
MATLAB EXPO 2018
KOREA

MATLAB EXPO 2018

비정형 데이터의 숨어있는 가치 창출을 위한 Text Analytics

송완빈
Application Engineer
MathWorks Korea

Agenda

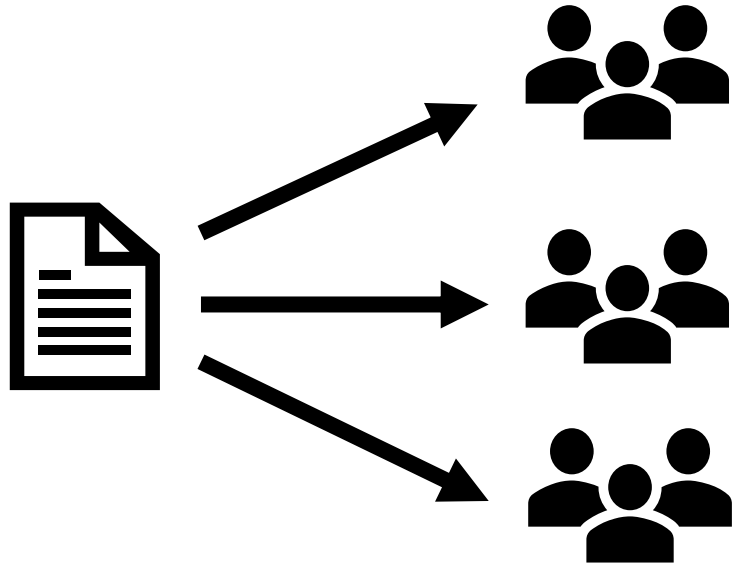
- What is Text Analytics?
- Text Analytics Workflow
- Simple Demo for Text Analytics Technique
- Sentiment Analysis with Text Analytics
- Key Takeaways

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

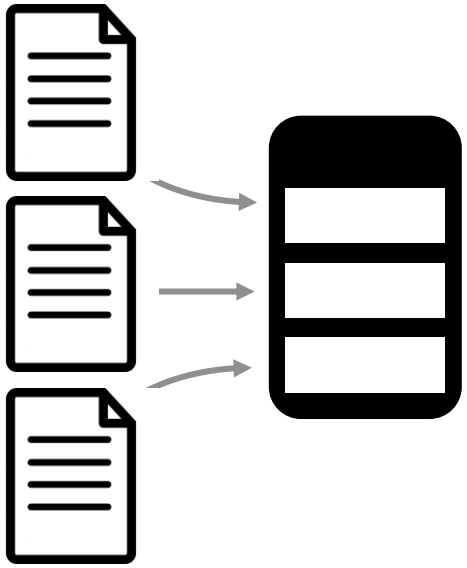


Text Analytics Toolbox

R2017b

Access and Explore Data → **Preprocess Data** → **Develop Predictive Models**

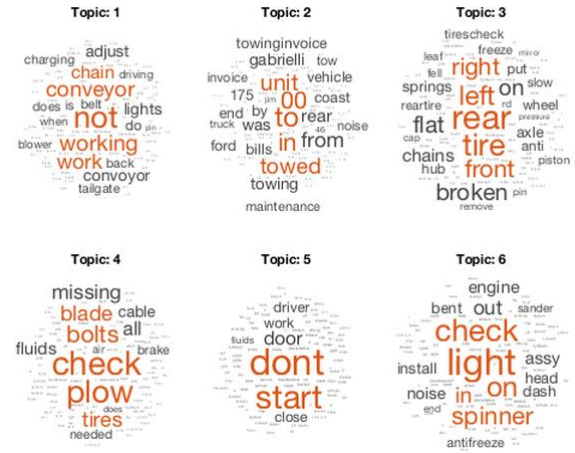
Clean-up Text → **Convert to Numeric**



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

media report two tree blown down i40 old fort area

	cat	dog	run	two
doc1	1	0	1	0
doc2	1	1	0	1



- Word Docs
- PDF's
- Text Files
- HTML Files

- Stop Words
- Stemming
- Tokenization

- Bag of Words
- TF-IDF
- Word Counting

- Word-Embedding
- Machine Learning
  - LDA
  - LSA
- Deep Learning
  - LSTM

Strings

The better way to work with text

R2016b

- 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")`

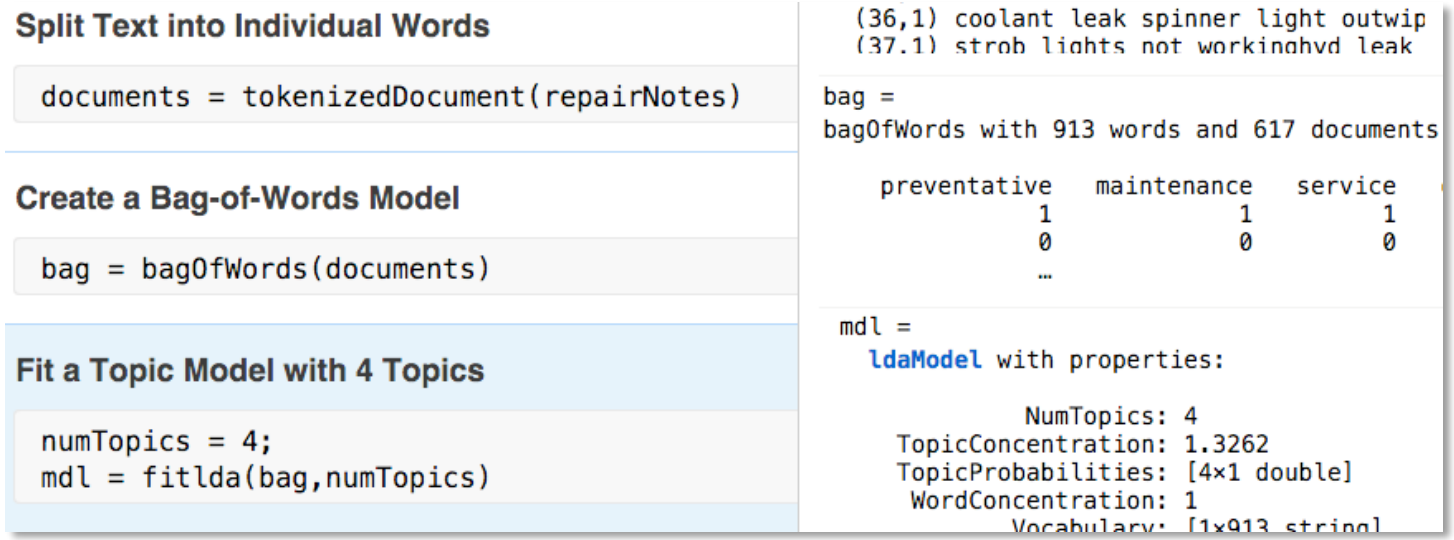
- Performance improvement
  - Up to 50x faster using **`contains`** with **`string`** than **`strfind`** with **`cellstr`**
  - Up to 2x memory savings using **`string`** over **`cellstr`**

Brief overview on Toolbox with Simple Demo

Text Analytics Toolbox

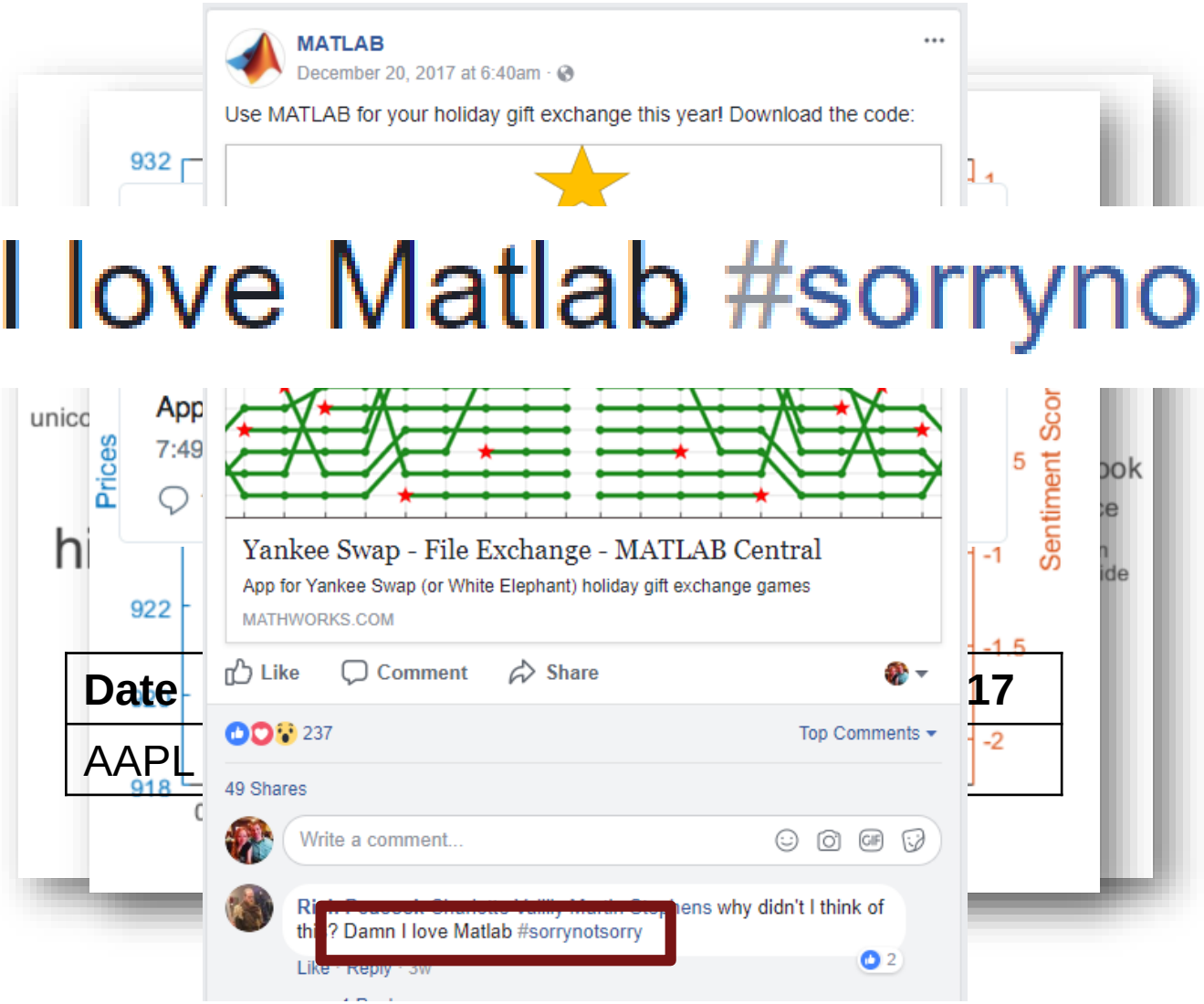
R2017b

- 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 (eg. for triaging, routing, etc.)

What is Sentiment Analysis?



Demo: Sentiment Analysis of Tweets

Goal:

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

```
ans = 508×1 string array
    "Walmart: "you wanna destroy Amazon?" Google: "bet" $WMT $GOOG
    "$WMT wants next level customer service w/highly personalized
    "Ironic prelude to $DIS buying $TWTR soon IMO $AAPL $GOOG $SPY
    "$AMZN the $WMT threat grows each and every day  https://t.co/
    "MU Investments Co. Ltd. Sells 30 Shares of Alphabet Inc. $GOO
    "Ad $ are going to $GOOG and $FB away from wppgy  #Advertising
    "Big bullish unusual option activity detected: $SPX, $GOOG, $C
    "REPORT: Apple to build data center in Iowa: https://t.co/jwH6
    "RT @theflynews: REPORT: Apple to build data center in Iowa: h
```

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

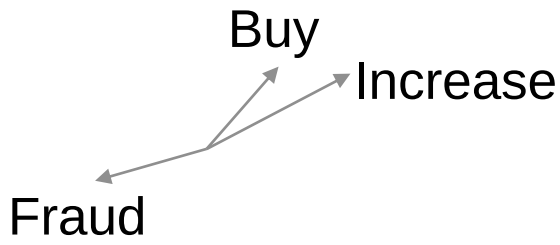
Access Tweets → **Preprocess Tweets** → **Develop Model** → **Predict Sentiment**

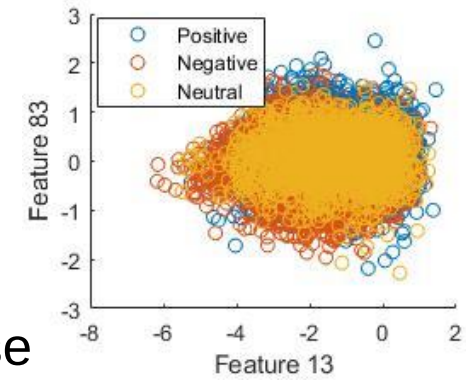
Preprocess Tweets: **Clean-up Text** → **Convert to Numeric**

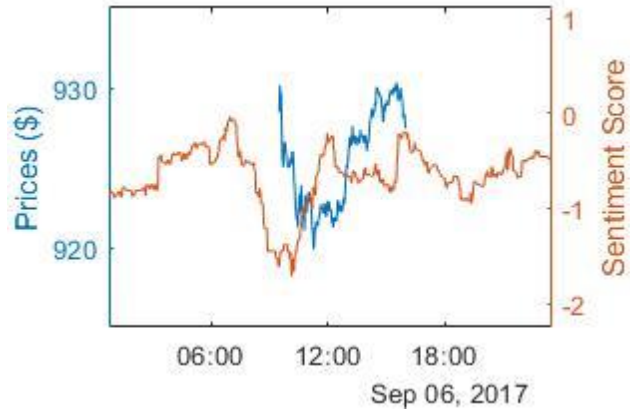
Apple's iPhone 8 to Drive 9.1% Increase in Shipments Per IDC https://t.co/n085F65upk $AAPL $GRMN $GOOG

apples iphone drive increase shipments per idc

	appl e	ipho ne	incr ease	sell
tweet1	1	1	1	0
tweet2	1	0	0	1







Demo: Workflow

Tweets

```
ans = 508×1 string array
    "Walmart: "you wanna destroy Amazon?" Google: "bet" $WMT $GOOG
    "$WMT wants next level customer service w/highly personalized
    "Ironic prelude to $DIS buying $TWTR soon IMO $AAPL $GOOG $SPY
    "$AMZN the $WMT threat grows each and every day  https://t.co/
    "MU Investments Co. Ltd. Sells 30 Shares of Alphabet Inc. $GOO
    "Ad $ are going to $GOOG and $FB away from wppgy  #Advertising
    "Big bullish unusual option activity detected: $SPX, $GOOG, $C
    "REPORT: Apple to build data center in Iowa: https://t.co/jwH6
    "RT @theflynews: REPORT: Apple to build data center in Iowa: H
```

Preprocess text

Word Embedding

```
wordEmbedding with properties:

    Dimension: 100
   Vocabulary: [1×1193514 string]
```

Positive + Negative Word List

pos: 2006x1 string

	1
1	a+
2	abound
3	abounds
4	abundance
5	abundant
6	accessable
7	accessible

neg: 4783x1 string

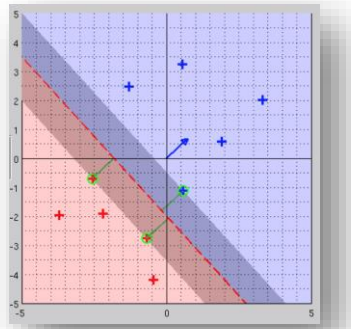
	1
1	2-faced
2	2-faces
3	abnormal
4	abolish
5	abominable
6	abominably
7	abominate

Training data

Test data

Trained model

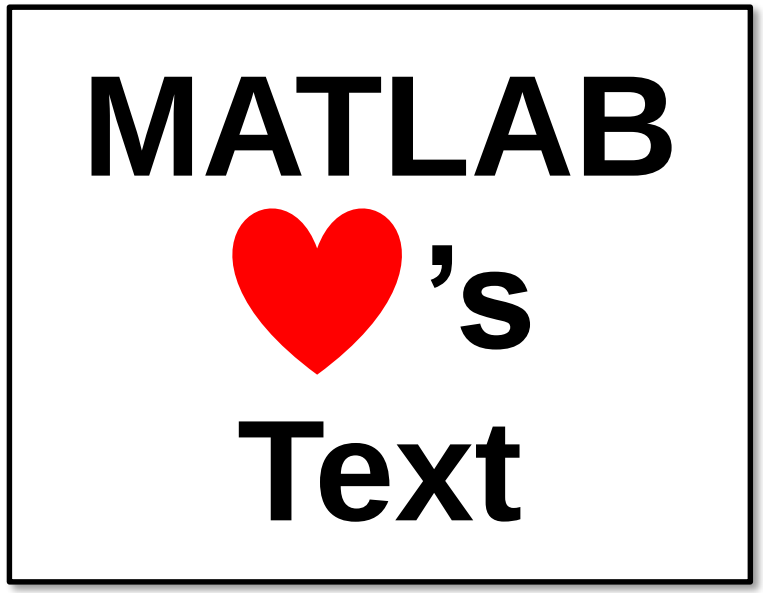
Machine Learning Model



Score

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