Problem statement:

Identification of chats whether given sample is a similar to other sample short text .

This dataset contains text pairs extracted from chat logs from 50+ bots with a similar or not similar label. Note that both texts of the pairs are from human users.

1. The texts in the pairs are text messages written by humans to get some task done with a task oriented chat bot.

2. Unlike ideal datasets which are mostly generated from news, emails, etc., and contains well structured English sentences, this dataset contains anomalies in sentence structure, spelling mistakes, etc. which are common on messaging/voice interface.

3. It can be best used to evaluate domain agnostic word representations for semantic similarity downstream task.

The file `train.csv` has 4 columns

- `pid` - Unique id for pairs.

- `sentence1` - One of messages sent by some user to one of the bots.

- `sentence2` - Another one of message sent by some user to the same bot as `sentence1`.

- `label` - Either 0 or 1. `0` means `sentence1` and `sentence2` are not similar. `1` means `sentence1` and `sentence2` are similar

Instructions:

1. Pre-process your data, split it to test and evaluation.
2. Build your model , you can use pre-trained models
3. Predict on test data

**Note**: Only the fast.ai library should be used to build the model; for other tasks (preprocessing etc.) you may use any library

If you face any labs related issues please mail to [support@nuvepro.freshdesk.com](mailto:support@nuvepro.freshdesk.com)

Evaluation process:

Evaluation will be in **two stages**

1. need to get at least 70% F1-score to get your name reflected on evaluation board
2. then later your code files(.zip file) will be evaluated by SME and graded according to the given rubric below.

30% marks will be from evaluation board and 70% will be form SME evaluation according to the given rubric.

You need 70% marks to pass the assessment and get the relevant credits.

Submission I :

Upload the prediction file in the ‘Upload’ folder in your lab instance

Submit your prediction file with format ***name\_employeeid\_problemX.csv* (X is your set number)**

**e.g. :**  mike\_ F02839\_problem2.csv

\*please remember to add **problemX** keyword otherwise it will be rejected.

See your score in evaluation board link here at <http://fractal.cloudloka.com/>

Submission II :

|  |  |
| --- | --- |
| Submission includes all files required for grading **(to be sent to a FAA member)** | * Includes NLPset1.ipynb displaying output for all executed cells * Includes NLPset1.html, which is an HTML copy of the notebook showing the output from executing all cells |

Submit both files in zip named ***employee\_name\_employee\_ID.zip.***

E.g. **mike\_F02839.zip**

Evaluation Rubric:

Code Functionality (10 marks)

All the cells should work giving desirable output when executed.

Preprocessing step (20 mks)

1. Normalize case(5 mks)

2. Tokenize (using word\_tokenize from NLTK) or spacy (5 marks)

3. POS tagging using the NLTK pos tagger or spacy (5 marks)

4. Lemmatize or use stemming the corpus & Remove stop words and punctuation (if there are any at all after the POS tagging)(5 marks)

Hyperparameters (40 marks)

|  |  |
| --- | --- |
| Number of epochs | * The number of epochs is chosen such the network is trained well enough to accurately make predictions but is not overfitting to the training data. |
| Learning rate | * The learning rate is chosen such that the network successfully converges but is still time efficient. * Try different learning rates. Your program should demonstrate use of different learning rate and choose the learning rate which gives best accuracy. |
| Final Results | * The training loss is below 0.4 and the validation loss is below 0.5. |