Project Update #1

Group Members:

Jay Lohokare (111492930)
Revati Damle (111461639)
Nittin Aggarwal (111401512)

Our project aims to anonymize users by all means possible while they use the internet. Privacy is a in limelight these days as search engines/social media websites are using data from any sources possible to profile users and target advertisements. Though advertisements are often alright for most internet users, someone storing data about them on servers if not as comforting. In this project, we are developing a tool that will anonymize all these data points about users by masking them or sending data that can't be mapped to the sender. After a through literature review, we concluded that the 3 most important data points used to profile users are IP address, Cookies, and search queries. Other factors like account information (Linked to individual websites due to signups) can be easily avoided when users sign out from the individual websites (again related to cookies/IP address). In our 1st project update, we present the consolidated problem statement that we aim to solve, preliminary work on exploring NLP approaches for query anonymization and study of approaches for IP/Cookies masking.

1. Cookies and IP masking:

   Cookies are data files stored on computers, containing data related to user's preferences, history and accounts. Queries on search engines are often paired with such cookies to understand who the user is and create a user profile. We explored various options for achieving anonymization of cookies and IP address. As mentioned in the project proposal, we had 2 approaches in mind – VPN and proxy servers. After trying out various open source tools to achieve either of these, we have finalized to base our tool on TOR browser. The primary reason for this is that TOR browser helps us achieve IP masking (thanks to TOR network) and cookies masking out of the box. This allows us to focus on the query masking part of this project (Which is our main focus). To conclude, we are creating a Tor browser (Firefox) plugin so as to get the VPN and cookie masking feature out of the box.

2. Query masking:
   This feature is what we are focusing our project on. The idea is that queries made on search engines often revel details of users and their data profiles. Most existing tools handle IP/cookie masking, but no tools handle masking of search engine queries. We explored various methods to anonymize the queries using NLP. Primarily, we classify our approaches into 2 parts – Statistical/Machine learning approaches, Deep learning-based approaches.
   a. Machine learning (Traditional NLP approaches):
      Here, we use traditional top-down NLP approaches for detecting named entities in user queries. Using libraries like Spacy/NLTK, we are able to detect named entities. Detecting Nouns (Proper, common), pronouns, verbs can be helpful for finding personal data points in the query. The idea is to change the nouns with pre-decided pronouns. This approach is simple to use, but effective and a simple python code achieve the functionality

b. Deep Learning:
   This approach involves using neural network architectures for text anonymization. We studied various machine translation approaches that can be used to achieve this. We also have created various codes involving different architectures like CNN, LSTM, BLSTMs in order to achieve the machine translation. For this project update, we have not completed the implementation of the machine translation, and just have explored the various possible ways to achieve it.

Approaches to Machine translation for anonymization:
Machine translation is conversion of text by a computer, with no human involvement. Pioneered in the 1950s, machine translation can also be referred to as automated translation, automatic or instant translation. Machine translation is generally used to convert text from one language to another, while we want to convert from English to well, English. Luckily, such amazing work has been done in this paper. Even though Bulyko and Ostendorf [1] introduced the concept of using weighted finite state transducers to create multiple input sentences for a synthesiser, that doesn't seem to work well in our scenario as it deals with domain specific queries and what we want is a generalized synthesizer. For the first project update we have explored thoroughly on various methods to convert from one English query to another with a similar meaning. To train an English to English machine translation system, a we would need a parallel monolingual English text corpus. As same-language machine translation corpora for synthesis are not available, we will develop a small corpus suitable for a this project. We would be using The ARCTIC [2] corpus, as a starting point for the MT corpus.

The next part is to develop a machine translation technique. As suggested in the paper mention above, we would be using a SMT approach. Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The concept presented in the paper is to use the N-best hypotheses from the machine translation system as input to the synthesiser, which we also intend to do.

References

[1] I. Bulyko and M. Ostendorf, "Efficient integrated response generation from multiple targets using weighted finite state transducers," Computer Speech and Language, vol. 16, no. 3, pp. 533–550, 2002.
[2] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI03-177 http://festvox. org/cmu arctic, 2003.