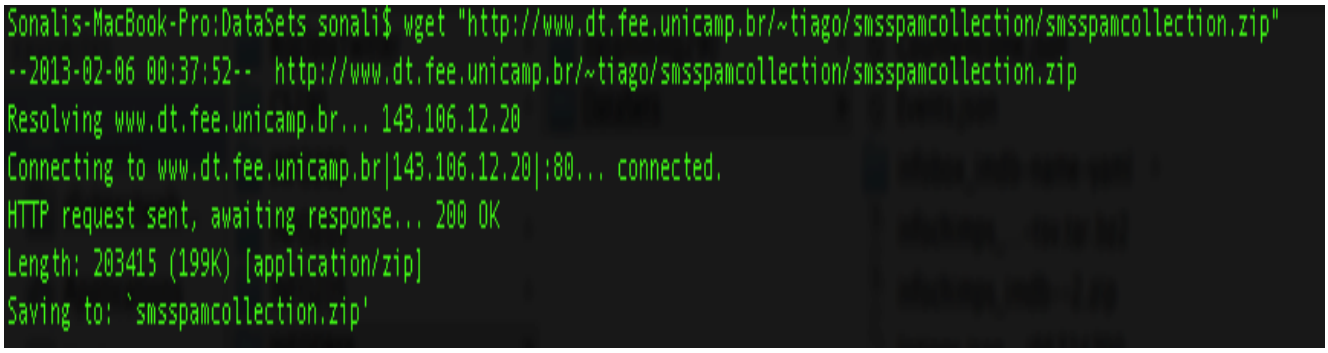
Assignment 1

Dataset: SMS Collection Dataset

Description:

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research and obtained from http://bitly.com/bundles/hmason/1 It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.
The dataset was directly downloaded as a zip file from the website using wget command



The file was unzipped

Format:

By looking at the file on the editor using less and more commands, the following format was found. 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:-.

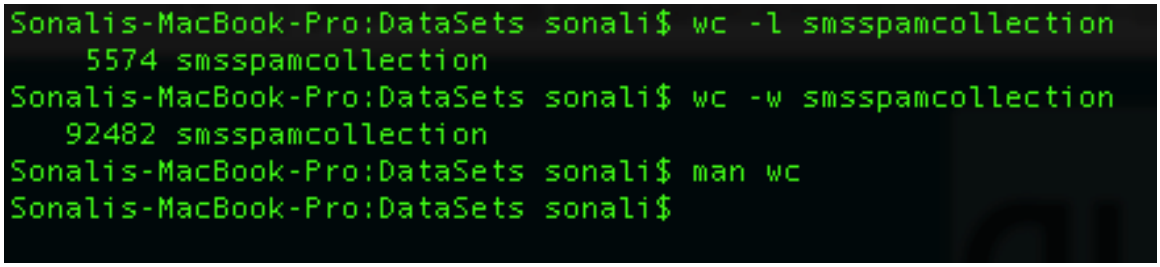
ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.

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

No of records available:
Using the wc command the number of lines available were
5575 and the number of words in the file are 92482
This is a small sized dataset.



Interesting facts:
This entire dataset is very interesting because just by looking at the dataset one can make out why the messages have been marked as Spam. Below are some interesting observations:

1) Most of the spam messages have frequent use of the word reward, win, free, text. Using the grep command this can be

proved easily. It gives a sense that SPAM target customers by making promises of rewards and prizes.

2) A lot of spam message start with bold letters and have words like 'Important, 'Urgent'. Looks like this is done to catch the attention of people and make them open the SMS.

   spam *important information 4 orange user . today is your lucky day!2find out why log onto http://www.urawinner.com THERE'S A FANTASTIC SURPRISE AWAITING YOU!*

   spam *URGENT. Important information for 02 user. Today is your lucky day! 2 find out why , log onto http://www.urawinner.com there is a fantastic surprise awaiting you !*

3) On an average Spam messages are much longer in length than the ham (legitimate messages)

Three questions that can be answered using data set.

1) Most frequently terms used in SMS messages, what sort of words are prevalent in spam messages. As described above From the dataset we get a good answer for this.

2) The messages are all labeled as Spam and Ham (legitimate), Hence, this can be used a good training set to build a classifier to classify SMSs. We can also get a good statistics in percentage about the number of spam messages vs. no of legitimate messages for a given set.

3) We can get the url used in Spam messages. These url can then be traced to a lot of fake websites and non secure websites.

What we don't get from this particular dataset is the location and timestamp details. If these details were available we could have extracted a pattern of conversation between two people and other statistics like at what time are people most active on SMSs and from which locations do SMS come the most.