Framework for Using Social Media to Better Understand Health

A Preprint

Curtis Murray
The University of Adelaide
Adelaide, SA 5005
curtis.murray@adelaide.edu.au

Lewis Mitchell
The University of Adelaide
Adelaide, SA 5005
lewis.mitchell@adelaide.edu.au

Jonathan Tuke
The University of Adelaide
Adelaide, SA 5005
simon.tuke@adelaide.edu.au

Mark Mackay
The University of South Australia
Adelaide, SA 5000
mark.mackay@unisa.edu.au

May 30, 2021

Abstract

Health care discourse on social media presents an opportunity to gain a unique perspective on increasingly desired patient-reported experiences, complementing traditional survey data. These social media reports often appear as first-hand accounts of patients' journeys through health care, whose details extend beyond the confines of structured surveys, and at a far larger scale than focus groups. Natural language processing equips us with the tools to analyse data of this kind, providing us with methods to gain insights into social media discourse. These kinds of analyses have demonstrated the benefits of using social media as a tool to monitor health care. However, there is a need to develop a general reproducible framework to guide further analyses. In this paper we propose the Design-Acquire-Process-Model-Analyse-Visualise (DAPMAV) framework to meet this need.

***K*eywords** Digital Health · Social Media · Patient Experiences · Natural Language Processing · Topic Modelling

1 Introduction

[[CM: Introduce social media data]]

Social media is a vibrant forum and a ubiquitous platform hosting the discussion of topics related to social issues. These discussions are information-rich, and can provide us with insight into how complex issues are perceived at individual and community levels. Social media is rapidly being adopted by researches as a tool to explore these social issues [1, 2].

Twitter, a social media giant, is also often the subject of research use [3]. On Twitter, users make posts online in the form of **tweets**. Prior to October 2018, these tweets were limited to 140 characters, and now are limited to 280 characters. Tweets may contain **hashtags** that allow users to tag keyword phrases that correspond to certain topics or issues. Reddit, colloquially known as the "front page of the internet", is another social media platform that receives attention from researches [4]. Discussion on Reddit occurs in a different form to that on Twitter. On Reddit, users post to communities that cover certain topics known as **subreddits**. A user may create a new post to a subreddit, or may post reply to another user's post or reply. Twitter and Reddit's widespread use and data policies make them suitable platforms for research over other platforms such as Facebook, whose data policies are more prohibative for research [1]

Twitter is more frequently the subject of research than Reddit (7,790,000 hits for "Twitter" on Google Scholar 1,760,000 for "Reddit", equalling 4.4 times more Twitter articles). One possible mechanism for this may be Twitters greater

[1] https://www.redditinc.com/policies/privacy-policy, https://twitter.com/en/privacy, [[CM: Facebook?]]

popularity. In 2020 there were 192 million daily Twitters users, compared to the 52 million daily Reddit users (3.7 times more Twitter users) [2]. However, Twitter has no prescribed community structure for the discussion of specific topics. Instead, topics are optionally discussed through the use hashtags. This approach allows posts to be clearly identified as being about multiple topics at the same time. Reddit on the other hand has no clear method for cross-topic discussion. Reddit's structure instead necessitates posting to exactly one subreddit, which usually hosts most of the discussion around that topic, creating a sense of community. One study noticed that in the context of the "Me too" movement, discussion on Twitter was less personal than on Reddit, and instead focussed on encouraging others to speak up [5].

Patient-reported experiences are vital in providing quality healthcare that matches the public's expectations. Surveys and focus groups are the traditional tools of choice for capturing patient-reported experiences. However, surveys are costly and infrequent, and focus groups are time-consuming, and hence difficult to scale to get a population-level understanding of patient experiences.

Discussion around healthcare is widespread on social media. Here, these patients have the scope to talk about what is important to them in as much detail as they want. Due to its anonymity and lack of direct face-to-face interactions, social media has been evidenced to elicit high levels of self-disclosure in people, who tend to reveal themselves more intimately than in face-to-face interactions. [6]. This makes social media a useful source of information that can reveal personal patient-reported experiences to researchers, and as such to improve health-care systems. There have been calls made for more researchers to start mining social media data as a complementary tool to improve health-care systems [7].

A significant obstical that obstructs this the lack of structure inherent in social media data. In particular, free-text data lacks obvious structure that allows for simple analysis. Natural language processing is a field that aims to address this lack of structure by using statistical and machine learning models to uncover latent structure. These models allow us to analyse text data.

These tools are often applied by researches who have a technical background. However, many of these tools are inaccesible to a non-techinal person, and as such many health care researchers and practioners without the background neccesary to apply these tools are unable to mine social media text to uncover patient-reported experinces. This motivates the need for a generalised reproducible framework. To meet this need, we introduce the DAPMAV framework. The DAPMAV framework outlines the complete process from design of problem through to visualisation of results for text analysis from social media discourse.

[[MM: The book "healthcare data analytics" by Reddy and Aggarwal devotes a chapter to social media analysis. But provides no framework for doing this work. And this is the missing piece which the framework fills.]]

[[CM: Introduce prostate cancer]]

Prostate cancer is type of cancer that can effect persons with a prostate, although data is typically limited to cis men. Aside from non-melanoma skin cancers, prostate cancer is the most commonly diagnosed form of cancer in Australia (21,808 diagnoses in 2009). It is also the fourth leading cause of death amongst Australian men[[CM: CITE: AIHW]] (3,294 deaths in 2011).

2 DAPMAV Framework

2.1 D - Design

The design stage is the first stage in the DAPMAV process. In the design stage, we choose a relevant subreddit (or other possible social media data source). We then choose the time-frame to collect the posts.

One potential way to select the subreddit would be use the Reddit search bar to search for the overall theme of the discussion, and select a relevant community from the communities tab. For example, say we are interested in looking at discussion on Prostate Cancer. We type "Prostate Cancer" into the search bar, and navigate to the communities tab. We notice that there is a specific prostate cancer community, called "/r/ProstateCancer". We can click on this community and explore the posts to ensure that they are relevant to the conversation we want to capture. If so, we move on to specifying our timeline.

We can see when the subreddit was created by observing the top box on the right hand side of the screen when viewing the subreddit. We might find that the subreddit we are looking at is too new for our purposes, so we could consider different subreddits.

[2] [[CM: How can we cite this kind of stuff?]] `https://www.wsj.com/articles/reddit-claims-52-million-daily-users-revealing-a-key-figure-for-social-media-platforms-11606822200`

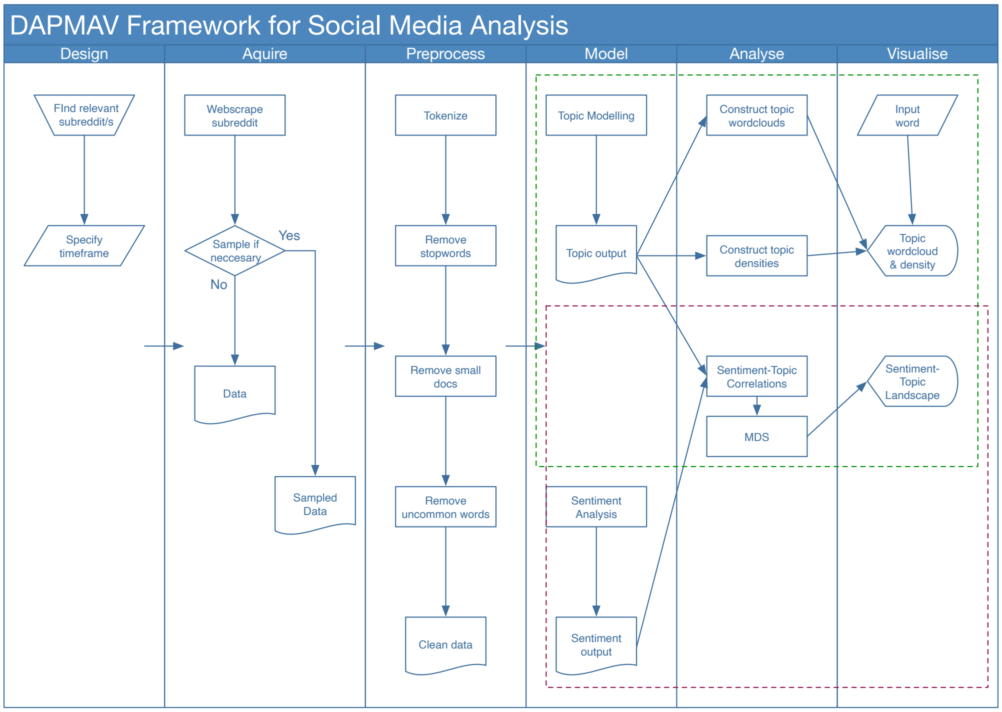

Figure 1: DAPMAV framework for analysis of social media discussion relating to patient-reported experiences.

2.2 A - Acquire

Data acquisition is the second stage of the DAPMAV process, where we retrieve the Reddit posts in the subreddit and time specified in the design phase that we wish to analyse[3]. We obtain the Reddit posts using the Pushshift API [8], an archiving platform that provides researchers with Reddit data.

2.2.1 Sampling and Filtering

Computationally demanding algorithms inhibit rapid exploration of large text datasets. To avoid lengthy processing, we can thin out our text data to reduce the size of the dataset, if necessary. As a means of thinning out our dataset we can sample the documents in our corpus without replacement. The number of samples we take will determine the speed of our analysis. As a rule of thumb, we can reduce our documents down to [[CM: Insert number here]].

Reddit posts also may be tagged in certain ways. As an example, the subreddit "/r/COVID19Positive" can have posts tagged with the flairs; Question-for medical research, Tested Positive - Family, Presumed Positive - From Doctor, Tested Positive - Me, Question-to those who tested positive, and finally Tested Positive. These flairs give us additional data about that post that we may want to make use of. In particular, if we are only in posts related to a particular flair, we can filter out the other flairs to reduce the run time of the models, and potentially improve the specificity of the results, although it should be noted that in cases with limited data it may improve the models to keep the other flairs in.

2.3 P - Preprocess

[[CM: Preprocessing or Processing?]]

2.3.1 Tokenize

In order to model the data we have acquired, we must preprocess it to be in a useable form. **Tokenization** is the process of deconstructing text into a list of smaller units, called **tokens**. A token may be a unigram (a single word), a bi-gram (two consencutive words), more generally an n-gram (n consecutive words), a sentence, and many other things. Most commonly we will use unigrams as tokens.

[3] [[CM: Should we have the Sampling in this step (Acquire) or preprocessing? I would have thought preprocessing but that box is getting a bit big in Figure 1]]

2.3.2 Remove Stopwords

When conducting text analysis, we may wish to remove a list of words, denoted **stopwords** from the corpus. Words that are often considered stopwords are common words that add little contextual information. Examples of stopwords include 'the', 'and', 'of', and 'in'.

2.3.3 Remove Small Docs

Excessively small documents contain no valuable information, and dilute the effectiveness of our results, while increasing the computational complexity. As such we can remove them from our analysis to improve speed and reduce the noise, which is incurred by treating small documents as being as important as larger documents.

2.3.4 Remove uncommon words

Words that appear rarely in a corpus add little information to our data, especially when we are concerned with looking at a global overview of what is being talked about. Word frequencies are often very heavily tailed distributions[4], meaning that there are often a large number of rare words. As these words are rare, we have little information about the ways in which they are used, and as such, results that we draw relating to them will be often contain substantial noise. Therefore, as a method to reduce the run time of the models being employed at little perceivable cost, we can reduce the size of our vocabularies by filtering out particularly rare words.

2.4 M - Model

2.4.1 Topic Modelling with hSBM

Large amounts of unstructured text data pose a challenge to researchers as their complexity is too high for easy digestion. Topic modelling is a natural language processing tool that allows us to summarise large amounts of text data. We can apply it to social media data to get an overview of what topics are discussed at global and individual post levels. It does so without the need for any domain knowledge, acting in what is known as an unsupervised way. This allows the automatic detection of important topics of discussion; telling us what words are in the topics, and how the topics are dispersed through the posts.

2.4.2 Sentiment Analysis

[[CM: Include this?]]

Sentiment analysis is another natural language processing tool, that like topic modelling reduces the complexity of documents, here instead by breaking them down into the emotions that are conveyed in the text. This is important as it allows the the identification of how people feel about the issues that they talk about.

[[CM: Explain sentiment analysis]]

2.5 A - Analyse

[[CM: Not sure if this should be here - maybe we just need MDS]]

2.5.1 Construct Topic Wordclouds

2.5.2 Construct Topic Densities

2.6 Dimension Reduction

2.7 Visualise

2.7.1 Topic wordcloud and density

To produce a topic wordcloud we make use of the topic distributions found from topic modelling. These tell us the probability that a word is in a topic, $P(\text{word}|\text{topic})$. From this, we can construct a wordcloud as a visualisation tool that shows the words contained in the topic, with their sizes being related to probability that the word is selected in the topic.

[4] [[CM: cite]]

2.7.2 Sentiment-Topic Landscape

3 Appling the DAPMAV framework to discussion on prostate cancer

[[CM: Outline for what this section might look like.]]

- We can understand how this disease affects people by observing their discussion online.
- To do so we make use of the DAPMAV framework and show how it allows us to understand this issue.
- Step through the stages of DAPMAV? Or jump straight to results?

4 Results

[[MM: Results from Prostate Cancer data]]

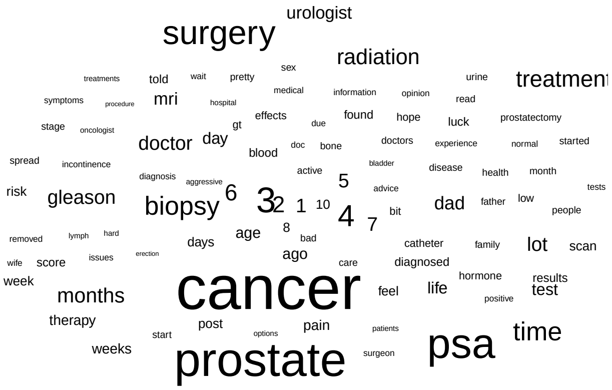

Figure 2:

Overall word frequencies in text give significant insight into what the text is about. A wordcloud is an effective tool for visualising word frequencies. Figure 2 is a wordcloud that captures the overall conversation on /r/ProstateCancer, with the size of the words given as their empirical frequency[[CM: Actually its $p^{(}1/0.7)$ as per ggwordcloud's default - not sure if should mention this or not]].

[[CM: Talk about age/people in posts (Figure 3 and Figure 4]]

[[CM: Topic modelling overview]]

Topic modelling reveals the hierarchical structure of topics discussed in the prostate cancer data. Figure 5 displays this structure in a radial tree, where each root node is a word whose size is given by its empirical use, and hierarchical topic membership is defined by connection to parent nodes. An interactive version of this is available in the supplementary materials[5].

[[CM: 2d projection of topics, similar to interactive version but improved in part because it shows the topic-interactions at each level in the hierarchy (once it's working properly). Useful as it gives an overview of what is being talked about]]

[5] https://curtis.pw/apps/topic_network.html

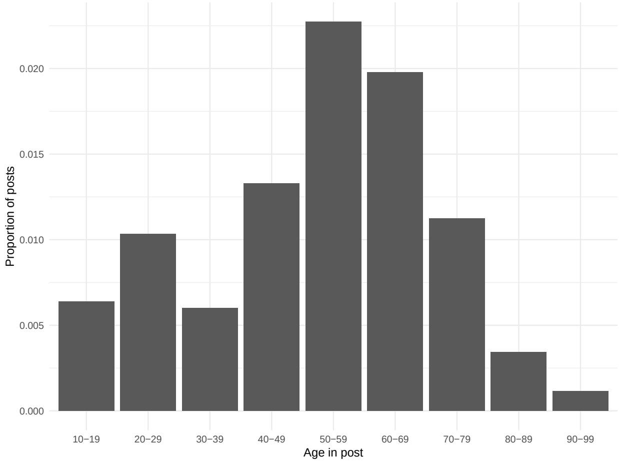

Figure 3: Proportions of time each age range is mentioned. [[CM: If one post mentions 20, and another mentions 20 and 30, the total counts will be 20: 2, 30: 1. We should probably have this as 20: 1.5, 30: 0.5. Need to change code to do this.]] [[CM: I'm not sure if these Figures are worth including as they don't actually come from the framework, it's just a bit of REGEX.]]

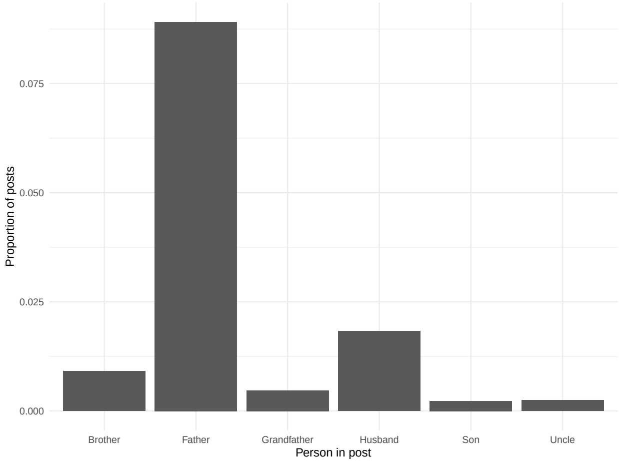

Figure 4: Proportions of time male figures are mentioned. [[CM: Same as ages post, need to get age probability instead of count before aggregating]]

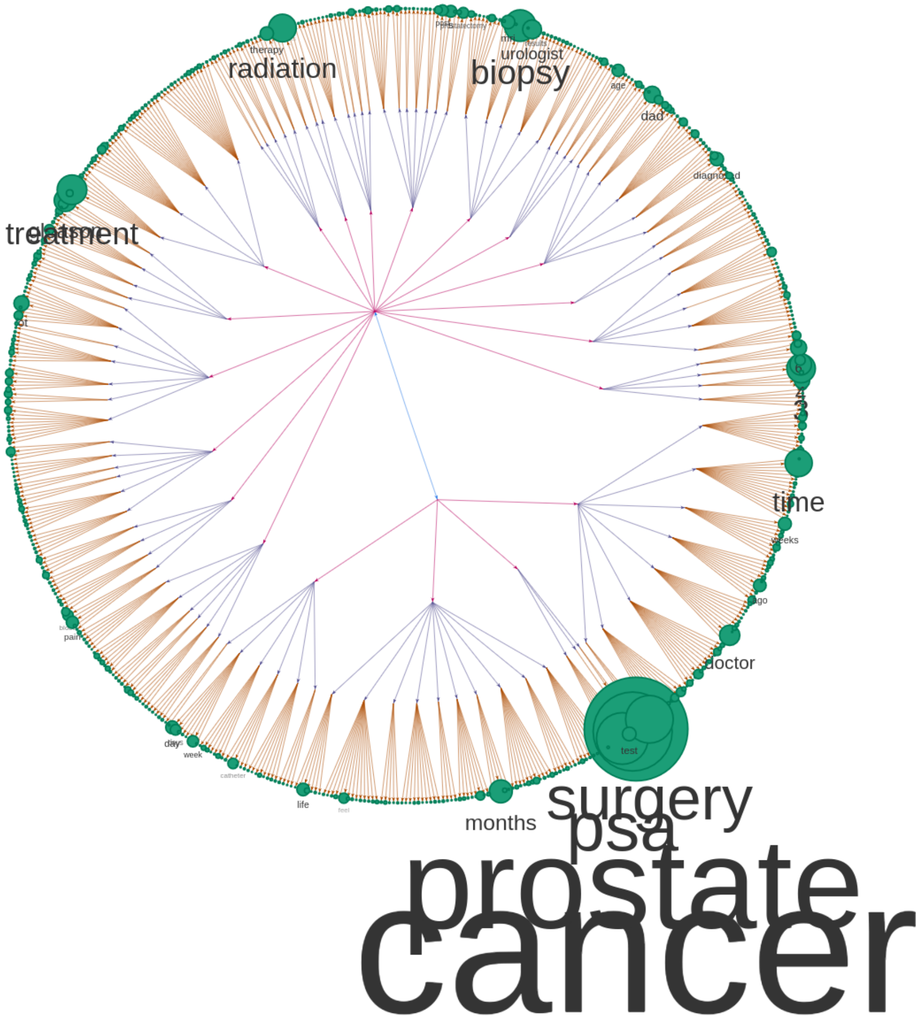

Figure 5: Topic structure of prostate cancer data revealed in a radial tree.

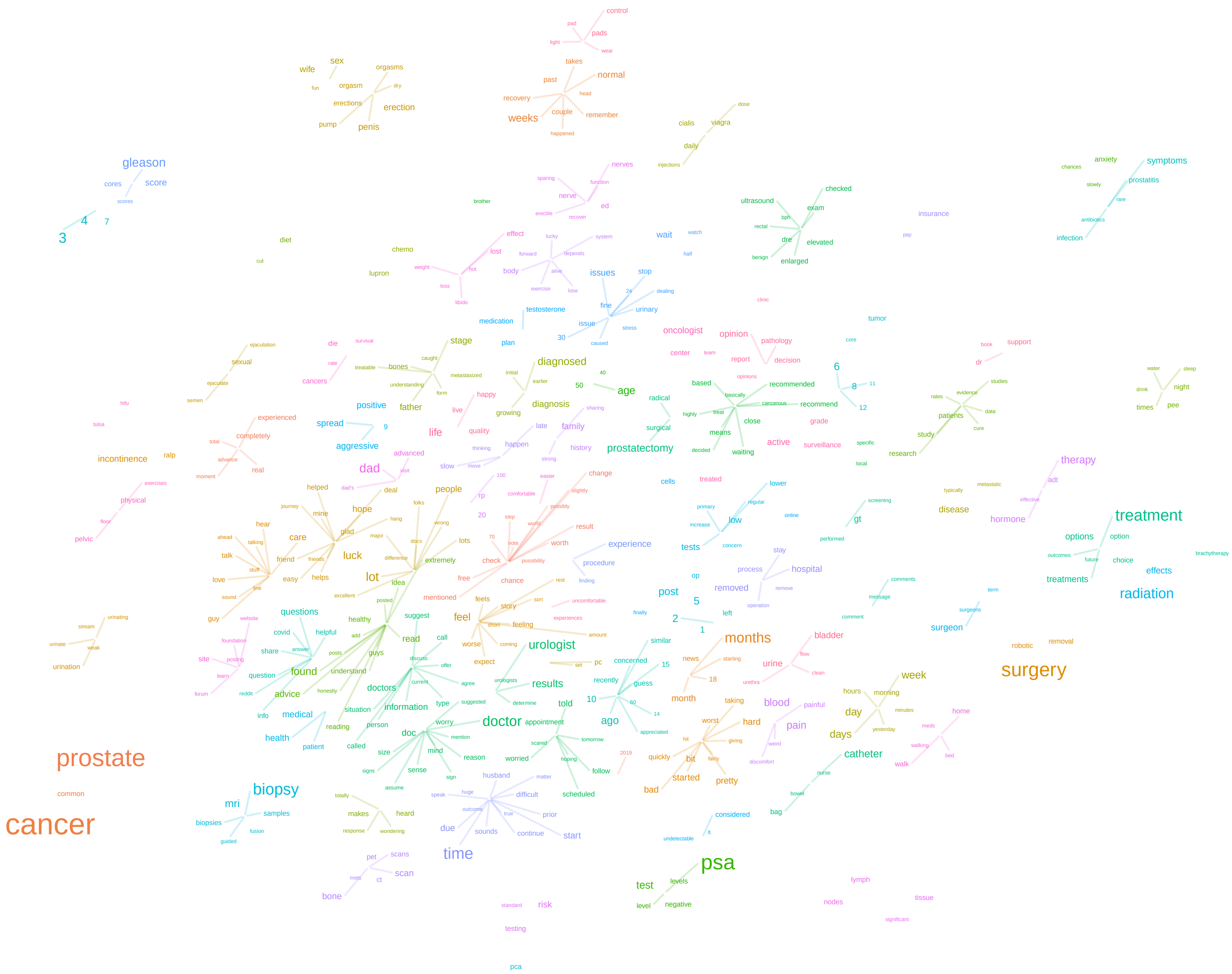
control
pad
pads
light
wear
takes
sex
wife
orgasms
normal
past
orgasm
dry
fun
head
recovery
erections
erection
couple
remember
weeks
pump
penis
happened
dose
cialis
viagra
daily
injections
gleason
cores
score
scores
nerves
sparing
function
nerve
ed
brother
erectile
recover
3
4
7
ultrasound
checked
exam
bph
rectal
dre
elevated
benign
enlarged
anxiety
symptoms
chances
prostatitis
slowly
rare
antibiotics
infection
insurance
pay
diet
cut
chemo
lupron
effect
lost
lucky
system
forward
depends
wait
watch
half
weight
hot
body
alive
exercise
lose
loss
libido
issues
stop
24
dealing
fine
urinary
testosterone
medication
issue
stress
30
caused
clinic
tumor
oncologist
opinion
pathology
center
team
report
decision
core
book
support
dr
ejaculation
survival
die
sexual
rate
cancers
treatable
bones
caught
stage
metastasized
understanding
form
initial
earlier
diagnosed
plan
40
50
age
happy
growing
diagnosis
sharing
radical
based
recommended
basically
opinions
cancerous
recommend
highly
treat
close
6
8
11
12
water
sleep
drink
night
times
pee
studies
rates
evidence
patients
data
cure
study
research
ejaculate
semen
hifu
tulsa
experienced
completely
total
advance
moment
real
incontinence
ralp
positive
father
spread
9
aggressive
life
quality
late
family
thinking
happen
history
strong
advanced
dad
visit
slow
move
surgical
prostatectomy
means
decided
waiting
grade
active
surveillance
specific
local
therapy
adt
metastatic
typically
effective
disease
hormone
8
exercises
physical
floor
pelvic
helped
dad's
deal
people
100
easier
comfortable
change
slightly
possibly
rp
20
step
worry
result
worth
note
70
check
possibility
free
chance
experience
procedure
finding
cells
treated
lower
primary
regular
low
online
increase
concern
tests
stay
process
hospital
removed
remove
operation
op
screening
gt
performed
treatment
options
option
outcomes
future
choice
brachytherapy
treatments
effects
radiation
journey
mine
hope
hang
folks
wrong
lots
ahead
hear
talking
care
glad
major
docs
talk
friend
friends
luck
difference
extremely
stuff
love
easy
helps
lot
idea
fine
sound
excellent
posted
mentioned
feels
rest
sort
story
uncomfortable
experiences
guy
website
healthy
suggest
feel
short
feeling
amount
questions
covid
helpful
add
read
call
worse
coming
urologist
foundation
answer
share
posts
guys
discuss
offer
expect
set
pc
site
posting
found
understand
current
agree
urologists
learn
question
advice
honestly
doctors
results
forum
reddit
information
type
suggested
determine
told
info
medical
situation
person
worry
doctor
appointment
health
patient
called
reading
doc
mention
scared
tomorrow
size
mind
reason
worried
hoping
follow
signs
sense
sign
husband
matter
scheduled
assume
speak
huge
difficult
outcome
true
prior
due
sounds
continue
start
totally
makes
heard
response
wondering
finally
post
5
2
1
left
similar
concerned
15
recently
guess
10
60
14
ago
appreciated
2019
comments
message
comment
term
surgeons
surgeon
months
bladder
news
starting
flow
18
urine
urethra
clean
month
taking
blood
painful
worst
hard
giving
weird
hit
bit
fairly
quickly
started
pretty
bad
discomfort
pain
robotic
removal
surgery
week
hours
morning
minutes
day
yesterday
days
home
meds
walking
bed
walk
catheter
nurse
bowel
bag
considered
undetectable
lt
urinating
stream
urinate
weak
urination
prostate
common
cancer
mri
biopsy
biopsies
samples
fusion
guided
pet
scans
mets
ct
scan
bone
time
standard
risk
testing
pca
psa
test
levels
level
negative
lymph
nodes
tissue
significant

Figure 6:

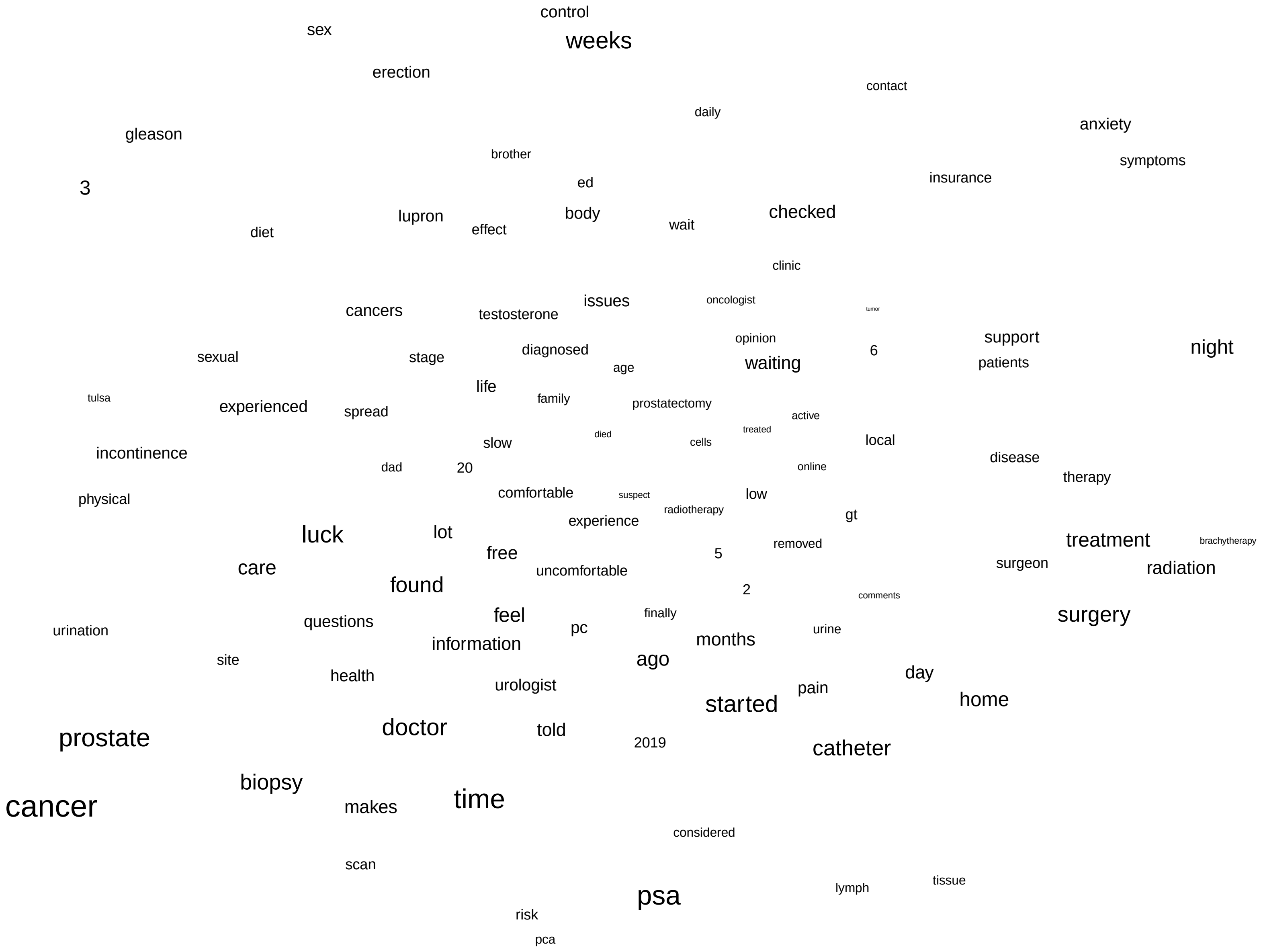

Figure 7: Topic landscape founds using t-SNE with dissimilarity based on the negative co-occurrence rate between topics [[CM: Need to actually define what I mean here]]

[[CM: Specific topic discussion, not sure which to use?]]

4.1 Diagnosis

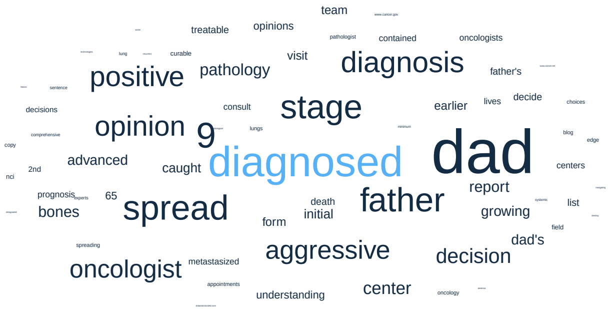

Figure 8:

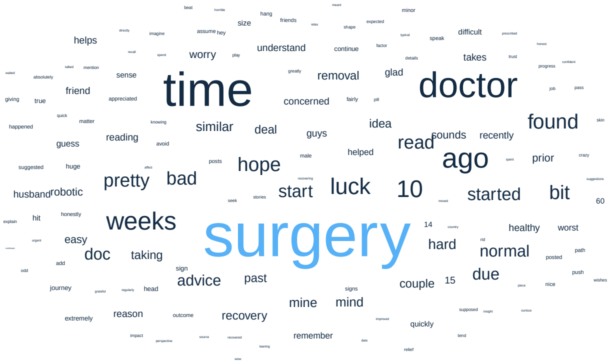

Figure 9:

5 Discussion

[[MM: We find that the framework is useful.]]

6 Conclusion

In this paper we motivated the need for a generalised repurposeable framework for conducting social media analysis for discussion of specified topics. To meet this need, we propose the Design-Acquire-Process-Model-Analyse-Visualise (DAPMAV) framework. We explore the use of this framework in the context of discussion on prostate cancer.

References

[1] Nick Anstead and Ben O'Loughlin. Social media analysis and public opinion: The 2010 uk general election. *Journal of computer-mediated communication*, 20(2):204–220, 2015.

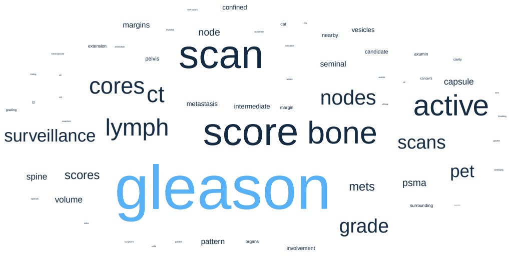

Figure 10:

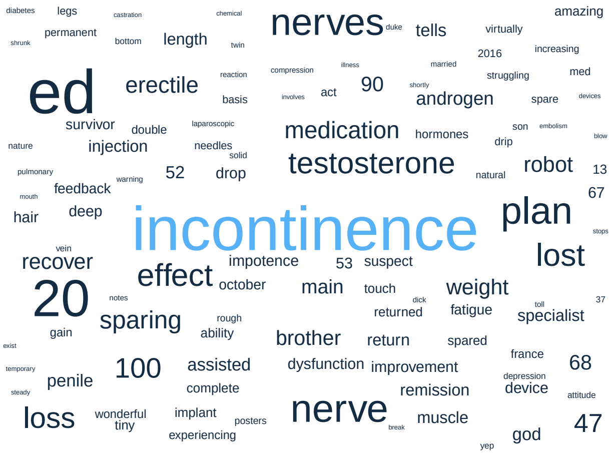

Figure 11:

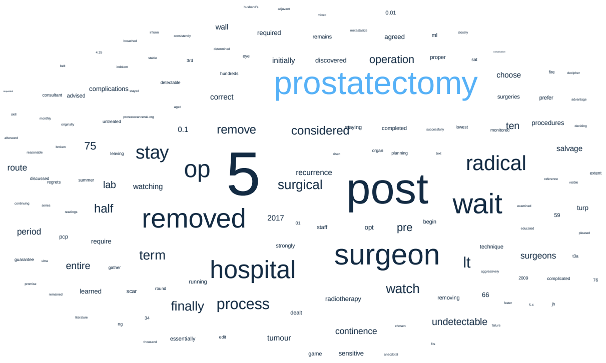

Figure 12:

[2] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter,

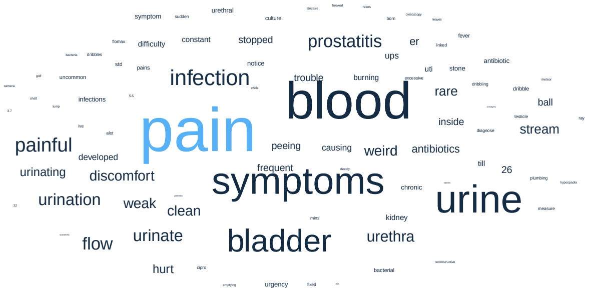

Figure 13:

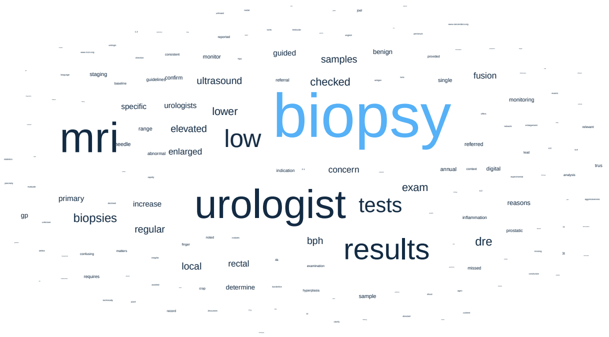

Figure 14:

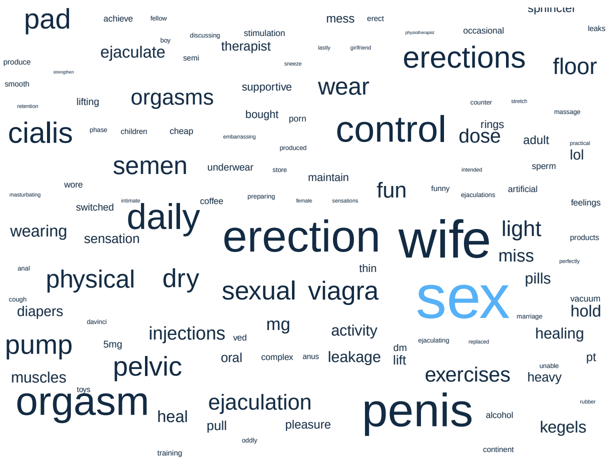

Figure 15:

2011.

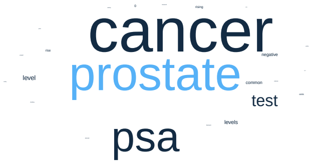

Figure 16:

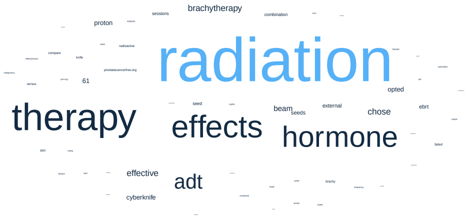

Figure 17:

[3] Amir Karami, Morgan Lundy, Frank Webb, and Yogesh K Dwivedi. Twitter and research: a systematic literature review through text mining. *IEEE Access*, 8:67698–67717, 2020.

[4] Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. The anatomy of reddit: An overview of academic research. In *Dynamics on and of Complex Networks*, pages 183–204. Springer, 2017.

[5] Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. Twitter for sparking a movement, reddit for sharing the moment:# metoo through the lens of social media. *arXiv preprint arXiv:1803.08022*, 2018.

[6] Adam N Joinson and Carina B Paine. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology*, 2374252, 2007.

[7] Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ quality & safety*, 22(3):251–255, 2013.

[8] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839, 2020.

Table 1: Table showing most prevalent word in topic (topic identifier) and the overall usage of the topic in the /r/ProstateCancer discourse. [[CM: TODO: might need to change the highlighted words in top topic wordclouds to align with this]]

Topic identifier	Figure	Usage density
surgery	9	0.138
prostate	16	0.071
prostatectomy	12	0.052
sex	15	0.051
biopsy	14	0.051
incontinence	11	0.050
diagnosed	8	0.042
pain	13	0.031
gleason	10	0.028
radiation	17	0.023