1. What kind of model would you use? What is your process for selection?
   1. After following steps:
      1. Read the Data
      2. Create Dependent and Independent Datasets based on our Dependent and Independent features
      3. Split the Data into Training and Testing sets
   2. And considering following points
      1. Model Selection at different Scale
      2. Model Evaluation
      3. If data is not an issue
      4. Prediction speed vs accuracy
      5. Learning curves
      6. K-fold Cross validation
   3. Model selection: (due to time constraint, we did not get to experiment with all of the below models)
      1. Naive Bayes
      2. svm.SVC
      3. DecisionTreeClassifier
      4. RandomForestClassifier
      5. Xgboost
      6. LightGBM
      7. Random Forest
      8. Extra Trees
      9. k-NN
      10. Logistic Regression
      11. Neural Networks
   4. Evaluation and Hyperparameter tuning
   5. making pipeline
   6. Jenkins
2. What other signals would you use? Why?
   1. (what do you mean by a signal anyway?) In this solution, we used existing unique tags without performing any complex feature engineering. Just made sure that we are not mixing upper/lower cases. In a more timely way, one would like to perform a text analysis and cluster keywords in order to group tags which refer to the same context to include most of the data.   
      We also did not use the ‘title’ and the ‘description’. It is also possible to relate tags to the description and the title using text analytics (NLP).
3. What is your process for ensuring data quality?
   1. Rigorous data profiling and control of incoming data
   2. Careful data pipeline design
   3. Accurate gathering of data requirements
   4. Integration of data lineage traceability into the data pipelines
   5. Automated regression testing as part of change management
4. How would you evaluate your model?
   1. **Accuracy**: for whatever data described, it needs to be accurate.
   2. **Relevancy**: the data should meet the requirements for the intended use.
   3. **Completeness**: the data should not have missing values or miss data records.
   4. **Consistency**: the data should have the data format as expected and can be cross reference-able with the same results.
      1. Data format and data patterns
      2. Data consistency on each record
      3. Data value distributions and abnormalies
      4. Completeness of the data
5. When is your model ready for production?
   1. EDA
   2. Feature Selection
   3. Feature Engineering
   4. Train and Test set selection
   5. Model selection
   6. Model Tuning and evaluation
      1. GridsearchCV
      2. RandomizedSearchCV
      3. Classification evaluation
         1. Confusion Matrix
         2. F1-Score
         3. ROC-AUC
         4. …
      4. Regression evaluation
         1. RMSE
         2. MSE
         3. MAE
         4. …
   7. Making needed pipeline
   8. Model Training Options
      1. Batch Trainng
      2. Real-time Training
   9. Model Deployment Options
      1. Batch Model Scoring
      2. Operational Database
   10. Realtime Model Serving
   11. Web Service API
   12. Monitoring Model Performance
   13. Dockerizing
   14. jenkins
6. If you had to prioritize certain petitions over the other what approach would you choose and why?
   1. No more time to answer
7. Can you think of ways for getting better data that would improve your model?
   1. No more time to answer