|  |  |
| --- | --- |
| Description | In the final phase of our class project, you will use the datasets you created in phase 2 to build a text categorization model. |
| Instructions | The goal of this final phase of the project is to build a text categorization model on your primary dataset, and to evaluate it on both your primary and your secondary dataset.  Follow the steps below.  1. **Data partitioning**: Create a Training set for your model by randomly selecting 70% of the texts in your PRIMARY dataset.  Use the remaining 30% of texts from the PRIMARY dataset as your Test (PRIMARY) set.  Designate 100% of your SECONDARY dataset as the Test (SECONDARY) dataset.  So you should have one Training set (drawn from the PRIMARY data), and two different Test sets (one from PRIMARY and one from SECONDARY).  2. **Baseline model training**: Train a simple bag-of-words classifier on your Training dataset.  If your data comes from the *stance t*ask, you will build a multiclass model (one which can assign one of three labels - *pro-mitigation, anti-mitigation,* or *unclear*).  If your data comes from the *topic* task, choose only one of the topics (*masking and distancing, lockdowns, vaccination*) to model as a binary classification task.  (You should avoid topics with low numbers of positive examples.)  An example of how to use scikit-learn to build a simple text categorization model is [**here**](https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html).  The course sample notebooks [here](https://blackboard.iit.edu/bbcswebdav/courses/14544_14545_14979.202210/Text%20Categorization%20-%20Logistic%20Regression.ipynb) and [here](https://blackboard.iit.edu/bbcswebdav/courses/14544_14545_14979.202210/Sentiment%20analysis%20examples.ipynb) may also be useful as examples.  3. **Model evaluation 1**: Calculate your baseline model's accuracy for your model's predictions on the Test (PRIMARY) set, and on the Test (SECONDARY) set.  Enter these values in the answer boxes provided.  4. **Feature engineering**: In order to try to improve your model, think about what features of the text might be associated with the category you are trying to predict.  What attributes of a text besides the presence of individual words might be good predictors (for example, regular expression patterns or specific word sequences)?  Create at least three new features that represent attributes of the text.  Add them to your model and retrain.  An example of how to add a set of features (defined as a vector of 1/0 values indicating whether the attribute is present or absent for a given text) is shown [**here**](https://gist.github.com/DerrickHiggins/20c77745b080e3d493231424d7da9a2f).  5. **Model evaluation 2**: Calculate overall model accuracy for your new model's predictions on the Test (PRIMARY) set, and on the Test (SECONDARY) set.  Enter these values in the answer boxes provided.  6. **Reflection**: Answer the questions on model performance |