DataMites™ Project Mentoring PR-0019

Data Science PROJECT

Al-ClientProject-4

© Copyright 2018 DataMites™ Solutions Pvt Ltd. All rights reserved. Doc Ref: PM-PR-0019 This document is proprietary and confidential. No part of this document may be disclosed in any manner to a third party without the prior written consent of DataMites™ Solutions Pvt Ltd. Contact <u>license@datamites.com</u>

Business Case:

The goal of this project is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labeling is also a 'noisy' process, and reasonable people will disagree. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labeling. We believe the labels, on the whole, to represent a reasonable consensus, but this may often not be true on a case by case basis for individual items in the dataset.

PROJECT GOAL:

1. Create a model that can predict which of the provided pairs of questions contain two questions with the same meaning.

Feature Details:

id - the id of a training set question pair qid1, qid2 - unique ids of each question (only available in train.csv) question1, question2 - the full text of each question is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

Link to dataset:

https://www.kaggle.com/competitions/quora-question-pairs/data?select=train.csv.zip

© Copyright 2018 DataMites™ Solutions Pvt Ltd. All rights reserved. Doc Ref: PM-PR-0019 This document is proprietary and confidential. No part of this document may be disclosed in any manner to a third party without the prior written consent of DataMites™ Solutions Pvt Ltd. Contact license@datamites.com