비정형 데이터의 숨어있는 가치 창출을 위한 Text Analytics
송완빈
Application Engineer
MathWorks Korea
Agenda

▪ What is Text Analytics?
▪ Text Analytics Workflow
▪ Simple Demo for Text Analytics Technique
▪ Sentiment Analysis with Text Analytics
▪ Key Takeaways
What is Text Analytics?

- **Text Analytics**
  - The goal of deriving information from text data

- **Text Mining**
  - Older phrase for ‘Text Analytics’

- **NLP (Natural Language Processing)**
  - A method which leverages human language (syntax, semantics, discourse, speech)
Text As Data

- Document Classification
  - Field reports
  - Bug reports

- Sentiment Analysis
  - Survey data
  - Trial notes
  - Social media

- Predictive Maintenance
  - Equipment log notes

+ Numeric Data
Media reported two trees blown down along I-40 in the Old Fort area.

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>dog</th>
<th>run</th>
<th>two</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>doc2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Strings
The better way to work with text

- Manipulate, compare, and store text data efficiently
  
  ```
  >> "image" + (1:3) + ".png"
  1×3 string array
  "image1.png"   "image2.png"   "image3.png"
  ```

- Simplified text manipulation functions
  - Example: Check if a string is contained within another string
    - Previously:  `if ~isempty(strfind(textdata,"Dog"))`
    - Now:        `if contains(textdata,"Dog")`

- Performance improvement
  - Up to 50x faster using `contains` with `string` than `strfind` with `cellstr`
  - Up to 2x memory savings using `string` over `cellstr`
Brief overview on Toolbox with Simple Demo
Text Analytics Toolbox

- **Sources of Text Data**
  - Maintenance Logs
  - News/Social Media
  - Customer Surveys
  - Field Reports
  - Research Papers

- **Applications**
  - **Sentiment Analysis:** Determine if news about a product is positive/negative
  - Maintenance: Identify hidden groups of issues in maintenance logs
  - Document Classification: Tag unread documents (e.g. for triaging, routing, etc.)
What is Sentiment Analysis?

Damn I love Matlab #sorrenotsosorry
Demo: Sentiment Analysis of Tweets

Goal:
- Analyze the sentiment of tweets for several tech companies and compare the sentiment scores to stock prices

Approach
- Access data from Twitter
- Preprocess to clean-up text and deal with domain-specific terms
- Predict sentiment from word embedding
- Compare sentiment scores to prices
Sentiment Analysis with Twitter Data

1. **Access Tweets**
2. **Preprocess Tweets**
   - Clean-up Text
   - Convert to Numeric
3. **Develop Model**
4. **Predict Sentiment**

---

Apple's iPhone 8 to Drive 9.1% Increase in Shipments Per IDC
https://t.co/n085F65up k $AAPL $GRMN $GOOG

<table>
<thead>
<tr>
<th>Tweet</th>
<th>apple</th>
<th>iphone</th>
<th>increase</th>
<th>sell</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>tweet2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

apples iphone drive increase shipments per idc

Buy | Increase | Fraud
Demo: Workflow

Tweets

Preprocess text

Test data

Trained model

Score

Word Embedding

Positive + Negative Word List

wordEmbedding with properties:
Dimension: 100
Vocabulary: [1x1193514 string]

Training data

Machine Learning Model

MATLAB EXPO 2018
Sentiment Analysis Demo
Key Takeaways

- **Text data is everywhere, and contains valuable information**

- **Text Analytics Toolbox has tools to help you extract the signal from the noise**

- **Combine text with other data sources to take advantage of all your data**