COS720 – Cybersecurity

2018 – Assignment

Prof Jan Eloff

Total marks: 80
Bonus marks: 20

Due dates:

14 May 2018 Submit your assignment on COS720 web portal and sign anti-plagiarism agreement. Upload all code/ screenshots/ visualisations of data/ project design documentation.

Book for a hands-on practical demonstration – schedule and demonstration date to be confirmed. Check the CS web portal.

Background:

The practice of sending emails and the use of Internet are ever-present in the current day working environment. As such, an employee spends a fair amount of the workday responding to and sending out emails from his/ her company email account. It is also known that amidst those "business as usual" emails sent by employees, are non-company related mail and perhaps malicious activities and behaviours. These threats can originate inside the company itself or from external parties. Employees might, for example, be using their company email accounts to rant about the company and thereby negatively impact on the culture of the company, partake in insider trading, violate laws (such as the Protection of Personal Information Act in South Africa), not comply with company policies, or by mistake send a corporate email to a non-secure public account. The company might also fall victim to phishing, malware, and ransomware attacks from both inside or outside the company. This project aims to identify these abnormal, malicious activities by performing authorship analysis on a large dataset of emails pertaining to a large organization. Authorship analysis for this project will be done on the metadata that is related to emails.

You will design and implement a solution where a large dataset of emails will be used from a publicly available set like the Enron employee email dataset. You can focus on any one specific threat found in emails to date such as impersonation and insider threats. Remember that the email content / body data is not to be used for this project as we are performing authorship analysis on the metadata related to emails. You can use any Big Data Science (Big Data & Data Science) tools or language for the task at hand.

Background reading:

- What information can be found in the header of an email:
    - https://www.howtogeek.com/108205/htg-explains-what-can-you-find-in-an-email-header/
- Link on how emails were manually generated to look like it came from Hillary Clinton:
    - http://www2.cs.uh.edu/~gnawali/papers/phishing-asiaccs2017.pdf
- Links to email datasets and past research:
    - https://www.cs.cmu.edu/~enron/
        - Note that the Enron email dataset has been used in lots of past research and in online competitions like Kaggle, see the URL given earlier on in this assignment. Please do not copy or re-use any of the existing approaches. Do your own work.
    - http://ieeexplore.ieee.org/abstract/document/8002481/
    - https://link.springer.com/chapter/10.1007/978-3-540-30115-8_22
    - http://www.aclweb.org/anthology/W17-2408
- Privacy modelling:
    - https://www.inderscienceonline.com/doi/abs/10.1504/IJBDI.2016.073904
- Authorship Analysis
    - https://link.springer.com/article/10.1007/s11063-017-9593-7
- Machine learning
    - https://dl.acm.org/citation.cfm?id=1299021

Main objective of this assignment:

You need to design, implement test and demonstrate a Big Data Science & Cybersecurity solution. This solution should identify cyber-threats such as malicious behaviour; impersonation or any other insider threats that can be discovered from a large email dataset.

In this assignment, you will be performing the following main tasks:

1. Data Acquisition / Generation / Fabrication: Use an email header meta-data structure, such as what is available for Gmail, and acquire from a publicly available email dataset, such as the Enron email dataset, at least 200,000 emails. Alternatively you can fabricate your own email dataset as long as the emails are created according to a standard metadata email structure and you are able to eventually discover some cyber threats in it. (10 marks)

2. Data Cleaning: Your data needs to be reliable. Dirty data is data which has values which are missing, incorrect or inconsistent (Krishnan et al., 2015). Do data cleaning on your dataset with a program / script. (10 marks)

3. Do exploratory data analysis on the email dataset keeping in mind you want to do authorship analysis so to eventually detect any kind of malicious or abnormal behaviour such as impersonation.

   Look at these kernels available on Kaggle as a start:
   https://www.kaggle.com/wcukierski/enron-email-dataset/kernels
   (10 marks)

1. Identify attributes from emails using only email header information that can play a role in detecting any abnormal or malicious behaviour in the email dataset you are working with. Keep in mind you are working towards conducting authorship analysis on the email dataset from a cybercrime point of view.

   (10 marks)

2. Do feature engineering by using the attributes identified in step 2 above that will assist you in authorship analysis for the purposes of cybercrime.(10 marks)

3. Use machine learning tools and techniques, minimum of two, to detect any abnormal behaviour in the email dataset. (10 marks)

4. Do cyber criminal profiling of an individual or a group of people. You can only use data from attributes and features originating from the metadata of email headers.

   (10 marks)

5. PPDM: As it is illegal to use personal data of a person in South Africa without the person's consent, (Luck, 2014), the email dataset has to be anonymised. Xu et al. (2016) maintains that it is possible to preserve privacy when performing machine learning by means of Privacy Preserving Data-Mining (PPDM). Investigate techniques to anonymize the email dataset you have used. Create a script/program to anonymize your email dataset. Please bear in mind that the data needs to maintain its machine learning utility / value without having enough information to uniquely identify a person.

   (10 marks)

References

1. Luck, Russel, "POPI-Is South Africa keeping up with international trends?", May (2014).
2. Xu, Lei; Jiang, Chunxiao; Chen, Yan; Wang, Jian & Ren, Yong, "A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining", Computer (2016), 54--62.
3. Krishnan, Sanjay, et al. "SampleClean: Fast and Reliable Analytics on Dirty Data." IEEE Data Eng. Bull. 38.3 (2015): 59-75.

General

1. Work in groups of 3 or 4 persons. Each person must be able **to clearly identify** his/her contribution. Please ensure that your name is linked to a group not later than 6 March. I will circulate a list in class the evening of 6 March for this purpose.

2. The objective of this assignment is to let you explore with big data and data science technologies within the cyber-security domain. Some of the tasks will be easy to accomplish whilst other are more difficult. It is important that you can show in the assessment meeting what you have learned – you can still learn a lot without getting everything 100% accomplished!

3. The assignment counts 80 marks.

4. An additional 20 marks can be earned as follows: (1) demonstrate your solution in full on the CS Data Science cluster (10 marks) and (2) demonstrate an extended solution on the CS Data Science cluster showing the discovery of any other cyber-threat(s) wrt email communications that is not part of your normal project submission as detailed in this document.

5. Assessments will be face-to-face and each group member must be present at the evaluation session to get a mark and each person must be able to clearly demonstrate his/her individual contribution to the group effort. Not all members in a group will necessarily get the same marks. A schedule for booking a time-slot will be posted on the CS COS720 web-portal for this purpose.

--o0o--