SMS Classification

Background

Classify short messaging service (SMS) text as:

- REGULAR/HAM: usual conversations between two persons
- INFORMATIONAL: from establishments like banks, airlines, e-commerce business, insurance companies, etc. that has a value to the user
- SPAM: All other SMS that need to be discarded

Examples of REGULAR/HAM messages:

- Hi, how r u?
- Can we go for a movie
- Let's have dinner together

Examples of INFORMATIONAL messages:

- Your amazon package will be delivered today at 12pm. You can track here ….
- Your credit card bill is dispatched
- Your electricity bill is due on 20/Sep/2017

Examples of SPAM messages:

- 50% off sale: ….
- Thanks for donating blood ..
- The big online home fest is now live. Buy this coupon & get minimum saving of 7.2%. Hurry! …..

Bonus problem

Besides above the three classes, further classify INFORMATIONAL messages into sub-classes like DELIVERY, TICKETS (BUS, TRAIN, FLIGHT), MOVIE, HOTEL, APPOINTMENT, PAYMENT etc. (We will publish the final classes during the contest).

Competition

As part of this challenge we need you to develop/use a machine learning algorithm to classify incoming SMS into above three classes REGULAR, INFORMATIONAL & SPAM.

You can also take up a stretch challenge to solve the bonus problem besides solving the main problem. Please note the training data is highly sparse for these categories.

Evaluation

The completion is evaluated on weighted average of f_1 score (https://en.wikipedia.org/wiki/F1_score) of each class, given by:

$$F_1 = \frac{\sum_{c \in C} f_1^c \times |c|}{N}$$

where :

$|c|$*is the no. of records of a given class*
C *is set of all classess and*
N *is total size of the test set.*

The formula for computing f_1 score for given class will be

$$f_1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

Higher F_1 score is better classification.

Submission file format

The file should contain a header and have the following format

- Dataset

We have split the dataset into three datasets (TRAIN, DEV and TEST dataset).

The TRAIN dataset (TRAIN_SMS.csv) as expected would be used for training and building the ML model and is available during development phase. There are two columns, Label and Message, in csv format as follows:

```
Label,Message
ham,oh how abt 2 days before Christmas
info,"Welcome to OVATION HOLD R.No. 184, 114, 395, 378 Ch.In 2014-10-21 3:53 Ch.out
2014-11-01 12:00."
spam,OTP is 817453 for the txn of INR 8262.00 at SPICE JET on your SBI bank Debit card
ending with 1385. Valid till 20:32:54. Do not share the OTP with anyone for security
reasons
```

DEV dataset (DEV_SMS.csv) containing 7500 no. of records is available as a part of development phase as well. DEV dataset will be used during DEV round to evaluate F1 score. There are two columns, RecordNo and Message, in csv format as follows:

```
RecordNo,Message
```

```
1,oh how abt 2 days before Christmas
2,"Welcome to OVATION HOLD R.No. 184, 114, 395, 378 Ch.In 2014-10-21 3:53 Ch.out
2014-11-01 12:00."
3,OTP is 817453 for the txn of INR 8262.00 at SPICE JET on your SBI bank Debit card
ending with 1385. Valid till 20:32:54. Do not share the OTP with anyone for security
reasons
```

The LeaderBoard data set (LeaderBoard_DEV_SMS _label.csv) will be used for evaluation of output of the DEV set during the contest to maintain the leader board. This set is actually DEV set with correct labels. There are also two columns, RecordNo and Label, in csv format as follows:

```
RecordNo,Label
1,ham
2,info
3,spam
```

TEST dataset (TEST_SMS.csv) will be used to generate the final submissions of participants. However we will release TEST dataset in the final 30 minutes. The format is same as DEV dataset, there are two columns, RecordNo and Message, in csv format as follows:

```
RecordNo,Message
1,oh how abt 2 days before Christmas
2,OTP is 817453 for the txn of INR 8262.00 at SPICE JET on your SBI bank Debit card
ending with 1385. Valid till 20:32:54. Do not share the OTP with anyone for security
reasons
3,"Welcome to OVATION HOLD R.No. 184, 114, 395, 378 Ch.In 2014-10-21 3:53 Ch.out
2014-11-01 12:00."
```

The evaluation set (Evaluation_SMS_label.csv) is the labels of TEST dataset and will be used for evaluation of final submissions and final scoring of participants. The evaluation set is secret. However we will release Evaluation set after final scoring is finished. The format is same as LeaderBoard data set, there are two columns, RecordNo and Label, in csv format as follows:

```
RecordNo,Label
1,ham
2,spam
3,info
```

The same format is also maintained for the dataset of the bonus problem.

Dataset Details:

For main problem:

1. TRAIN_SMS.csv --- 30000
2. DEV_SMS.csv --- 7500
3. LeaderBoard_DEV_SMS _label.csv --- 7500
4. TEST_SMS.csv --- 7500
5. Evaluation_SMS_label.csv (hidden) --- 7500

For Bonus problem:

1. TRAIN_info_SMS.csv --- 12000
2. DEV_info_SMS.csv --- 3000
3. LeaderBoard_DEV_info_SMS _label --- 3000

4. TEST_info_SMS.csv --- 3000
5. Evaluation_SMS_label.csv (hidden) --- 3000

You may use additional data for training purposes, but remember that the distributional properties of the additional data must match those of the training data (so that they may match the distributional properties of the test data).

----------XXXXXX----------XXXXXX----------