Review

Detecting and Recovery of Private Information in Large Scale Social Network using mixed Machine Learning Techniques

Pasquale De Luca [1]

[1] University of Naples "Parthenope", Centro Direzionale C4; deluca@ieee.org

Abstract: The violation of privacy, others people or personal, is a very current problem, which concerns not only on the web but also in private life. In the years 1990 it was expected that nowadays, that any routine operation was carried out "manually", and it would be performed through mobile phones or personal computers. The problem pertains the distribution network that allows to share and bring together information and as result the network becomes unsafe, if subjected to attacks. Noway we put personal information on web because otherwise we are seen as "weak". This work aims to measure and analyze how much information are shared by users of a pre-established social network and it is carried out through a set of algorithms techniques of machine learning.

Keywords: Privacy; security; Machine Learning; K-Means; Natural Language Processing; Twitter; Private Information Retrieving.

1. Introduction

Whatever it is the target of *social network* to analyze, the aim is always to **share** the own experiences and what happens during a simple day. Unfortunately over time it has arrived at sharing of contents which highly **private** background that make possible, using several techniques and analysis process, a reconstruction of user's profile more detailed than it is on the social network. Two of these **complex** techniques are used in this project. The aim of this project is analyze and estimate the amount of private information shared on **Twitter**. In this manuscript will be described the techniques used to extrapolate the *start dataset* and the relative algorithms used to analyze the latter. It will shows the process to analyze the private information using **K-means**[6] and **LDA**[8] algorithms.

2. Results

At first time the choice of topics number was equal to **5** as there is not present a **privacy corpus** to train the algorithms in efficient way. But after obtained the **Privacy Tool**[1][2] or rather a tool to analyze the privacy into text, the topics number is changed to **20**, because are extracted a topics to build *pseudo-corpus*. Fig.1 shows the results using **K-Means** on results obtained by LDA computed on dataset:

```
articoloPrivacy/no_tool.jpg
```

Figure 1. K-Means using 5 Topics clustering

articoloPrivacy/kmeans_20_topics.png

Figure 2. K-Means using 20 Topics clustering

We note that in Fig. 2 how the clustering of topics change form because the high number of topics categorizes the word in relative topic. We do not show the *label* word because the figure would be confused but it is possible try to show this result using the link of project in relative section.

2.1. Formatting of Mathematical Components

To perform the exactly clustering of the words to relative topics it is used a **Perplexity formula** which it is computed by every run of LDA [7]:

$$P_{\mathrm{Dt}} = \exp - \frac{\sum_{d=1}^{Dt} \log p(w_d|\alpha, \beta)}{\sum_{d=1}^{Dt} N_{\mathrm{d}}}$$

Where:

- **P** = Perplexity;
- **Dt** = Dataset;

- **Summation** is the total number of Token obtained from Tokenization;
- **log** represents the probability of a word belonging to a topic.

Lower perplexity scores represent a more robust model.

The Table 1 shows several topics obtained from K-Means after LDA results:

Topic	**Top 10 terms**
Sports	years champions football win winner lose number player games game
People	https person people love hate boy girl fuck age man woman
Emotions	https hate love funny friends angry strange bad annoy play
Personal	people life things make hate https find person age mother
News	https wtf fuck kill blurred video high news sports school

Table 1. Top 5 topics

3. Discussion

We have noted how overtime the people share a lot of private information using more words. The analysis has been computed on dataset of April 2018 which dimension equal to **41 GigaByte** or rather over **220k** tweets and we have compared with result obtained in 2011 [7] but the result are similar.

4. Materials and Methods

In this section will be described the step to reach the goal and in particular the techniques used.

4.1. Extrapolation of Dataset

The dataset has been download in April 2018 using NVIDIA nvParse or rather a parser of csv[1] file using the GPU to speed up this process. From dataset only the body of tweet and the sex of user were extracted.

4.2. LDA and K-means

The first technique used is the **Latent Dirichlet Allocation** is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities [8]. While the second technique used is **K-Means** [9] that has in input the data computed by LDA then transformed in numerical data as the result will be more exactly.

We have used **Jupyter** environment so using the **Python libraries** so:

- scikit;
- nltk;

The first library has been used to compute machine learning algorithms as K-Means [5], while the second library used to compute *natural language processing* routines.

The steps of analysis have been:

1. Make a *corpus* on Privacy target;
2. Processing the text to perform with LDA using NLP process or rather Tokenization and Lemmatization;
3. Compute LDA algorithm, choosing the exact number of topics, on Text cleaned by NLP routines;

[1] Comma Separated Value.

4. Transform in fitting data the result of LDA, in particular transform in numerical data;
5. Execute K-means on precedent data;
6. Plot and shows the results.

During the building of LDA model we have used a **Bayesian** method of learning to have a result more efficient and exact. So we have chosen Bayesian classifier to compare this result with another results but we have note the similar final results.

5. Conclusions

A good idea would be use a Supervised Machine Learning method because both techniques used are **unsupervised**, but this is difficult so the privacy model is a subjective and build a corpus would go again more rules.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
GPU	Graphics process unit

References

1. Vasalou, A., Gill, A. J., Mazanderani, F., Papoutsi, C., Joinson, A. (2011). Privacy dictionary: A new resource for the automated content analysis of privacy. Journal of the American Society for Information Science and Technology (JASIST), 62(11), 2095-2105. https://doi.org/10.1002/asi.21610
2. Gill, A., Vasalou, A., Papoutsi, C. and Joinson, A. (2011)Privacy Dictionary: A Linguistic Taxonomy of Privacy for Content Analysis. Proceedings of CHI, ACM Press, Vancouver, Canada.
3. K. Nalini andL. Jaba Sheela. Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter.
4. Guan, P. (2016). K-means Document Clustering Based on Latent Dirichlet Allocation.
5. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12 (November 2011), 2825-2830.
6. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation," in IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 24, no. , pp. 881-892, 2002. doi:10.1109/TPAMI.2002.1017616 keywords:Pattern recognition; machine learning; data mining; k-means clustering; nearest-neighbor searching; k-d tree; computational geometry; knowledge discovery., url:doi.ieeecomputersociety.org/10.1109/TPAMI.2002.1017616
7. Caliskan, A., Walsh, J., Greenstadt, R. (2014). Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network. WPES.
8. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003), 993-1022.
9. Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.
10. A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1524–1534. Association for Computational Linguistics, 2011
11. K. Thomas, C. Grier, and D. M. Nicol. unfriendly: Multi-party privacy risks in social networks. In M. J. Atallah and N. J. Hopper, editors, Privacy Enhancing Technologies, volume 6205 of Lecture Notes in Computer Science, pages 236–252. Springer, 2010

12. H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, pages 1–12. ACM, 2011
13. Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. ”i regretted the minute i pressed share”: A qualitative study of regrets on facebook. In Proceedings of the Seventh Symposium on Usable Privacy and Security, SOUPS ’11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.

Sample Availability: https://colab.research.google.com/drive/1y9nmKtW0M3R4Oeu5KYgEptY6lP-QCggp