR765_Fall2022_TeamDataScience

2.1 FINDINGS FOR UNI-GRAMS

Figure 2.1: Bar Plot of Top 20 Most Mentioned Uni-grams

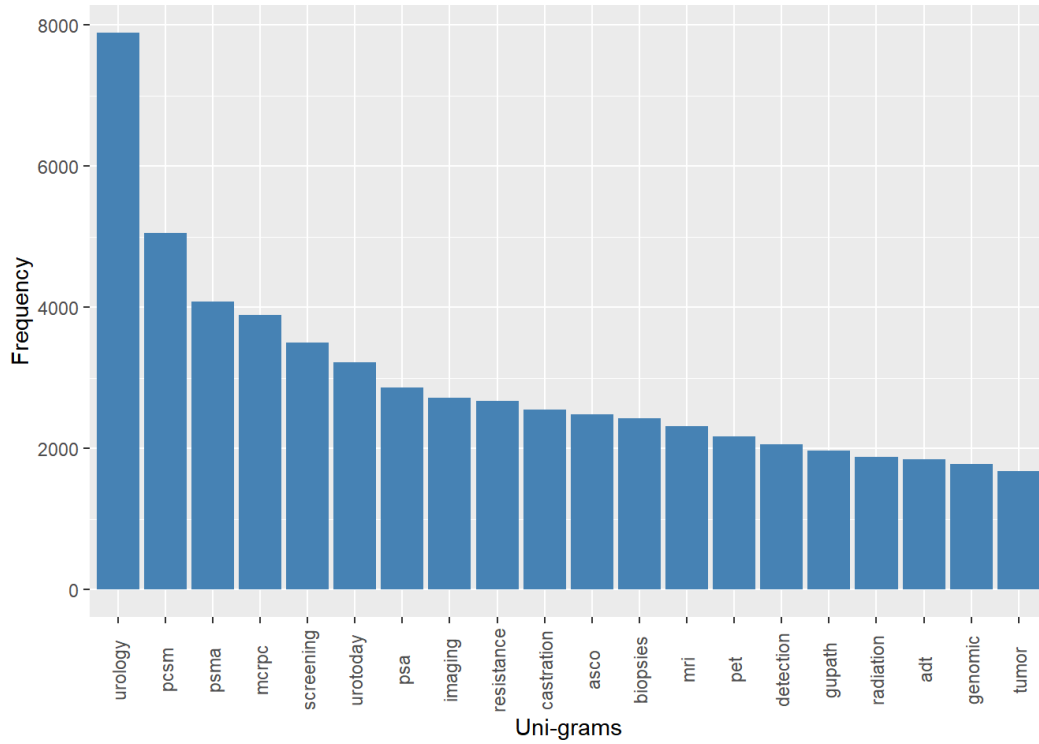


Table 2.1 contains the 30 most frequently mentioned uni-grams. The first column contains the root form (stemmed), which allows finding inflected variants of a word, of the uni-grams. The second column is the original word. The third column contains the frequency of each term. The following tables for bi- and tri-gram follow the same structure. To compare the frequency of the phrases, we create a bar plot of the top 20 most frequently mentioned term given in Figure 2.1. The x-axis shows the uni-grams. The y-axis is the frequency of each phrase. The bar plots for bi- and tri-grams is drawn similarly. From Figure 2.1, we can easily see that the term “urology,” with a frequency of 7889, stands out compared to the other uni-grams. Even compared to the second most mentioned uni-gram, “pcsm,” with a frequency of 5050, there is a huge drop. For the other terms we see a gradual decrease in the frequencies.

Figure 2.2: Wordcloud of Uni-grams



The results for uni-grams are visualized as a wordcloud as shown in Figure 2.2. In the wordcloud, larger font indicates that the term appears more commonly. The colors of the texts does not have any meaning, but provide visual separation of the phrases. In the wordcloud, “urology” is again shown as the most frequently mentioned topic. Other highly mentioned uni-grams are abbreviations of commonly used terminologies in the field of prostate cancer, such as “PCSM” (prostate cancer–specific mortality) and “PSMA” (prostate-specific membrane antigen PET imaging scan). Looking at the figure, we can easily see many terms are related to PSMA, such as “screening,” “imaging,” “psa,” “detection,” and “genomic.” This shows that there is an interest in the screening, detection, and monitoring of prostate cancer. We also see that “mCRPC,” metastatic castration-resistant prostate cancer, is highly mentioned. Terms related to mCRPC, “resistance” and “castration,” also appear in the figure. Terms like “UroToday,” which is an online community for urologists, and “ASCO,” American Society of Clinical Oncology, provide insight into possible interested invitees for a conference. Other terms on this list are possible treatments and medication for prostate cancer; “prostatectomies,” “radiotherapies,” and “enzalutamide.” The frequency of uni-grams provides insight into general and specific topics of interest and possible invitees to a conference.

2.2 FINDINGS FOR BI-GRAMS

Figure 2.3: Bar Plot of Top 20 Most Mentioned Bi-grams

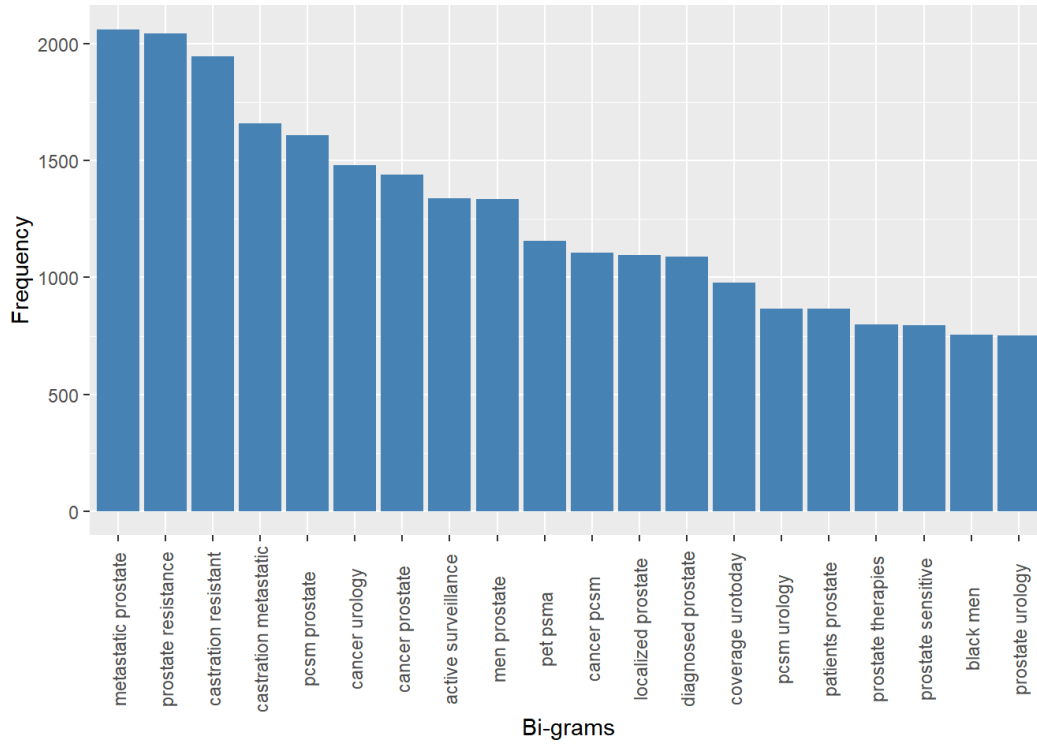


Table 2.2 and Figure 2.3 contain the result for bi-grams (two words that appear together in either order). When looking at the bi-grams note that the words have been rearranged in alphabetical order. Tri-grams are treated similarly. The result of bi-grams adds support to our findings from the uni-gram as most are similar topics. Looking at Figure 2.3, three phrases appear most frequently with around 2000 mentions. All three, ‘metastatic prostate,’ “prostate resistance,” and “castration resistance,” are related to “mCRPC.” The next two commonly mentioned phrases have a similarly high frequencies around 1600. The term “castration metastatic” is also related to “mCRPC.” The other term, “pcsm prostate,” supports our finding about PCSM from the results of uni-grams. After, we see a gradual decrease in the frequencies of the terms.

Figure 2.4: Wordcloud of Bi-grams

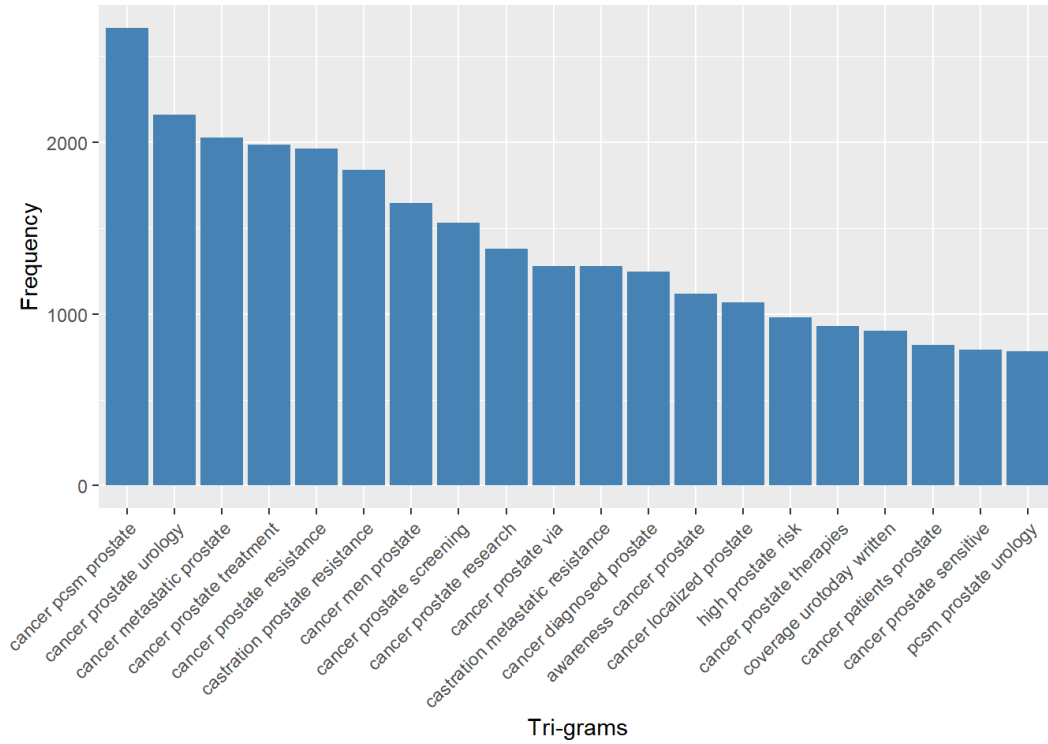


Figure 2.4 is the wordcloud for bi-grams. As discussed above, the terms related to “mCRPC” appear in larger fonts. There are a few interesting findings from the wordcloud. The first is that while urology was the most mentioned term with uni-grams, it is no longer prevalent in the wordcloud for bi-grams. This can be interpreted that the single word “urology” is a topic itself. Therefore, there are no set combination of two words, including “urology,” that is frequently mentioned.

Secondly, the result of bi-grams identifies topics that were not captured using uni-grams. For instance, the term “active surveillance,” which refers to close monitoring of prostate cancer [1], is a topic that is not identified with uni-grams. Another is “localized prostate,” which refers to localized prostate cancer, cancer that is only inside the prostate gland and has not spread to other parts of the body [2]. The term “localized prostate” indicates a strong discussion of either hormone- or castration- localized cancer. The bi-gram “prostatectomies radical” shows that radical prostatectomy treatment is frequently mentioned. Such cases show that the combination of words is an important factor to consider when using text data and allows us to identify specific topics in prostate cancer.

2.3 FINDINGS FOR TRI-GRAMS

Figure 2.5: Bar Plot of Top 20 Most Mentioned Tri-grams



The results for tri-grams, three words that appear together in any order, are given in Table 2.3 and Figure 2.5. From Figure 2.5, the tri-gram “cancer pcrsm prostate” is significantly highly mentioned compared to other tri-grams. The second group of four phrases with frequencies around 1800 contains more general terms related to urology and mCRPC; “cancer prostate urology,” “cancer metastatic prostate,” “cancer prostate treatment,” “cancer prostate resistance”. We see a similar pattern in the gradual decrease in frequencies with the following phrases as with uni- and bi-grams. While, the results further substantiate our findings in the previous two method, most of the findings here are redundant with our findings from uni- and bi-grams.

3 DATA AND METHODOLOGY

3.1 DATA

For this study, we use data provided by ZS Associates. It is a collection of Twitter data across three medical fields: prostate cancer, asthma, and IBD. The entire data has 272397 tweets. In this study, we focus on just the prostate cancer data, which contains 70159 tweets. The other two medical fields of data is used to eliminate general phrases from prostate cancer data that do not provide meaningful insight.

The data includes tweets that were written between January 1st, 2017 to May 30th, 2022. The data set includes original posts, replies, and retweets (sharing the original post of another person). However, as in this study, the goal is to create a dictionary for possible trending topics, we believe that replies and retweets indicate interest in the topic by the users. Therefore, we treat the different types of tweets as the same and count each appearance of a phrase in any type of tweet as a single frequency.

3.2 DATA PRE-PROCESSING

To use natural language, applying the appropriate feature extraction method is a crucial step since lines of text can not be congregated to identify trending topics. Before we can apply the chosen feature extraction method we need to pre-process text data [5].

Table 3.1 is an example to help understand the steps of pre-processing text data. The first row shows the raw scraped text. In the second row, we see the partially cleaned text data provided by ZS Associates. In this stage they removed mentions (using the @), tags (using #), and URL links. Based on the ZS Associates' cleaned data, we further clean the text by removing numbers, punctuation, special characters, whitespace, and stopwords. We also change all text to lowercase. This step is shown in the third row. Note that in this step hyphenated words like "castration-resistance" would be cleaned as "castration resistance" and treated as two words.

The final step is stemming, which is shown in the last row. In the English language, a single word can take different forms. For instance, the word "change" could appear in the text in forms like "changing," "changed," and "changer." However, in feature extraction, all these forms should be counted as a single term as they have the same root and meaning. Therefore, we would stem all these variations in the form of "chang." The same process is repeated over all words in the text. For the pre-processing stage, we use the `tm` and `stringr` package in **R**. The `clean.up` function is defined in the **R** code provided, which can be used to clean and stem text data. For a deeper

understanding of text pre-processing, one can refer to the paper Vijayarani et al. (2015)[5].

Original Text	What is clinical relevance of new bone lesions that appear during AR inhibitor therapy in men with metastatic prostate cancer? See our latest study! https://t.co/eWooK1pMss in @JMAOnc @DukeGUCancer
Partially Cleaned Text	What is clinical relevance of new bone lesions that appear during AR inhibitor therapy in men with metastatic prostate cancer? See our latest study! in
Cleaned Text	clinical relevance new bone lesions appear ar inhibitor therapy men metastatic prostate cancer see latest study
Stemmed Text	clinic relev new bone lesion appear ar inhibitor therapy men metastat prostat cancer see latest studi

Table 3.1: Steps of Pre-Processing Data

3.3 *Bag of n-Gram*

For this project, we decide to use the *bag of n-grams* method to extract topics. The reason why we use n-gram is to capture words that appear together. For instance, with topics like "castration resistant," having one word like "castration" and "resistant" would not have much value in topic detection. Therefore, we need to look at word combinations. The *bag of n-grams*, which is the count of n words together, allows us to capture word combinations. For instance, taking the example given in Table 3.1, for the last row if we set $n = 2$ (bi-grams), we get n-grams like "clinic relev," "relev new," and "new bone." and the frequency of each term is one. As there are 16 words in this text, we have 15 unique bi-grams. After repeating this process for all the rows, we combine them to create one frequency table for all the tweets. For a detailed explanation of this method refer to Jurafsky and Martin (2021) [4].

After looking at the given list of topics from ZS Associates, most of the terms were less than three words. Therefore, we decide to look at uni-, bi-, and tri-grams. The extraction is done using the ngram package in **R**. For these three methods, the preliminary results shows the top 30 most frequently mentioned terms for uni-, bi-, and tri-grams as shown in Tables 4.1, 4.2, and 4.3. We presented the preliminary results to ZS Associates, and based on some discussions, we made a few adjustments. With the finalized result, we visualize our findings into wordclouds using the wordcloud2 package (shown in Figure 2.2, 2.4, and 2.6).

3.4 ADJUSTMENTS

The preliminary result had some issues that needed to be adjusted for. First, the preliminary top frequent n-grams contained terms that were too general and did not provide any insight. Some of these terms were general words used in the English language and others were commonly used words in the medical field. This was most common with uni-grams. The frequent terms that appeared were words like “people,” “patient,” and “trial.” These words do not provide meaningful insight into the topics that are of interest to the client. Hence, to remove such general terms, we use the other two data sets provided by ZS Associates in the medical field of asthma and inflammatory bowel disease (IBD). We repeat the data pre-processing and n-gram extraction process for the two sets of data. After creating a frequency table for each topic, we only keep the terms that appear in both the asthma and IBD lists. We remove the top 50 most commonly mentioned terms from that list to focus on frequent terms for prostate cancer.

The second issue is with stemmed versions of the word. When conducting analysis on text data, it is customary to work with a stemmed version of the word as we want the words with the same root to be counted as the same. However, it does have a disadvantage in that it hurts readability. For instance, for terms like ‘prostat” or “diagnosis,” it is difficult to understand what these words mean. Therefore, in the results, we also provide the unstemmed, original version, of the n-grams.

Third, after conferring with the client, we decide to ignore word order and discard repeating terms. For this project, ZS Associates confirmed that the word order did not provide additional information. Some instances of this were prevalent with tri-grams. For instance, “resist prostate cancer” and “prostate cancer resist” could be regarded as the same topic. However, disregarding the order of words should be done with caution when working with text data. Typically for n-grams, word order contains essential information. For example, looking at two very simple sentences “I like you” and “You like me,” if we disregard the order of the words, the two sentences would be viewed as the same despite having different meanings. For this project, as we are only interested in topic extraction and as the client confirmed that information would not be lost when we disregard word order, we rearrange each bi- and tri-gram in an alphabetical order. As for the example above, it is rearranged as “cancer prostate resist.” In addition, we also discard repeating terms. In the pre-processed data, as stopwords are removed and the words are stemmed, there are instances where the same word would appear repeatedly. The most common is with the word “urology” (i.e, “urology urology”). As we believe, that repeating words would not provide meaningful topics, we discard these terms.

4 CONCLUSION

This project successfully identifies widely discussed topics in prostate cancer using Twitter data. Looking at the topics identified by the bag of n-grams, some of the topics do overlap with the list of topics provided by the client, which shows that this method does capture the important topics in prostate cancer. In addition, this project also identifies topics that are not included in the given list that the client may choose to expand into as possible topics for seminars or talks. If the client chooses to implement this method into data sets for asthma and IBD, it can be easily replicated using the **R** code provided in the repository. The pre-processed data is used in the other two reports by Alexander Murph and Taebin Kim. With a pre-processed text, the client would be able to implement other forms of feature extraction methods and analysis if desired.

REFERENCES

- [1] Observation or active surveillance for prostate cancer. <https://www.cancer.org/cancer/prostate-cancer/treating/watchful-waiting.html#:~:text=Active/20surveillance/20is/20often/20used,to/203/20years/20as/20well>. Accessed: 2022-11-02.
- [2] Treating localized prostate cancer. <https://effectivehealthcare.ahrq.gov/products/prostate-cancer-therapies-update/consumer>. Accessed: 2022-11-02.
- [3] Ilkhanian M Chowdhury-Paulino, Caroline Ericsson, Randy Vince Jr, Daniel E Spratt, Daniel J George, and Lorelei A Mucci. Racial disparities in prostate cancer among black men: epidemiology and outcomes. *Prostate Cancer and Prostatic Diseases*, 25(3):397–402, 2022.
- [4] D Jurafsky and J Martin. N-gram language models. *Speech and Language Processing*, 2021.
- [5] S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

Uni-gram	Phrase	Frequency
urolog	urology	7889
pcsm	pcsm	5056
psma	psma	4080
mcrpc	mcrpc	3893
screen	screening	3496
urotoday	urotoday	3224
psa	psa	2867
imag	imaging	2716
resist	resistance	2673
castrat	castration	2548
asco	asco	2482
biopsi	biopsies	2431
mri	mri	2320
pet	pet	2175
detect	detection	2058
gupath	gupath	1974
radiat	radiation	1881
adt	adt	1844
genom	genomic	1775
tumor	tumor	1677
androgen	androgen	1644
stage	stage	1585
oncolog	oncology	1526
surveil	surveillance	1515
grade	grade	1505
prostatectomi	prostatectomies	1462
written	written	1434
hormon	hormone	1407
radiotherapi	radiotherapies	1337
enzalutamid	enzalutamide	1237

Table 2.1: Top 30 Most Frequent Uni-grams

Bi-Gram	Phrase	Frequency
metastat prostat	metastatic prostate	2062
prostat resist	prostate resistance	2044
castrat resist	castration resistant	1947
castrat metastat	castration metastatic	1659
pcsm prostat	pcsm prostate	1609
cancer urolog	cancer urology	1481
cancer prostat	cancer prostate	1440
activ surveil	active surveillance	1338
men prostat	men prostate	1337
pet psma	pet psma	1158
cancer pcs	cancer pcs	1107
local prostat	localized prostate	1096
diagnos prostat	diagnosed prostate	1090
coverag urotoday	coverage urotoday	978
pcsm urolog	pcsm urology	866
patient prostat	patients prostate	865
prostat therapi	prostate therapies	800
prostat sensit	prostate sensitive	796
black men	black men	755
prostat urolog	prostate urology	753
prostat via	prostate via	752
prostatectomi radic	prostatectomies radical	752
prostat treatment	prostate treatment	732
cancererawarenessmonth prostat	cancererawarenessmonth prostate	711
androgen depriv	androgen deprivation	689
prostat stage	prostate stage	686
ct pet	ct pet	660
lu psma	lung psma	643
mcrpc prostat	mcrpc prostate	582
depriv therapi	deprivation therapies	566

Table 2.2: Top 30 Most Frequent Bi-grams

Tri-Gram	Phrase	Frequency
cancer pccsm prostat	cancer pccsm prostate	2666
cancer prostat urolog	cancer prostate urology	2160
cancer metastat prostat	cancer metastatic prostate	2030
cancer prostat treatment	cancer prostate treatment	1988
cancer prostat resist	cancer prostate resistance	1962
castrat prostat resist	castration prostate resistance	1838
cancer men prostat	cancer men prostate	1645
cancer prostat screen	cancer prostate screening	1531
cancer prostat research	cancer prostate research	1382
cancer prostat via	cancer prostate via	1281
castrat metastat resist	castration metastatic resistance	1279
cancer diagnos prostat	cancer diagnosed prostate	1248
awar cancer prostat	awareness cancer prostate	1118
cancer local prostat	cancer localized prostate	1069
high prostat risk	high prostate risk	979
cancer prostat therapi	cancer prostate therapies	932
coverag urotoday written	coverage urotoday written	902
cancer patient prostat	cancer patients prostate	821
cancer prostat sensit	cancer prostate sensitive	790
pccsm prostat urolog	pccsm prostate urology	782
cancer mcrpc prostat	cancer mcrpc prostate	774
cancer diagnosi prostat	cancer diagnosis prostate	742
cancer prostat treat	cancer prostate treatment	720
cancer detect prostat	cancer detection prostate	685
cancer prostat stage	cancer prostate stage	679
cancer dispar prostat	cancer disparities prostate	615
cancer prostat risk	cancer prostate risk	587
androgen depriv therapi	androgen deprivation therapies	564
cancer prostat recurr	cancer prostate recurrence	522
cancer imag prostat	cancer imaging prostate	508

Table 2.3: Top 30 Most Frequent Tri-grams

Uni-gram	Phrase	Frequency
men	men	10190
patient	patients	9941
urolog	urology	7889
treatment	treatment	7486
risk	risk	6639
studi	studies	6357
trial	trial	5717
new	new	5581
dr	dr	5564
therapi	therapies	5408
pcsm	pcsm	5056
metastat	metastatic	4859
discuss	discuss	4609
present	presentation	4538
research	research	4465
year	year	4180
psma	psma	4080
clinic	clinical	3991
mcrpc	mcrpc	3893
work	work	3854
use	use	3772
high	high	3586
can	can	3557
screen	screening	3496
will	will	3440
gu	gu	3263
health	health	3257
advanc	advanced	3228
urotoday	urotoday	3224
great	great	3023

Table 4.1: Preliminary Top 30 Most Frequent Uni-grams

Bi-gram	Phrase	Frequency
risk prostat	risk prostate	2460
urolog urolog	urology urology	2227
advanc prostat	advanced prostate	2106
metastat prostat	metastatic prostate	2062
resist prostat	resistance prostate	2044
cancer patient	cancer patients	1990
castrat resist	castration resistance	1947
metastat castrat	metastatic castration	1659
pcsm prostat	pcsm prostate	1609
high risk	high risk	1577
cancer urolog	cancer urology	1481
cancer research	cancer research	1468
cancer prostat	cancer prostate	1440
cancer treatment	cancer treatment	1396
written coverag	written coverage	1361
activ surveil	active surveillance	1338
men prostat	men prostate	1337
cancer screen	cancer screening	1298
psma pet	psma pet	1158
cancer pcs	cancer pcs	1107
local prostat	localized prostate	1096
diagnos prostat	diagnosed prostate	1090
cancer awar	cancer awareness	1016
coverag urotoday	coverage urotoday	978
et al	et al	924
clinic trial	clinical trial	891
urolog pcs	urology pcs	866
patient prostat	patients prostate	865
therapi prostat	therapies prostate	800
sensit prostat	sensitive prostate	796

Table 4.2: Preliminary Top 30 Most Frequent Bi-grams

Tri-gram	Phrase	Frequency
risk prostat cancer	risk prostate cancer	2409
advanc prostat cancer	advanced prostate cancer	2074
metastat prostat cancer	metastatic prostate cancer	2030
resist prostat cancer	resistance prostate cancer	1962
castrat resist prostat	castration resistance prostate	1838
prostat cancer patient	prostate cancer patients	1822
pcsm prostat cancer	pcsm prostate cancer	1580
prostat cancer urolog	prostate cancer urology	1445
men prostat cancer	men prostate cancer	1313
metastat castrat resist	metastatic castration resistance	1279
prostat cancer treatment	prostate cancer treatment	1278
prostat cancer research	prostate cancer research	1217
prostat cancer screen	prostate cancer screening	1157
prostat cancer pcs	prostate cancer pcs	1086
local prostat cancer	localized prostate cancer	1069
diagnos prostat cancer	diagnosed prostate cancer	1067
cancer urolog urolog	cancer urology urology	1027
high risk prostat	high risk prostate	979
cancer prostat cancer	cancer prostate cancer	946
prostat cancer awar	prostate cancer awareness	946
written coverag urotoday	written coverage urotoday	902
prostat cancer prostat	prostate cancer prostate	829
urolog urolog pcs	urology urology pcs	828
patient prostat cancer	patients prostate cancer	821
sensit prostat cancer	sensitive prostate cancer	790
urolog pcs prostat	urology pcs prostate	782
therapi prostat cancer	therapies prostate cancer	753
cancer awar month	cancer awareness month	743
via prostat cancer	via prostate cancer	743
urolog prostat cancer	urology prostate cancer	715

Table 4.3: Preliminary Top 30 Most Frequent Tri-grams