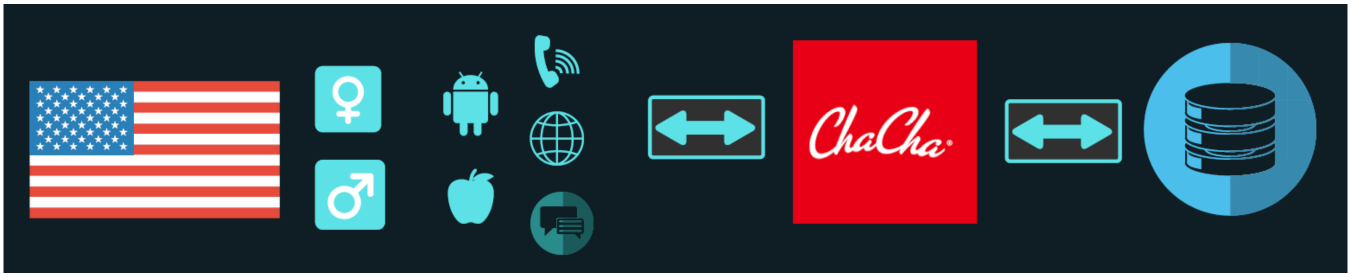
1Text-Mining of User-Generated Queries on Menstrual Pain   Menstrual Pain – What You Need to Know!

Naimesh Chaudhari, Siddarth Shankar, Gautham Nagendra Kamatchi and Sponsor: Dr. Chen X. Chen



Abstract— ChaCha a question-answer service which provided answers to questions posed by users over phone, web or voice. Between Jan 2009 to Nov 2012, there were close to 2 Billion queries posed by people in the US. Menstrual pain is highly prevalent among women of reproductive age. This paper shows how text-mining of user-generated queries, was used to understand the public's information needs and concerns related to menstrual pain. The information gained was used to develop interventions to support menstrual pain management. Specific goals on this paper are to visualize the text data; cluster and categorize questions and summarize patterns and themes. From the 2 billion questions, available 1.9 billion questions had complete information. These queries were from both females and males across a wide range of topics. From these 1.9 billion, only queries related to menstrual pain/dysmenorrhea were filtered. This resulted in a data set of 620,000 questions. There were 507,000 questions from females and 114,000 were from males.

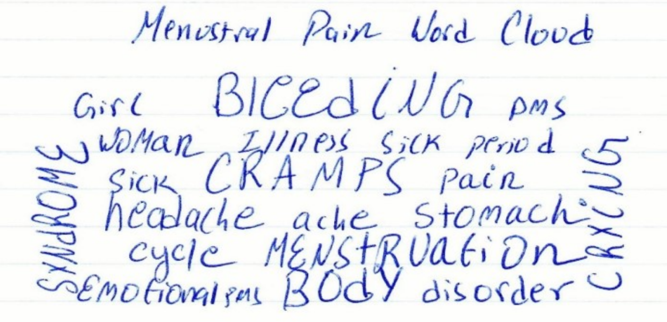
INTRODUCTION

Menstrual cramps are throbbing or cramping pains in the lower abdomen. Many women have menstrual cramps just before and during their menstrual periods. The dataset that we recived highlight the quireies during mensutral period by both men and women. The received dataset has 2 billion questions available out which 1.9 billion questions had complete information. These queries were from both females and males across a wide range of topics. From these 1.9 billion, only queries related to menstrual pain/dysmenorrhea were filtered. This resulted in a data set of 620,000 questions. There were 507,000 questions from females and 114,000 were from males. To further analyze these questions we have identified some relevant visualizations.

1.1 Identify Relevant Visualizations:

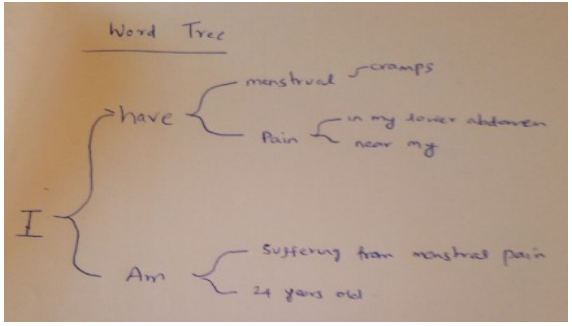
1.1.1 Word Cloud

a. We focused on creating a few word clouds of some of the common words that users are typing.
b. Create word clouds based on the gender of the user posing the queries.
c. Pre-processing Techniques [k d]:
   - To do this we created a document-term matrix
   - Applied lower casing to all words
   - Separated the word blob into individual words. (Tokenization)
   - Common or low content prefixes and suffixes were removed to identify the core concept. (Stemming)
   - Low content words like of, in, etc. are removed. (Stop Words)
   - The below is a sample sketch of a word cloud that we planned on creating.



1.1.2 Word Tree

a. To identify the sequence of the words used.
b. Ability to select different root words and view the subsequent sequences of words to the selected root word. It provides a good idea of the frequency of word usage and co-occurrences of words.
c. The below is a sample sketch of a word cloud that we are planning to create.

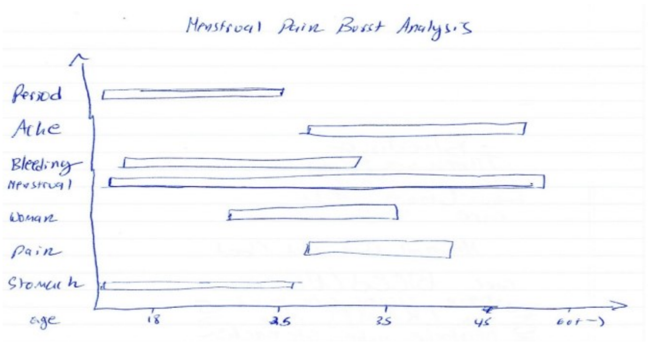


1.1.3 Temporal Analysis:

a. Pain experienced by women by day – before onset, 1st, 2nd day, 3rd day, etc.

b. If we have enough data, it might also be interesting to do a temporal burst analysis of the words as age gets bigger.

c. Below is a sample burst analysis that we planned to visualize.



1.1.4 Choropleth and Cartograms:

a. Geospatial Analysis: We wanted to see if there are any differences or variances in these words as the location of a user changes.

1.1.5 Topical Analysis:

a. We wanted to see if there are any differences or variances in these words as the gender of user changes.

b. Tree Maps can be generated and show the differences in the questions posed by males/females.

1.1.6 Cluster Analysis:

a. We attempted to run different cluster analysis techniques – K-Means Clustering etc on the dataset.

b. Identify broad clusters and the way we can classify the questions.

1.2 Tools:

1.3 Why is this important:

a. Dysmenorrhea (Menstrual Cramps) is a very common occurrence in 60-70% of young women j e. This condition can affect the quality of life for women and thereby affecting family or social relationships.

b. The aim of the study is to understand the user needs in terms of information and their concerns.

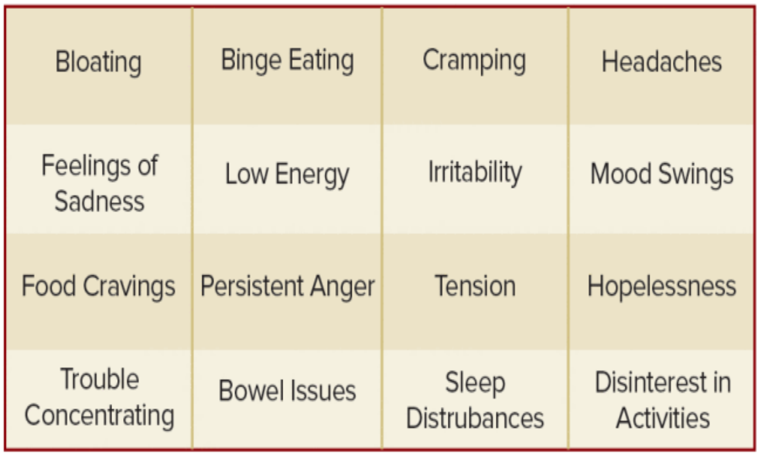
c. This study will enable develop better interventions thereby improving women's quality of life.

2 Discussion of related work

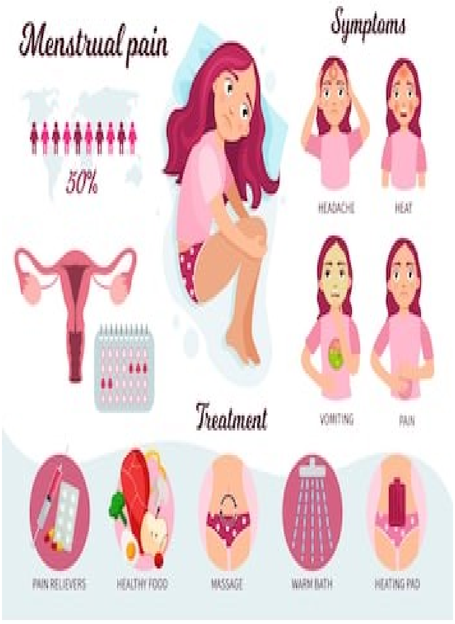
a. Since this is a very common problem that has existed for a very long time there are already great visualizations that have been published. Some are table based, some are simple infographics, and some are more in-depth word clouds. Examples are below. Our focus is going to be using information already available online alongside our own research to create a visualization that is both meaningful and usable.

stomach
cramp
pain
ache
illness
period
nutrition
stomachache
constipation
diarrhea
lifestyle
body
sofa
diet
poisoning
problem
griping
menstrual
food cramps
disease
upset
menstruation
suffering
hurting
hungry
health
hurt
belly
bloating
abdomen
adult
hair
touching
tummy
nausea
medicine
expression
stress
natural
person
gas
indigestion
cycle
unhealthy
painful
eating
grimacing
young
discomfort
teenager
virus
weakness
sickness
medical
sick
ill
injury
aching
bedroom
pressure
bowel
suffer

COMMON SYMPTOMS OF PMS



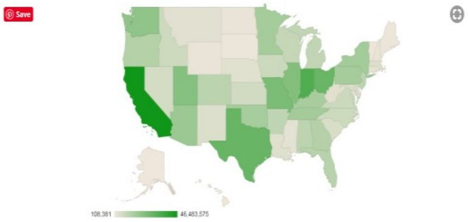
Bloating	Binge Eating	Cramping	Headaches
Feelings of Sadness	Low Energy	Irritability	Mood Swings
Food Cravings	Persistent Anger	Tension	Hopelessness
Trouble Concentrating	Bowel Issues	Sleep Distrubances	Disinterest in Activities



shutterstock.com • 1188742360

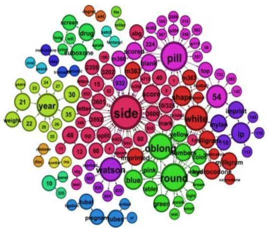
b. The related paper - Finding the Patient's Voice Using Big Data: Analysis of Users' Health-Related Concerns in the ChaCha Question-and-Answer Service (2009-2012) -- also has relevant visualizations which can be used as a building block for advanced insights. As an example, the below choropleth can be further broken down at the district level of the information available.

Figure 1



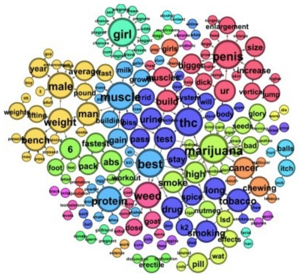
Number of queries posted to ChaCha by user location within the United States, 2009-2012.

Figure 6



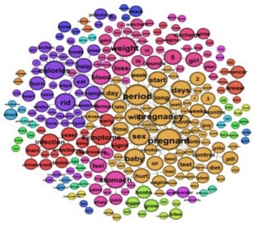
The most prevalent 2-word phrases submitted to ChaCha by users aged 20â€"39 years.

Figure 2



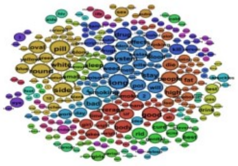
The most prevalent 2-word phrases submitted to ChaCha predominately by male users.

Figure 3



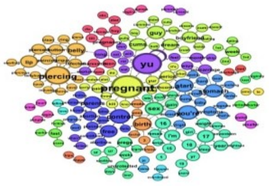
The most prevalent 2-word phrases submitted to ChaCha predominately by female user.

Figure 4



The most prevalent 2-word phrases submitted to ChaCha by both males and females.

Figure 5



The most prevalent 2-word phrases submitted to ChaCha by users aged 13â€"19 years.

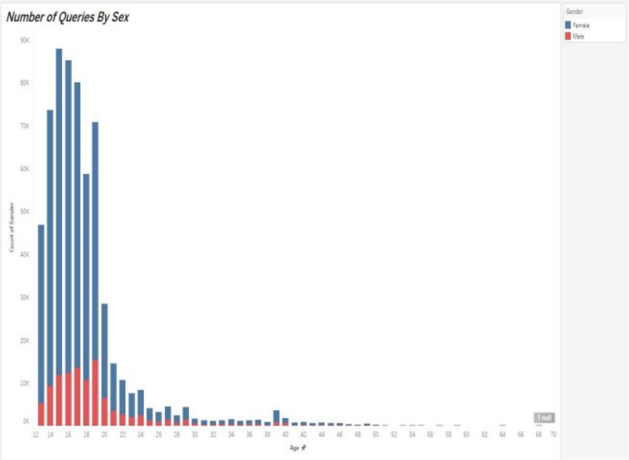
2.1 Simple Statistics on The Dataset

The data provided is from a United States company called ChaCha. ChaCha collected all queries asked by its users whether the queries were from online services or text. In total, they collected 1.5 billion queries. When the data set was provided to us it was already filtered to only queries that were related to menstrual pain. In total, we had **113,889** queries from men and **507,328** queries from women. The website also had the ability to collect user information if they had created an account with them. The attributes collected are below.

Attribute	Description
Query ID	Unique id given to each query asked
Created at	Time of asking the query
Category Path	The category to which the query relates to. (I think it's selected by the user from the options provided).
Primary Category	Derived from Category Path
Sub Category	Derived from Category Path
Source	The method by which the user has posted the question – SMS, ANDROID, IPAD, etc.
System	
City	City of the user
State	State of the user
Region	Region of the user
Country	Country of the user
Area Code	The area code of the user
Zip Code	The zip code of the user
Gender	Gender of the user
Age	Age of the user
User ID	A unique ID assigned to each user.
Question Text	The actual text of the query raised by the user.

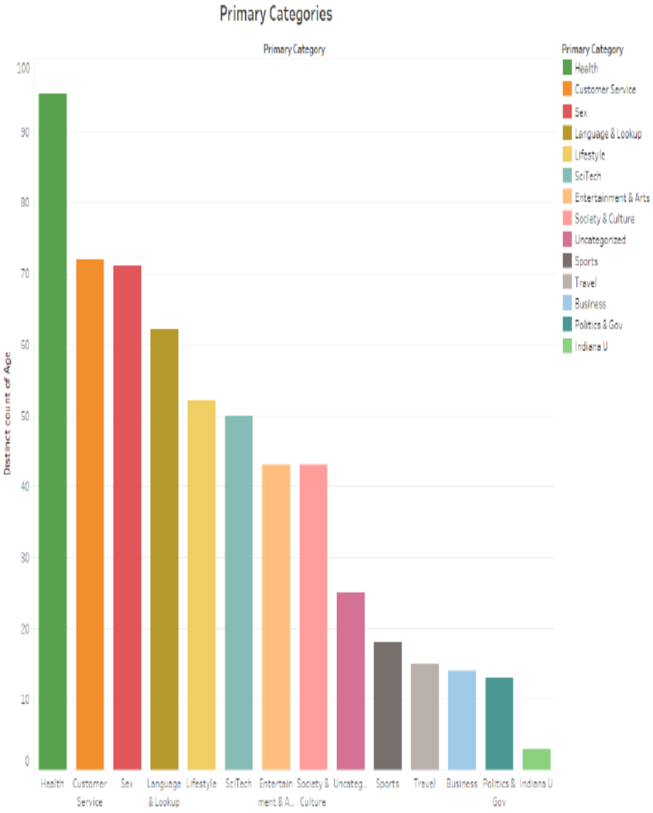
In our research, we focused primarily on the age, gender, and queries asked by the user. They have been highlighted in the above table. Below is a bar graph that identified the total queries by age. The coloring resembles the user's gender.

Queries By Sex



Below is another chart that identifies the total queries for both men and women by the primary category of the website.

Primary Category Graph

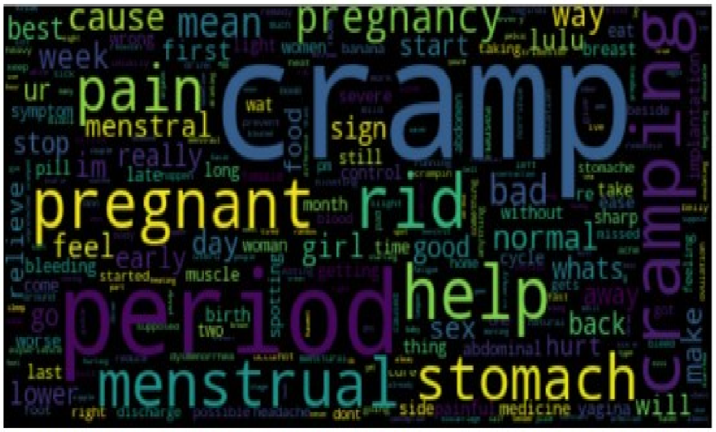


Primarily **99%** of the queries collected came via SMS.

2.2 Data Analysis

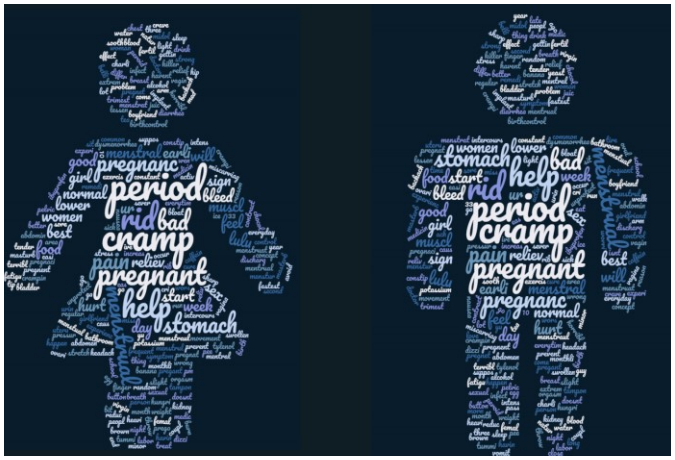
Our main goal was to derive valuable information from the data provided. Considering that this was a text related problem we knew that we were going to have to take the queries provided and split them up into individual word tokens. After we do this, we should be able to analyze the data and get valuable insight out of it. The programs we used to do this were both JMP and Python. Within Python, there is a fantastic library called Scikit Learn that allows us to do data mining tasks quickly and efficiently. We used the tfidfvectorizer object to split the question text column into individual tokens. While tokenizing we ensured that we were stemming and removing stop words from our dataset. After we were able to tokenize the words we used the WordCloud library to come up without initial word cloud.

Overall WordCloud



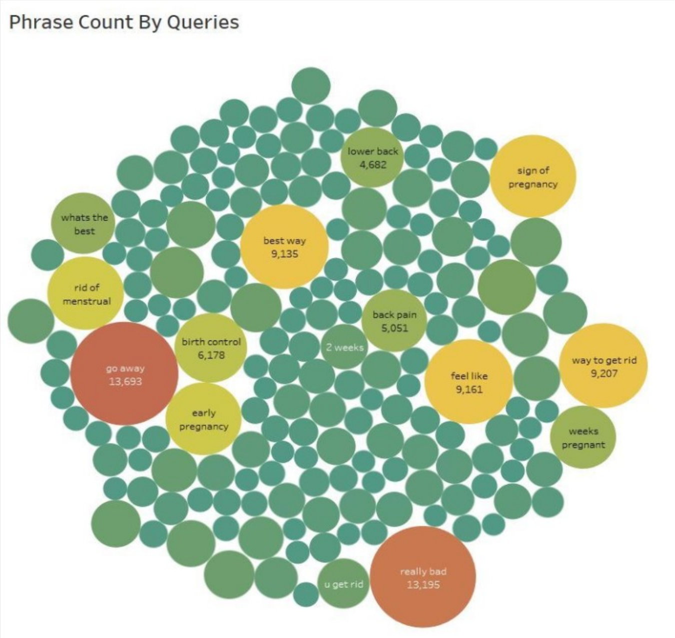
This would cloud gave us a general idea of what the general population was querying when they had questions regarding menstrual pain. To understand if different genders were asking different questions or top keywords, we split the data into Male and Female and analyzed the word clouds. The results are below.

Male Female WordCloud



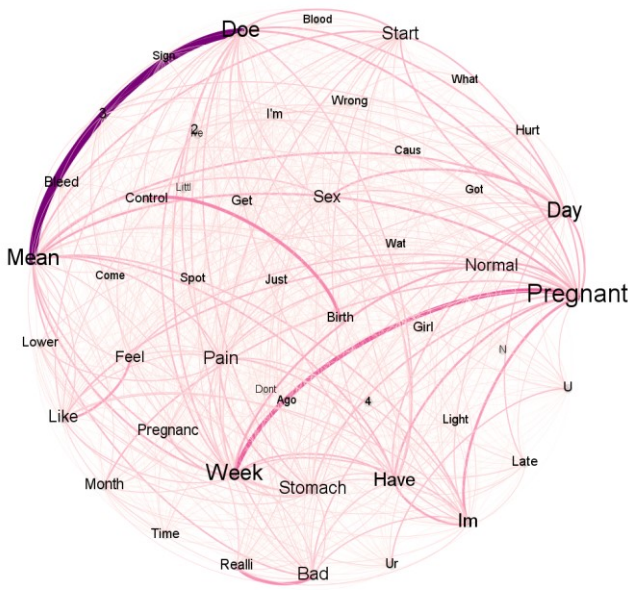
Along with the above word clouds we wanted to try and understand if there are any specific topics in words we can find. By topics, we mean queries that have a common theme. To do this we used a tool called JMP via SAS that was quickly able to help us analyze the data. We did similar tasks as we did with the initial word cloud, by removing stopwords and lemmatizing the words. We took this a bit further and removed words that we knew were common in all queries. Words such as cramp, pregnant, menstrual, etc. Then we tried to see what some of the most common phrases are. The results are below.

Phrases



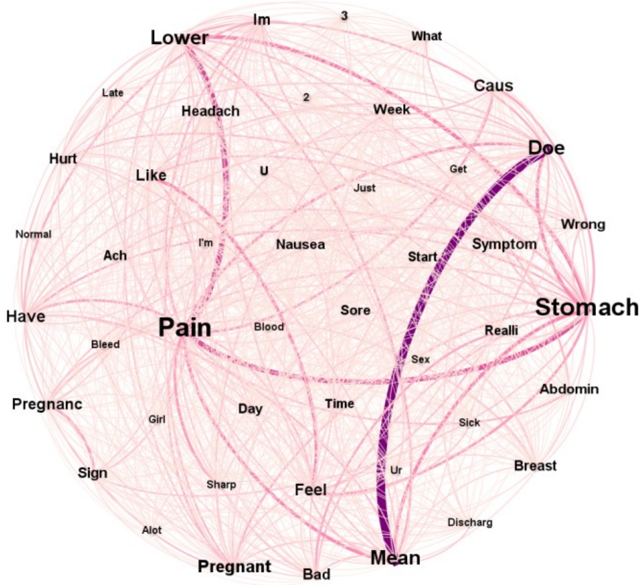
From the phrase counts, we moved on the trying to understand the individual topics. We use LDA (Latent Dirichlet Allocation) to see if can get any insight. We found two common themes within the queries. Theme 1 was primarily related to questions when women are pregnant or think they have become pregnant. Theme 2 primarily revolved around the symptoms women were feeling when they have menstrual pain. We created a co-word occurrence network graph using Gephi to understand these two themes. The graphs below represent the most common words in their themes, the size of the word represents how often they were used and the thickness of the line connecting them represent how often they were used together

Theme 1: Sex/Pregnancy



As we can see from the above image many of the words displayed are from times when women are having sex or are feeling pregnant. We see common relationships between words like pregnant, week, birth, control, mean, does, sex, etc. Theme 2 was more related to symptoms and its network graph is below.

Theme 2: Symptoms



In theme two we can see words related to the symptom's women are feeling. Words such as pain, stomach, lower, feel, lower, nausea, bleed, etc.

2.3 Key insights gained from analysis

One of the first things we learned from the dataset that there is a male population trying to understand the symptoms and solutions to relieve menstrual pain. Our initial assumptions were that all queries would be from women. When we analyzed the data, we realized 1/6th of the queries are from men! This could imply that young men are trying to understand the symptoms that their significant others are feeling to be in better touch with them. Along with this we also learned that 90% of the queries that were asked are from men and women under the age of twenty. This primarily does make sense to us since this is a monthly occurrence for women and as they age, they are better in tune with their bodies and now what actions they need to take. Young teenagers, on the other hand, have not had much experience in having these sorts of biological changes and are undereducated in the actions they need to take to help alleviate or resolve the issue. We also learned that the data was not entirely cleaned, there were queries that entered from categories that do not make sense. Some of the categories that came in were Government, Sci-Tech, Travel, and sports. In big data, we need to be aware of these sorts of conditions and get better at removing these anomalies. We tried to do this as best as we can by restriction the words in the word clouds and trying to see the most common topics. Our assumptions in the analysis above is that one we need to do a better job in educating our teenagers about menstrual pain. We could use the charts we have created above to ask them if they are feeling any of the symptoms if they feel enough of them we need to be prepared to help them through the changes their bodies are feeling.

2.4 Problems Faced During Validation

The major problem we faced during validation was getting the information across in a presentable manner. We had very good information on many different charts, but the audience from our initial feedback review was trying to understand how everything was connected. After discussing with our audience, we discovered that we had opportunities to combine multiple charts into one and show the information in a much more organized and consolidated way. Some of the steps we have taken are, combining bar graphs for number of queries by age and number of queries by gender into one and designing word clouds Also, although we thought that our initial network graph was interesting, it really identified what we already knew, most of the words on the original graph are words that were initially used to filter the queries. We tried to redesign the network graph but with many of the common words removed and see if it adds any additional value. This led us to the topic analysis which provided great results.

2.5 Discussion of Challenges & Opportunities

One of the major challenges we faced during analysis was how to handle the raw amount of data. We have over 600K comments. When converting these comments to a document-term matrix our machines have a very difficult time processing the data. When loading the information into sci2 to design a co word occurrence network, it took many hours to validate and complete. We resolved this by attempting to process the data in python which took significantly less time. We were then able to save the word co-occurrence matrix into a CSV file and use Gephi to do the final visualization. We were still challenged as to how we will be presenting the final information, on one side, we want to create a beautiful visualization that holds a lot of the word co-occurrence networks and on the other hand, we ask ourselves how useful that really is? Would a simple word cloud split into the proper categories provide more value than to create a large network graph where the information is hard to follow? We faced a similar question when we attempted to create a dendrogram via hierarchical clustering when we looked at the results, we saw groups were being formed but again it was not adding much value to the questions we are trying to solve. After reviewing these questions, we decided to take the information we have and create an information sheet that laid all the information out in a readable manner. We thought this would be the best solution to give our audience as much relevant information as possible.

2.6 Conclusion

Overall this has been a fantastic project to work on. We were able to learn new techniques in data mining that helped us analyze text data. A special thanks to Dr. Chen for providing us with the dataset to explore. From our findings, we have concluded that it is extremely important for us as a society to educate our teenagers when it comes to menstrual pain. We have created a few visualizations that help us identify what sort of information or questions these individuals are asking. We should take this information and focus on creating programs that in middle school and high schools that answer these questions in an informative manner. In doing so we will educate out teenagers to be better prepared when they do have menstrual pain/cramps.

Acknowledgments

We would again like to thank Dr. Chen for providing us the data to analyze and giving us hints on topics we should explore. We would also like to thank Katy Borner and her team in guiding us in properly creating visualizations that are able to tell a clear story in a compact manner.

References

[1] *Home Remedies for Menstrual Pain - Say Bye to Cramps*. 12 Aug. 2018, famoday.com/home-remedies-for-menstrual-pain-say-bye-to-cramps

[2] Ibreakstock. “Menstruation Animated Word Cloud.” *Shutterstock*, www.shutterstock.com/video/clip-29450050-menstruation-animated-word-cloud-text-design-animation

[3]

[4] “IndianaUniversity.” *Canvas*,iu.instructure.com/courses/1799279/pages /when-temporal-data-theory?module_item_id=18138735

[5]

[6] Najafi, Nastaran, et al. “Major Dietary Patterns in Relation to Menstrual Pain: a Nested Case-Control Study.” *BMC Women's Health*, BioMed Central, 21 May 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5963185

[7]

[8] Priest, Chad, et al. “Finding the Patient's Voice Using Big Data: Analysis of Users' Health-Related Concerns in the ChaCha Question-and-Answer Service (2009-2012).” *Journal of Medical Internet Research*, JMIR Publications Inc., 9 Mar. 2016, www.ncbi.nlm.nih.gov/pubmed/26960745.