Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 56-60

Removing Spam from Web Corpora Through Supervised Learning Using FastText

Vít Suchomel
Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czech Republic
xsuchom2@fi.muni.cz

Abstract

Unlike traditional text corpora collected from trustworthy sources, the content of web based corpora has to be filtered. This study briefly discusses the impact of web spam on corpus usability and emphasizes the importance of removing computer generated text from web corpora.

The paper also presents a keyword comparison of an unfiltered corpus with the same collection of texts cleaned by a supervised classifier trained using FastText. The classifier was able to recognise 71 % of web spam documents similar to the training set but lacked both precision and recall when applied to short texts from another data set.

1 Web Spam in Text Corpora

It has been shown that boilerplate, duplicates, and spam skew corpus based analyses and therefore have to be removed, see nonsense examples of word use in an application for English learners based on a web corpus in figure 1. While the first two issues have been successfully addressed, e.g. by (Marek et al., 2007; Pomikálek, 2011; Versley and Panchenko, 2012; Schäfer and Bildhauer, 2013), spam might be still observed in web corpora as reported by (Kilgarriff and Suchomel, 2013). It was spam that represented the main difference between their 2008 and 2012 corpora crawled from the web. That is why a spam cleaning stage should be a part of the process of building web corpora.

The traditional definition of web spam is *actions intended to mislead search engines into ranking some pages higher than they deserve* (Gyöngyi and Garcia-Molina, 2005). The Google document 'Fighting Spam'[1] describes the kinds of spam that Google finds, and what they do about it.

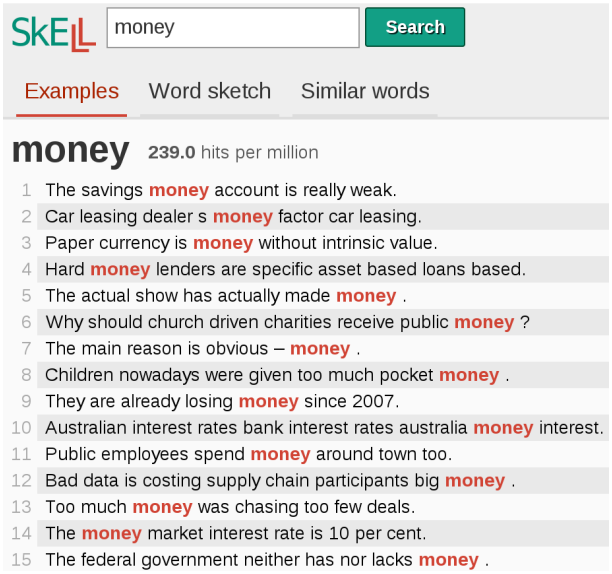

Figure 1: Web spam in examples of use of word 'money' at skell.sketchengine.co.uk – see lines 2, 4 and 10.

Text alteration techniques consist in changing the frequency properties of a web page content in favour of spam targeted words or phrases: *repetition of terms related to the spam campaign target, inserting a large number of unrelated terms, often even entire dictionaries, weaving of spam terms into contents copied from informative sites, e.g. news articles, glueing together sentences or phrases from different sources* as reported by (Gyöngyi and Garcia-Molina, 2005).

Automatically generated content does not provide examples of authentic use of a natural language. Nonsense, incoherent or any unnatural texts such as the following short instance have to be removed from a good quality web corpus: *Edmonton Oilers rallied towards get over the Montreal Canadiens 4-3 upon Thursday.Ryan Nugent-Hopkins completed with 2 aims, together with*

[1] https://www.google.com/insidesearch/howsearchworks/fighting-spam.html

the match-tying rating with 25 seconds remaining within just legislation.[2]

The following types of automatically generated content are examples of documents penalised by Google:[3] *Text translated by an automated tool without human review or curation before publishing. Text generated through automated processes, such as Markov chains. Text generated using automated synonymizing or obfuscation techniques.* These kinds of spam should certainly be eliminated from web corpora while the other two examples given by Google may not present a harm to the corpus use: *Text generated from scraping Atom/RSS feeds or search results. Stitching or combining content from different web pages without adding sufficient value.*

In contrast to the traditional or search engine definitions of web spam, the corpus use point of view is not concerned with intentions of spam producers or the justification of the search engine optimisation of a web page. A text corpus built for NLP or linguistics purpose should contain coherent and consistent, meaningful, natural and authentic sentences in the target language. Only texts created by spamming techniques breaking those properties should be detected and avoided. The unwanted non-text 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. stock market reports, weather forecast, sport results – all of the same kind very similar), and finally any incoherent text. Varieties of spam removable by existing tools, e.g. duplicate content, link farms (quite a lot of links with scarce text), are only a minor problem.

Avoiding web spam by selecting trustworthy corpus sources such as Wikipedia, news sites, government and academic webs works well: (Baisa and Suchomel, 2014) show it is possible to construct medium sized corpora from URL whitelists and web catalogues. (Spoustová and Spousta, 2012) used a similar way of building a Czech web corpus. Also the BootCaT method (Baroni and Bernardini, 2004) indirectly avoids spam by relying on a search engine to find non-spam data. Despite the avoiding methods being successful, it is doubtful a huge web collection can be obtained just from trustworthy sources.

Furthermore, language independent methods of combating spam might be of use. (Ntoulas et al., 2006) reported web spamming was not only a matter of the English part of internet. Spam was found in their French, German, Japanese and Chinese documents as well.

2 Removing Spam Using a Supervised Classifier

This section describes training and evaluation of a supervised classifier to detect spam in web corpora.

We have manually annotated a collection of 1630 web pages from various web sources from years 2006 to 2015.[4] To cover the main topics of spam texts observed in our previously built corpora, we included 107 spam pages promoting medication, financial services, commercial essay writing and other subjects. Both phrase level and sentence level incoherent texts (mostly keyword insertions, n-grams of words stitched together or seemingly authentic sentences not conveying any connecting message) were represented. Another 39 spam documents coming from random web documents identified by annotators were included. There were 146 positive instances of spam documents altogether.

The classifier was trained using FastText (Joulin et al., 2016) and applied to a large English web corpus from 2015. The expected performance of the classifier was evaluated using a 30-fold cross-validation on the web page collection. Since our aim was to remove as much spam from the corpus as possible, regardless false positives, the classifier confidence threshold was set to prioritize recall over precision. The achieved precision and recall were 71.5 % and 70.5 % respectively. Applying this classifier to an English web corpus from 2015 resulted in removing 35 % of corpus documents still leaving enough data for the corpus use.

An inspection of the cleaned corpus revealed the relative count of usual spam related keywords dropped significantly as expected while general words not necessarily associated with spam were affected less as can be seen in table 1.

Another evaluation of the classifier was performed by manually checking 299 random web documents from the cleaned corpus and 25 ran-

[2] `http://masterclasspolska.pl/forum/`

[3] Google quality guidelines – `https://support.google.com/webmasters/answer/2721306`

[4] The collection is a part of another classification experiment by the same authors not covered by this paper.

dom spam documents removed by the classifier. The achieved precision was 40.0 % with the recall of 27.8 %. The error analysis showed the classifier was not able to recognise non-text rather than spam. 17 of 26 unrecognised documents were scientific paper references or lists of names, dates and places, i.e. *Submitted by Diana on 2013-09-25 and updated by Diana on Wed, 2013-09-25 08:32* or *January 13, 2014 January 16, 2014 Gaithersburg, Maryland, USA*. Such web pages were not present in the training data since we believed it had been removed from the corpus sources by a boilerplate removal tool and paid attention to longer documents. Not counting these 17 non-text false negatives, the recall would reach 52.6 %.

To find out what was removed from the corpus, relative counts of lemmas[5] in the corpus were compared with the BNC[6] in figures 2 and 3. A list of lemmas in the web corpus with the most reduced relative lemma count caused by removing unwanted documents is presented in 4.

The inspection showed there were a lot of spam related words in the original web corpus and that spam words are no longer characteristic of the cleaned version of the corpus in comparison to the BNC.[7]

3 Conclusion

We view computer generated text as the main kind of spam decreasing the quality of web corpora. A classifier trained on spam documents was applied to remove unwanted content from a web corpus. Although the classifier significantly decreased the presence of spam related words in the corpus, it was not able to recognise short non-text documents. That remains to be addressed in the future.

Acknowledgments

This work was partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

[5]Corpora in the study were lemmatised by TreeTagger.

[6]The tokenisation of the BNC had to be changed to the same way the web corpus was tokenised in order to make the counts of tokens in both corpora comparable.

[7]The comparison with the BNC also revealed there are words related to the modern technology (e.g. *website*, *online*, *email*) and American English spelled words (*center*, *organization*) in the 2015 web corpus.

References

[Baisa and Suchomel2014] Vít Baisa and Vít Suchomel. 2014. Skell: Web interface for english language learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 63–70, Brno. Tribun EU.

[Baroni and Bernardini2004] Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *LREC*.

[Gyöngyi and Garcia-Molina2005] Zoltan Gyöngyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*.

[Joulin et al.2016] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[Kilgarriff and Suchomel2013] Adam Kilgarriff and Vít Suchomel. 2013. Web spam. In Paul Rayson Stefan Evert, Egon Stemle, editor, *Proceedings of the 8th Web as Corpus Workshop (WAC-8) @Corpus Linguistics 2013*, pages 46–52.

[Marek et al.2007] Michal Marek, Pavel Pecina, and Miroslav Spousta. 2007. Web page cleaning with conditional random fields. In *Building and Exploring Web Corpora: Proceedings of the Fifth Web as Corpus Workshop, Incorporationg CleanEval (WAC3), Belgium*, pages 155–162.

[Ntoulas et al.2006] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM.

[Pomikálek2011] Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk University.

[Schäfer and Bildhauer2013] Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*, volume 6. Morgan & Claypool Publishers.

[Spoustová and Spousta2012] Johanka Spoustová and Miroslav Spousta. 2012. A high-quality web corpus of Czech. In *LREC*, pages 311–315.

[Versley and Panchenko2012] Yannick Versley and Yana Panchenko. 2012. Not just bigger: Towards better-quality web corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 44–52.

Table 1: Comparison of the corpus before and after spam removal using the classifier. Corpus sizes and relative frequencies (number of occurrences per million words) of selected words are shown. Reducing the corpus to 55 % of the former token count, phrases strongly indicating spam documents such as “cialis 20 mg”, “payday loan” or “essay writing” were almost removed while innocent phrases from the same domains such as “oral administration”, “interest rate” or “pass the exam” were reduced proportionally to the whole corpus.

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

Figure 2: Relative wordcount comparison of the original 2015 web corpus with British National Corpus, top 26 lemmas sorted by the keyword score. Score = $\frac{fpm_1+100}{fpm_2+100}$ where fpm_1 is the count of lemmas per million in the focus corpus (3rd column) and fpm_2 is the count of lemmas per million in the reference corpus (5th column).

	Original English Web 2015		*British National Corpus*		
Lowercase lemma	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
download	32,877,718	992.0	35	0.3	10.9
pdf	30,658,156	925.0	37	0.3	10.2
online	23,683,595	714.6	596	5.3	7.7
program	20,333,705	613.5	5,814	51.8	4.7
website	9,586,380	289.2	0	0.0	3.9
center	9,903,586	298.8	573	5.1	3.8
essay	11,563,807	348.9	2,317	20.6	3.7
viagra	7,620,095	229.9	0	0.0	3.3
url	7,168,836	216.3	0	0.0	3.2
ebook	6,969,380	210.3	0	0.0	3.1
web	7,206,520	217.4	729	6.5	3.0
internet	6,248,400	188.5	97	0.9	2.9
student	24,584,996	741.8	22,133	197.1	2.8
cialis	5,816,475	175.5	0	0.0	2.8
blog	5,110,812	154.2	0	0.0	2.5
email	5,074,946	153.1	43	0.4	2.5
cheap	9,787,744	295.3	6,649	59.2	2.5
epub	4,761,306	143.7	0	0.0	2.4
video	10,278,042	310.1	7,672	68.3	2.4
free	20,406,767	615.7	21,963	195.6	2.4
u.s.	4,976,297	150.1	458	4.1	2.4
post	13,400,787	404.3	12,576	112.0	2.4
outlet	5,501,465	166.0	1,375	12.2	2.4
color	4,553,463	137.4	143	1.3	2.3
click	5,326,832	160.7	1,273	11.3	2.3
your	95,303,049	2875.4	134,413	1197.0	2.3

Figure 3: Relative wordcount comparison of the cleaned web corpus with British National Corpus

	Cleaned English Web 2015		*British National Corpus*		
Lowercase lemma	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
program	14,384,115	782.9	5,814	51.8	5.8
center	7,509,618	408.8	573	5.1	4.8
website	4,792,518	260.9	0	0.0	3.6
student	16,973,541	923.9	22,133	197.1	3.4
online	4,753,580	258.7	596	5.3	3.4
u.s.	4,225,425	230.0	458	4.1	3.2
project	14,949,773	813.7	21,742	193.6	3.1
university	12,182,707	663.1	18,899	168.3	2.8
community	15,164,485	825.4	26,564	236.6	2.7
global	4,585,347	249.6	3,529	31.4	2.7
web	3,322,320	180.8	729	6.5	2.6
download	3,011,631	163.9	35	0.3	2.6
email	2,901,189	157.9	43	0.4	2.6
dr.	3,290,385	179.1	1,215	10.8	2.5
internet	2,753,028	149.9	97	0.9	2.5
our	39,914,081	2172.6	93,457	832.3	2.4
click	3,144,338	171.2	1,273	11.3	2.4
focus	6,345,601	345.4	9,538	84.9	2.4
technology	7,397,599	402.7	12,865	114.6	2.3
organization	5,944,514	323.6	9,240	82.3	2.3
research	12,854,262	699.7	27,567	245.5	2.3
update	3,452,461	187.9	2,814	25.1	2.3
datum	7,682,640	418.2	14,212	126.6	2.3
network	5,810,016	316.2	9,291	82.7	2.3
video	5,202,487	283.2	7,672	68.3	2.3
photo	3,054,229	166.2	2,036	18.1	2.3

Figure 4: Relative wordcount comparison of the original web corpus with the cleaned version

	Original English Web 2015		*Cleaned English Web 2015*		
Lowercase lemma	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
pdf	30,658,156	925.0	1,851,347	100.8	5.1
download	32,877,718	992.0	3,011,631	163.9	4.1
essay	11,563,807	348.9	623,760	34.0	3.4
viagra	7,620,095	229.9	62,899	3.4	3.2
ebook	6,969,380	210.3	265,781	14.5	2.7
cialis	5,816,475	175.5	45,613	2.5	2.7
url	7,168,836	216.3	509,596	27.7	2.5
buy	17,364,124	523.9	2,867,958	156.1	2.4
cheap	9,787,744	295.3	1,180,506	64.3	2.4
online	23,683,595	714.6	4,753,580	258.7	2.3
epub	4,761,306	143.7	203,405	11.1	2.2
prescription	4,646,919	140.2	280,013	15.2	2.1
outlet	5,501,465	166.0	651,024	35.4	2.0
book	29,921,305	902.8	7,889,796	429.5	1.9
generic	3,594,096	108.4	257,090	14.0	1.8
ugg	3,022,464	91.2	93,591	5.1	1.8
loan	5,512,504	166.3	888,181	48.3	1.8
jersey	4,873,552	147.0	836,729	45.5	1.7
insurance	7,150,681	215.7	1,588,816	86.5	1.7
pharmacy	2,941,876	88.8	290,211	15.8	1.6
sex	6,452,251	194.7	1,502,817	81.8	1.6
de	10,572,331	319.0	2,986,557	162.6	1.6
mg	2,776,320	83.8	298,031	16.2	1.6
you	195,234,032	5890.4	68,409,350	3723.6	1.6
binary	4,226,875	127.5	839,475	45.7	1.6
levitra	1,873,011	56.5	34,646	1.9	1.5