GitHub Comments Processing

GitHub

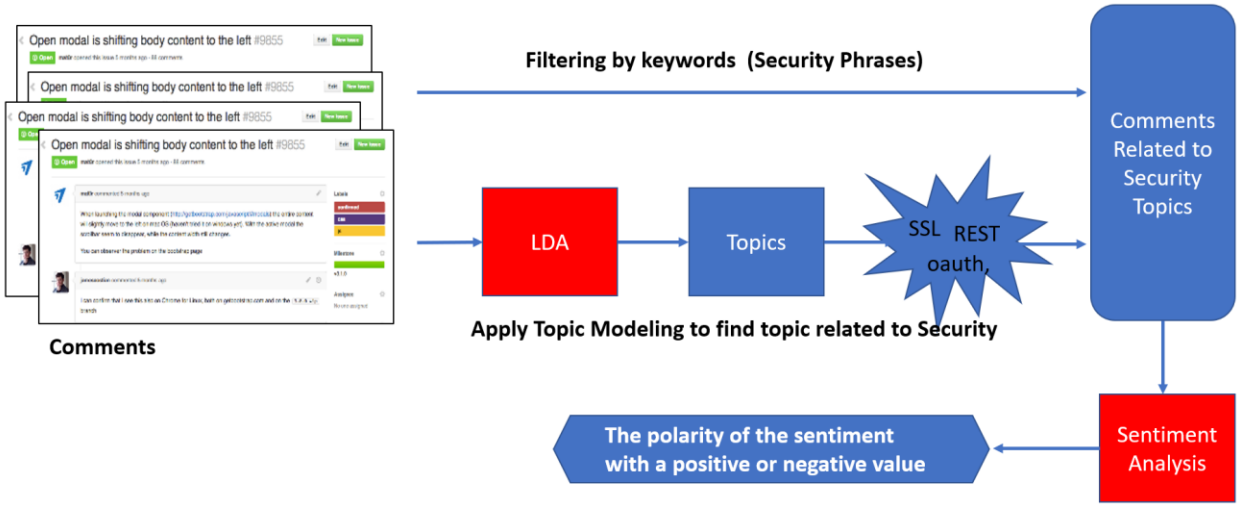
Part 2 (Due March 19, 2020)

Introduction:

In this task, we are looking to analyze the extracted comments in task 1. The initial stage we need to filter the text of any comments/issues/pull requests that contain a security phrase. This task composes of two part

1- Straightforward filtration by filtering those comments if it contains a security phrase from a give key word list. E.g. list of words (virus, trojan, etc.)
2- Apply Topic Modelling - is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents. Based on the return topics from the model, we will consider the security related topics.

Finally, we will apply sentiment analysis to measure the polarity of the comments that related to a security subject.



How to process this task:

- The first step is you need to filter the extracted comments one by using direct filtering text matching technique such as contains () method in Java. And the other one by applying *Topic Modeling* Latent Dirichlet Allocation (LDA). For LDA task you need to clean and prepare the text by remove stop

word punctuation and lower casing. Then train LDA model to get the expected result.
Useful resource to start implementing LDA step by step using Python
https://github.com/kapadias/mediumposts/blob/master/nlp/published_notebooks/Introduction%20to%20Topic%20Modeling.ipynb

- The second step is to apply sentiment analysis on comments that related to security subject. You can use Stanford CoreNLP to be performing sentiment analysis on the comments to return polarity of the text. Useful resource to start implementing Sentiment Analysis
https://stanfordnlp.github.io/CoreNLP/tutorials.html

Please store comments that has been filtered by keyword in a CSV file (e.g. **Filtered_by_words.CSV**) and for comments that has been filtered by LDA in a CSV file (e.g. **Filtered_by_LDA.CSV**). Moreover, the result after sentiment should be store in new file with their sentiment value (**Sentiment_by_words.CSV** and **Sentiment_by_LDA.CSV**)

Please feel free to contact me for any clarification:

Phone: 786 556 4167

Email: Faisalb@umich.edu