Understanding Privacy: Clustering of Privacy Related Questions from Question-Answer Forum Quora

Sayantan Mukhopadhyay

Project Goals:

The goal of the project was:
First analysis :
Create topic clusters from list of questions containing work privacy
- We will test a variety of clustering and topic detection techniques on a corpus of questions from quora that use the word "privacy".
- try 2-3 techniques of each
    Topic Detection through NLP
    Clustering the topics using unsupervised learning.
- Have privacy experts (Nathan, Jen and perhaps deirdre) look through clusters and comment on them.

(Referenced from: NLP Proposal Modified (https://docs.google.com/document/d/1UKx49FS1A0sCDHXFdNjMbzmEKYoFfZu24sOKwcynye8/edit))

Original intent:

The original intent of this project was that, privacy researchers are interested in understanding what topics and concepts are being discussed when people use the term "privacy". A grounded understanding of consumers use of the word in practice is important to policy makers and designers as they seek to understand how perceptions and expectations of privacy is static and/or evolves over time. Previous work by privacy researchers has looked at manually categorizing consumer comments and questions that contain the word "privacy" in them. This is time consuming and difficult, and these barriers prevent researchers from acquiring a data-based understanding of the usage and expectations of privacy over time and with regards to various topics. One hope is that NLP technologies and topic mapping and detection could bootstrap and provide means for researchers to explore the larger scope of public opinion, thereby providing a stronger database for understanding the discourse or privacy that occurs on the web, as well as the various overlapping topics and concerns that arise in the discussion or privacy, as well as to what extent these notions evolve and or static over time.

A first step in this direction is to perform an exploratory analysis of a corpus of Q&A text mentioning privacy" using a variety of topic detection and clustering techniques to determine what techniques can assist experts in this space in understanding how the word privacy is used in question and answers. Answers gained from this preliminary step will provide us a means to create future models that look at privacy changes over time, by topic and also provide points for future exploration for privacy researchers looking at peoples opinions online.

How far you got:

I have reached the goals I aimed for as part of this project.
I have done exploratory analysis on the question corpus retrieved from Quora which shows interesting results. After that I have done clustering analysis to group the questions in the platform and found some good clusters. Clustering involves manual evaluation and that is completed by me already and will be done by Nathan, Deirdre and Jen also.

What future work would be if this were continued:

As part of future work I am going to meet Deirdre’s research group later this week to decide the road map.

Distinctly I have below plans in mind:
First, in technical way I want to deploy more grammatical methods and tree based methods to understand the sentence structure and derive sense from the questions. Second, we (Nathan and me) want to look into the answers too. The answers have much more information and signals.
Also I want to work on creating a classification model for the coding template Jen and Deirdre shared with me about the privacy complain corpus.

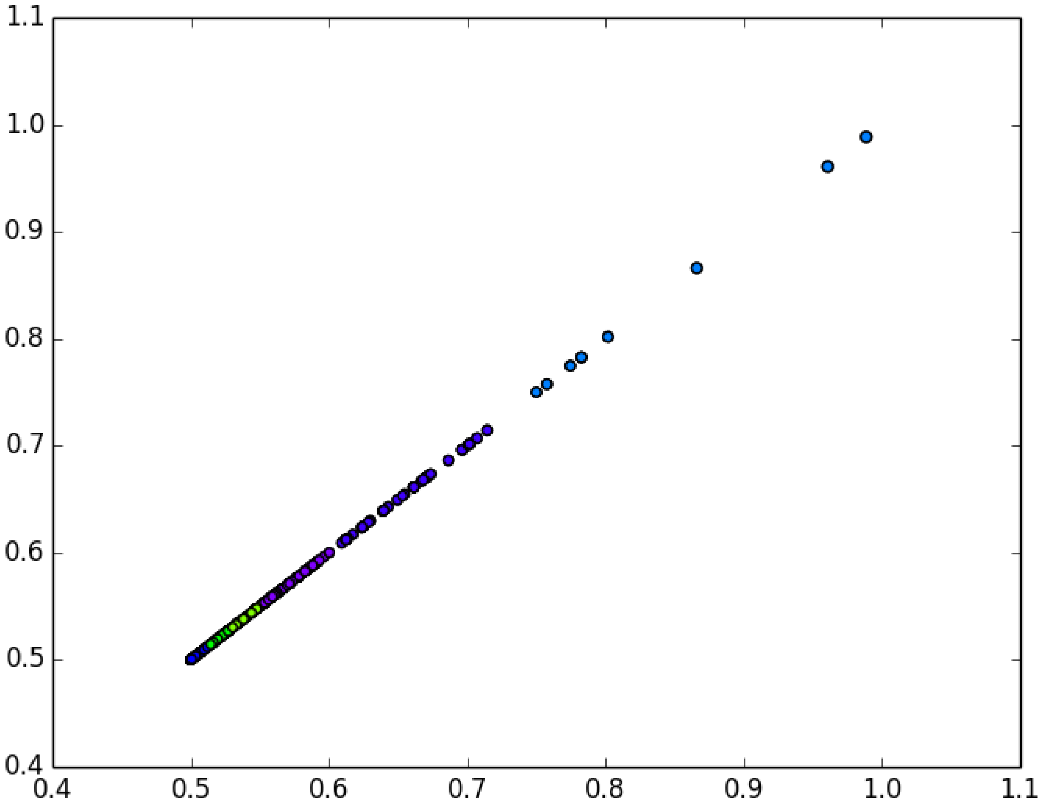
Results / Evaluation:

I used dictionary method initially to understand the meaning of the sentences also used couple of grammatical method such as investigating the verbs and nouns before and after the word Privacy but I did not get very good results.

For the Clustering I first used Cosine similarity between each pair of questions. Cosine similarity provides similarity between two documents and treats the documents as vectors. Higher cosine value suggested better similarity between the sentences.

```
e.g. (0.7745966692414834, 'What is privacy', 'What is too much privacy')
(0.8660254037844387, 'What is privacy', 'What good is privacy')
(0.7745966692414834, 'What is too much privacy', 'What is privacy')
(0.8017837257372731, 'Are there any privacy concerns for Siri', 'Are there any privacy concerns for Amazon Silk')
(0.8660254037844387, 'What good is privacy', 'What is privacy')
(0.7826237921249264, 'What are the privacy implications of Facebook Questions', 'What are the privacy implications of using
BranchOut on Facebook')
(0.7826237921249264, 'What are the privacy implications of Facebook Questions', 'What are the privacy implications of using
BranchOut on Facebook')
(0.7499999999999999, 'What are the privacy implications of Facebook Questions', 'What are the privacy implications of installing
Truecaller')
(0.8017837257372731, 'Are there any privacy concerns for Amazon Silk', 'Are there any privacy concerns for Siri')
(0.7826237921249264, 'What are the privacy implications of using BranchOut on Facebook', 'What are the privacy implications of
Facebook Questions')
(0.7826237921249264, 'What are the privacy implications of using BranchOut on Facebook', 'What are the privacy implications of
Facebook Questions')
(0.7499999999999999, 'What are the privacy implications of installing Truecaller', 'What are the privacy implications of Facebook
Questions')
(0.960768922830523, 'Are the privacy policies of the online shopping stores enough 1', 'Are the privacy policies of the online
shopping stores enough')
(0.960768922830523, 'Are the privacy policies of the online shopping stores enough', 'Are the privacy policies of the online
shopping stores enough 1')
(0.9885710532241613, 'So under EU law most android apps would need unambiguous consent to collect personal data Whats
the best way to build that requirement into an App with a privacy policy eg pop up link to website etc 1', 'So under EU law most
android apps would need unambiguous consent to collect personal data Whats the best way to build that requirement into an App
with a privacy policy eg pop up link to website etc')
(0.7576160802966969, 'When you hide a Facebook post on your timeline does your friend who posted it still see the post on your
timeline Or can they see that you hid the post', 'When you hide a Facebook post on your timeline does it still show up to the poster
when viewing your timeline Or will he she see that you hid it')
(0.7576160802966969, 'When you hide a Facebook post on your timeline does it still show up to the poster when viewing your
timeline Or will he she see that you hid it', 'When you hide a Facebook post on your timeline does your friend who posted it still
see the post on your timeline Or can they see that you hid the post')
(0.9885710532241613, 'So under EU law most android apps would need unambiguous consent to collect personal data Whats
the best way to build that requirement into an App with a privacy policy eg pop up link to website etc', 'So under EU law most
android apps would need unambiguous consent to collect personal data Whats the best way to build that requirement into an App
with a privacy policy eg pop up link to website etc 1')
```

After that based on their Cosine score I divided them into buckets which is a similar approach to K-Means clustering and therefore I used K-Means. In my analysis I found that while using K=8 I had meaningful clusters.



But K-Means clustering did not provide me the topics mentioned in the questions in concern therefore I decided to use NMF. Using NMF through using SVD it's very possible to extract the topics. I removed the stopword and the punctuations and then I used NMF for topic extraction. As for the particular case I am looking at a small dataset therefore it was appropriate to limit down the topics and the words related to the topic to a certain limit. In this particular case I found 7 topics and 5 words per topic gives much better result than using large number of topics or containing words.

For 20 topics and 10 words in each topic the result is overlapping not meaningful:

```
['facebook', 'set', 'chang', 'ad', 'new', 'photo', 'feat', 'group', 'mak', 'account']
['us', 'imply', 'facebook', 'brows', 'many', 'on', 'system', 'viol', 'affect', 'tel']
['inform', 'person', 'phon', 'ident', 'collect', 'websit', 'remov', 'off', 'lik', 'company']
['internet', 'saf', 'mak', 'today', 'rel', 'good', 'remov', 'explain', 'big', 'surf']
['dat', 'sel', 'company', 'collect', 'big', 'leg', 'country', 'eu', 'cloud', 'own']
['search', 'engin', 'nam', 'en', 'result', 'account', 'graph', 'set', 'hid', 'allow']
['protect', 'onlin', 'best', 'way', 'book', 'good', 'resourc', 'anonym', 'keep', 'stor']
['peopl', 'websit', 'many', 'much', 'car', 'com', 'gen', 'group', 'dont', 'know']
['googl', 'regard', 'glass', 'contact', 'brows', 'chrome', 'stor', 'track', 'gmail', 'result']
['concern', 'phon', 'might', 'rais', 'market', 'amazon', 'key', 'glass', 'freedom', 'regard']
```

```
['policy', 'term', 'websit', 'serv', 'startup', 'writ', 'diff', 'good', 'draft', 'condit']
['post', 'friend', 'see', 'hid', 'someon', 'timelin', 'shar', 'chang', 'tag', 'pag']
['soc', 'network', 'med', 'shar', 'right', 'respect', 'sit', 'google', 'twit', 'mobl']
['publ', 'leg', 'address', 'avail', 'permit', 'access', 'med', 'nam', 'hom', 'record']
['law', 'get', 'help', 'country', 'right', 'much', 'eu', 'doesnt', 'regard', 'ad']
['import', 'govern', 'individ', 'id', 'stat', 'sec', 'thought', 'what', 'explain', 'plac']
['quor', 'quest', 'anonym', 'answ', 'view', 'viol', 'ask', 'becom', 'mess', 'opt']
['ap', 'nee', 'android', 'sit', 'iphon', 'mobl', 'im', 'might', 'stat', 'link']
['email', 'act', 'work', 'read', 'provid', 'address', 'gmail', 'mail', 'comprom', 'text']
['sec', 'issu', 'serv', 'rel', 'nat', 'pot', 'reason', 'bas', 'advert', 'intern']
```

For 7 clusters and 5 words per category:

```
['facebook', 'set', 'chang', 'post', 'friend']
['us', 'imply', 'facebook', 'quor', 'soc']
['inform', 'person', 'internet', 'publ', 'remov']
['policy', 'term', 'websit', 'serv', 'ap']
['googl', 'search', 'engin', 'nam', 'issu']
['concern', 'peopl', 'ap', 'phon', 'much']
['dat', 'protect', 'law', 'best', 'sec']
```

Description of Data

The current data is collected from the question answer forum Quora. The questions which came through the keyword search “privacy” are included for the current analysis.
I looked into 880 questions for the current analysis.
(https://github.com/tantanm/NLP-Final-Project/blob/master/quora.txt)

Descriptions of Algorithms:

In my analysis so far I have used dictionary method and algorithm methods.

I found through my research that there are two different ways of classifying questions. Through content word and through grammar. The grammar structure was not consistent in this corpus and it was a small corpus so I focussed more in the content words.
In the algorithm methods so far I have focussed on the content words in the sentence. There are different approaches to look into the structures and the content words of the sentence to conduct clustering analysis. I have done my research in both the areas and tried to employ both of them. I have so far achieved success analyzing the content words in the sentences while clustering. For understand the mutual similarity of the question sentences I used standard TF-IDF [1] and CoSine similarity methods[2].
To cluster the documents I used initially K-Means Clustering[3] with the Cosine scores which gave me the results as mentioned above. I used NMF[5] for my purpose after that which utilizes SVD in the background to cluster documents and also provides the content words which are also known as topic words. In this case I used the TF-IDF score of the keywords in the document to identify the topics. The results were pretty useful.

In the pursuit of grammatical structure I am particularly interested to do some kernel based tree analysis[9], [11] and also going to look into Predicate-Argument Structures[10].

For better understanding of question analysis I also looked into a number of papers specially in the areas of question analysis and question-answer analysis including the papers referred in the class[7]. Then I looked into the papers about analysis of questions based on the grammatical structures [8] .
Further I looked into the papers related to sense disambiguation[6] because the word Privacy has

multi-dimensional meaning in many different situation.

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.
2. http://en.wikipedia.org/wiki/Cosine_similarity
3. Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979): 100-108.
4. Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." *Discourse processes* 25.2-3 (1998): 259-284.
5. C. Boutsidis, E. Gallopoulos: SVD based initialization: A head start for nonnegative matrix factorization - Pattern Recognition, 2008
6. Yarowsky, David. "One sense per collocation." *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1993.
7. Dang, Hoa Trang, Diane Kelly, and Jimmy J. Lin. "Overview of the TREC 2007 Question Answering Track." *TREC*. Vol. 7. 2007
8. Selecting Features for Paraphrasing Question Sentences Noriko Tomuro and Steven L Lytinen
9. Kambhatla, Nanda. "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations." *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004.
10. Surdeanu, Mihai, et al. "Using predicate-argument structures for information extraction." *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003.
11. Suzuki, Jun, et al. "Question classification using HDAG kernel." *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*. Association for Computational Linguistics, 2003.

Contributions of Each Team Member:
In the current version of coding and analysis all are done by myself. I had very detailed inputs from Nathan in the project proposal phase but all technical research, implementation and coding are done by me.

Code (either a zip file or a link to the code online): https://github.com/tantanm/NLP-Final-Project (It’s a public repo)