PROfOUND NETWORKS

The Domain Append Match Score

The match score isn't a measure of confidence in the result, but a quantitative measure revealing how much of the input was found on the domain.

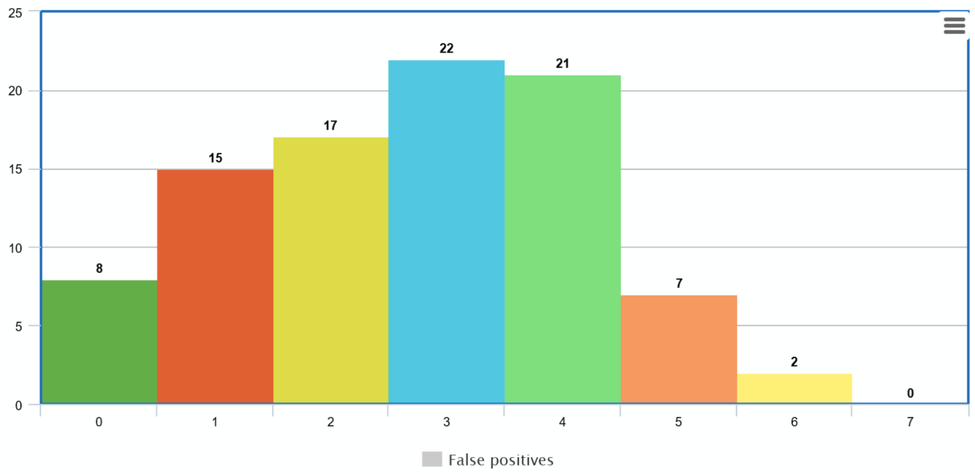
The score strongly depends on how much company data the owner decides to publish on the website in the first place, and on how reliable the input record is. While a high score is indicative of a confident match, a valid match can also have a low score for a variety of legitimate reasons.

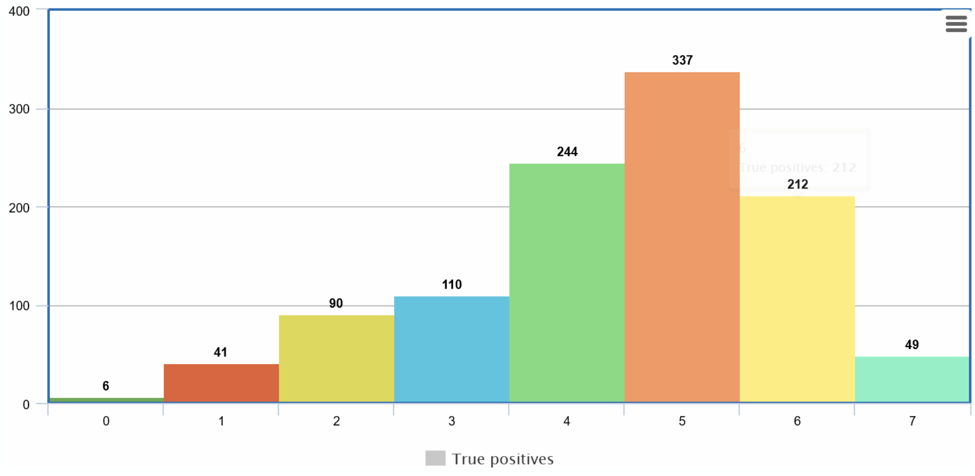
This situation is inherent to the problem itself, simply because:

- Some sites don't publish as much data as others (if the website itself doesn't publish much contact data, it will have a low match score.)
- The input record itself may contain outdated info.

Below, see histograms side by side for the count of "**T**" flags in the matching strings for:

- False positives (above)
- True positives (below)





There can be true positive matches that, from a data standpoint, look exactly the same as false positives (similar matching string, similar score). It is not unexpected for true positive matches to populate the whole range of values.

There are even false positives with a matching string that indicates that almost everything from the input record was found on them, but the domain is still not (or no longer) the correct one.
Two examples:

- **linksmedicalclinic.com** (TFTTTTT) is still listed in internal resources as a healthcare website, but it now redirects to a Chinese portal. This site was recently operational but the domain has since been hijacked.

- **ndag.org** (TTTTPTT) is the website for an agricultural association that the input company is a member of, that lists all the contact data of that company (MAPLE VALLEY AG PRODUCTS LLC).

Other signaled problems were problems of analysis.

The domain **jamessmithhealthclinic.com** for CANADA JAMES SMITH HEALTH STN was listed as redirecting to **starcityniteriders.ca**. But the **www** version of the site, **www.jamessmithhealthclinic.com**, loads a medical website.

The input data in the column "Website address".

One source that accounts for up to half of the spurious results is the data from the input file, under the column "Website address".

When unable to find a match during normal operation, DA can be configured to use the domain provided as input which it treats as a high-trust candidate, it inspects fully, and validates the presence of certain flags.

Because the input is treated as a high-trust candidate, the quality of the output will greatly depend on the quality of the input.

Broken down by label, the "Website address" domain input accounted for the following outputs:

#	Class	Count	Out of	%
1	incorrect	39	92	42.39%
2	broken	34	80	42.50%
3	unsure	54	104	51.92%

Broken entries

We see from the above table that **42.50%** of entries were records where the input domain was treated as a high-trust candidate.

Another **50%**, 40 out of 80, are cached results that had been previously validated through internal processes, but have since then closed. Many still show up in the Google search engine, when performing the query site: DOMAIN or cache: DOMAIN.

Some cases are websites that don't open when visited on **non-www**, but work on **www**:
- https://www.hmms.on.ca/
- https://www.northernlightshealthfoundation.ca/
- http://www.bigcountryenergy.com/
- https://www.trackentertainment.com/
- http://www.heilind.com/
- https://www.firstchoice.bank/
- https://www.ci.oswego.or.us/
- http://www.kibois.org/

Others had a temporary outage:
- https://www.leoniaschools.org/ (works now, is listed as a 403 error during manual review)
- https://www.honeoye.org/ (works now, is listed as a 403 error during manual review)