Readme

I have submitted below 3 artifacts as per submission guidelines:

1. Jar file name is SparkMLAssignment-0.0.1-SNAPSHOT.jar
2. SparkMLAssignment.zip contains entire project
3. 2017CBDE639_AmitGoel_C2BDE_SparkMLAssignment_Report.pdf is the entire project report.

Steps to execute jar file:

1. Put jar file and input file in the same directory on ec2 instance or cloudera VM.
2. Put input file in hdfs using below command:

   ```
   hadoop fs -put gender-classifier-DFE-791531.csv .
   ```

3. Run below commands:

   a. Command to run jar file for **main program** to see performance metrics of finalized models:

   ```
   spark2-submit --class com.upgrad.ml.sparkmlassignment.SparkMLClassificationAssignment --master local --deploy-mode client SparkMLAssignment-0.0.1-SNAPSHOT.jar gender-classifier-DFE-791531.csv
   ```

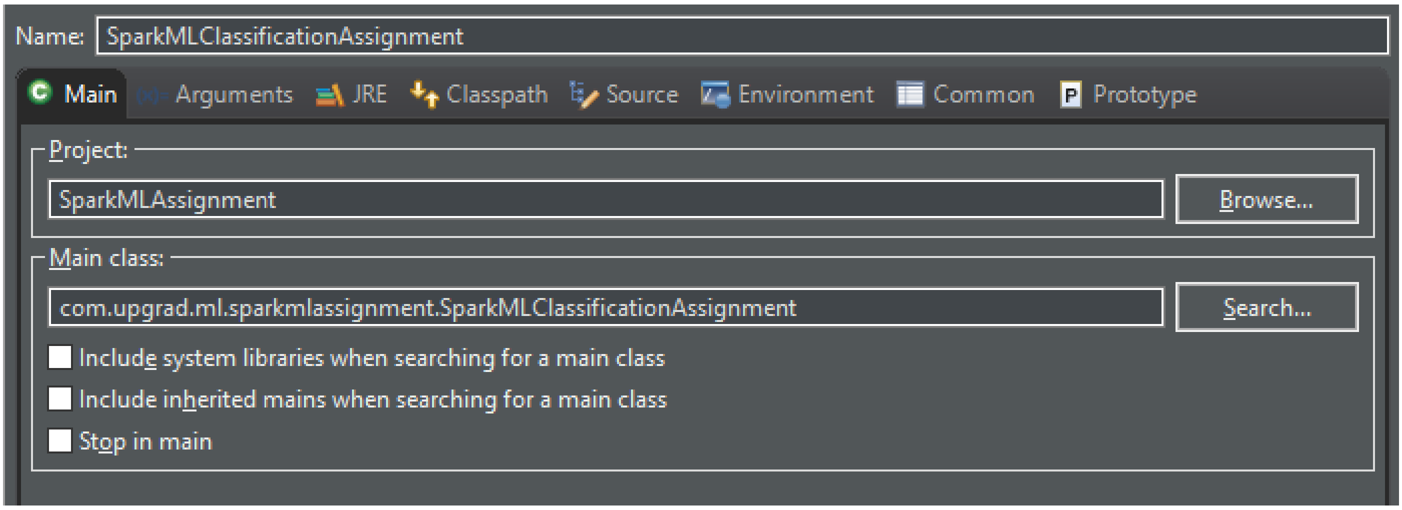
   b. Command to run jar file for **testing program** to see the output of various test models:

   ```
   spark2-submit --class com.upgrad.ml.sparkmlassignment.SparkMLClassificationAssignmentModelTesting --master local --deploy-mode client SparkMLAssignment-0.0.1-SNAPSHOT.jar gender-classifier-DFE-791531.csv
   ```

   c. Command to run jar file for **data pre-processing program** to see data during pre-processing steps:

   ```
   spark2-submit --class com.upgrad.ml.sparkmlassignment.SparkMLClassificationAssignmentDataPreProcessing --master local --deploy-mode client SparkMLAssignment-0.0.1-SNAPSHOT.jar gender-classifier-DFE-791531.csv
   ```

4. In order to run the program in eclipse itself, we can setup run configurations for each program and add argument as path to the input file. Something like this:



Name: SparkMLClassificationAssignment

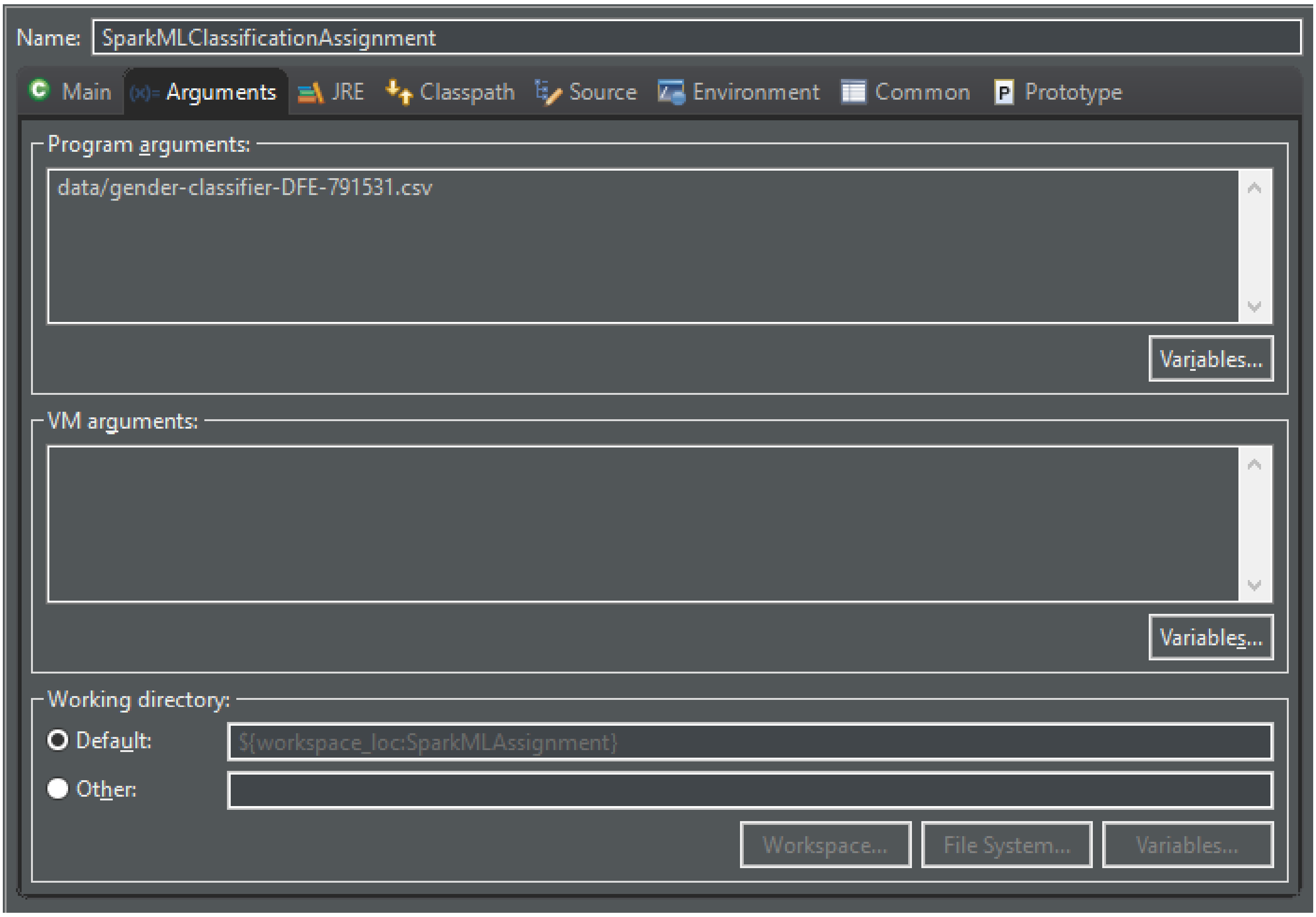
Project: SparkMLAssignment

Main class: com.upgrad.ml.sparkmlassignment.SparkMLClassificationAssignment

□ Include system libraries when searching for a main class

□ Include inherited mains when searching for a main class

□ Stop in main



Name: SparkMLClassificationAssignment

Program arguments:

```
data/gender-classifier-DFE-791531.csv
```

VM arguments:

Working directory:

● Default: ${workspace_loc:SparkMLAssignment}

○ Other: