Event Focused URL Extraction from Tweets

By: Chris Bridges, Carter Tat, David Chun
CS 4624: Multimedia, Hypertext, and Information Access
Instructor: Edward A. Fox
Client: Liuqing Li
April 24, 2018
Virginia Tech, Blacksburg VA 24061
Outline

1. Project Goal
2. Overall Design
3. Testing /Evaluation
4. Demo
5. References
6. Acknowledgements
Project Goal

- Link existing Twitter collections and Event Focused Crawler (EFC)
- Classify and rank relevance of URLs in Tweets to collection using deep learning and natural language processing techniques
- Provide client with program that ties it all together
Overall Design
Testing/Evaluation

- 80% Training and 20% Testing
- Classifiers
  - Decision Tree
  - Random Decision Forest
  - Support Vector Classifier (SVC)
  - Gaussian NB
- Cross-Validated using 10 subsamples
## Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>Support Vector (SVC)</th>
<th>GaussianNB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Accuracy</strong></td>
<td>0.970967</td>
<td><strong>0.974193</strong></td>
<td>0.969354</td>
<td>0.790322</td>
</tr>
<tr>
<td><strong>Cross Validation Accuracy</strong></td>
<td>0.94 (+/- 0.06)</td>
<td>0.95 (+/- 0.06)</td>
<td>0.95 (+/- 0.06)</td>
<td>0.75 (+/- 0.29)</td>
</tr>
</tbody>
</table>
Optimal Parameters

```python
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None,
min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto',
max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False,
n_jobs=1, random_state=None, verbose=0, warm_start=False, class_weight=None)
```

---

0.98064516129

---

0.972580645161

---

0.970967741935

---

0.974193548387

---

Demo
“Future Florida Gators Softball Prodigy Is the Youngest NCAA Commit of All Time”
“Kentucky school shooting: 2 students killed, 18 injured”
References


Acknowledgements

Project Client: Liuqing, Li

Instructor: Edward A. Fox

Global Event and Trend Archive Research (GETAR) is supported by NSF (IIS-1619028 and 1619371)
Questions?