Removing Spam from Web Corpora Through Supervised Learning and Semi-manual Classification of Web Sites

Vít Suchomel

Natural Language Processing Centre
Masaryk University, Brno, Czech Republic
xsuchom2@fi.muni.cz
https://nlp.fi.muni.cz/en/

Lexical Computing, Brno, Czech Republic
https://www.lexicalcomputing.com/

RASLAN
2020-12-08

What Is Wrong with this Text?

Now on the web stores are very aggressive price smart so there genuinely isn't any very good cause to go way out of your way to get the presents (unless of course of program you procrastinated).

What Is Wrong with this Text?

Now on the web stores are very aggressive price smart so there genuinely isn't any very good cause to go way out of your way to get the presents (unless of course of program you procrastinated).

Web spam, computer generated text – Not a good evidence of natural language phenomena

Web Spam Definition – Search Engines vs. NLP

Search engine:

- The goal: serve relevant, importnant and original texts
- Users: people searching for websites, products, information
- Spammers: interested in search engine results

Web Spam Definition – Search Engines vs. NLP

Search engine:

- The goal: serve relevant, importnant and original texts
- Users: people searching for websites, products, information
- Spammers: interested in search engine results

NLP:

- The goal: words/phrases/sentences/linguistic phenomena in a context in a natural language
- Users: linguists, translators, teachers, NLP scientists and NLP applications
- Spammers: not interested in text corpora

Web Spam Definition – Non-text Types

Good content: fluent, natural, consistent text (regardless its purpose)
Bad content – computer generated text

- machine translation
- keyword stuffing
- phrase stitching
- synonym replacement
- automated summaries
- any incoherent text

Varieties of spam removable by existing tools dealt with by other means

- duplicate content
- link farms
- redirection

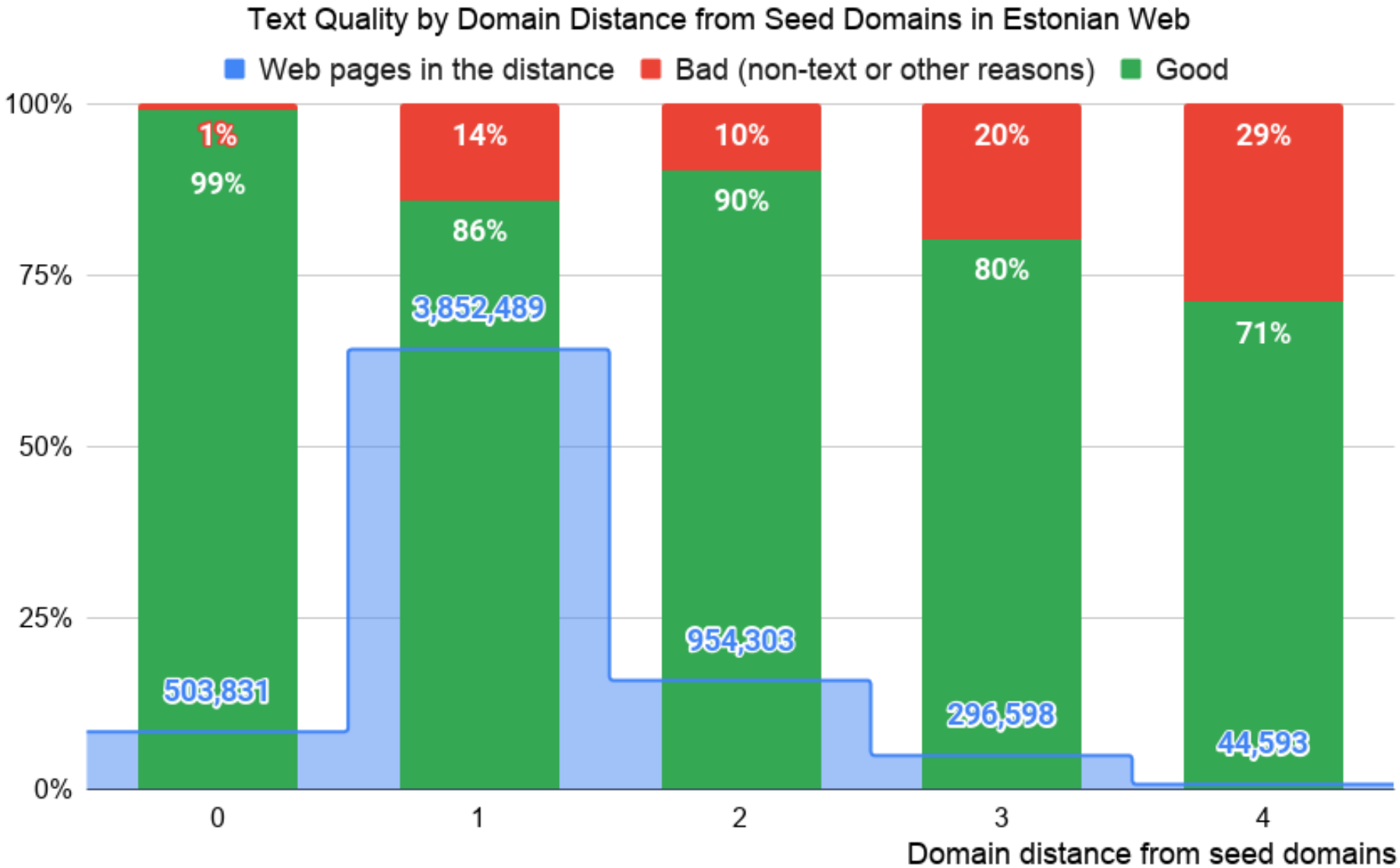
Approaches to Web Spam Removal

1. trustworthy websites only
2. website rules in the crawler: distance from the seeds, hostname
3. supervised classification
4. semi-manual filtering of websites

① Trustworthy Websites Only

- works well but not perfect
- limited amount/size of trustworthy sources ⇒ unsuitable for small languages

② Website Distance from the Seed (Trustworthy) Websites



③ Supervised Classification – Data & Method

- 146 spam pages of 1630 manually classified web pages
- various web sources, 2006 to 2015
  - phrase and sentence level incoherency
  - frequent spam topics: medication, financial services, essay writing
  - other non-text, various techniques
- FastText supervised classifier (Mikolov, 2016)
- applied to a large English web corpus from 2015
- 35 % most 'spam-like' documents removed
- recall: 70.5 %
- precision: 71.5 %

Supervised Classification – Evaluation – Wordlist

	Original corpus	Clean corpus	Kept
Document count	58,438,034	37,810,139	64.7 %
Token count	33,144,241,513	18,371,812,861	55.4 %

Phrase	Original hits/M	Clean hits/M	Kept
viagra	229.71	3.42	0.8 %
cialis 20 mg	2.74	0.02	0.4 %
aspirin	5.63	1.52	14.8 %
oral administration	0.26	0.23	48.8 %
loan	166.32	48.34	16.1 %
payday loan	24.19	1.09	2.5 %
cheap	295.31	64.30	12.1 %
interest rate	14.73	9.80	36.7 %
essay	348.89	33.95	5.4 %
essay writing	7.72	0.32	2.3 %
pass the exam	0.34	0.36	59.4 %
slot machine	3.50	0.99	15.8 %
playing cards	1.01	0.67	36.8 %
play games	3.55	3.68	53.9 %

Supervised Classification – Evaluation – Collocates/Lexicography: Objects of 'Buy'

Top collocate objects of verb 'buy' before and after spam removal

Original corpus				**Cleaned corpus**		
lemma	**frequency**	**score**		**lemma**	**frequency**	**score**
viagra	569,944	10.68		ticket	52,529	9.80
ciali	242,476	9.56		house	28,313	8.59
essay	212,077	9.17		product	37,126	8.49
paper	180,180	8.93		food	24,940	8.22
levitra	98,830	8.33		car	20,053	8.18
uk	93,491	8.22		book	27,088	8.09
ticket	85,994	8.08		property	17,210	7.88
product	105,263	8.00		land	15,857	7.83
cialis	71,359	7.85		share	12,083	7.67
car	75,496	7.75		home	22,599	7.63
house	70,204	7.61		item	12,647	7.40
propecia	55,883	7.53		good	9,480	7.37

Supervised Classification – Evaluation – Collocates/Lexicography: Modifiers of 'House'

Top collocate modifiers of noun 'house' before and after spam removal

Original corpus				**Cleaned corpus**		
lemma	**frequency**	**score**		**lemma**	**frequency**	**score**
white	280,842	10.58		publishing	20,314	8.63
opera	58,182	8.53		open	39,684	8.47
auction	41,438	8.05		guest	13,574	7.94
publishing	41,855	8.02		opera	9,847	7.67
geisha	38,331	7.95		old	32,855	7.64
open	37,627	7.78		haunted	9,013	7.58
old	73,454	7.52		auction	8,240	7.40
guest	28,655	7.44		manor	7,225	7.28
country	26,092	7.07		bedroom	7,717	7.26
stone	18,711	6.77		country	9,926	7.20
dream	17,953	6.77		coffee	8,171	7.18
coffee	18,336	6.74		wooden	6,803	6.96

④ Semi-manual Website Filtering

Data:

- 1,000 Estonian 2019 web sites, manually checked by Kristina Koppel (Tartu University)
- 16 % marked as computer generated non-text, mostly machine translated, 6 % marked as poor quality

Method:

- FastText supervised classifier
- probability threshold set to aim for a high recall

Evaluation:

- 100 positive & 100 negative random pages for manual evaluation
- recall: 97.1 %, precision: 66.7 %
- quite efficient method – just several man-days of manual work

Conclusion: Approaches to Web Spam Removal Combined

- ✓ Trustworthy websites only
- ✓ Website rules in the crawler: distance from the seeds, hostname
- ✓ Supervised classification
- ✓ Semi-manual filtering of websites