Website Properties in Relation to the Quality of Text Extracted for Web Corpora

Vít Suchomel[†‡], Jan Kraus[‡]

†Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czechia
xsuchom2@fi.muni.cz

‡Lexical Computing
Brno, Czechia
{vit.suchomel,jan.kraus}@sketchengine.eu

Abstract. In this paper we present our research concerning the relation between two properties of websites and the quality of the text extracted from a website in the context of crawling the web and building large web corpora. A manual classification of text quality of 18 thousand websites from 21 European languages was used to verify our assumption that certain web domain properties can be used to identify potential sources of bad quality content.
The first property is the distance of a web domain from the seed domains in a web crawl. The second property studied in this work is the length of the website name. Although these properties were recommended to help identify good quality websites in our previous work, in this paper we show there is only a small difference between the quality of text-rich web domains with various seed distances or name lengths. This conclusion holds for the post-crawling text processing when starting the web crawl with a large amount of seed domains.

Keywords: Web crawling · Web spam · Text corpus · Text processing

1 Introduction and Motivation

Large web corpora are used in many linguistic, lexicographic and NLP applications. Although the web is a large and easy-to-use source of texts, there is a lot of low quality content. We defined the good and bad content with regards to a linguistic use of text corpora in [1, p. 72]: *A fluent, naturally sounding, consistent text is good, regardless of the purpose of the web page or its links to other pages. The bad content is this: computer generated text, machine translated text, text altered by keyword stuffing or phrase stitching, text altered by replacing words with synonyms using a thesaurus, summaries automatically generated from databases (e.g. weather forecast, sport results – all of the same kind very similar), and finally any incoherent text. This is the kind of non-text this work is interested in.*

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 167–175, 2021. © Tribun EU 2021

To get a fluent, naturally sounding and consistent text in the corpus, one should avoid downloading websites providing low quality content and – since that is only partially possible [1, p. 64] – filter out poor quality text from the crawled data as a post-processing procedure. Since the nature of a significant part of non-text is to look like a human-produced text, a human intervention is needed.

We proposed a semi-manual approach consisting in manually checking the largest sources of data and training a non-text classifier, using this data, for the rest of the corpus in [1, p. 85]: *Our assumption in this setup is that all pages in a web domain are either good – consisting of nice human produced text – or bad – i.e. machine generated non-text or other poor quality content. Although this supposition might not hold for all cases and can lead to noisy training data for the classifier, it has two advantages: Much more training samples are obtained and the cost to determine if a web domain tends to provide good text or non-text is not high.*

This paper presents the process of the manual check of text quality of large websites in the corpus in chapter 2.

Furthermore, we were interested in the usefulness of web domain properties for assessing the quality of the text yielded by the site. Some properties are evaluated on-the-fly by web crawler SpiderLing [2] that is used by us to crawl the web. Selected web domain characteristics are described in chapter 3. The relation of these metrics to the website quality is dealt with in chapter 4. This research broadens the evaluation reported in [1, p. 90] to 18 thousand websites from 21 European languages.

2 Checking Website Text Quality in Large Web Corpora

Here follows the procedure of checking website text quality in TenTen web corpora [3] we build for text corpus management system Sketch Engine [4].

The number of websites to be checked is proportionate to the size of the domains in tokens. If a domain contains more than 10 million tokens, a higher priority will be given to such domain. On the other hand, if a domain contains less than 2 million tokens, there will be a lower priority during the checking process and this often creates the threshold, i.e. smaller websites will not be manually checked, since their impact on the corpus quality is marginal.

On average it is possible to manually check about 50 to 70 domains per hour, depending on the familiarity with the language, language script, etc. The size of the language also plays a role. Languages like English, Spanish, German etc. are much more extensive in content (tens of billions of tokens) and that is why a larger number of domains will be manually checked, usually 2,000 to 5,000. For smaller corpora (billions of tokens), the number of websites to check will be usually about 300 to 500.

The first step in web domain checking is to pick the largest domains that make up the majority of the corpus, usually that is at least 50 % of the corpus, depending on the total size of the corpus and language. The second step is to check random concordances of three consecutive sentences from the selected web domains. This concordance usually consists of 50–70 lines.

The third step involves a manual checking of the random concordances. One of the most important things when determining whether to keep a specific domain in our corpora is the genuineness of the texts. After the web domains are downloaded, there might be a certain percentage of spam and other texts of lesser quality impacting the corpus quality and such texts must be removed from the corpus. During this phase of checking, each domain is either labelled as *ok* or *bad*. The domains labeled as bad contain either spam (generated text without any meaning) or machine translated texts, which might be difficult to spot in languages we do not know in depth. In such cases the website source code, domain name or the live website will usually give clues.

Apart from this, there might be other criteria for keeping web domains in corpora. If a certain domain contains a large amount of lists, square brackets, angle tag brackets or other non-text elements, these domains will be tagged as bad and thus removed from the corpus. Sometimes this decision will depend on the language and corpus size. Especially if the corpus is rather small, for instance no more than one billion words, such texts might be preserved for the sake of having some linguistic data and meeting the first condition of the text being a spam or not will suffice.

After this phase of checking is completed, there might be other ways to identify the bad content. Since some of the bad domains were already identified in the previous step, we can use some of the words present in bad domains to run a concordance search to find other bad domains. This step usually works for spam. If spam contains words like „porn“, „xxx“, „viagra“ etc., other bad domains might be identified this way.

3 Selected Web Domain Properties

The data is obtained from the internet through crawling – starting from seed URLs (or domains), downloading web pages (or other documents) and following links found in these pages. We selected two web domain properties evaluated on-the-fly by web crawler SpiderLing [2]: The distance of a web domain from the seed domains and the length of the website name. In addition to text yield ratio, these characteristics are used by the crawler to determine which sources to focus on.

Assuming the web is an oriented graph with web pages being the nodes and links being the vertices, the lowest graph distance from the seed (initial) web pages to a web page in a website is the *domain distance* of the web domain. The domain distance is measured by the crawler. The distance of a domain is heavily dependent on the seed domains and it can vary for different runs or settings of the crawler. The crawler is set to download more often from websites with a short distance.[1]

The *hostname length* is the length of the name of the website, i.e. the hostname character count. The crawler is set to ignore sites with hostname length greater than 40 and to download more often from websites with a short name.[1]

4 The Quality of Text in Relation to Website Properties

The quality of text in web domains human-labelled by *ok* or *bad* is shown in relation to hostname length and domain distance in the following tables and charts. In our corpus building projects, the crawling is usually started from all URLs known to us in the target language, including the previous versions of the corpus. Thus not only trustworthy domains (such as news sites, government webs and site whitelists [5]) are in distance 0. That means we care less for avoiding bad sites and identify them in the post-processing phase to discover as many links to good parts of the web (hopefully) as possible.

Note this is an evaluation of the largest text sources in a particular language (i.e. from a website containing documents in the language) that were downloaded by the crawler already giving priority to domains with a short distance or a short hostname.

The text quality by domain distance for 18 thousand websites from 21 European languages is shown in Fig. 1. The same data is evaluated with regards to hostname length in Fig. 2. A zero distance or a very short name is somewhat indicating a good content. Based on this findings, we do not recommend using the domain distance in decisions about text quality in post-processing when the crawler started with all URLs available rather than a trustworthy seeds. That is also the main difference from conclusions based on the chart in [1, p. 90].

A detailed breakdown of the counts of good and bad domains grouped by the domain distance or the hostname length can be found in Table 1.

The detailed figures for selected separate languages are presented in Table 2 for Czech, in Table 3 for Slovene, in Table 4 for Polish, in Table 5 for German and in Table 6 for Latvian.

[1] This measure has an impact just for crawls with a large number of domains in the download queue, mainly the English web, since all domains are scheduled for download anyway in case there is less domains to choose from.

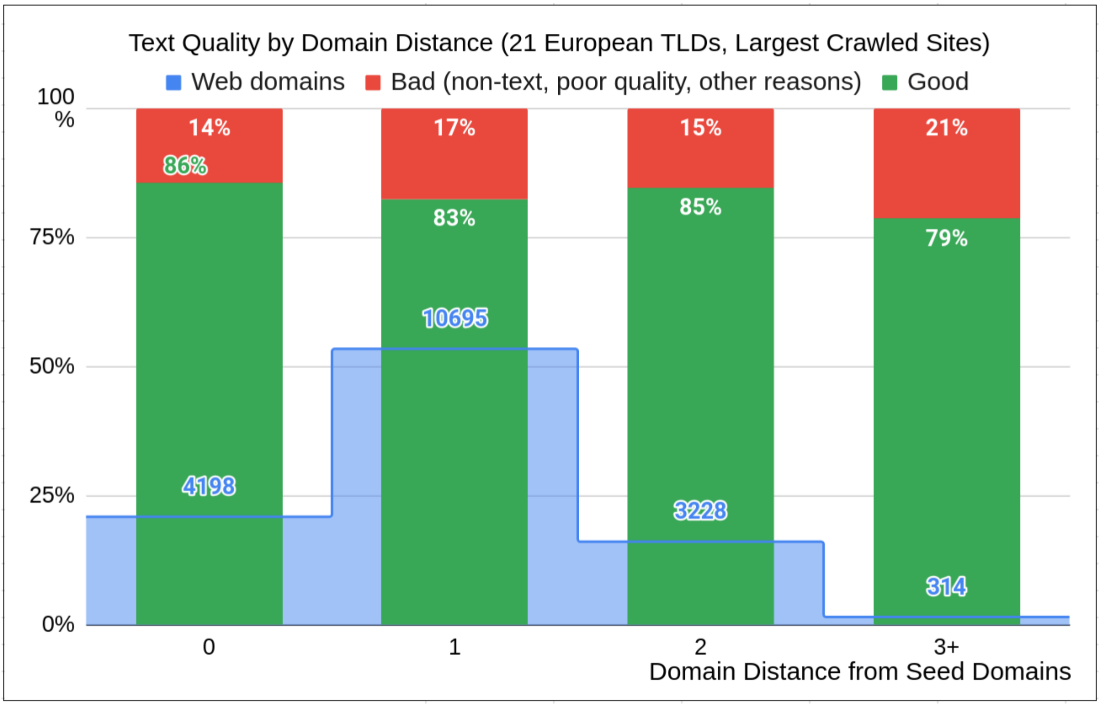

Fig. 1: Text quality by domain distance, all data from this report together. The proportion of good and bad domains is shown in green and red, respectively. The number of web domains in each band is displayed by the blue stepped chart.

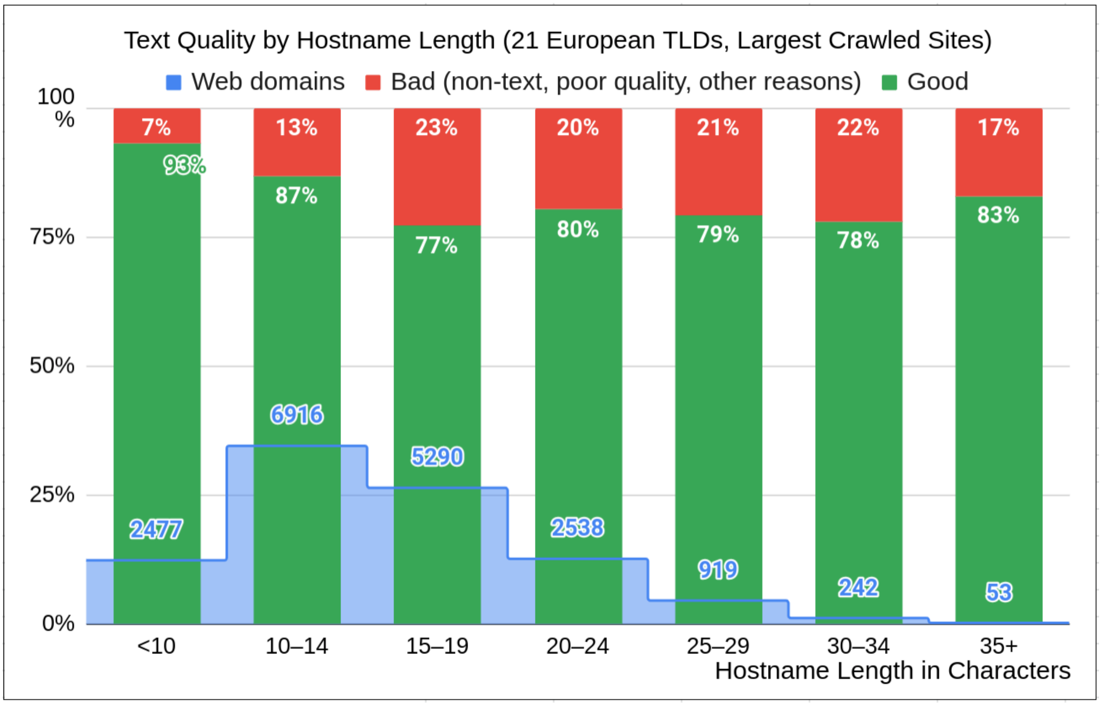

Fig. 2: Text quality by hostname length, all data from this report together. The proportion of good and bad domains is shown in green and red, respectively. The number of web domains in each band is displayed by the blue stepped chart.

Table 1: Domain count analysis for all data in Fig. 1 and Fig. 2.

21 European languages	**domains**	**ok**	**bad**
domains	18529	83%	16%
median distance		1	1
median name length		14	16

distance	**domains**	**ok**	**bad**
0	4239	85%	14%
1	10738	82%	17%
2	3238	84%	15%
3+	314	79%	21%

name length	**domains**	**ok**	**bad**
<10	2482	93%	7%
10–14	6953	86%	13%
15–19	5323	77%	23%
20–24	2552	80%	20%
25–29	924	79%	21%
30–34	242	78%	22%
35+	53	83%	17%

Table 2: Domain count analysis for a 2019 crawl of Czech. The domain distance is unrelated to data quality. The hostname length is somewhat related to data quality.

Czech Web 2019	**domains**	**ok**	**bad**
domains	878	91%	9%
median distance		2	2
median name length		12	14

distance	**domains**	**ok**	**bad**
0	244	94%	6%
1	154	84%	16%
2	426	92%	8%
3+	54	91%	9%

name length	**domains**	**ok**	**bad**
<10	181	96%	4%
10–14	396	90%	10%
15–19	238	90%	10%
20–24	56	89%	11%
25–29	5	60%	40%
30–34	2	100%	0%

Table 3: Domain count analysis for a 2020 crawl of Slovene. The measures are almost unrelated to data quality here.

Slovene Web 2020	**domains**	**ok**	**bad**
domains	250	91%	9%
median distance		1	1
median name length		13	14

distance	**domains**	**ok**	**bad**
0	65	95%	5%
1	155	93%	7%
2	29	72%	28%
3+	1	100%	0%

name length	**domains**	**ok**	**bad**
<10	40	95%	5%
10–14	107	91%	9%
15–19	77	90%	10%
20–24	23	91%	9%
25–29	2	100%	0%
30–34	0		
35+	1	100%	0%

Table 4: Domain count analysis for a 2019 crawl of Polish. The measures are unrelated to data quality here.

Polish Web 2019	**domains**	**ok**	**bad**
domains	762	91%	9%
median distance		1	0
median name length		14	13

distance	**domains**	**ok**	**bad**
0	299	87%	13%
1	431	94%	6%
2	31	94%	6%
3+	1	100%	0%

name length	**domains**	**ok**	**bad**
<10	124	90%	10%
10–14	318	92%	8%
15–19	223	91%	9%
20–24	79	94%	6%
25–29	17	88%	12%
30–34	1	0%	100%

Table 5: Domain count analysis for a 2020 crawl of German. The measures are unrelated to data quality here.

German Web 2020	**domains**	**ok**	**bad**
domains	2398	94%	4%
median distance		1	1
median name length		14	15

distance	**domains**	**ok**	**bad**
0	592	89%	7%
1	1614	96%	3%
2	189	97%	3%
3+	3	100%	0%

name length	**domains**	**ok**	**bad**
<10	326	97%	2%
10–14	893	94%	5%
15–19	753	92%	6%
20–24	299	97%	2%
25–29	100	95%	2%
30–34	26	92%	8%
35+	1	100%	0%

Table 6: Domain count analysis for a 2019 crawl of Latvian. The domain distance is rather negatively related to data quality, it seems like the crawler found a better content then was yielded by the initial sites. The hostname length is related to data quality well.

Latvian Web 2021	**domains**	**ok**	**bad**
domains	453	46%	54%
median distance		1	0
median name length		12	18

distance	**domains**	**ok**	**bad**
0	198	34%	66%
1	235	56%	44%
2	17	53%	47%
3+	3	0%	100%

name length	**domains**	**ok**	**bad**
<10	57	91%	9%
10–14	125	85%	15%
15–19	254	17%	83%
20–24	14	36%	64%
25–29	3	33%	67%

5 Conclusions

In this paper we have described the website checking part of the process of extraction and cleaning text from the Internet for building large web corpora in Sketch Engine. The relations of web domain seed distance and hostname length to the quality of the website content were studied using 18 thousand websites from 21 European languages.

We found there is none or a small difference between the content quality of text-rich web domains and the domain distance. The host name length is somewhat related to the domain text quality. Both relations depend on the particular crawl setup.

Although the studied website properties may be helpful for the crawler's scheduler to decide which small domains to visit more frequently, they are not related much to the text quality of the largest websites when starting the web crawl with a large amount of seed domains.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

1. Suchomel, V.: Better Web Corpora For Corpus Linguistics And NLP. PhD thesis, Masaryk University (2020)
2. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Proceedings of the seventh Web as Corpus Workshop (WAC7). (2012) 39–43
3. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster (2013)
4. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1** (2014)
5. Baisa, V., Suchomel, V.: Skell: Web interface for english language learning. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, Brno (2014) 63–70