NLP Applications in Security: Spamming and Phishing

Definitions

Email spam, also known as junk email or unsolicited bulk email (UBE),is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Many email spam messages are commercial by nature but email spam messages may also contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware.

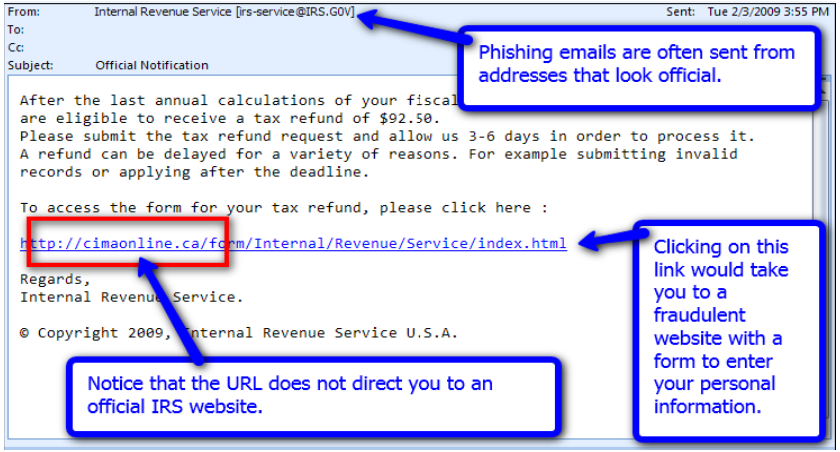
Phishing is the act of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication.

SPAMMING

Spamming is when a cyber criminal sends emails designed to make a victim spend money on counterfeit or fake goods.

PHISHING

Phishing attacks are designed to steal a person's login and password details so that the cyber criminal can assume control of the victim's social network, email and online bank accounts.



Anti-Phishing Working Group Q1 2016 Report

The Retail/Service sector remained the most- targeted industry sector during the first quarter of 2016, with 42.71% of attacks.

The number of brands targeted by phishers in the first quarter remained constant – ranging from 406 to 431 brands each month.

The United States continued its position at top on the list of nations hosting phishing websites.

In Q1 2016, 20 million new malware samples were captured.

The world's most-infected countries are led by China, where 57.24% of computers are infected, followed by Taiwan (49.15%) and Turkey at 42.52%.

The proportion of emails with malicious attachments grew by 0.1 percentage points compared to Q4 2011 and averaged 3.3%. The share of phishing emails averaged 0.02% of all mail traffic.

Spam-filtering techniques

Two classes of methods have been shown to be useful for classifying e-mail messages. The rule based method uses a set of heuristic rules to classify e-mail messages while the statistical based approach models the difference of messages statistically, usually under a machine learning framework.

Rule based approach is fruitful when all classes are static, and their components are easily separated according to some features.(However this is not the case most of the time.)

Simple searching strategy

Many spam-filtering techniques work by searching for patterns in the headers or bodies of messages. To defeat such filters, the spammer may intentionally misspell commonly filtered words or insert other characters, often in a style similar to leetspeak, as in the following examples: V1agra, Via'gra, Vi@graa, vi*gra,\/iagra. This also allows for many different ways to express a given word, making identifying them all more difficult for filter software.

Naive Bayes spam filtering

Naive Bayes spam filtering is a baseline technique for dealing with spam that can tailor itself to the email needs of individual users and give low false positive spam detection rates that are generally acceptable to users.

Bayesian spam filtering

Particular words have particular probabilities of occurring in spam email and in legitimate email. For instance, most email users will frequently encounter the word "Viagra" in spam email, but will seldom see it in other email. The filter doesn't know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user must manually indicate whether a new email is spam or not. For all words in each training email, the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database. For instance, Bayesian spam filters will typically have learned a very high spam probability for the words "Viagra" and "refinance", but a very low spam probability for words seen only in legitimate email, such as the names of friends and family members.

After training, the word probabilities (also known as likelihood functions) are used to compute the probability that an email with a particular set of words in it belongs to either category. Each word in the email contributes to the email's spam probability, or only the most interesting words. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the email's spam probability is computed over all words in the email, and if the total exceeds a certain threshold (say 95%), the filter will mark the email as a spam.

Dealing with rare words

The words that were encountered only a few times during the learning phase cause a problem, because it would be an error to trust blindly the information they provide. A simple solution is to simply avoid taking such unreliable words into account as well.

Applying again Bayes' theorem, and assuming the classification between spam and ham of the emails containing a given word ("replica") is a random variable with beta distribution.

Bayesian filtering

As Bayesian filtering has become popular as a spam-filtering technique, spammers have started using methods to weaken it. To a rough approximation, Bayesian filters rely on word probabilities. If a message contains many words that are used only in spam, and few that are never used in spam, it is likely to be spam. To weaken Bayesian filters, some spammers, alongside the sales pitch, now include lines of irrelevant, random words, in a technique known as Bayesian poisoning. A variant on this tactic may be borrowed from the Usenet abuser known as "Hipcrime"—to include passages from books taken from Project Gutenberg, or nonsense sentences generated with "dissociated press" algorithms.

Markovian discrimination

Markovian discrimination in spam filtering is a method used in CRM114 and other spam filters to model the statistical behaviors of spam and nonspam more accurately than in simple Bayesian methods. A simple Bayesian model of written text contains only the dictionary of legal words and their relative probabilities. A Markovian model adds the relative transition probabilities that given one word, predict what the next word will be. In essence, a Bayesian filter works on single words alone, while a Markovian filter works on phrases or entire sentences.

Markov models

There are two types of Markov models; the visible Markov model, and the hidden Markov model or HMM. The difference is that with a visible Markov model, the current word is considered to contain the entire state of the language model, while a hidden Markov model hides the state and presumes only that the current word is probabilistically related to the actual internal state of the language.