Yea-Wen Yang, Fall 2003

Identity Theft – Mining Social Security Numbers (SSNs) from Online Resumes & Curriculum Vitae

Abstract

Identity theft is a serious problem here in the United States. The overuse of personal information (SSNs, date of birth, etc.), ease of obtaining such information, and lack of public awareness has contributed to the growing problem. This project had 2 focuses: [1)] developing a program that could mine online resumes and then determine those web pages that were real resumes with Social Security Numbers, [2)] demonstrating the ease at which SSNs can be mined and determine how many people put themselves at higher identity theft risk because they include their SSN on their resumes/CVS. The method I used was to design a Java program using the Google API to gather statistics & web page samples. Then developing & training a Boolean function to determine if a given webpage is a real resume/CV with a SSN. The results show [1)] the boolean function is very accurate more than 90% correct classification [2)] over 300 Social Security Numbers were mined from various online resumes and curriculum vitae.

Background

Identity theft is "when someone uses your personal information (name, SSNs, credit card number, or other identifying info) without your permission (1)". The consequences can be very severe and often victims spend inordinate amounts of time fixing/cleaning up the mess that is left behind. Loss of money, job opportunities, and loans, and accusations of committing crimes can be just some of the consequences from identity theft.

According to the Federal Trade Commission 2003 report, ~10 million Americans discovered that they were victims of some form of ID theft last year and total loss to business from "new accounts" ID theft equaled $33 billion. Also $50 billion were lost last year due to credit card theft and over ½ of ID theft victims didn't know how their personal information was stolen.

Anyone can be at risk for identity theft because of the overuse of personal info (SSNs), ease of obtaining such information, and lack of public awareness has contributed to the growing problem. The few pieces of personal information needed for identity theft, like SSN, date of birth, and mother's maiden name can sometimes be found from a variety of different databases and online information. However, people who include their SSNs on their online resumes/CVs are even more vulnerable because they've given key pieces of personal information to a thief who now can use that information on different databases to find other personal info on the victim. So in general, the more personal information an individual puts online, the easier it is to find other personal information, and the more likely that the individual will be a victim of ID theft.

Method

For the first part of the project focus, I implemented a program to gather web pages given a particular query, examine the web pages, and determine those pages that were real resumes with SSNs. The process is as follows:

1. I used Sweeney's Google API, the Google interface, and Java to obtain the web page results (html tags and all) from my queries.
2. Stripping the web pages (obtained in the step above) of HTML tags was actually more difficult than I first originally thought i.e. '<' & '>' do not always donate HTML tags sometimes appear in JavaScript as less than & greater than. I used a combination of Swing functions and my own HTML parser to do the processing.
3. With only the text left, the Boolean function runs through the file and then determines if it is a non-resume, resume, or a resume with SSN. I defined non-resumes as sample resumes or pages on how to write resumes. Basically all pages that are not real resumes.

For the Boolean function, I found it much easier to first determine if the page was a non-resume. So before the Boolean function evaluation, I assumed that I have a non-resume page. The Boolean function is then called and it looks for key word and phrases from the text. If the function finds a non-resume word/phrase, then it stops evaluation and concludes the pages is not a resume otherwise it continues to the end of the document. Each "resume word" is counted and at the end of the Boolean function, if a non-resume word/phrase was not found and more than one "resume word" was present, the page is determined to be a real resume. If the page is a real resume, then the function will also determine if a SSN was present.

Non-resume phrases: resume writing, resume posting, resume guide, resume sample, career service, career center, action word, action verb, help guide, sample resume, job search, job hunt.
Non-resume keywords: should, must, login, log-on, log on
Resume words: objective, experience, education, university, college, director, manager, supervisor, position, award, GPA

Using key phrases was also much more effective than just using key words. For example, if the word "sample" was found it doesn't necessary mean that the page is not a resume. It could be a title of a research paper or description about "samples collected". However, the phrase "sample resume" or "resume sample" is more likely indication of a non-resume page.

I trained the function on 330 web pages: 71 non-resume, 79 resume, and 181 resumes with SSNs. Note that the majority of SSNs found were from PDF CVs. The program can't save PDF files, however Google does cache the pages as html and then I manually saved each page.

For the second part of the project, I used my program to gather more statistics and samples. I ran a broad query (vitae ssn or "social security number"), collected 266 html pages and ran those files through my program to get an estimate of how many people put their SSN online resume/CVs i.e. how many people put themselves at higher risk for ID theft.

Results

I gather over 330 web pages to test and train my Boolean function. I ran the program and then manually verified my results.

	Total	# Identified as non resumes	# Identified as resumes	# Identified as resumes w/SSN
Non resumes	71	65	5	1
Resumes	79	3	76	0
Resume w/SSN	181	8	0	173

The program correctly identified 65/71 (92%) non-resumes, 76/79 (96%) resumes, and 173/181 (96%) resumes w/SSNs. Another analysis shows that 11/65 (17%) files were incorrectly identified as non-resumes, 5/76 (7%) resumes, and 1/173 (.6%) as resumes with SSNs.

As for the second part of the study, Google estimates that there are 10,300 hits for the query (vitae ssn or "social security number"). I collected 266 html files from that query and the results from my program show 1.5% have real SSNs. Also I collect 187 PDF files earlier that also contained real SSNs.

Discussion

The Boolean function is fairly accurate. It makes sense that the function has a higher rate of error when it incorrectly identifies a page as a non-resume given my earlier reasoning to put more weight on non-resume words and to initially assume the given page is a non-resume before the Boolean evaluation. However, this also means that the percentage for error is less when classifying resumes & resumes with SSNs which the results show.

It was also a good decision to use more non-resume phrases versus key words in the Boolean function. In reviewing the log files, I found several instances where if I had used key words such as "sample" instead of "sample resume" I would have more incorrect classifications then my final results.

In manually checking the errors made by the Boolean function, I found in most cases all the errors were understandable. The majority of incorrectly identified files contained my non-resume key word/phrase when they were, in fact, resumes. However as the statistics show those incorrectly classified files were small and in general I still believe the word/phrases I choose served as good indications of non-resume pages. In one of the file, the program did not pick up the phrase "sample faculty vita/resume" because it was looking for "sample vita" or "sample resume". Also in another example, an internship resume application was not recognized as a non-resume page because it didn't contain any of my non-resume key word/phrases. In these two cases, I could have modified my Boolean function to correctly classify these files however, it's important not to over train or fit the Boolean function exactly to the data set. Over training will result in a decline of

performance when the same function is tested on different data sets because the function was made to work specifically to a particular type of data.

In the second part of the project, I ran my Boolean function on the samples collect on the Google query (vitae ssn or "social security number"). Out of the 266 files, 34 files had real SSNs. That means that 1.5% were more vulnerable to ID thefts. 1.5% of 10,300 hits are approximately 150 SSNs. This is a slight overestimation given that Google shows the more relevant materials first and the end results are usually not what the user was looking for. However other than going through the entire list of hits, this was the best way to determine how many online pages (resumes/CVs) contained SSNs. The above results do not include the PDF files. Those PDF files I collected earlier to train my Boolean function totaled to 181, which means the total number of SSNs I collected was roughly 330.

My results show that there were more PDF than html files that contained SSNs. I'm not really sure why that is, although I'm guessing that printed PDF files are nicer to view than html files. Another interesting observation during my studies was the discrepancies on instructional pages in writing resumes & CVs. Many resume-writing pages geared towards industry-working professionals urged NOT to include SSNs on online resumes. However, instructional CV-writing pages do just the opposite. These pages encouraged professors/researchers to include their SSNs. I'm not sure why but the majority of the web pages I collected with SSNs were from professors'/researchers' CVs. Perhaps this incorrect perception of not protecting SSNs has led some professors to post online grades using students SSNs (another **bad** thing to do).

Considering the number of SSNs I collected, I'm sure many of these individuals are not aware of how much more risk they are putting themselves by including their SSNs online. A future extension to this project is to collect the names & emails and send these individuals notification about online resumes with SSNs and identity theft. It would also be interesting to observe and note the recipients' response to the email; this could make for future studies on the public's awareness & response to identity theft.