Microsoft Bing Hackathon!

Microsoft IDC Bangalore, 16 January, 2015

1 THE STORY SO FAR

While traveling in your spaceship, you intercept an interstellar communication of a highly interesting nature. Your cryptolinguist friend tells you that you've discovered a collection of documents from the planet Klauti, containing discussions of a highly technical and academic nature. The documents appear to be about three technical disciplines the Klautinians care very deeply about: Psychohistory, Computational Neuropharmacology, and Sociophilosophy. As we all know, these topics are very closely related to each other, so much so that it is sometimes difficult to tell them apart.

This is your chance! As a data scientist, you have the tools to help your cryptolinguist friend decipher the documents. Well, maybe not decipher, but at least figure out what they are about. You are given tokenized representations of the documents (after all, you can't read Klautinian, can you?). You must build a classifier that categorizes documents into the three different topics. As an additional service to cryptolinguistics, maybe you could also figure out roughly when the document was written?

2 THE DATA

You are given information about text books on three topics; each row is a tab-separated list of columns.

(1) Record ID,
(2) Topic,
(3) publication year,
(4) authors (semicolon-separated),
(5) title,
(6) summary (a sequence of sentences, separated by periods).

All words and name tokens have been replaced by IDs. There are two separate files for training and testing, both following the same 6 column format. In the test data, however, the topic and publication year are all zeros.

3 THE GOAL

For each record in the test set, predict the Topic (***challenge 1***), and the year of publication (***challenge 2***). You can submit your results on the submission website in the prescribed format.

Format: For challenge 1, please submit a tab-separated text file, with each row having the columns "record id" and "topic" (no quotes). For challenge 2, the submission format is similar: rows containing tab-separated values of "record ID", and "publication year".

Shortly after submitting your predictions, you will receive a *score* measuring the performance of your submission. Submissions on the first challenge will be graded on *prediction accuracy*, and the second challenge on *root mean squared error*.

Note: Each team can make a *maximum of 3 submissions on Day 1, and 2 additional submissions on day 2.* The team's score in each challenge is the best-of-k, among the k submissions the team has made.

Before the end of the hackathon, please upload all code you have written, to the prescribed website. Your code submission should include the following:

(1) A general outline of your solution, with technical details,
(2) Any additional code or frameworks that are needed to make your code work,
(3) Instructions on how to run your code to produce your best-performing result in each challenge.

4 BONUS CHALLENGE

Now that you've learned enough about Klautinian academics, maybe you can masquerade as one? If you were to send a manuscript on one of the Topics to a Klautinian publisher, and it was a really convincing one, they might just invite you to Klauti to sign a book contract. So now you have to write and submit a manuscript; one that talks authoritatively about one of the 3 Topics. Warning! Try too often too poorly, and they might blacklist you. How do you get past this hurdle?

We'll add some constraints on this *generative* challenge, to make it more concrete. Please submit the following columns in a single, tab-separated row. Please submit only one row at a time.

- Year of publication: up to you.
- Authors: only from authors in training data; not more than 3 authors; semicolon-separated.
- Title: not more than 10 words (token IDs), from the existing vocabulary
- Summary: only from the set of sentences in all the training data summaries; not more than 8 sentences. A sentence is a sequence of token IDs, space separated, and followed by a period.

Submission & Scoring: You can submit your data to the specified website, where our topic classifier will work on the submitted document and give you normalized scores for the 3 Topics (s1, s2, s3), adding up to 1. The scores indicate how closely the submitted document matches each topic. Your submission will be a file containing a single row with these above columns, tab-separated in a manner exactly like the training data. You will get to submit up to 10 documents (separate submissions) to the classifier. Your goal is to maximize any one of the scores of your choice; for each submission we will take the max of s1-s3 as score, and your best score across submissions will be taken into consideration.

5 EVALUATION

Teams will be evaluated both on their performance in the prediction challenges, as well as on the innovative aspects of their methodology. We will announce the top 5 teams shortly after the termination of the hackathon. These top 5 teams will be invited to present a short sketch of their ideas & methods in a 5-minute presentation to the hackathon participants—the presentations will immediately follow the announcement of winners.