Thesis Title: TBD

Introduction (5-6 pages)

Introduction

Data science has become a popular discipline in extracting useful information from lots of data. Data science involves the use of scientific methods, mathematics, algorithms and visualization techniques to gain insight into data [citation needed]. Data is generally divided into two categories namely structured and unstructured. Computer systems are designed with protocols and formats to structure data for storage, retrieval, transmission and routing. The data itself may provide further structure or be regarded as unstructured.

As an example, consider a database of financial transactions. The database may provide structure in the form of transaction history and the outcome of associated calculations of the transactions. Source documents like scanned invoices may also be attached as source documents for the calculations. Structure is provided by the database up to associated invoice level, but the details of the invoice can be regarded as unstructured because it requires a human to interpret the data in the invoice.

It is possible to apply machine learning techniques to capture the data in the invoice by means of optical character recognition, interpret the transaction and compare it to the captured transaction. In such a case scientific methods and algorithms that rely on statistical relationships are used to automatically capture and structure the unstructured data i.e., the information in the invoice. It can save a lot of time in e.g. audits of the transactions by extracting more information from associated documentation and flag the necessary transactions to the auditor which requires human intervention.

Similarly, an email server stores emails in an ordered way so that the email user can search for emails, often by sender, receiver, date of arrival etc. Emails also contain an email header which is used by the transmitting protocols to route the email. The email header is partially displayed as information to the email user e.g. the To, From, CC and Subject fields, but some information is hidden to the user, e.g. the receiving server. The email header is structured with a strict set of rules [cite RFC] governing detailed aspects of header construction. The email body is largely unstructured except perhaps for the definition of attachments and their encoding, or in some cases some structure is provided by embedding html code for email display. The actual email text is, from the point of view of the email standard, largely unstructured.

Criminal investigations involve the reconstruction of the most likely event that led to a crime. The reconstruction of the sequence of events rely on data available to the investigator e.g. witness statements, financial transactions, camera footage correspondence and forensic data [1]. In attempting to identify the profiles of suspects, reliance on statistical data is often used [1]. One can therefore argue that investigators apply data science to reconstruct events and prove guilt or innocence of a suspect. It is possible to draw an analogy with the database example presented earlier. A case file is similar to a database. The investigator classifies information relevant to the reported crime in the form of witness statements, forensic evidence, intelligence and interviews. This provides structure to the case data. The investigator is responsible to develop multiple lines of enquiry [1], gather and order more data and interpret the unstructured data to extract more information, ultimately proving guilt or innocence.

There are three main questions that the investigator deals with [1]:

* What crime was committed?
* Who are the suspects?
* What evidence exist to support a line of enquiry?

A major source of forensic evidence is digital forensic evidence [2]. Digital forensics has its roots in the observation that digital devices such as computers are utilized in committing crimes [2], and often the crime only takes place in the virtual or cyber world. Examples of such crimes are child pornography, theft of personal data, ransomware encrypting a person’s data which will only be released for a fee, etc. From these examples it is clear that digital forensics is a highly specialized field which often requires technical knowledge of the inner workings of digital devices, communication protocols and networks to find and extract evidence leading to catching the perpetrator [2]. A digital forensic investigator is therefore a specialist investigator supporting general investigators in investigating crimes, similar to e.g. a forensic pathologist or a forensic financial investigator [2]. Some of specialist aspects the digital forensic investigator is well equipped to deal with are [2]:

* Reconstructing network communication events through log analysis to determine how a hacker gained access to a computer system.
* Identifying and analyzing deleted files with the purpose of reconstructing such files as evidence.
* Analyzing malware to determine the purpose of the malware, how it was installed and what possible crime was committed by executing the malware.
* Extracting information from files, metadata, communications like email to assist the investigator in finding evidence for lines of enquiry.

A special relationship exists between an investigator and a forensic investigator. Each has its own focus and is complimentary. The forensic investigator must be able to present evidence to the investigator and the court in an understandable manner. The investigator needs to understand all relevant detail presented to him or her without becoming a forensic investigator. Investigators should not hesitate to question what the forensic investigator presents to him/her [3].

During an investigation as much as possible evidence material is collected that might be applicable to the investigation [3]. This allows for the formulation of many lines of enquiries. During the investigation some lines of enquiry are eliminated which lead to an attrition of relevant material for prosecution [3]. In fact it may be argued that an investigation starts with lots of material or data, gradually reducing it to the core evidence by means of illimitation. All material gathered has to be preserved even if it is not used in the prosecution.

Legislation in some countries require the police service to record all relevant evidence material gathered, retain the material and reveal all material relevant to a prosecution to the defense even if such material is not used in the prosecution [3]. Evidence presented in court needs to stand up to scrutiny of all alternative hypothesis so that there is no reasonable doubt the presented evidence points to the accused [3].

During an investigation new evidence may become available which opens more lines of enquiry. The investigators will therefore have to revisit evidence previously examined with new lines of enquiry in mind [3].

Increasingly crime is committed through digital channels and it often results in large volumes of data that needs to be analyzed. The Enron case is one such case where prosecution was successful due to evidence uncovered in over 200 000 emails that was seized on company servers [2]. Emails contain both metadata in the form of email headers and routing information, attachments in the form of photographs, invoices and other documents, and natural language information in the form of the email itself. The processing of such a large number of emails requires skills from both the investigator as well as the digital forensic investigator.

Problem statement

Natural language processing is one possible solution from data science that can be used in processing large amounts of emails, looking for evidence. There are a number of challenges to applying it in practice:

* It is not necessarily possible to train the solution as a supervised training model, as that would require labelling of the data beforehand.
* The natural language processing solution should provide significant time or cost savings compared to using keyword search instead.
* Reconstruction of events and timelines cannot come only from the natural language processing solution. Instead, it should be combined with the metadata in the email headers.
* The natural language processing solution should not introduce bias into an investigation potentially leading to a situation where an investigator overlooks evidence which the defense may discover to cast reasonable doubt in a court case.

This list is not exhaustive. It shows that there are a number of questions that arise when one wants to apply natural language processing to emails successfully to aid investigations.

This work addresses the following research questions:

* What is a suitable natural language processing model to use on emails?
* What are the required processing steps to perform on the emails to extract useful information for the investigators?
* How can the results of the email processing be presented to the investigators in a useful manner to aid in their investigation?

Motivation

Objectives/methodology

Layout

Conclusion