CIS5560 Term Project Tutorial

Authors: Gregory D. Mamoyac, Tejas Agara Chandrakumar, Nitesh Kamboj, Julia Stachurska

Instructor: Jongwook Woo

Date: 05/06/2018

Lab Tutorial

jstachu (jstachu@calstatela.edu)

nkamboj(nkamboj@calstatela.edu)

afnu3(afnu3@calstatela.edu)

gmamoya(gmamoja@calstatela.edu)

05/06/2018

Azure ML and Spark ML Analysis of Stack 0verflow

For R Language Question and Answers

Azure LDA

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Get data manually
- Create Spark cluster
- Train NLP system
- LDA Clustering of topic/ tags from the “Body” column of Dataset using table features

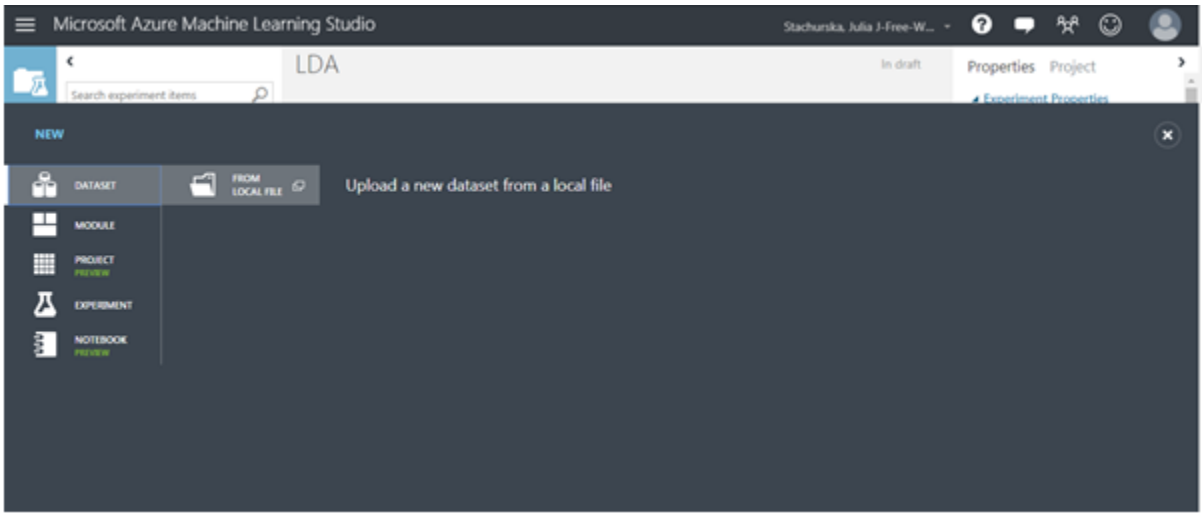
- Visualization
- https://gallery.cortanaintelligence.com/Experiment/LDA-2

Platform Spec

- Microsoft Azure ML
- CPU Speed: ~3.4GHz
- # of nodes: 1
- Total Memory Size: 10GB

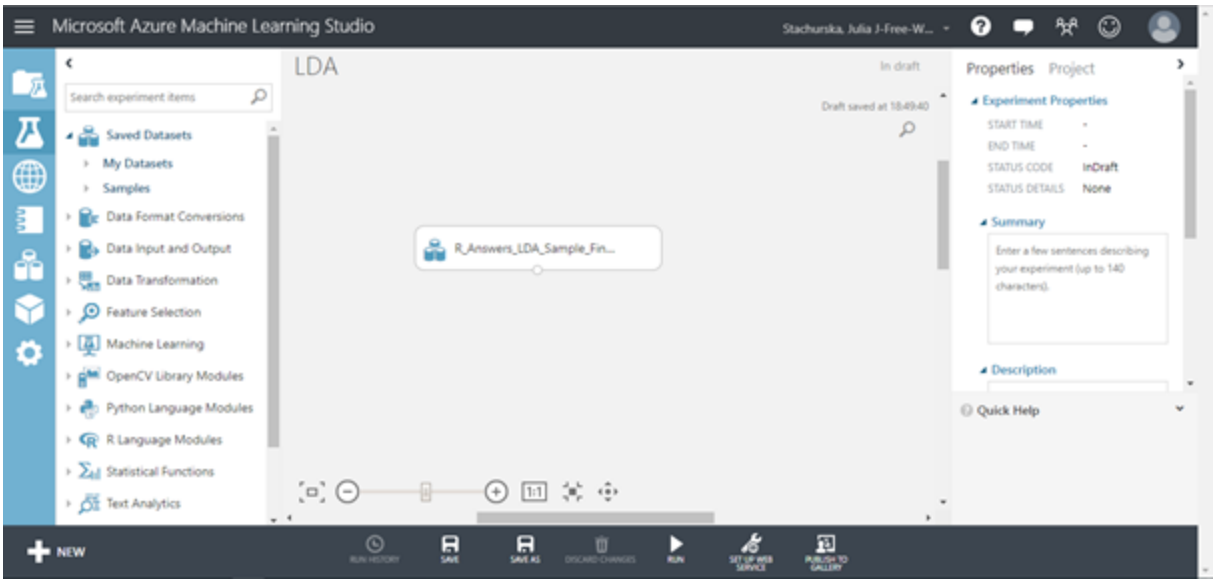
Step 1: Upload the Data Set from local file

This step is to get upload Data Set R_Answers_LDA_Sample_Final.csv from local file.



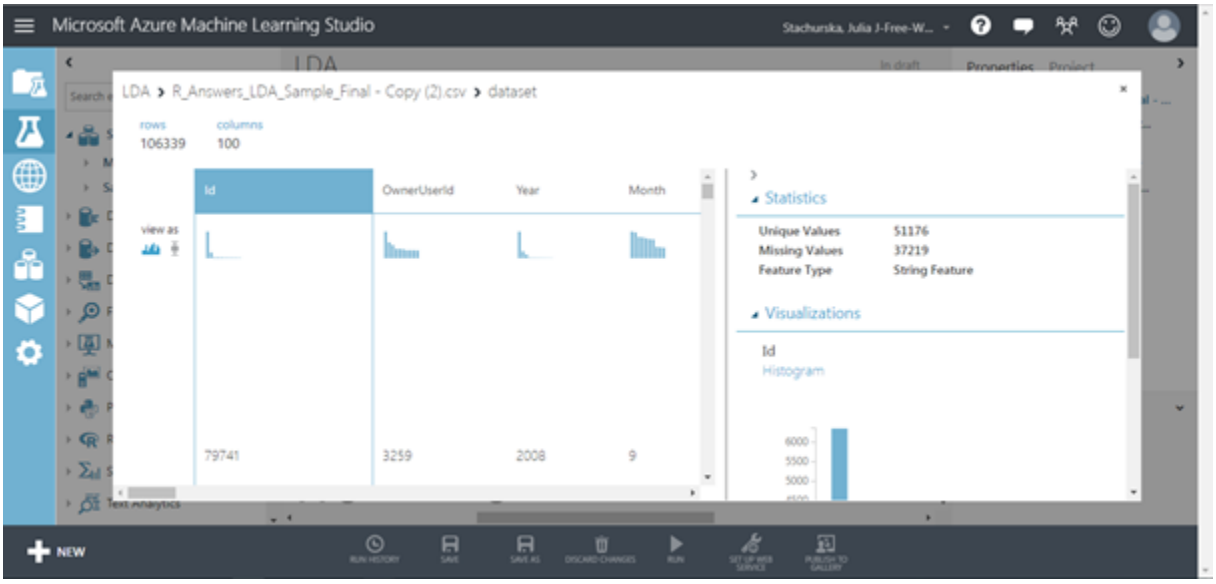
Properties:

- This is a new version of an existing dataset: Unselected
- Enter a name for the new dataset: R_Answers_LDA_Sample
- Select a type for the new dataset: Generic CSV file with a header(.csv)
- Provide an optional description: R Answers RDA.



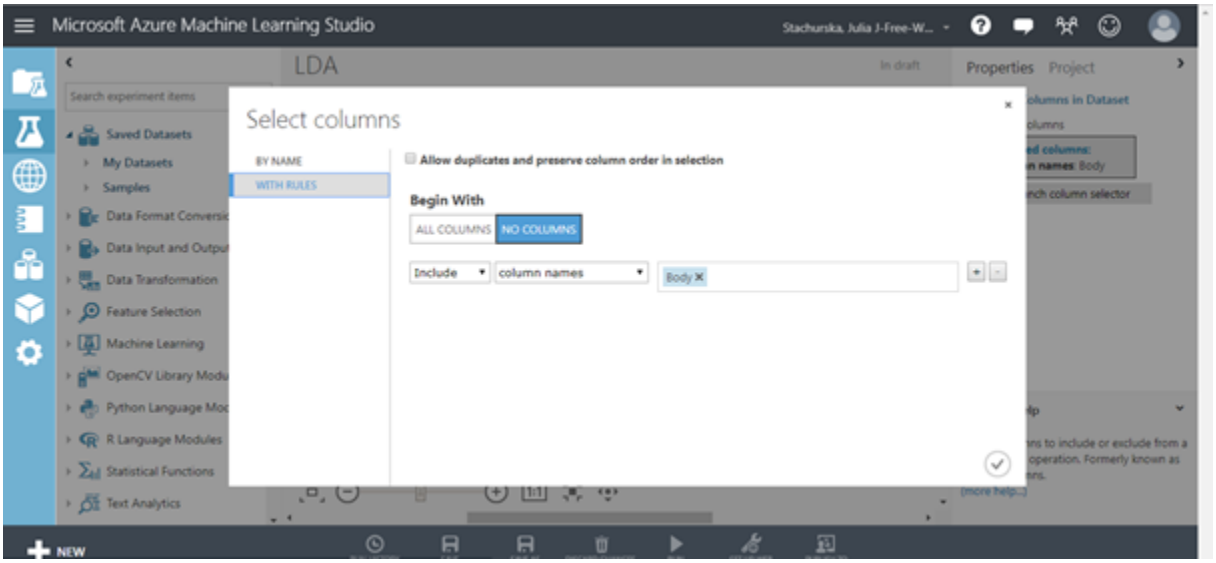
Step 2: Visualize the Dataset in Azure ML

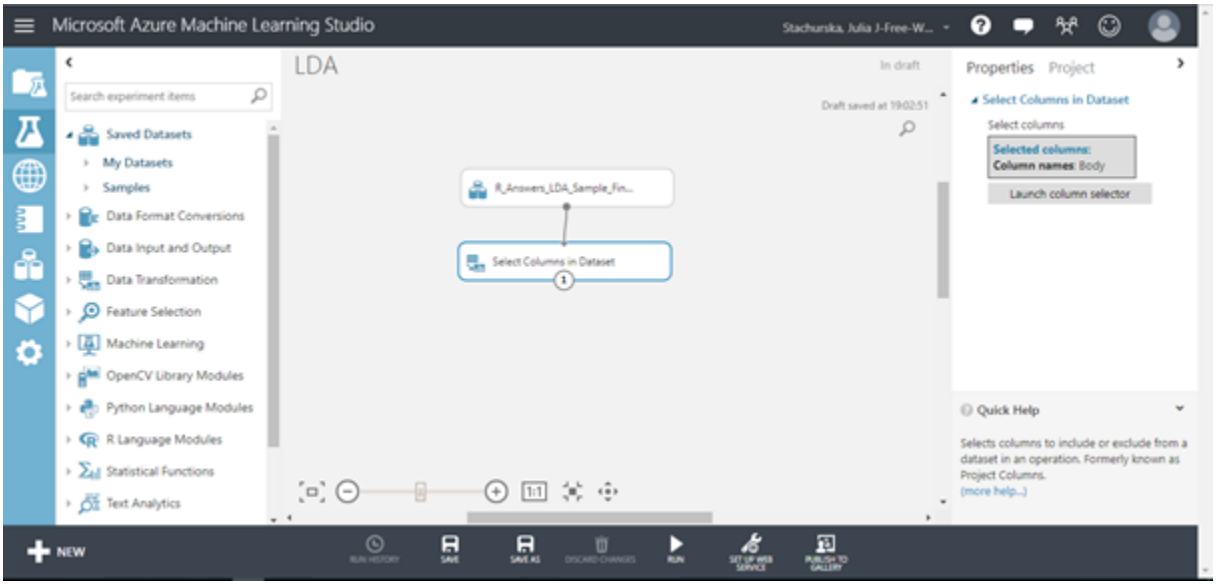
This step is to verify if the data set uploaded contains all the data from the source file.



Step 3: Select columns in Data Set

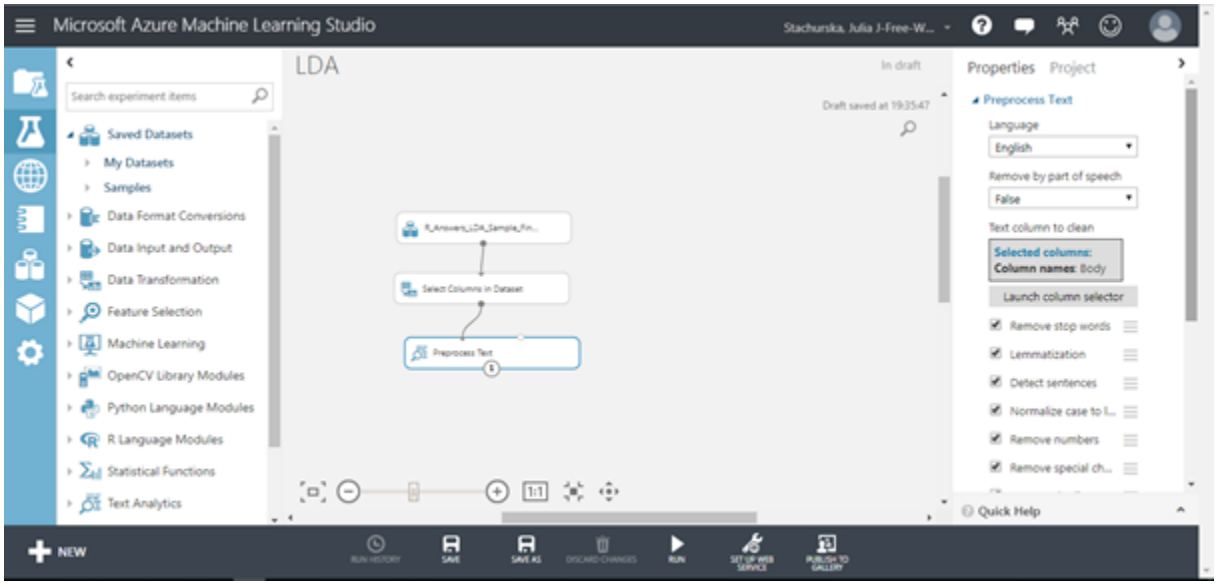
This is a common user interface element in Azure ML modules to enable selecting the columns you want to use in the module, in this case column “Body”. In the Select columns dialog box, select option With Rules to begin with no columns, and include “Body” column name.





Step 4: Preprocess text

This step is to prepare data for analysis.

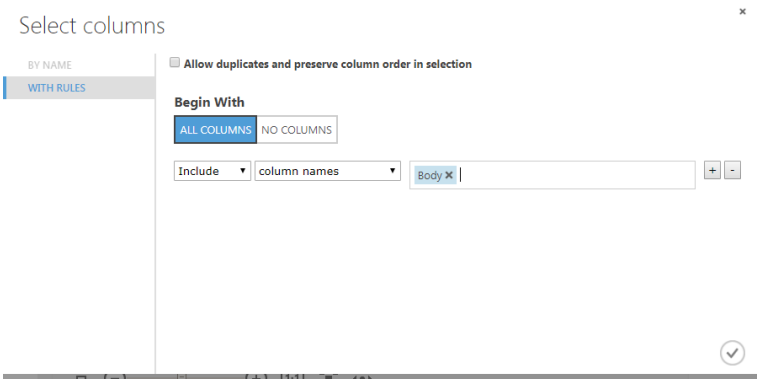


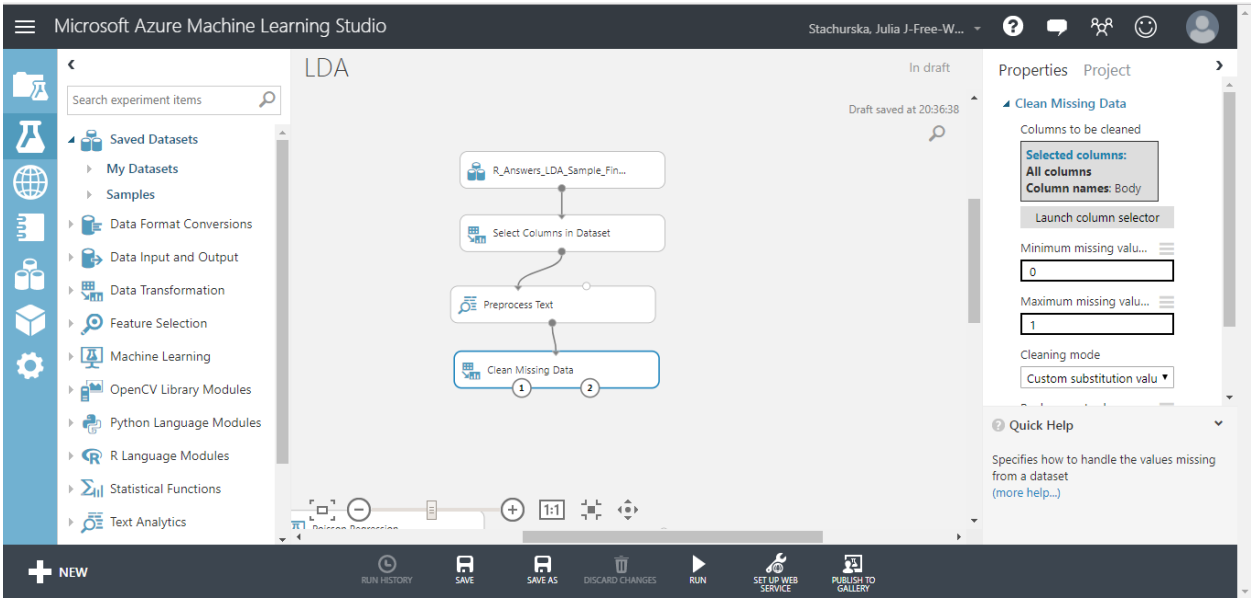
Properties:

- Remove stop words: checked
- Lemmatization: checked
- Detect sentences: checked
- Normalize case to lowercase: checked
- Remove numbers: checked
- Remove special characters: checked
- Remove duplicate characters: checked
- Remove email addresses: checked
- Remove URLs: checked
- Expand verb contractions: checked
- Normalize backslashes to slashes: checked
- Split tokens on special characters: checked

Step 5: Clean missing data

This step is to clean missing data in the Data Set, with column “Body” specified in column selector.



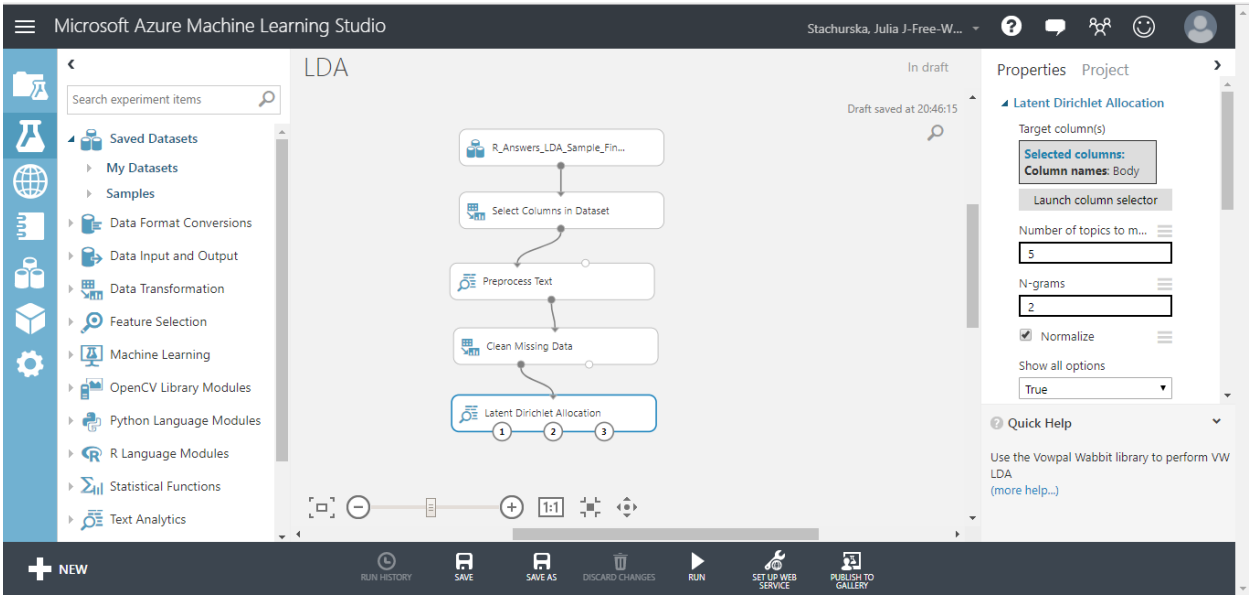


Properties:

- Minimum missing value ratio: 0
- Maximum missing value ratio: 1
- Cleaning mode: Custom substitution value
- Replacement value: „ „
- Generate missing value indicator column: not checked

Step 6: Latent Dirichlet Allocation

This step is to group unclassified text into a number of categories in column “Body” specified in column selector.

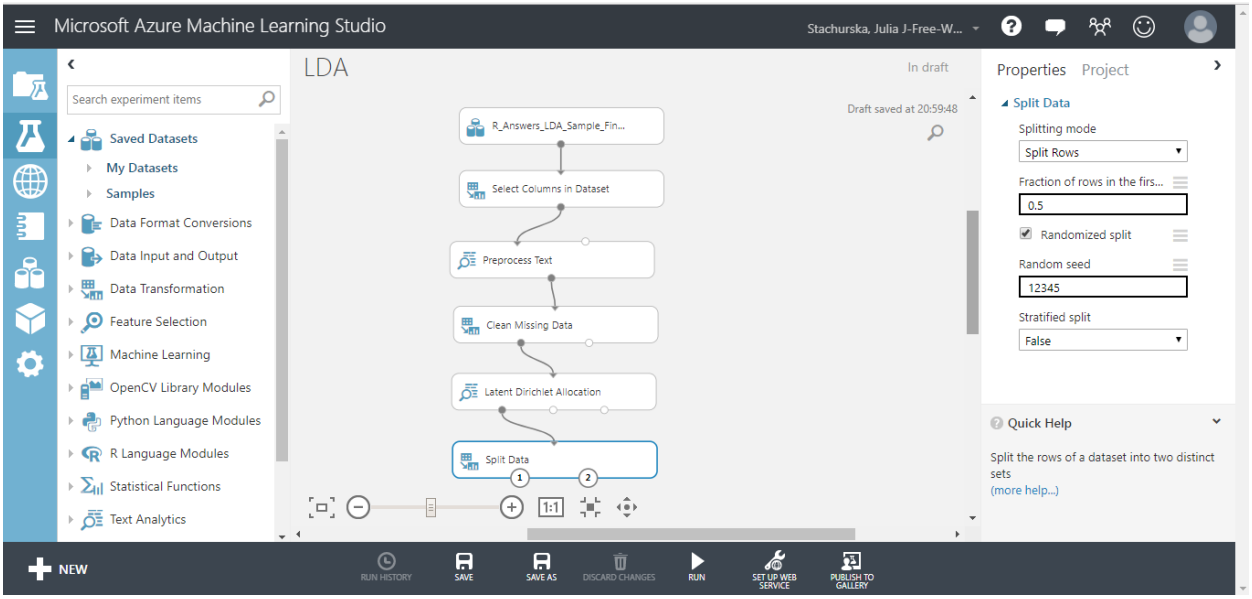


Properties:

- Number of topics to model: 5
- N-grams: 2 (Normalize: checked)
- Show all options: true
- Rho parameter: 0.01
- Alpha parameter: 0.01
- Estimated number of documents: 1000
- Size of the batch: 32
- Initial value of iteration used in learning rate update schedule: 0
- Power applied to the iteration during updates: 0.5
- Number of training iterations: 25
- Build dictionary of ngrams prior to LDA: True
- Maximum number of ngrams in dictionary: 20000

Step 7: Split Data

This step is to split data.

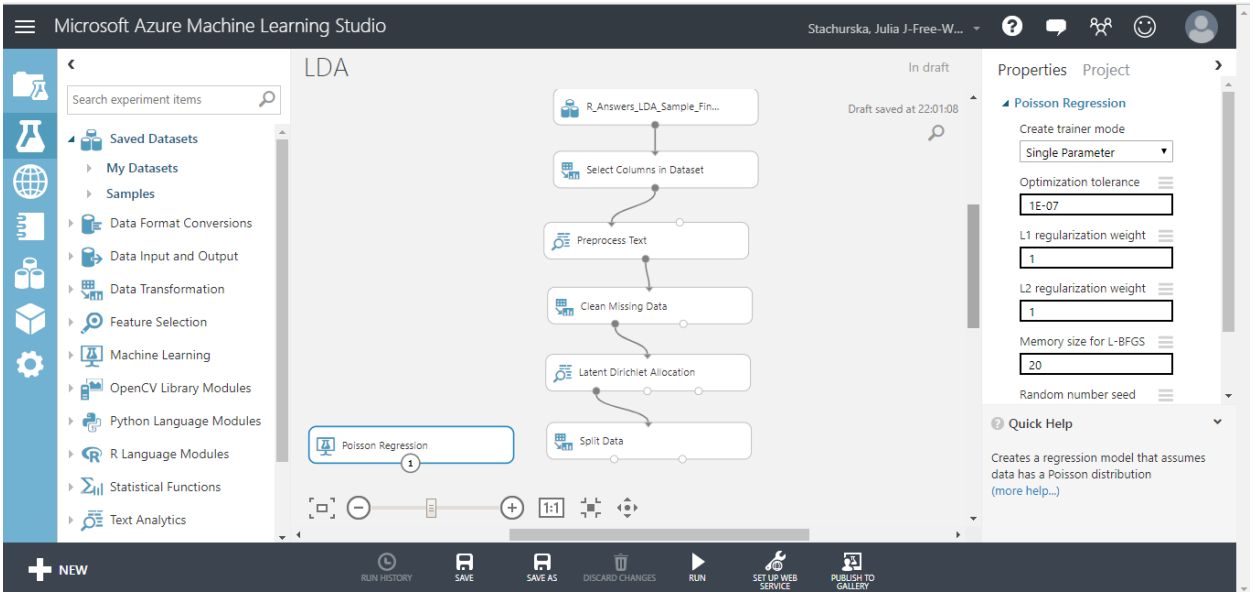


Properties:

- Splitting mode: Split Rows
- Fraction of rows in the first output dataset: 0.5
- Randomized split: checked
- Random speed: 12345
- Stratified split: False

Step 8: Poisson Regression

This step is to create regression model for data with Poisson distribution.

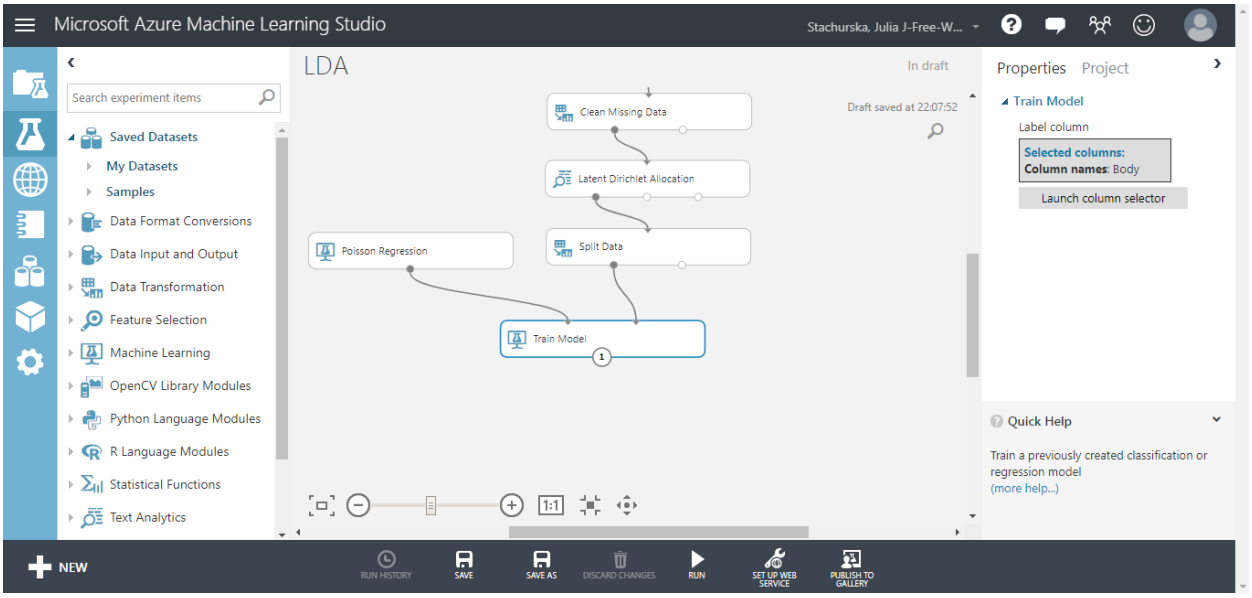


Properties:

- Create trainer mode: Single Parameters
- Optimization tolerance: 1E-07
- L1 regularization weight: 1
- L2 regularization weight: 1
- Memory size for L-BFGS: 20
- Random number seed: „ „
- Allow unknown categorical levels: checked

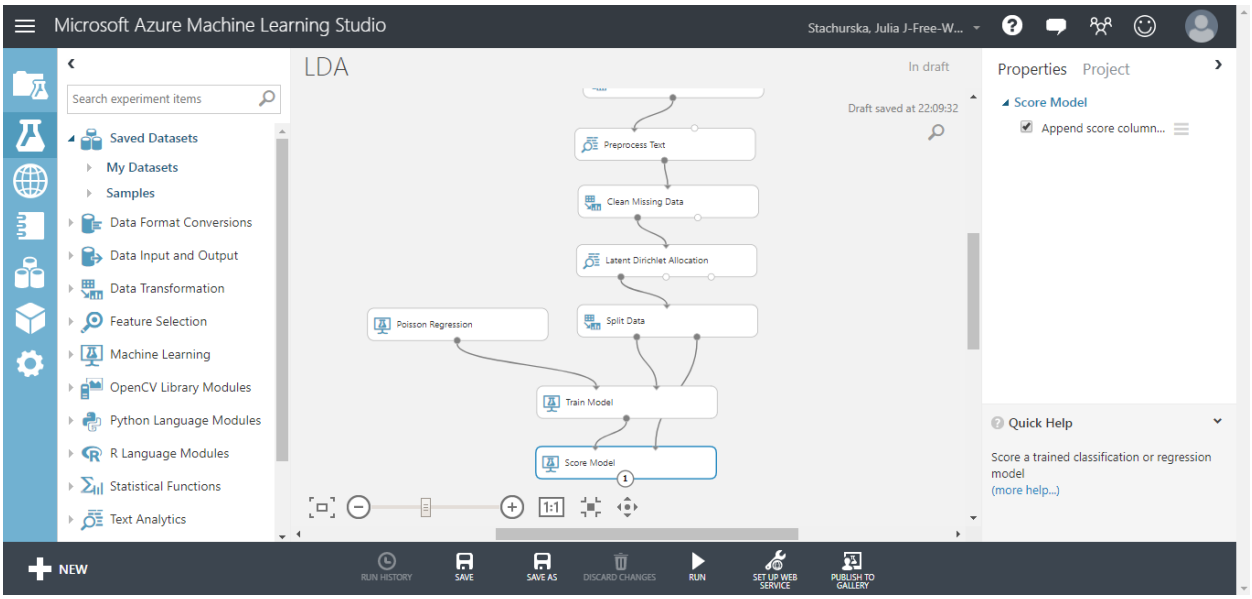
Step 9: Train model

This step is to train model on “Body” column, specified in Column Selector.



Step 10: Score

This step is to score trained model.

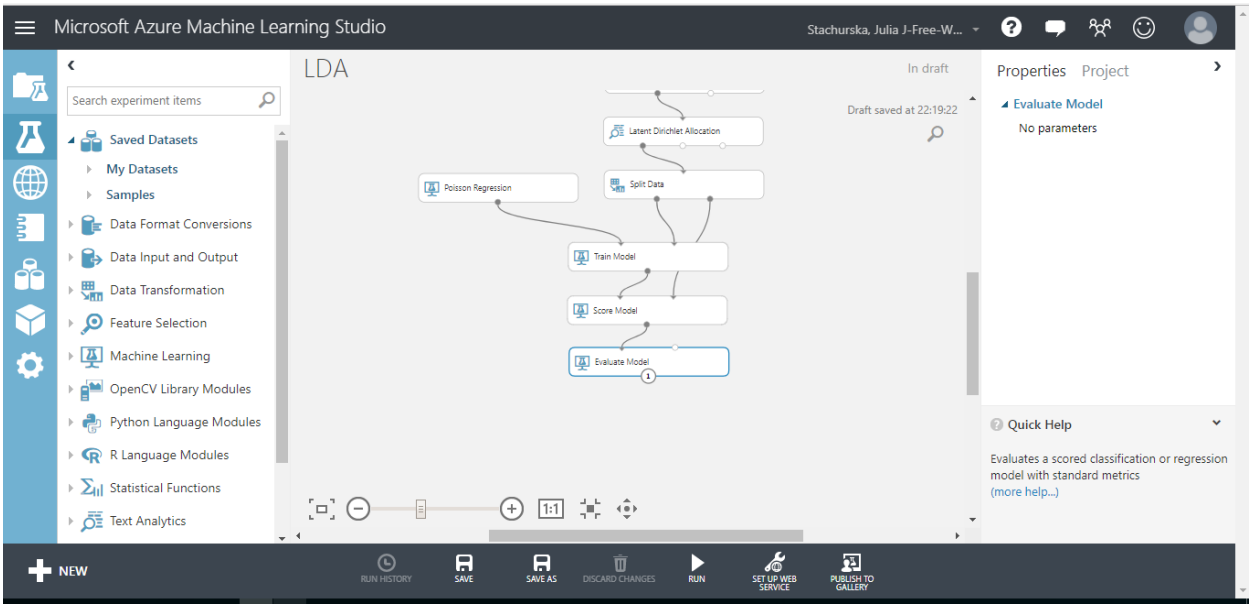


Properties:

- Append score columns to output: checked

Step 11: Evaluate model

This step is to evaluate a scored model with standard metrics.



References

- https://www.kaggle.com/stackoverflow/rquestions 2GB
- https://gallery.cortanaintelligence.com/Experiment/LDA-2