Privacy Not Found: A Study of the Availability of Privacy Policies on the Web

PennState College of Information Sciences and Technology

Soundarya Nurani Sundareswara, Shomir Wilson, Mukund Srinath, C. Lee Giles

The Human Language Technologies Lab at Penn State http://hltlab.org

Introduction

Users typically visit a website's landing page to read the privacy policy.

Some websites that purport to post the document do not actually provide them.

We investigate the availability of privacy policies derived from a large data set of company domains.

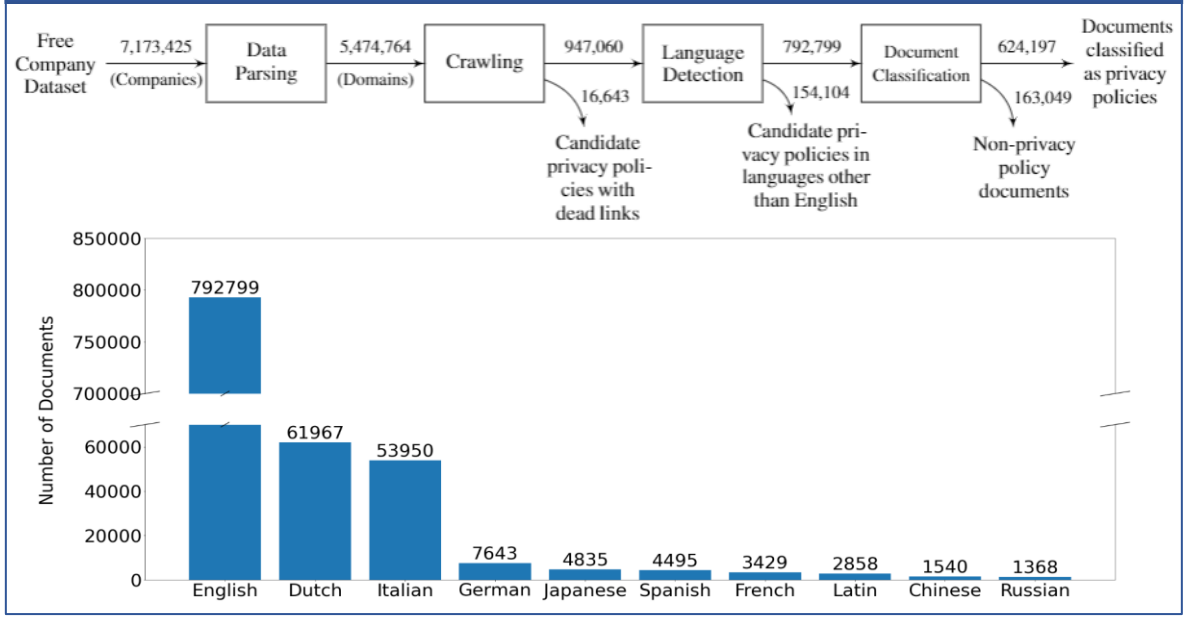
We estimate the frequencies of various anomalies found at each stage of privacy policy collection and overall unavailability of privacy policies.

Crawling Technique Evaluation

Estimating the accuracy of extracting privacy policies using particular keywords to crawl website landing pages.

Accuracy - 77.4% using a random sample of 500 websites.

Document Collection and Classification



Analysis

Dead Links:

- Dead links to candidate privacy policies from website landing pages.
- Error types indicating unavailability: HTTP, Connection refused and Value errors.
- 1.39% of total privacy policy retrieval attempts contained dead links.

Natural Language Discrepancies:

- Privacy policy unavailable in certain languages though website landing page offers selection, placeholder texts in place of actual content.
- 4.4% of 500 randomly sampled presented language inconsistency.
- 1.85% of documents crawled consisted of placeholder texts.

Non-privacy policies:

- Documents classified as non-privacy policy examined for unavailability.
- 6.8% of 500 documents sampled, observed as having empty content.

Overall Estimation

Assumptions: Each website contains only one hyperlink to privacy policy; a large percentage of candidate privacy policies are actual privacy policies.

Anomalies found in privacy policies	% of websites in data set
Dead links	0.28% to 0.41%
Natural language discrepancies	0.15% to 0.39%
Empty content	0.16% to 0.45%

Considering 10,000 websites, 1.62% to 3.38% of websites are estimated to have privacy policy unavailable due to different anomalies identified above.