Extractive Text Summarization in Dark Web: A Preliminary Study

Akanksha Joshi[1,2,3],E. Fidalgo[1,2], E. Alegre[1,2,] M. AL. NABKI
[1]Universidad de León, Spain
[2]INCIBE (Spanish National Institute of Cybersecurity), León, Spain
[3]CDAC, Mumbai, India
{ajosh,eduardo.fidalgo, enrique.alegre}@unileon.es

Abstract. The Tor (The Onion Router) network is a type of Dark Net which is part of Internet and allows its users to connect through virtual tunnels to make their communications private. The anonymity in Tor network gives rise to illegal activities hidden from the law and enforcement agencies. In this paper, we address the problem of automatic text summarization of the documents extracted from onion websites especially, the illegal text. The objective of this work is three-fold; first, to verify whether automatic summarization can be applied to onion web content effectively or not. Second, to identify the most appropriate approaches for the task while analyzing their pros and cons. Third, we introduce a new dataset named by us as "OWIDSumm", which contains manually created two sets of ground truth summaries for 60 onion web documents related to illegal hidden services. Next, we conducted an empirical analysis of six well-known extractive text summarization algorithms on this dataset and on DUC2002 summarization benchmark dataset. The study demonstrates that the accuracies of automatic text summarization approaches on onion web documents are similar to obtained on DUC2002. Also, TextRank algorithm is found to perform in more robust and effective way compared to other methods, based on ROGUE-1, ROGUE-2, ROGUE-L evaluation measures.

Keywords. Text summarization, Extractive text summarization, Dark web,

1. Introduction

The deep web is part of Internet or World Wide Web (WWW) which is not indexed and accessed through the search engines such as Google and Bing. As part of deep web, the dark web or darknet is an encrypted network which utilizes volunteer relays located around the world to provide anonymity to its users. Tor is one such darknet which is quite famous for illegal activities1 including black marketing, selling counterfeit credit cards, hacking etc. and can be accessed through special tools only. The hidden services hosted in Tor network are known as onion sites, so named because they end with ".onion". In this work, we refer onion web document a text which is extracted from hidden services hosted in Tor network. According to Bergman et al. [12], the deep web is much bigger than the surface web as it is assumed that there are around 1 billion documents in surface web and around 550 billion documents on the dark web.

Text summarization is an active field of Natural Language Processing that has been researched for more than half a century and has been applied to scientific articles, web

[1]http://www.telegraph.co.uk/technology/2016/02/02/dark-web-browser-tor-is-overwhelmingly-used-for-crime-says-study/

documents, news articles and emails. to provide quick information to the users in compressed and readable form. Text Summarization can be Extractive or Abstractive. Extractive Summarization aims at selecting the relevant and informative sentences from the document by ranking those sentences. Abstractive text summarization tries to capture the key concept in the document and produces summary using a vocabulary different from the input document. In this paper, we focus on the task of single-document extractive text summarization on the text extracted from the tor network.

The earliest techniques that have been used for summarization are based on features such as word and phrase frequency [14], position in the text [15] and key phrases [16]. Various supervised machine learning approaches such as SVM [2], Naive Bayes Classification [18] and unsupervised machine learning techniques like clustering [27], hidden Markov [1] have also been applied to text summarization. Several graph based techniques have also been proposed for text summarization [4, 5, 6, 7].

Our main contributions in this paper are: (i) creation of a ground truth dataset OWIDSumm consisting of six categories with two versions of gold summaries for each document, (ii) comparative evaluation of six unsupervised text summarization algorithms including TextRank [5], LexRank [4], Luhn [14], LSA [10], KLSum [9] and Sum-basic [8] on this dataset, (iii) comparative evaluation of these six algorithms on DUC 2002 benchmark[2].

2. Evaluation Dataset

We created a dataset OWIDSumm (Onion Web Illegal Documents) by selecting the documents (referred as domains) from Darknet Usage Text Addresses dataset (DUTA [13]). DUTA contains total 7000 documents arranged under 26 categories related to illegal as well as legal activities in Tor Darknet. For this study, we selected around 10 documents per category from total 6 categories; Cryptocurrency, counterfeit credit cards, counterfeit money, hacking, market place and drugs. We selected the categories and documents randomly first and then discarded some of them based on their content and information. More precisely, during document selection, we gave preference to illegal documents and to those documents which have two to three times more text (after pre-processing) than the desired length of the summary. We discarded those categories or documents which mainly have URLs in their documents, has redundant content, or their text is highly unstructured to create a summary like forums category. At this end, for all the documents selected, the two ground truth summaries of around 100 words are created manually by two different operators. We also used DUC 2002 dataset[2] for comparative evaluation of algorithms which consists of 567 news documents from 60 different categories.

3. Extractive Summarization Algorithms for Evaluation

We selected six well-known techniques from the literature for evaluation which are highly used and whose source-code is also available. Among the selected algorithms; Luhn [14] is the very first technique for text summarization based on the frequency of words in the document. TextRank [5] and LexRank [4] are graph based approaches that represent the document as a graph of sentences or keywords and try to find significant sentences from it. LSA [10] is topic focused summarization method which represents

[2] http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

Table 1. Accuracy of Text Summarization Algorithms on OWIDSumm

Cat.	**Luhn**			**TextRank**			**LexRank**			**LSA**			**KLsum**			**SumBasic**		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**
C1	34.3	17.5	22.3	**46.7**	**29.9**	**26.0**	36.5	17.6	20.0	38.0	0.20	23.5	39.1	23.9	24.6	40.5	22.6	21.6
C2	44.9	33.0	**33.1**	**46.7**	**34.1**	32.5	42.1	29.8	32.2	42.3	31.1	28.8	39.5	29.8	31.8	36.5	22.3	27.0
C3	36.3	20.3	19.6	**40.9**	24.0	23.6	33.0	17.2	21.2	38.0	20.2	22.6	38.2	**25.2**	26.5	**29.7**	10.7	16.0
C4	27.4	14.4	16.6	**33.7**	**21.9**	**20.0**	28.9	9.2	15.2	30.7	16.5	18.0	26.8	15.5	12.9	33.5	19.3	21.4
C5	47.6	32.3	34.1	**54.5**	**39.6**	**43.7**	52.6	36.9	40.5	54.4	38.1	38.7	42.4	28.4	34.4	51.7	33.9	38.1
C6	34.5	21.4	23.1	37.8	24.6	28.6	45.0	29.0	33.2	**46.9**	**33.4**	**37.4**	33.9	23.6	27.5	31.9	17.0	19.8
Avg.	37.5	23.2	24.8	**43.4**	**29.0**	**29.1**	39.68	23.28	27.1	41.7	23.3	28.2	36.7	24.4	26.3	37.3	21.0	24.0

C1-cryptocurrency C2-counterfeit money C3-counterfeit credit card C4-drugs C5-hacking C6-market place

Table 2. Accuracy of Text Summarization algorithms on DUC 2002 dataset

Data-base	**Luhn**			**TextRank**			**LexRank**			**LSA**			**KLsum**			**SumBasic**		
	R1	**R2**	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
DUC 2002	42.5	21.2	28.0	**43.0**	**21.5**	**28.3**	42.9	21.1	27.9	**43.0**	21.3	27.6	38.3	16.9	24.8	39.6	17.3	24.7

each document as an m × n matrix with m sentences containing n number of terms and applies singular value decomposition to this matrix to generate significant words or sentences. SumBasic [8] and KLsum [9] are generative probability based techniques that work on the probability distribution of the words in the documents.

4. Experimental Analysis and Results

4.1 Experimental Setup

We evaluate the six algorithms (section 3) on the dataset OWIDSumm (section 2). For each algorithm, the summary length is kept around 100 words. We kept the last sentence in summary without truncating it even if it has few more words than 100. For evaluation; we used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure proposed by Lin [11]. More precisely, we used, ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) measure computed using matches of unigrams, bigrams and longest common subsequence with gold summaries.

4.2 Results on OWIDSumm

The results are shown in Table 1, and the maximum ROUGE score between the two gold summaries is reported. The accuracies in Table 1 illustrates that for all the categories of OWIDSumm except ‘Market Place’, the graph based approach, Text Rank consistently outperforms all other summarization algorithms. The ROUGE scores of LSA shows that it gives comparable accuracy for the hacking category but it does not show good results when applied to other categories such as cryptocurrency, counterfeit-money.

4.3 Results on DUC 2002 Dataset

We report the ROUGE measures for DUC 2002[2] news dataset in Table 2. We can observe from the table that TextRank algorithm is performing better than all algorithms for all the ROUGE scores. However, the ROUGE scores of LSA, LexRank, and Luhn are in close proximity to the TextRank measures, they can also be considered as good approaches for summarizing the news datasets.

5. Conclusions and Future Work

In this paper, we have presented a study for text summarization applied to illegal Tor domain content. We manually created the dataset OWIDSumm with two sets of gold summaries for sixty documents under six categories and evaluated six well-known text summarization approaches. The evaluations are performed using well-known measures; ROUGE-1, ROUGE-2, ROUGE-L. The major findings of the study can be summarized as follows: i) the Text Rank consistently outperformed all the other summarization algorithms for tor network documents as well as on DUC 2002. ii) the best values of ROUGE-1, ROUGE-2, ROUGE-L on tor network domains are almost similar to the best values on DUC 2002 dataset, which indicates that despite being different domain and comparatively more unstructured data, the automatic summarization algorithm can perform well on the documents. In future, we would extend this study by including more categories and would try to understand why algorithms have different behaviour for documents of different categories.

References

[1] Conroy, J. M. and O'leary, D. P. (2001). Text summarization via hidden markov models. *In Proceedings of SIGIR '01,* New York, NY, USA.

[2] Fattah MA (2014) A hybrid machine learning model for multi-document summarization. 592–600. *doi:10.1007/s10489-013-0490-0.*

[3] Svore, K., Vanderwende, L., and Burges, C. (2007). Enhancing single-document summarization by combining RankNet and third-party sources. *In Proceedings of the EMNLP-CoNLL.*

[4] Gunes¸ Erkan and Dragomir R. Radev. 2004. LexRank: ¨Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research,* 22:457–479.

[5] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 25–26 July 2004, pages 404–411.

[6] Parveen, D.; Ramsl, H.-M.; and Strube, M. 2015. Topical coherence for graph-based extractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 19491954.

[7] Wan, X. 2010. Towards a unified approach to simultaneous single-document and multidocument summarizations. *In Proceedings of the 23rd COLING,* 11371145.

[8] A. Nenkova and L. Vanderwende. 2005. The impact of frequency on summarization. Technical report, *Microsoft Research*.

[9] A Haghighi and L Vanderwende. 2009. Exploring content models for multi-document summarization. *In Proc. of HLT: NAACL* 2009, pages 362–370, Boulder, USA.

[10] J. Steinberger and K. Jezek, ˇ "Using latent semantic analysis in text summarization and summary evaluation," *in Proc. ISIM '04,* 2004, pp. 93–100.

[11] Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. *In Proc. of the ACL Workshop on Text Summarization Branches Out* .

[12] Michael K. Bergman. 2001. White paper: the deep web: surfacing hidden value. *Journal of electronic publishing,* 7(1).

[13] M. AL NABKI, E. Fidalgo, E. Alegre and I. de Paz, "Classifying Illegal Activities on Tor Network Based on Web Textual Contents", *European Chapter of the Association for Computational Linguistics*, 2017.

[14] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development*.

[15] Baxendale, P. (1958). Machine-made index for technical literature - an experiment. *IBM Journal of Research Development,*

[16] Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM,* 16(2):264.

[17] Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. *In Proceedings SIGIR '95,* New York, NY, USA.

[18] Lin, C.-Y. and Hovy, E. (1997). Identifying topics by position. *In Proceedings of the Fifth conference on Applied natural language processing.*

[19] Yang L, Cai X, Zhang Y, Shi P (2014) Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization. *Inf Sci* 260:37–50.