SMS Spam Collection v.1

1 DESCRIPTION

The SMS Spam Collection v.1 (hereafter the corpus) is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged acording being ham (legitimate) or spam.

1.1 Compilation

This corpus has been collected from free or free for research sources at the Web:

- A collection of between 425 SMS spam messages extracted manually from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages. The Grumbletext Web site is: http://www.grumbletext.co.uk/

- A list of 450 SMS ham messages collected from Caroline Tag's PhD Theses available at http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

- A subset of 3,375 SMS ham messages of the NUS SMS Corpus (NSC), which is a corpus of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available. The NUS SMS Corpus is avalaible at: http://www.comp.nus.edu.sg/ rpnlpir/downloads/corpora/smsCorpus/

- The amount of 1,002 SMS ham messages and 322 spam messages extracted from the SMS Spam Corpus v.0.1 Big created by Jose Maria Gomez Hidalgo and public available at: http://www.esp.uem.es/jmgomez/smsspamcorpus/

1.2 Statistics

There is one collection:

$The SMS Spam Collection v.1 (text file : smsspamcollection) has a total of 4,827 SMS legitimate messa$

1.3 Format

The files contain one message per line. Each line is composed by two columns: one with label (ham or spam) and other with the raw text. Here are some examples:

ham What you doing?how are you?
ham Ok lar... Joking wif u oni...
ham dun say so early hor... U c already then say...
ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham Siva is in hostel aha:-.
ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam FreeMsg: Txt: CALL to No: 86888 claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B
spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Note: messages are not chronologically sorted.

2 USAGE

We offer a comprehensive study of this corpus in the following paper that is under review. This work presents a number of statistics, studies and baseline results for several machine learning methods.

1. Almeida, T.A., Gomez Hidalgo, J.M., Yamakami, A. Contributions to the study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (ACM DOCENG'11), Mountain View, CA, USA, 2011. (Under review)

3 ABOUT

The corpus has been collected by Tiago Agostinho de Almeida (http://www.dt.fee.unicamp.br/ tiago) and Josr Hidalgo (http://www.esp.uem.es/jmgomez).

We would like to thank Dr. Min-Yen Kan (http://www.comp.nus.edu.sg/ kanmy/) and his team for making the NUS SMS Corpus available. See: http://www.comp.nus.edu.sg/ rpnlpir/
He is currently collecting a bigger SMS corpus at: http://wing.comp.nus.edu.sg:8080/SMSCorpus/

4 LICENSE/DISCLAIMER

We would appreciate if:

- In case you find this corpus useful, please make a reference to previous paper and the web page: http://www.dt.fee.unicamp.br/ tiago/smsspamcollection/ in your papers, research, etc. - Send us a message to tiago@dt.fee.unicamp.br in case you make use of the corpus.

The SMS Spam Collection v.1 is provided for free and with no limitations excepting:

1. Tiago Agostinho de Almeida and Josr Hidalgo hold the copyrigth (c) for the SMS Spam Collection v.1.

2. No Warranty/Use At Your Risk. THE CORPUS IS MADE AT NO CHARGE. ACCORDINGLY, THE CORPUS IS PROVIDED 'AS IS,' WITHOUT WARRANTY OF ANY KIND, INCLUDING WITHOUT LIMITATION THE WARRANTIES THAT THEY ARE MERCHANTABLE, FIT FOR A PARTICULAR PURPOSE OR NON-INFRINGING. YOU ARE SOLELY RESPONSIBLE FOR YOUR USE, DISTRIBUTION, MODIFICATION, REPRODUCTION AND PUBLICATION OF THE CORPUS AND ANY DERIVATIVE WORKS THEREOF BY YOU AND ANY OF YOUR SUBLICENSEES (COLLECTIVELY, 'YOUR CORPUS USE'). THE ENTIRE RISK AS TO YOUR CORPUS USE IS BORNE BY YOU. YOU AGREE TO INDEMNIFY AND HOLD THE COPYRIGHT HOLDERS, AND THEIR AFFILIATES HARMLESS FROM ANY CLAIMS ARISING FROM OR RELATING TO YOUR CORPUS USE.

3. Limitation of Liability. IN NO EVENT SHALL THE COPYRIGHT HOLDERS OR THEIR AFFILIATES, OR THE CORPUS CONTRIBUTING EDITORS, BE LIABLE FOR ANY INDIRECT, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES, INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF GOODWILL OR ANY AND ALL OTHER COMMERCIAL DAMAGES OR LOSSES, EVEN IF ADVISED OF THE POSSIBILITY THEREOF, AND REGARDLESS OF WHETHER ANY CLAIM IS BASED UPON ANY CONTRACT, TORT OR OTHER LEGAL OR EQUITABLE THEORY, RELATING OR ARISING FROM THE CORPUS, YOUR CORPUS USE OR THIS LICENSE AGREEMENT.