Springboard

Karolina Bzdusek

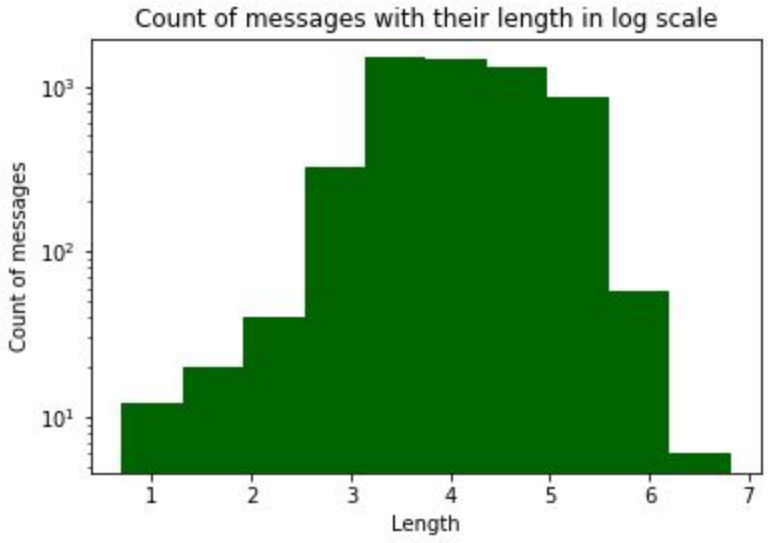
Capstone Project 1 : In Depth Analysis

Let’s dive into our intuition about data set and what assumption can we valide as good ones.
First let’s sum up some fact about SMS:
The maximum length of text message that you can send is 918 characters. However, if you send more than 160 characters then your message will be broken down into chunks of 153 characters before being sent to the recipient's handset.

Here is description of the length of whole dataset:

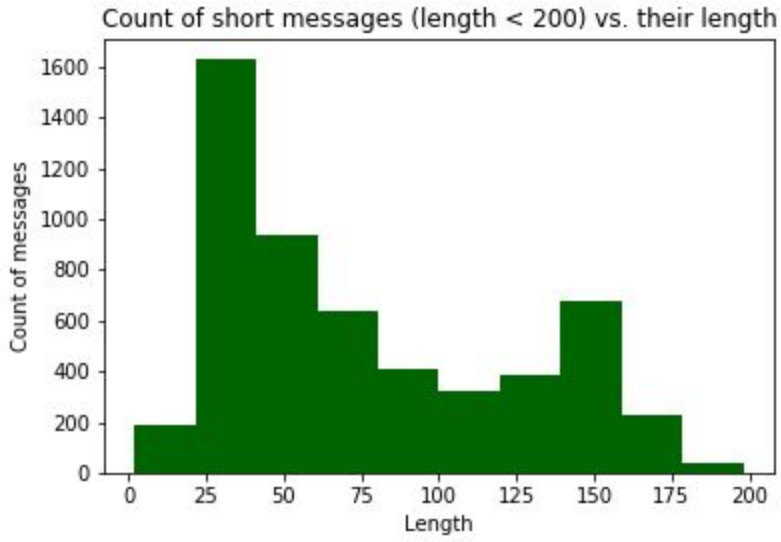
count	5572.000000
mean	80.489950
std	59.942907
min	2.000000
25%	36.000000
50%	62.000000
75%	122.000000
max	910.000000

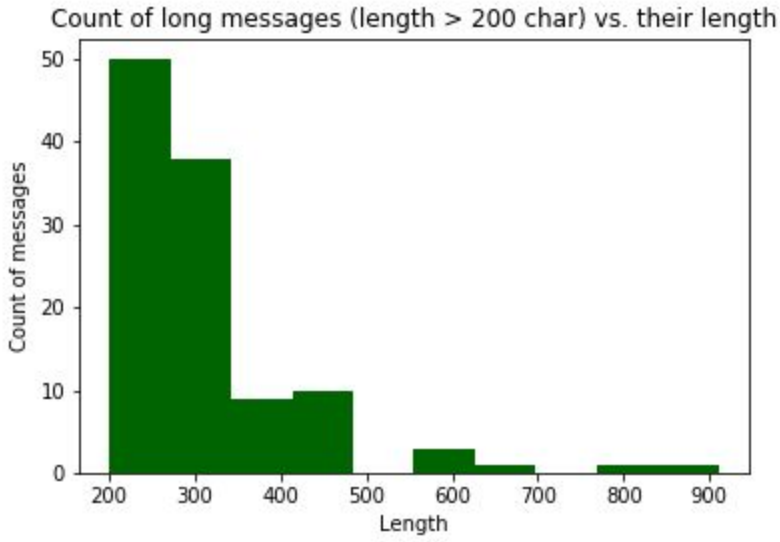
To visualize our distribution, look on the pictures below. First one is distribution in log scale.



The other two are in normal unscaled. We split the distribution into two pictures, to see first distributions of SMS which length is less than 200 characters and second with SMS's length more than 200 characters (we choose length = 200 only because of resolution reasons).

In our analysis we will split short and long messages whether they belong to the top 25% or not (length = 122).





1) Really short messages tends to be a 'ham'.

For example, someone texts you just 'Ok' as an answer to some question. As they do not contain a lot of information and these kind of messages are part of well-known conversation, then our intuition tells us that such a short messages have to be a 'ham'.

Let's explore this. We start with short messages (length = 122, 75% of all messages in our dataset) and then we look at 50 shortest messages, what is their contain and whether is 'spam' or 'ham'.

Number of short sms spam: 140 from 747 spam messages
Spam ratio = 0.18741633199464525 (short spam/all spam messages)
Number of short sms ham: 4025 from 4825 ham messages
Ham ratio = 0.8341968911917098 (short ham/all ham messages)

Unique content of 50 shortest messages:
'Ok', 'Yup', '645', 'Ok.', ':) ', 'Ok..', 'Okie', 'U 2.', 'Ok...', 'G.W.R', 'Y lei?', 'Yup...', 'ALRITE', 'Okie...', 'Where @', 'Oh ok..', 'Ok lor.', 'Nite...', 'Havent.', ':-) :-)', 'Thanx...', 'Thank u!', 'Beerage?', 'U too...', 'My phone', "I'm home.", 'Yup ok...', 'How come?'

2) *What about long messages? Are they spam or ham? I would guess, that too long messages would be 'ham' as well - spam usually tries to "sells something" and therefore too long messages would be big nuisance (although even short spam messages are nuisance). Therefore we will look at messages that are longer than two standard SMS (>300 characters).*

There is actually only 41 messages in our dataset being longer than 300 characters and all of them are 'ham'.

So according length of the message we can say that f message is too short or too long, it is most probably a 'ham'. However, if it is message having about 160 characters, we can't determine whether it is 'ham' or 'spam' using length as the only feature.

3) Messages containing more than 2 exclamation marks tends to be spam.

Intuitively we would say this so, although sometimes people are overusing exclamations marks when they are feeling strong emotions. After closer look at our data, we cannot say whether this is True or False as we have only few messages having more than 2 exclamations marks (N=78).

4) What about 2-grams such as “call” and “buy”? Can we say they tend to be a spam?

After filtering out messages containing “call” and “buy” we got 70 messages and all are labeled as a ‘spam’. Spam ratio for messages containing spam is 0.9. Although ration is small but we haven’t found any messages containing those two words and being ‘ham’.