Bug Triage with Natural Language Processing

Team: Frank Russo, Carlo Clarke, Ramya Raju

Client: Ning Yang, Anthony Escalona

Student-Faculty Research Day Conference
CSIS, Pace University

Agenda

- Introduction:
  - What is bug triage?
- Literature Review
  - word2vec
  - Long Short-Term Memory
  - Naive Bayes
- Methodology
- Results
  - word2vec/LSTM
  - Naive Bayes
- Conclusion

Introduction: What is bug triage?

- All software written has bugs
- Assigning bugs to the right developer is time consuming
- Machine Learning techniques can be used to assign bugs to developers

Literature Review

- Machine Learning/Natural Language Processing
  - Word2vec
  - Long Short-term Memory
  - Naive Bayes

Literature Review: word2vec

- Algorithms for unsupervised training of vectors of words.
- Model types:
  - Neural Bag-of-Words (NBOW)
  - Recurrent Neural Network (RNN)
  - Recursive Neural Network (RecNN)
  - Convolutional Neural Network (CNN)
- Highly accurate guesses

Literature Review: Long short-term memory

- RNN capable of learning order dependencies in sequences to predict problems.
- Hidden memory cell to update and expose the content
- Two types of LSTM:
  - Unidirectional
  - Bidirectional

Literature Review: Naive Bayes

- Set of classifiers
- Based on Bayes' theorem
- Text categorization
- Types of Naive Bayes Classifier
  - Multinomial Naive Bayes
  - Bernoulli Naive Bayes
  - Gaussian Naive Bayes

Methodology: Data

- 2 sets of Data:
  - Training Data
  - Test Data

Methodology: Cleanse Test Data


```
{
    "id" : 1,
    "issue_id" : 2,
    "issue_title" : "Testing if chromium id works",
    "reported_time" : "2008-08-30 16:00:21",
    "owner" : "",
    "description" : "\nwhat steps will reproduce the
      problem\n1\n2\n3\n\r\nwhat is the expected output
      what do you see instead\n\r\n\r\nplease use labels
      and text to provide additional information\n \n"
}
```

Methodology: Execute word2vec Method

- Continuous Bag of Words model
- Extract the vocabulary

Methodology: Cleanse Training Data


```
{
    "owner" : "amit@chromium.org",
    "issue_title" : "Scrolling with some scroll mice (touchpad, etc.)
      scrolls down but not up",
    "description" : "\nProduct Version      : <see about:
      version>\r\nURLs (if applicable) :0.2. 149.27\r\nOther browsers
      tested: Firefox / IE\r\nAdd OK or FAIL after other browsers
      where you have tested this issue:\nSafari 3:\n Firefox 3: OK\r\n
      IE7:OK\r\n\r\nWhat steps will reproduce the problem?\n1. Open
      any webpage on compaq 6715s running vista.\r\n2. Try scrolling
      with the touchpad\r\n3. Scrolling down will work , but up will
      not.\r\n\r\nWhat is the expected result?\nThe page to scroll
      up.\r\n\r\nWhat happens instead?\nThe page doesn't
      move.\r\n\r\nPlease provide any additional information below.
      Attach a screenshot if \r\npossible.\r\nOnly a minor bug .\n "
}
```

Methodology: Cross Validation

- Triaged bug dataset is divided into parts. eg: 3
- Iteration 1:
  - Train data: sub1
  - Test data: sub1
- Iteration 2:
  - Train data: sub1 + sub2
  - Test data: sub2
- Iteration 3:
  - Train data: sub1 + sub2 + sub3
  - Test data: sub3

Methodology: Deep Learning Model

- Looks at word sequence, forward and backward
- LSTM
- Solves Vanishing Gradient Problem

Methodology: Naive Bayes Model

- Multinomial Naive Bayes classifier type, eg word counts for text classification
- Create a classifier model
- OneVsRestClassifier strategy
- Probability estimates using frequency-based bag-of-words model
- Real data to test accuracy

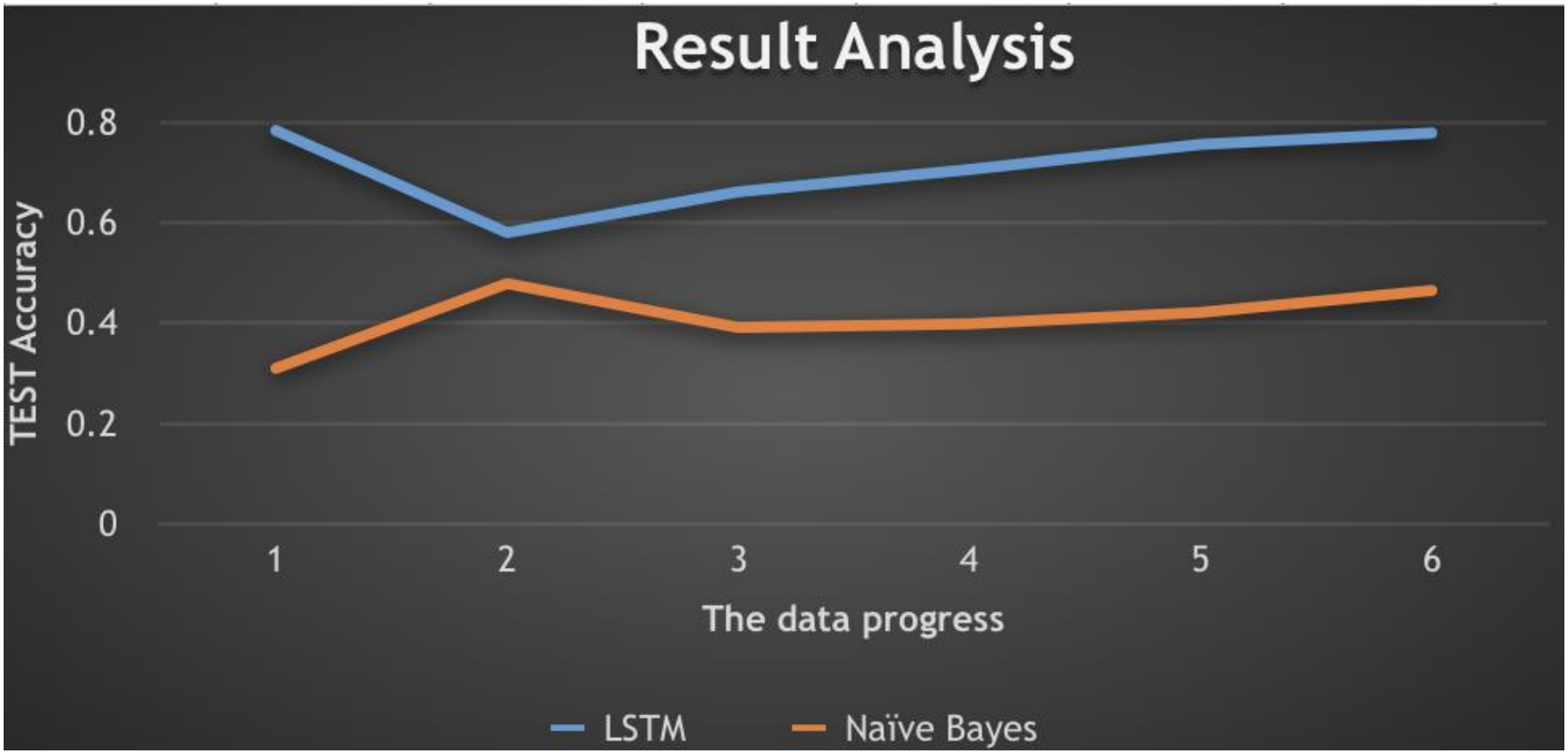
Results: word2vec with LSTM

Iteration	Epoch Accuracy	Test Accuracy
1	Epoch 1/6: 0.0588 Epoch 2/6: 0.1143 Epoch 3/6: 0.1571 Epoch 4/6: 0.2047 Epoch 5/6: 0.2518 Epoch 6/6: 0.3044	0.784385
2	Epoch 1/6: 0.0609 Epoch 2/6: 0.1031 Epoch 3/6: 0.1279 Epoch 4/6: 0.1514 Epoch 5/6: 0.1750 Epoch 6/6: 0.1960	0.580757
3	Epoch 1/6: 0.0566 Epoch 2/6: 0.0883 Epoch 3/6: 0.1047 Epoch 4/6: 0.1198 Epoch 5/6: 0.1299 Epoch 6/6: 0.1370	0.662245
4	Epoch 1/6: 0.0547 Epoch 2/6: 0.0827 Epoch 3/6: 0.0934 Epoch 4/6: 0.1026 Epoch 5/6: 0.1060 Epoch 6/6: 0.1073	0.706721
5	Epoch 1/6: 0.0514 Epoch 2/6: 0.0766 Epoch 3/6: 0.0848 Epoch 4/6: 0.0896 Epoch 5/6: 0.0926 Epoch 6/6: 0.0920	0.756831
6	Epoch 1/6: 0.0525 Epoch 2/6: 0.0759 Epoch 3/6: 0.0833 Epoch 4/6: 0.0878 Epoch 5/6: 0.0890 Epoch 6/6: 0.0920	0.779603

Results: Naive Bayes

Iteration	Accuracy
1	0.3100775
2	0.47913532
3	0.391366
4	0.398827244
5	0.42155314
6	0.4654118

Conclusion:


Result Analysis
0.8
0.6
0.4
0.2
0
TEST Accuracy
1
2
3
4
5
6
The data progress
LSTM
Naïve Bayes

Questions

Please unmute if you'd like to ask any questions.

The End…….