Data Science Methodology final assignment

Submitted on January 16, 2022

Shareable Link

PROMPT

Which topic did you choose to apply the data science methodology to? **(2 marks)**

I choose to apply the data science methodology to is **Emails**. I believe by automatically classifying emails, productivity can be increased intensely.

RUBRIC

Did the student pick one of the three topics proposed in the assignment overview?

- 0 points
  No
- **2 points** JA
  Yes

PROMPT

Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. **(3 marks)**

You are required to:

1. Describe the problem, related to the topic you selected.
2. Phrase the problem as a question to be answered using data.

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

Daily, we receive 100’s of emails every day and it may not be possible to look at all of them. We can determine which emails are worth taking a second look by organizing them into various categories like Promotions, Updates, Social, Order Receipts, Important/Not Important, Spam etc.

The Question would be: “**Is it possible to automatically determine the type/category of email based on the content of the email?**”

RUBRIC

The student is required to come up a problem related to the topic they selected and the problem must be phrased as a question that can be answered using data. Use your best judgement to rate the student's completion of the Business Understanding stage.

1 point
Poor. Some description is provided about the problem, but the question to be answered is missing.

2 points
Good. The problem to be solved is described and a question is submitted but the question does not match the problem described.

3 points JA
Excellent. The student gave sufficient description of the problem, and the question to be answered reflects the problem described.

PROMPT

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. **(5 marks)**:

1. Analytic Approach
2. Data Requirements
3. Data Collection
4. Data Understanding and Preparation
5. Modelling and Evaluation

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

1. Analytic Approach A Yes/No answer can be applied to this problem so we can use a classification model. 2. Data Requirements To create the model, we will require information regarding the sender including email address, domain, subject, language ,if the email has an attachment or not, and body of the email to see if it contains a list. 3. Data Collection We can gather all these data from email accounts from various email inboxes (Gmail, Hotmail, yahoo, outlook etc.). We can further merge the emails from the various inboxes to create a good dataset. Descriptive statistics & visualizations can be applied to the data set to assess the content quality and if we have the required information. 4. Data Understanding and Preparation We should remove the redundant data from our dataset. This could be two copies of the same email sent to different inboxes. Since we are working with text, we need to perform text analysis. We should ensure proper groupings to help classify the emails properly. These groupings should be done based on certain keywords present in the subject or content of the email. 5. Modeling and Evaluation We create the classification model. We evaluate the results of the model and see how much is classified correctly or incorrectly. Using this feedback we can tweak the model to add parameters and perform necessary changes to ensure that we're getting the intended results.

RUBRIC

The student is required to explain how they would complete each stage for the problem that they described in the Business Understanding stage. Use your best judgement to rate the student's description of each stage.

1 point
Poor. Many stages are missing and insufficient description is provided.

3 points
Good. At least three stages are described and the description is clear and applies to the question defined in the Business Understanding stage. However, some stages are missing.

5 points
Excellent. All stages are described appropriately and the description is clear and applies to the question that they defined in the Business Understanding stage.

JA