Phinding Phishes

A quick and dirty way to identify phishing domains and spell check the internet.

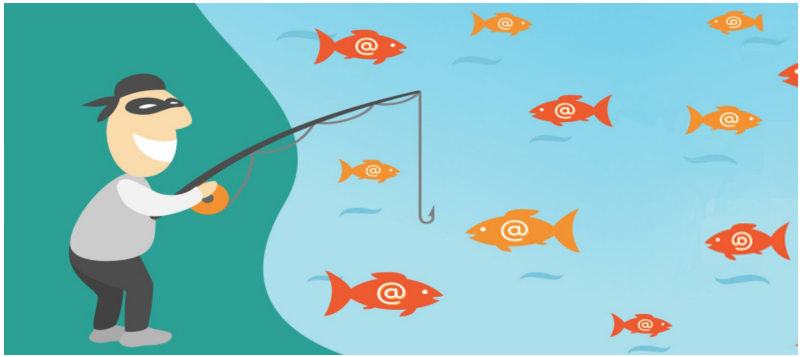
WHOAMI

ben smith.

tinkerer.

hater of powerpoint.

pcap junkie.



Intro

There had to be a way to identify various types of phishing domains based on misspelling. I wanted basically a spell check for the Internet.

➔ **Early Ideas**

It all started by wanting to identify malware masquerading as legit Windows processes..

➔ **Spell Check**

There are a few functions spell check uses to identify when words are wrong.

➔ **Apply to Phishing**

Why not apply the same idea and make a spell check for the Internet?

Types of Attacks

A few common examples are listed below. Phish Phinder works fairly well to detect each of these.

- **Typosquatting**
  Just a normal typo in the domain name.
- **Bitflipping**
  Change a single bit and you'll still hit a decent percentage of top site users.
- **Homoglyph**
  I'd click that. Does it look the same as an existing domain? Can you substitute a zero for an oh?

String Comparison Methods

Levenshtein Distance

A.k.a. **Edit distance**. This is how many changes it takes to get from one string to another.

Cosine Similarity

Create two vectors consisting of the counts of the unique letters in each string and calculate the cosine of the **angle between two vectors**.

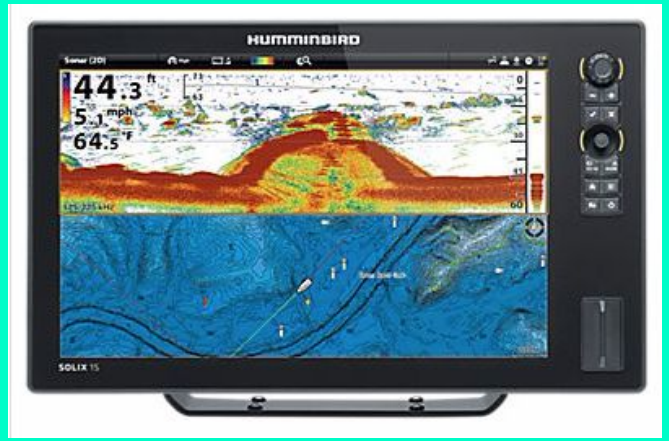
Others

Check out

https://en.wikipedia.org/wiki/Category:String_similarity_measures

Phish Phinder is born.

- Input: Potential Phishing Domain
- Input: Top 500 Domains
- Calculate Levensthein Distance
- Calculate Cosine Similarity
- Combine using scoring method
- Report anything above scoring threshold



Scores:

I tried several different methods to combine the two string comparison methods. I ended up dropping anything with edit distance less than 3 and then weighting them at a 3 to 1..

Testing

I ran a few tests to weed out false positives (there were lots).

- ➔ **dnstwist**
  I generated and ran lists of phishing domains for kickstarter (93%), amazon (92%), reddit (90%) with very good results..

- ➔ **Alexa Top 1 Million**
  I also kicked the tires against the alexa top 1 million (minus the top 500). This yielded 197 results. Many of these appeared to be **parked phishing domains**! This also identified some serious false positives, but still < 00.002%

```
[*] observed domain kickstarte4.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.955533085906; score: 18.5553308591
[*] observed domain kickstzrter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.95652173913; score: 18.5652173913
[*] observed domain kickst1rter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.95652173913; score: 18.5652173913
[*] observed domain kickstarrer.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.959166304663; score: 18.5916630466
[*] observed domain kickstaeter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.95652173913; score: 18.5652173913
[*] observed domain kickytarter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.95652173913; score: 18.5652173913
[*] observed domain k.ickstarter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain ki.ckstarter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kic.kstarter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kick.starter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kicks.tarter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kickst.arter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kicksta.rter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kickstar.ter.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kickstart.er.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] observed domain kickstarte.r.com looks a lot like kickstarter.com and may be a phish. edit distance: 1; cos similarity: 0.981432984131; score: 18.8143298413
[*] average phish score: 18.7079248796
[*] total found: 284, total rows: 305, percent: 0.931147540984
```

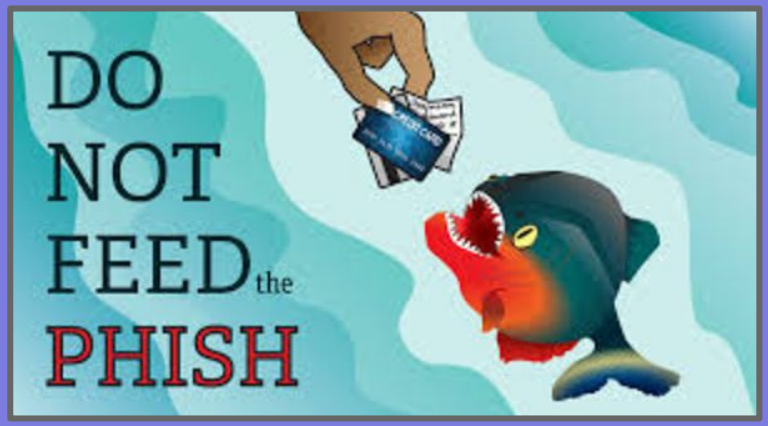
```
[*] observed domain varchive.org looks a lot like archive.org and may be a phish. edit: 1; cos: 0.970725343394; score: 18.7072534339; new_score: 0.029
[*] observed domain youtuibe.com looks a lot like youtube.com and may be a phish. edit: 1; cos: 0.968245836552; score: 18.6824583655; new_sc
[*] observed domain amazon.om looks a lot like amazon.com and may be a phish. edit: 1; cos: 0.968245836552; score: 18.6824583655; new_score:
[*] observed domain cricbuz.com looks a lot like cricbuzz.com and may be a phish. edit: 1; cos: 0.976187060184; score: 18.7618706018; new_sc
[*] observed domain aliyunc.com looks a lot like aliyun.com and may be a phish. edit: 1; cos: 0.964763821238; score: 18.6476382124; new_scor
[*] observed domain blogpot.com looks a lot like blogspot.com and may be a phish. edit: 1; cos: 0.971825315808; score: 18.7182531581; new_sc
[*] observed domain twiter.com looks a lot like twitter.com and may be a phish. edit: 1; cos: 0.98019605882; score: 18.8019605882; new_score
[*] observed domain iclud.com looks a lot like icloud.com and may be a phish. edit: 1; cos: 0.96698755683; score: 18.6698755683; new_score:
[*] observed domain exstratorrent.cc looks a lot like extratorrent.cc and may be a phish. edit: 1; cos: 0.984250984251; score: 18.8425098425
[*] observed domain gooogle.com looks a lot like google.com and may be a phish. edit: 1; cos: 0.989949493661; score: 18.8994949366; new_scor
[*] observed domain worldpress.com looks a lot like wordpress.com and may be a phish. edit: 1; cos: 0.974679434481; score: 18.7467943448; ne
[*] observed domain uyoutube.com looks a lot like youtube.com and may be a phish. edit: 1; cos: 0.981495457622; score: 18.8149545762; new_sc
[*] observed domain chaturebate.com looks a lot like chaturbate.com and may be a phish. edit: 1; cos: 0.979130048652; score: 18.7913004865;
[*] observed domain xhamaster.com looks a lot like xhamster.com and may be a phish. edit: 1; cos: 0.972305585328; score: 18.7230558533; new_
[*] observed domain spedtest.net looks a lot like speedtest.net and may be a phish. edit: 1; cos: 0.990043675898; score: 18.9004367759; new_s
[*] observed domain hostar.com looks a lot like hotstar.com and may be a phish. edit: 1; cos: 0.96896279025; score: 18.6896279025; new_score
[*] observed domain prnhub.com looks a lot like pornhub.com and may be a phish. edit: 1; cos: 0.964763821238; score: 18.6476382124; new_scor
[*] observed domain taoobao.com looks a lot like taobao.com and may be a phish. edit: 1; cos: 0.989949493661; score: 18.8994949366; new_scor
[*] observed domain witter.com looks a lot like twitter.com and may be a phish. edit: 1; cos: 0.98019605882; score: 18.8019605882; new_score
[*] observed domain youtrube.com looks a lot like youtube.com and may be a phish. edit: 1; cos: 0.968245836552; score: 18.6824583655; new_sc
```

Phound

Lots of possible phishing domains identified. Some may be preventatively registered by the actual company.

False Positives? Limitations?

- Google is too prolific in the top 1mil so I filtered it.
- Short domains can't be handled well
- With or without TLD? So far I included it.
- No way to look up domain via WHOIS
- Computationally Intensive



Future Ideas

- It would be neat to run this against dnstwist lists and see the types that have the best and worst detections.
- Add a feature to do a whois lookup or query the TLS cert to get info about domain age and registration.
- Tweak and refine the algorithm including trying different string comparisons.

El Phin. Questions?

Code

Code will likely be uploaded to GitHub at some point.

If you'd like to see the list of matches from the top 1 mil, feel free to get in touch with me.