UNIL | Université de Lausanne

Project 2 - Big-Scale Analytics

University of Lausanne, Professor: Michalis Vlachos

Deliverable due: Tue May 19, 23.59pm.

Project description

Real or Not? NLP with Disaster Tweets: In this project you are challenged to build a Machine Learning model that can predict which tweets are about a real disaster and which are not. The project topic is based around a Kaggle competition. You can find the link to the competition **here**. You will find more information about the project and the dataset in the competition page.

In this project, you will have the chance to compare your prediction results with your fellows (and also other Kaggle users). Make an account in Kaggle (if you don't have already) and join the competition. As soon as you make a submission you can see the prediction accuracy and your ranking in the leaderboard. Note that you can only make 5 submissions per day. To know more about the competition rules, check the competition page in Kaggle.

Submissions

As you build your model and train it on the training data, you can generate predictions for the (unlabelled) test data. Make sure that your submission file has the same format as the example submission file in the competition page. Remember that you have only 5 shots per day, so it is better to leave out part of your training data for testing. Once you are sure about your model and satisfied with the prediction accuracy you got (on your own test data), you can try to generate predictions for the actual test data and submit in the competition.

Deliverables

I. Create a **Python notebook** that explains every step of your pipeline, from loading the data and preprocessing to building the model. Your notebook also stands as your report as well, so make sure that you add sufficient explanations to it. Add appropriate plots and/or tables to your notebook in case you think it can make your notebook more comprehensive. It is very important also to document, the **progress** of your Kaggle submissions. Eg you should keep track of the reported accuracies of the different submissions and also what changes you introduced to have some improvement in your score.

II. Create a **short video** of your solution (duration is up to you) and report also in the video your best rank in the leaderboard. Post your video in slack.

III. During the last class (May 26) you will give a short **presentation in class** for 5-7 minutes, briefly explaining what you did and your best result.

Logistics and deadline

1. Make an account in Kaggle and join the competition. Under the team tab you can select a team name for yourself. The name should be of the type **UNIL_[team_name]**, eg UNIL_Google, etc. This name will be shown in the leaderboard.
2. Make sure to mention your team name in your notebook, so that we can match everyone with their team names in Kaggle.
3. There will be Gift for the Champion! (team with the highest ranks in the leaderboard) :)

   The deadline for the submission of your material is **19th of May 2020** at 23h59.

Grading

1. Presentation in the class: 1/3 of the grade
2. Notebook quality (clean code, sufficient explanations, etc): 1/3 of the grade
3. Having innovative ideas: 1/3 of the grade

Good luck with the project and the competition. We look forward to seeing your solution!