Data Scientist - Technical Assignment

Many thanks for your continued interest in joining Adthena. To understand if you are a good fit for Adthena and if we are a good fit for you, we would like to work together on a test assignment. Please complete this task as if you are already working as part of our team at Adthena. The assignment is as much about testing our ability to work together as it is your ability to implement features/products. Please take the opportunity to ask any questions during the assignment as you would in a normal working environment.

Technology Requirement

You should complete the test using Python

Duration

The assignment should be returned to us within 1 week. We believe it should not take you more than 4 to 6 hours working on it. If you need extra time due to other commitment please let us know.

Submission

Submit your code in any format you wish along with your source code. You could also submit your code using GitHub, BitBucket or similar platforms.

Documentation

Please provide a README.txt. This should contain details of the steps necessary to install the software that is required to train & test the model. Additionally, the README.txt should contain details of the steps necessary to train and test the model. These steps should be such that they can be called from the command line and it should be possible to specify the data for training/prediction.

Please save these predictions in a CSV that follows the same format as given in trainSet.csv, with any pre-processing occurring in the training/prediction process.

Contact

Please feel free to contact with any questions
Shaohong Bai (shaohong.bai@adthena.com.) or Vasanth John (vasanth.john@adthena.com)

Search Term Categorisation

At Adthena we analyse millions of search terms in the search advertising landscape. In order to enrich our in-house data set, we would like to associate metadata to the search terms in our data, such as assigning each search term with a category or categories. For example, the search term `cricket nets' could be assigned to the category: Sports & Fitness -> Sporting Goods -> Cricket Equipment

Data Introduction

A sample of labeled search term category data is provided in the CSV file, trainSet.csv. The file contains the two columns, the search term and the search term category. The search term category has been indexed. There are 606,823 examples in the data set with 1,419 different search term categories. There are roughly 427 examples in each category.

Test Files

https://s3-eu-west-1.amazonaws.com/adthena-ds-test/trainSet.csv
https://s3-eu-west-1.amazonaws.com/adthena-ds-test/candidateTestSet.txt

Problem Description

In this test we would like you to construct a classification model that can accurately categorise search terms. You are free to use whatever model that you feel is appropriate for the problem.

We would like you to provide us with a detailed description of the work that you have done in constructing this model. This document should include the following details:

1. A description of the type of model you selected for the problem, along with details explaining why you selected this type of model.
2. A description of any pre-processing steps that you took prior to training the model.
3. A description of the methods you considered when evaluating the performance of model.
4. A description of the runtime complexity of the model, as well as the memory overhead of the model at both during model training and predicting against new cases.
5. A description of the weaknesses of the selected model for the given problem, along with possible improvements that you would consider in the future.

The ability to write clean reusable code is an important part of this role. We would also like you to provide us with the code to train & test the model, including any pre-processing etc. which is necessary to train and test the model. We have constructed a hold-out set with which we will evaluate your model. The search terms in this hold-out set are provided in the file, candidateTestSet.txt. Please use your classification model to generate predictions for each of the search terms in the hold-out set.