Domain Identification based on message (text classification)

Difficulty: easy to medium

Required skill-set : ML Classification basics, text processing(basic), Deep Learning (Optional)

Some Useful Tools: Scikit Learn, libsvm, tensorflow, nltk, etc.

Problem Statement:

- For automation in conversational agents, it is very important to know the topic of the conversation. In Haptik, Domain Identification is one of the most critical parts of the automation system. The domain identification needs to be highly accurate as further processing depends on the domain identified.
- Provided a message from original conversation, the task is to identify the domain to which the message/conversation belongs. Possible domains: food, movies, nearby, recharge, reminders, shopping, support, travel.
- Attached are two files
  - **train.tsv** : Each line contains a message and corresponding domain (tab separated)
  - **test.txt**: Each line is a separate message
- The task is to train a supervised model using tagged data (*train.tsv*) and predict the domains/labels for each message in *test.txt* . Use of external data is allowed.
- The accuracy is calculated based on the results provided for test data.
- The messages provided are directly taken from the original conversations.

Submission:

- result.tsv (same format as that of train.tsv). This file contains messages in test file tab-separated by identified domain.
- A well commented script. Write modular code (separate functions/classes for training and testing).
- Trained models. Store all the trained models to disk.
- While testing, the code should read the model from disk and run it on test data.
- A document explaining approach (preprocessing, algorithm sued, etc.) and read for running code. You can also include the experiments which you tried but did not give good results (with explanations, if you can).

Notes:

- You are free to use any open source libraries for deep learning/machine learning/text processing.
- You can use any machine learning algorithm.
- We expect to see an accuracy of at least 85%.
- Use of NLP techniques and deep learning will be appreciated. If you tried using any such technique and didn’t produce good results, still mention it in the document.
- You are free to use external data from any source (mention in the document).