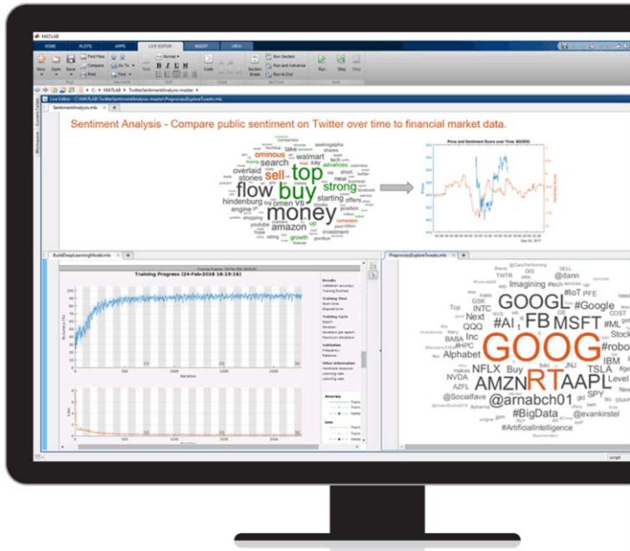
비정형 데이터의 숨어있는 가치 창출을 위한 Text Analytics

A hands-on MATLAB Workshop

June. 3rd, 2020



Sentiment Analysis - Compare public sentiment on Twitter over time to financial market data.

송완빈 과장

Wanbin Song

Application Engineer @ MathWorks

wsong@mathworks.com

For Online Workshop

About this workshop

- What this workshop is about
  - Text Analytics Toolbox overview – For Korean language
  - Hands-on introduction to the tools
    - We will not cover ***all*** the capabilities

- What this workshop is NOT about
  - Expertise in Text Analytics / Natural Language Processing
  - Expertise with Text Analytics Toolbox

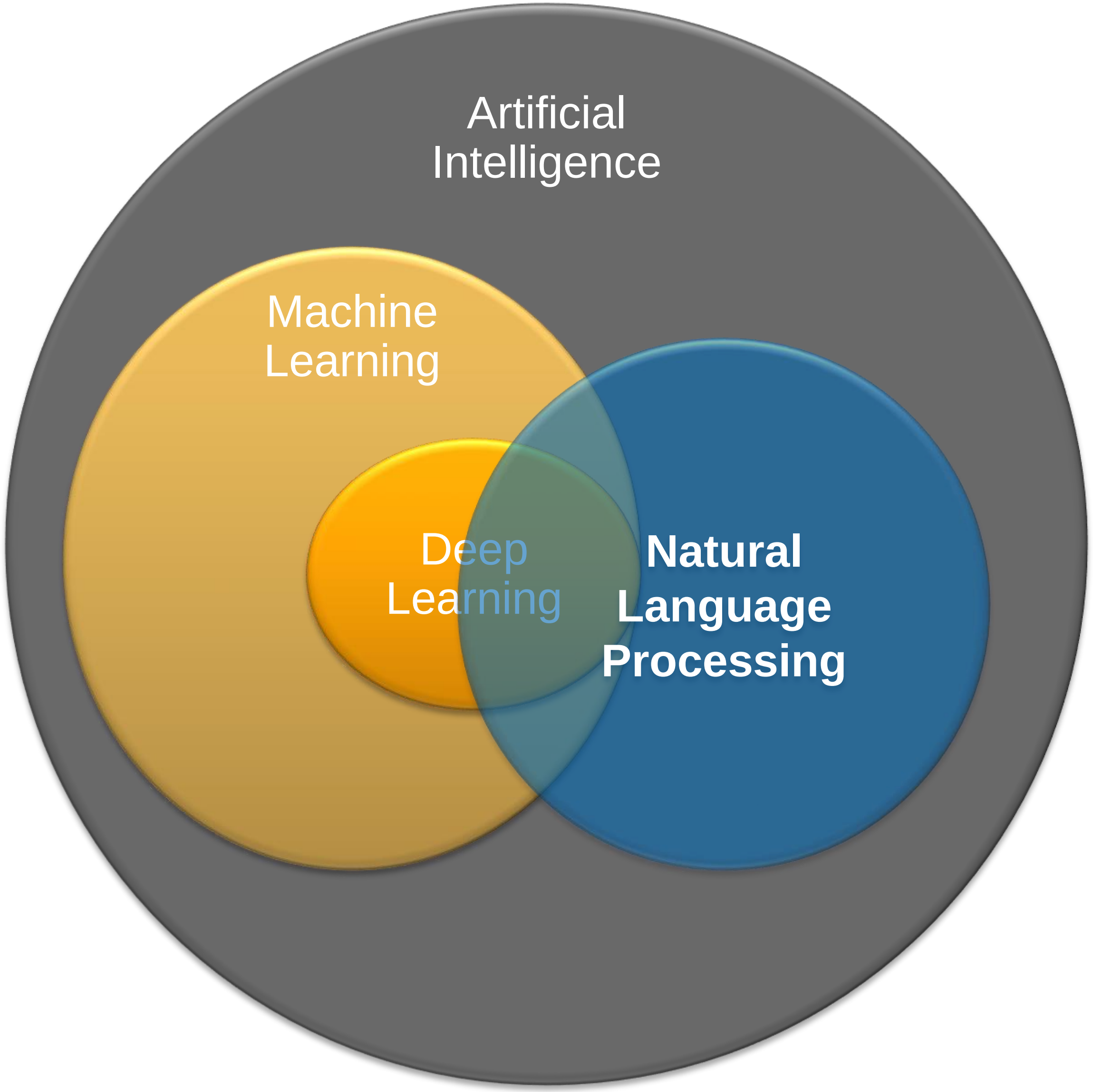
Way of working – Rules of the Road

- Slides to introduce topics
- Examples to demonstrate functionality
- Exercises to *get your hands dirty* and try something beyond what is demonstrated
  - Suggestion: do not modify the functions but save them with different names so that you can trace your steps back
  - This is intended for learning so…don't look at solutions unless you get behind
  - Each exercise starts anew, so don't worry if you get behind

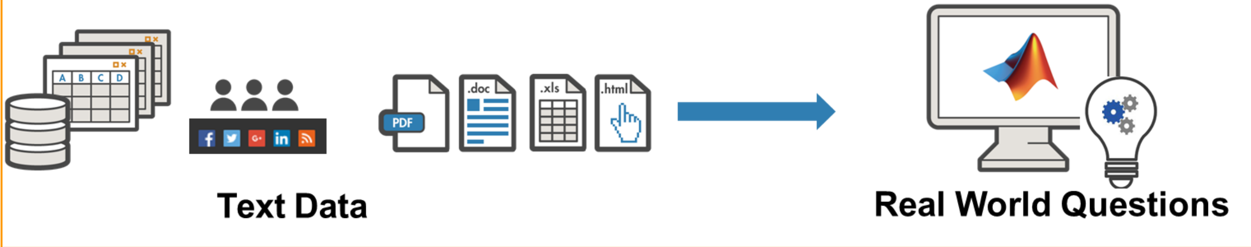
>> Be curious and ask all the questions you like!

Raise Your Hand!(Not the real one)

What is Natural Language Processing? Text Analytics?



Text Analytics

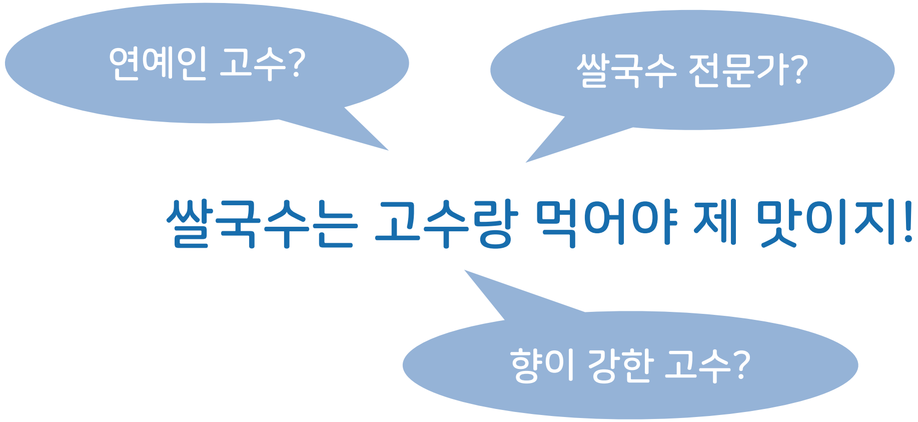


= Natural Language Processing + Machine Learning

What Makes It Difficult?

- Many words in a language, same word different meaning, dialects

- Machines understand logic.. Human beings not so logical



연예인 고수?

쌀국수 전문가?

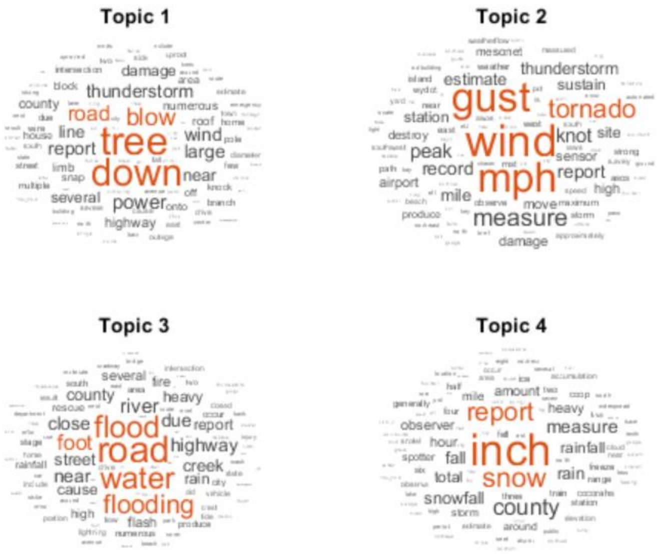
쌀국수는 고수랑 먹어야 제 맛이지!

향이 강한 고수?

- Ambiguity, emotion, subjectivity, personality, culture …

그래 너 잘났다~ 아주 대단하셔요
전문가 납셨어~

What Is Text Analytics Used for?

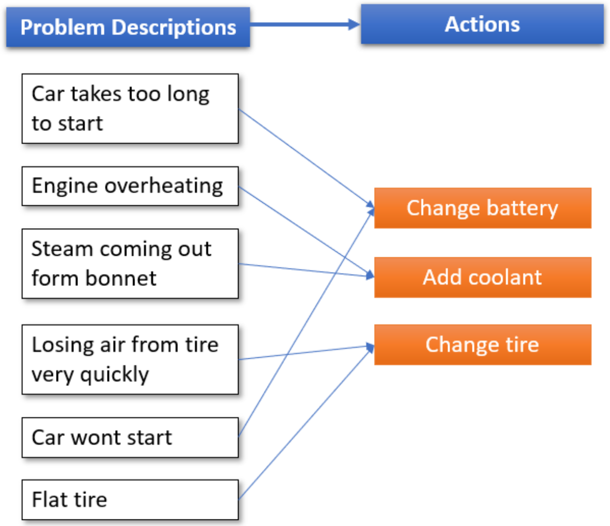


- **Topic Modeling**

  Identify topics from a collection of documents that shows underlying patterns and relationships in raw text data.

- **Text Classification**

  Classify documents into pre-determined categories for efficient information retrieval and prediction.



Problem Descriptions → Actions

Car takes too long to start

Engine overheating

Steam coming out form bonnet

Losing air from tire very quickly

Car wont start

Flat tire

Change battery

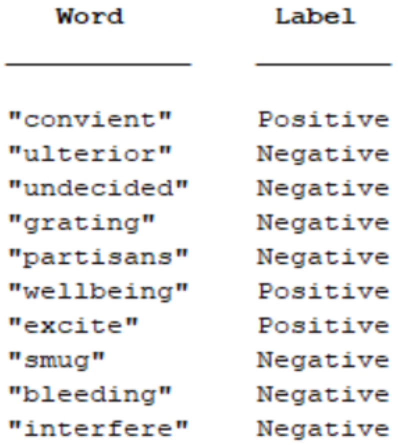
Add coolant

Change tire

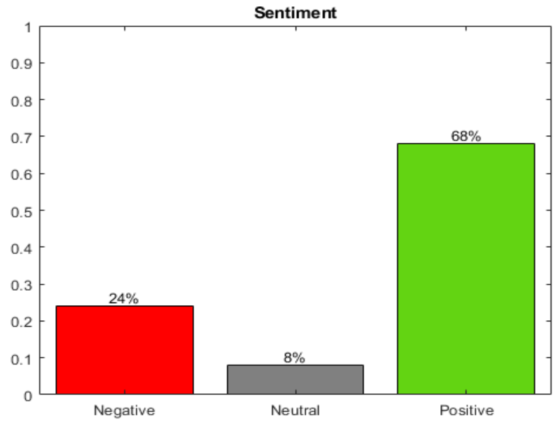
What Is Text Analytics Used for?

- ## Sentiment Analysis

  Identify and score sentiments expressed in text.

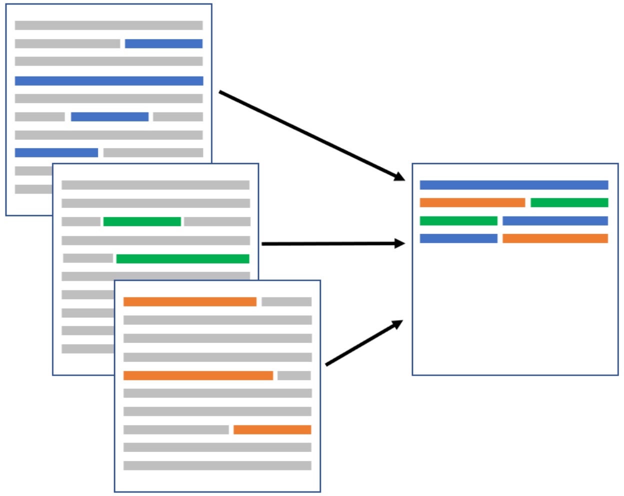


```
       Word             Label
    ___________       ________

    "convient"        Positive
    "ulterior"        Negative
    "undecided"       Negative
    "grating"         Negative
    "partisans"       Negative
    "wellbeing"       Positive
    "excite"          Positive
    "smug"            Negative
    "bleeding"        Negative
    "interfere"       Negative
```

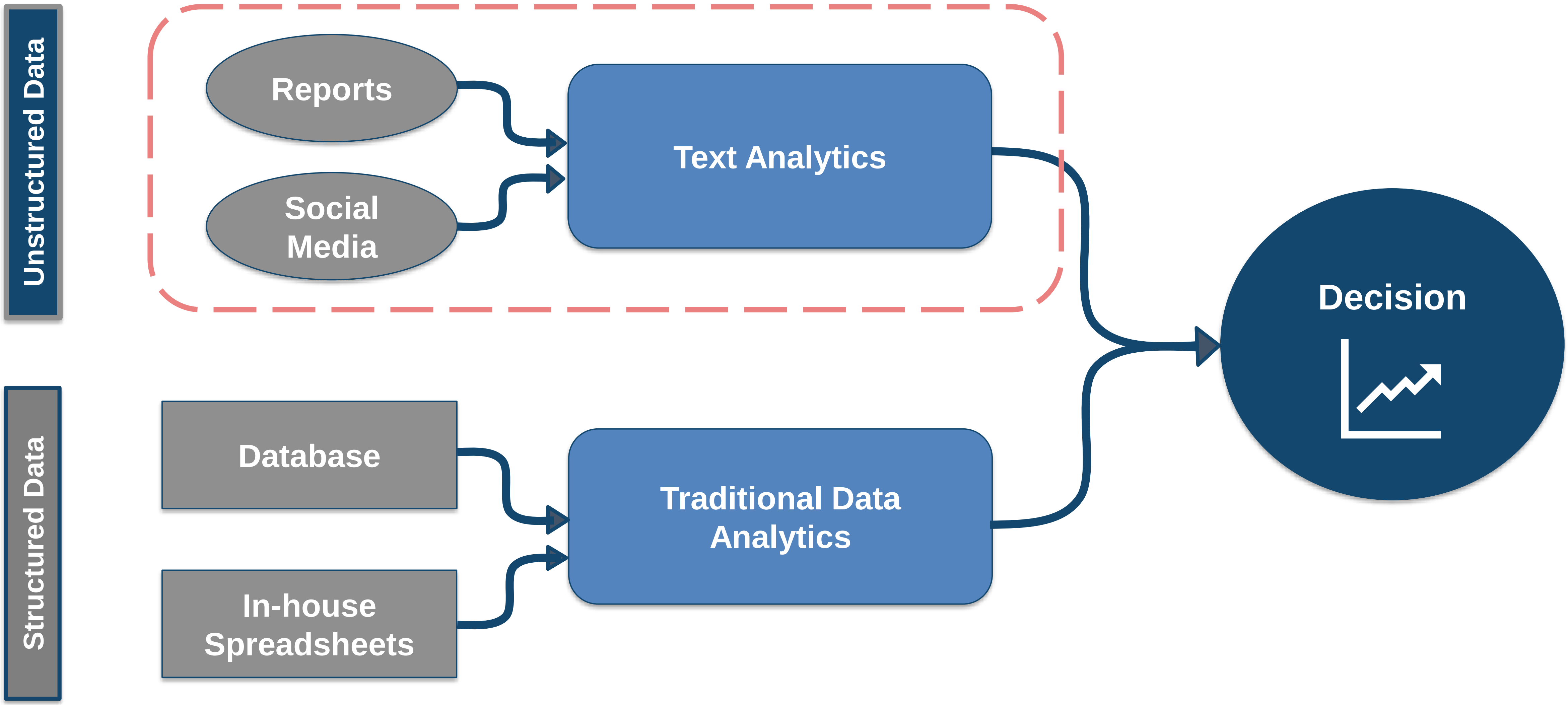


- ## Summarization

  Extract a summary from one or more documents automatically.



Big Picture


Unstructured Data
Reports
Social Media
Text Analytics
Structured Data
Database
In-house Spreadsheets
Traditional Data Analytics
Decision

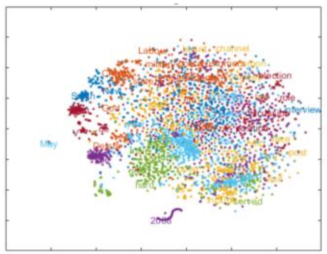
Text Analytics Workflow

1 Access and Explore → 2 Preprocess → 3 Develop Predictive Models → 4 Share Insights and Models

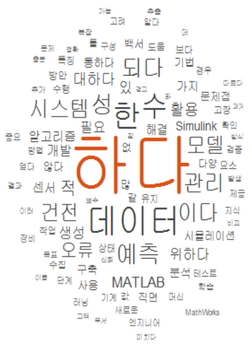
Data Sources

Languages

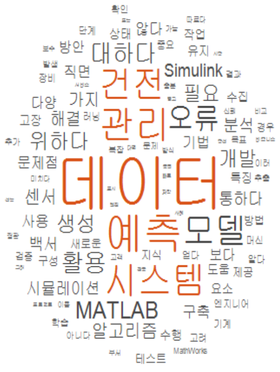
Visualize



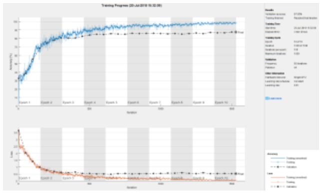
Raw Data



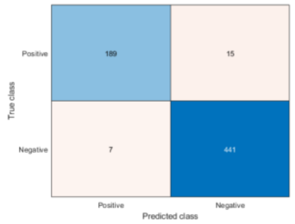
Clean Data



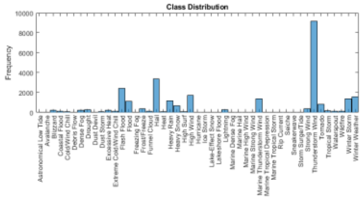
Deep Learning



Machine Learning



Statistics



Report, Publish, Interactive Notebook

Desktop & Web Apps

Production & Web App Servers

Strings

The better way to work with text

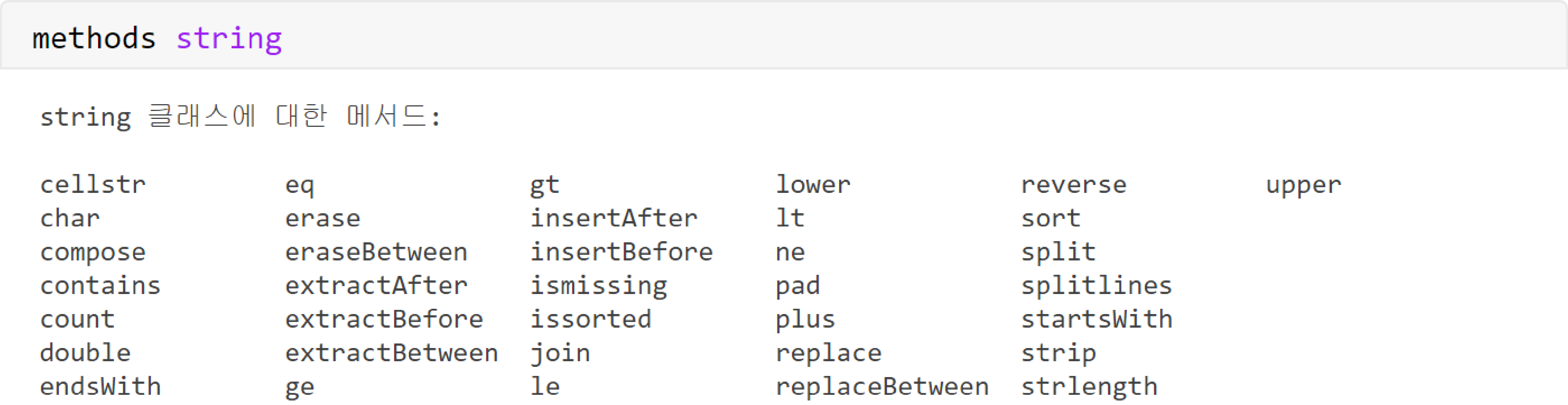
"Hello,world"

String
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



```
methods string
```

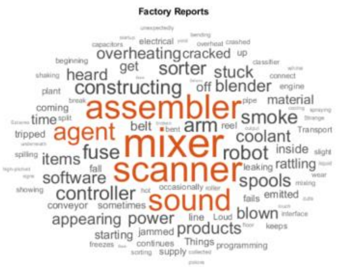
```
string 클래스에 대한 메서드:

cellstr         eq              gt              lower           reverse         upper
char            erase           insertAfter     lt              sort
compose         eraseBetween    insertBefore    ne              split
contains        extractAfter    ismissing       pad             splitlines
count           extractBefore   issorted        plus            startsWith
double          extractBetween  join            replace         strip
endsWith        ge              le              replaceBetween  strlength
```

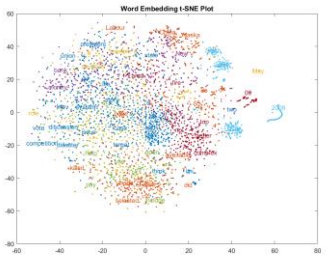
-

1 Access and Explore Text Data

- Import text from databases, social media, news feeds, equipment logs, reports, and surveys

- Visually explore data with word clouds & scatter plots





- Local language support for Japanese(18b), German (19a), and Korean (19b)

云う
主人
君
御
見る
思う
来る
人間
迷亭
吾輩

bad godesberg
godesberg
wurd
wird
sollt
hier
stadt
strass
bonn
viel
schon

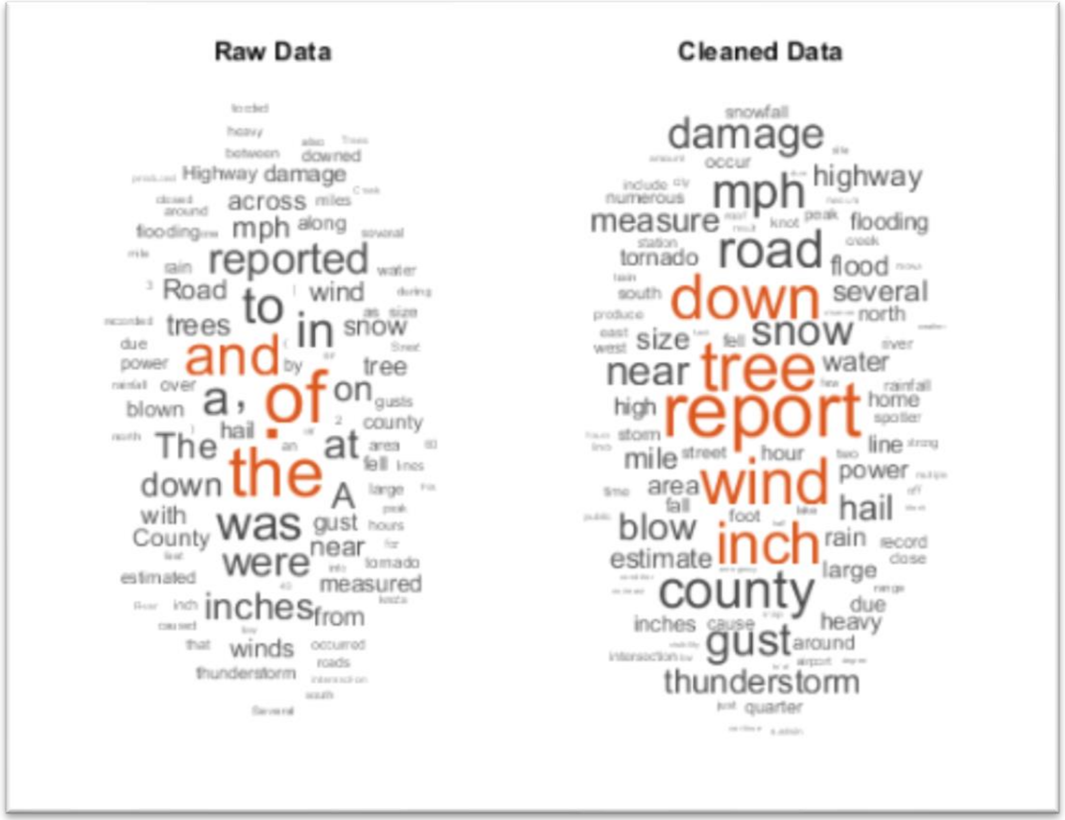
보다
영화
이다
하다
되다
한
정말
너무
좋

2 Preprocess Data

Prepare Text Data for Model Building

Text data may contain:

- Variations in case, for example "new" and "New”
- Variations in word forms, for example "walk" and "walking“
- Words which add noise, for example stop words such as "the" and "of“
- Punctuation and special characters
- HTML and XML tags



3 Build Predictive Models with Text

Convert Text to Numbers

- Bag of Words or N-grams
- Term Frequency-Inverse Document Frequency
- Word Embedding (FastText, Glove, train your own or read in someone else’s)
- Dimensionality Reduction

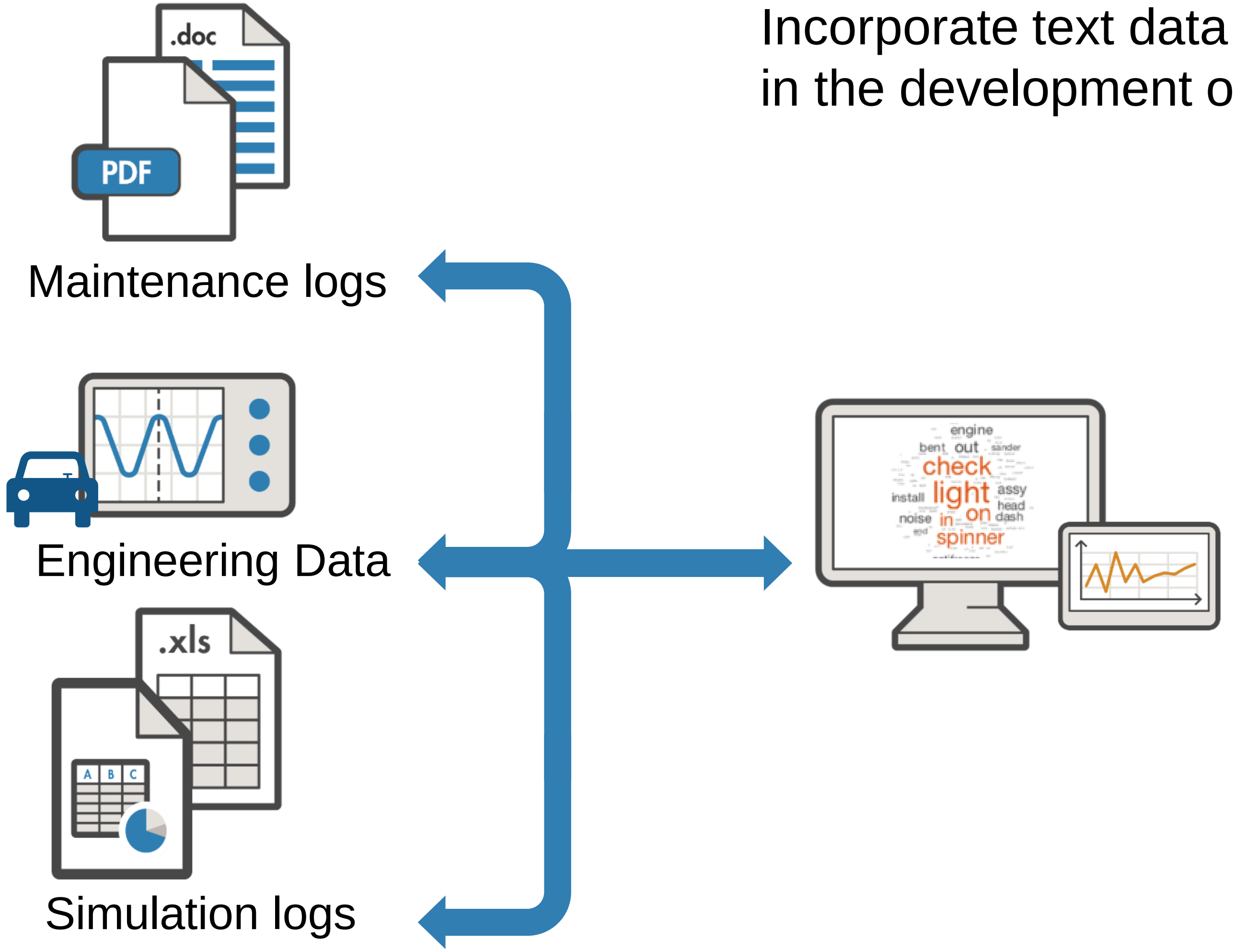
Build Models with Machine & Deep Learning Algorithms

- Choose an algorithm based on application
- State-of-the-art machine and deep learning algorithms available in MATLAB
  - Classification
  - Clustering
  - Descriptive Statistics
  - Regression

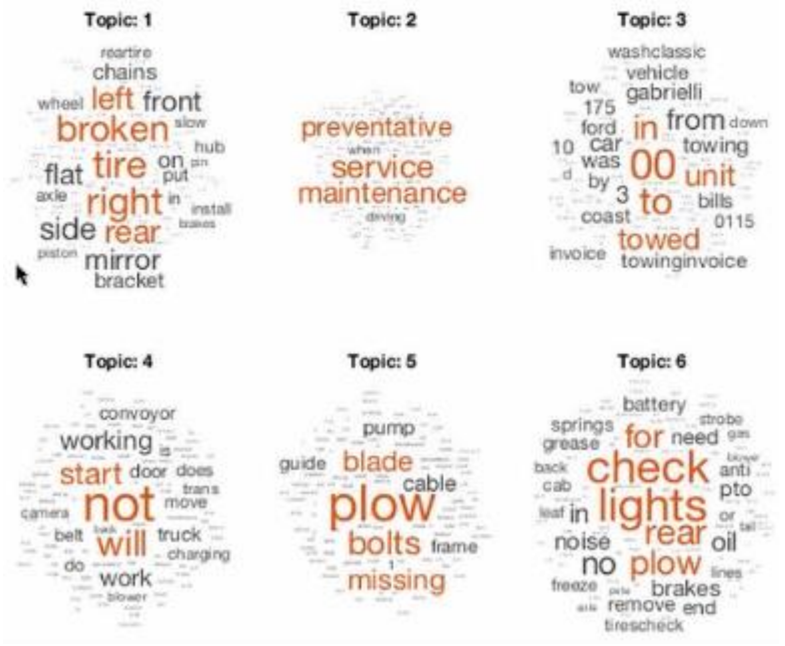
3 **DEMO:** Topic Modeling

Goal : Identify key topics in documentations.

Incorporate text data with other types of engineering data in the development of smart systems.



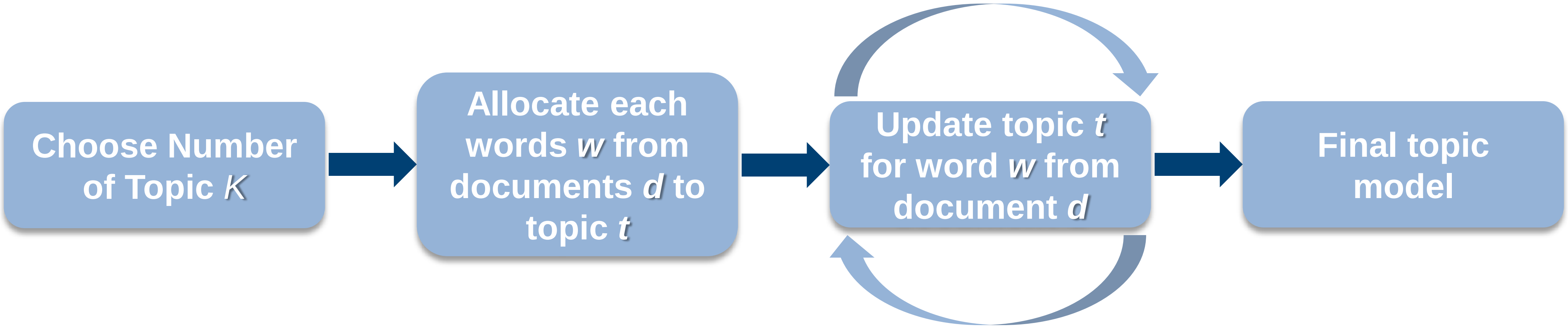
Sample Topics identified:



LDA (Latent Dirichlet Allocation)

- **Topic modeling** is a type of statistical **modeling** for discovering the abstract “**topics**” that occur in a collection of documents.

- Latent Dirichlet Allocation (**LDA**) is an example of **topic model** and is used to classify text in a document to a particular **topic**.



Choose Number of Topic *K*

Allocate each words *w* from documents *d* to topic *t*

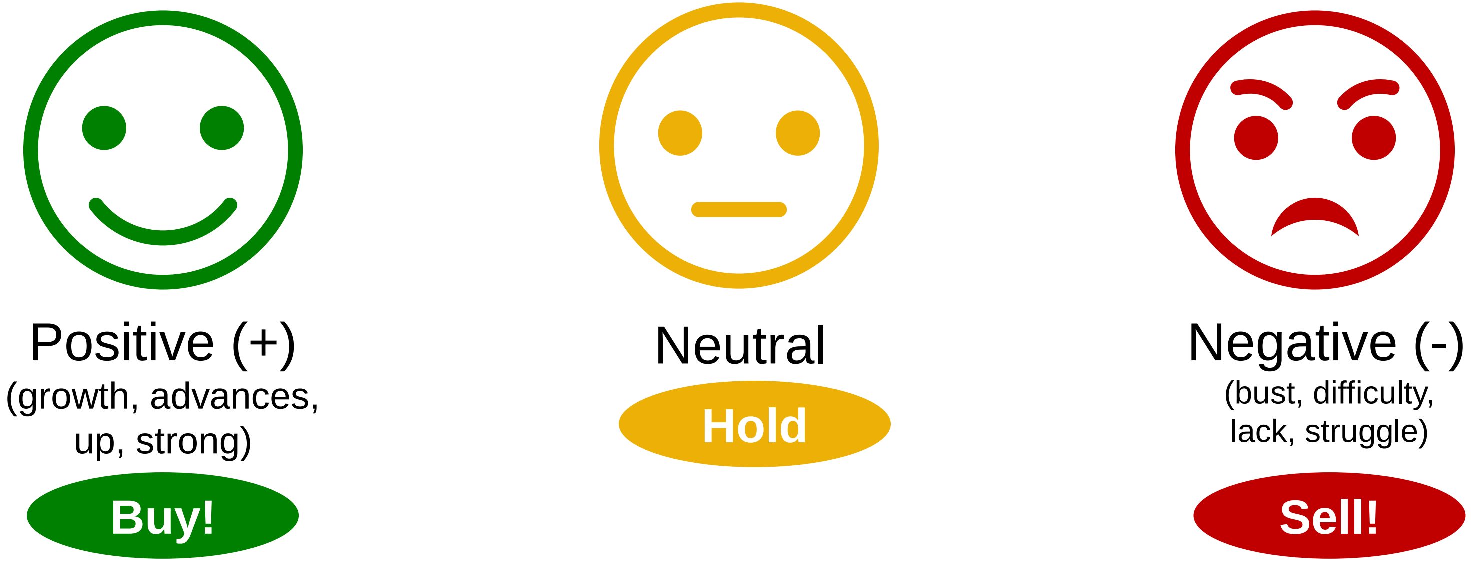
Update topic *t* for word *w* from document *d*

Final topic model

- $P(topic\ \boldsymbol{t}\ |document\ \boldsymbol{d})$
- $P(word\ \boldsymbol{w}\ |topic\ \boldsymbol{t})$

3 **DEMO:** Sentiment Analysis

Goal : Determining real-time sentiment scores for use in financial trading strategies



Positive (+)
(growth, advances, up, strong)
Buy!

Neutral
Hold

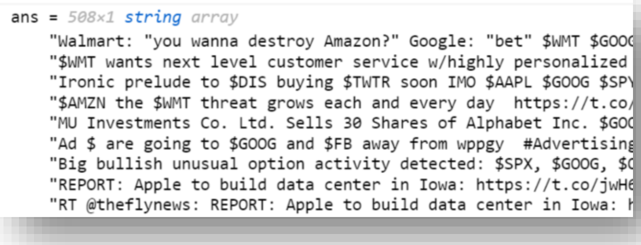
Negative (-)
(bust, difficulty, lack, struggle)
Sell!

Other Applications:

- Automating the classification of reviews, whether positive or negative
- Analyzing surveys to understand why customers are satisfied or dissatisfied
- Assessing counterparty credit risk

Demo: Workflow

Data



```
ans = 508×1 string array
    "Walmart: "you wanna destroy Amazon?" Google: "bet" $WMT $GOO
    "$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, $
    "REPORT: Apple to build data center in Iowa: https://t.co/jwH
    "RT @theflynews: REPORT: Apple to build data center in Iowa: 
```

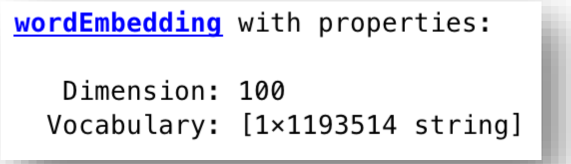
Preprocess text

Test data

Trained model

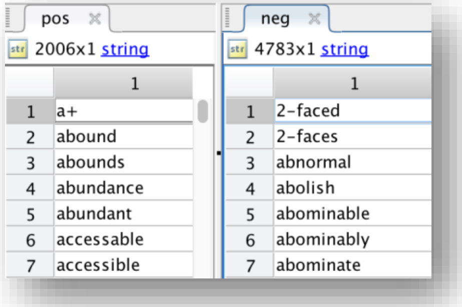
Score

Word Embedding



```
wordEmbedding with properties:

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

Positive + Negative Word List



pos — 2006x1 string

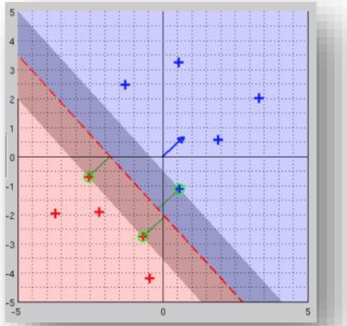
	1
1	a+
2	abound
3	abounds
4	abundance
5	abundant
6	accessable
7	accessible

neg — 4783x1 string

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

Training data

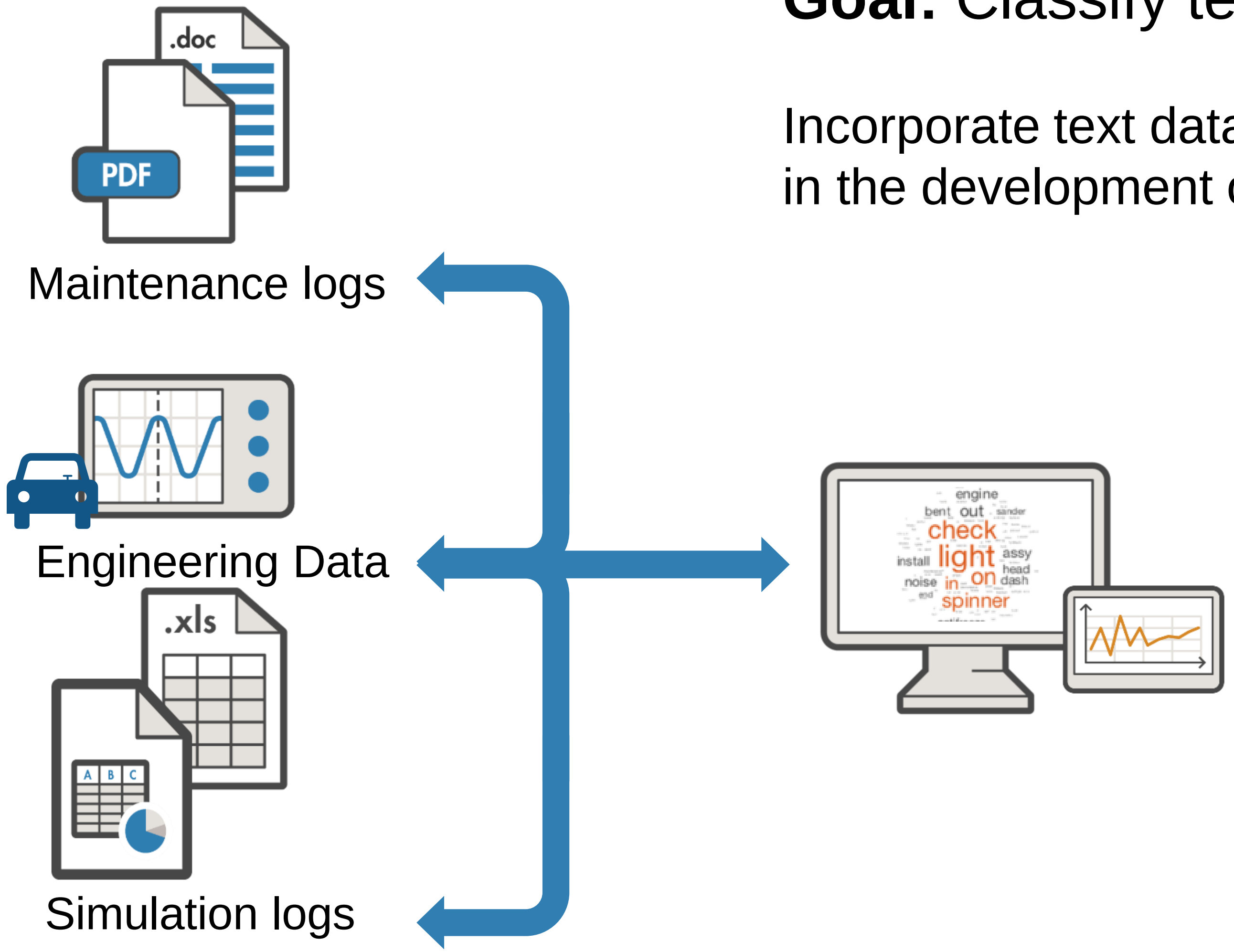
Machine Learning Model

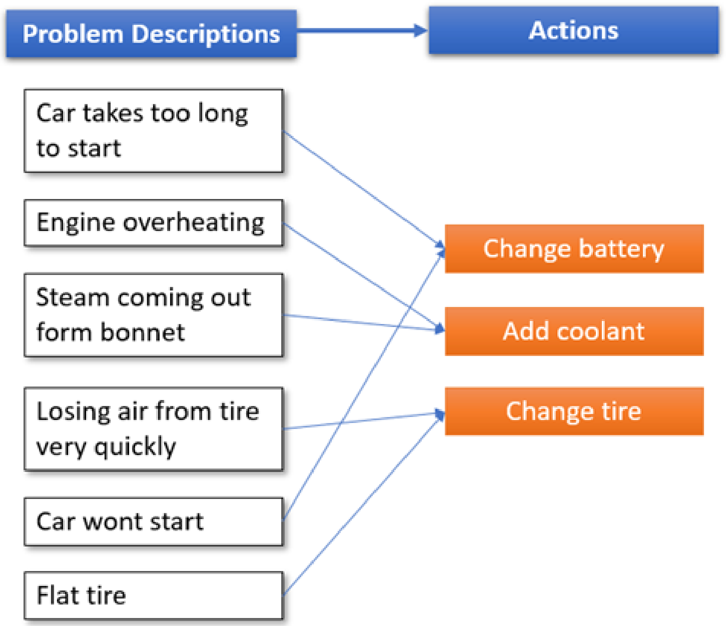


3 DEMO: Document Classification

Goal: Classify text based on historical data

Incorporate text data with other types of engineering data in the development of smart systems.




Problem Descriptions

Actions

Car takes too long to start

Engine overheating

Steam coming out form bonnet

Losing air from tire very quickly

Car wont start

Flat tire

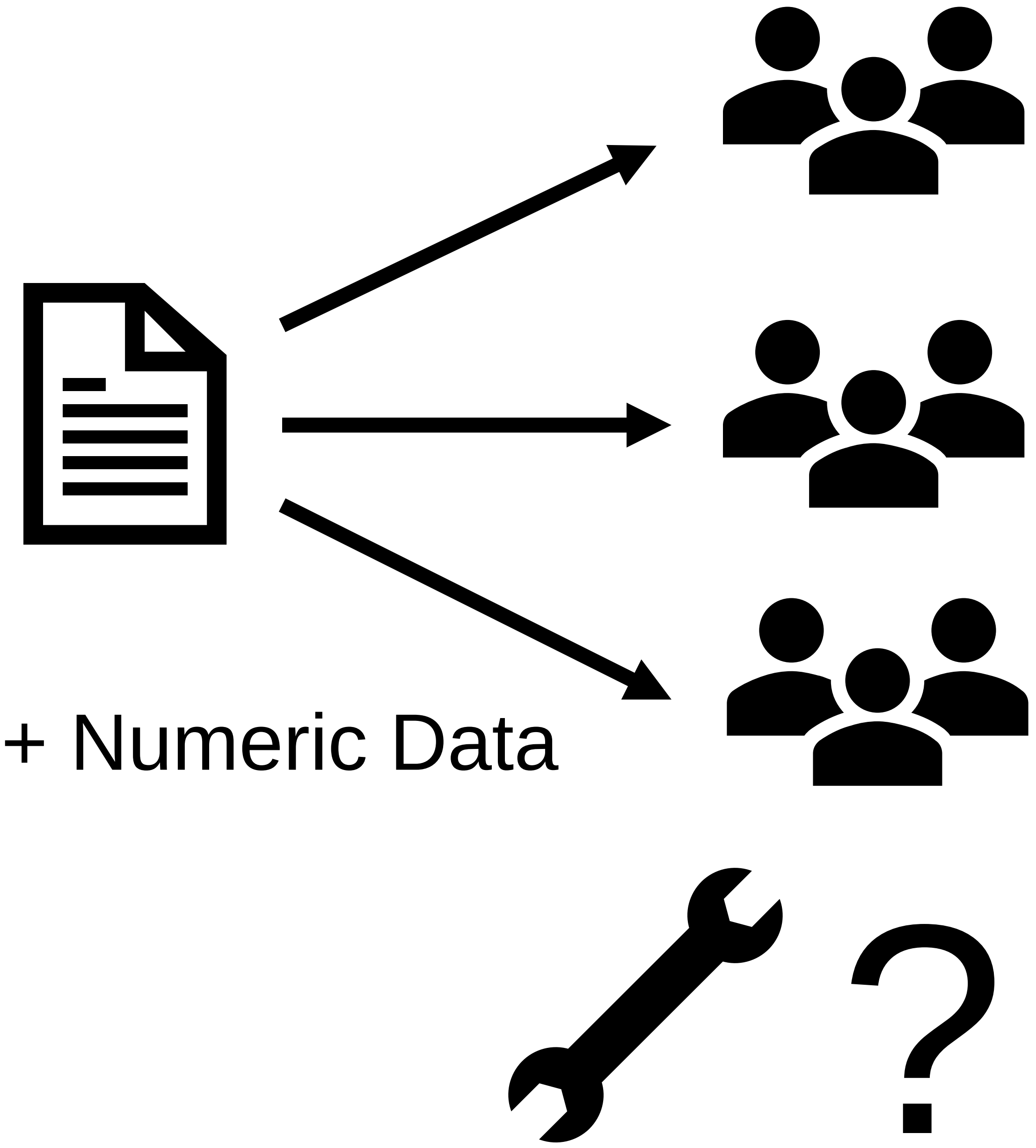
Change battery

Add coolant

Change tire

3 **DEMO:** Document Classification

- Document Classification
  - Field reports
  - Bug reports
- Predictive Maintenance
  - Equipment log notes



+ Numeric Data

4 Share Insights and Models

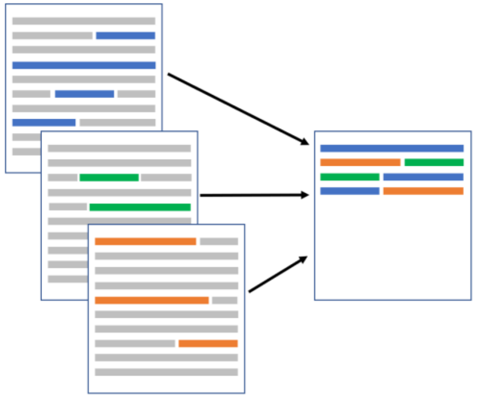
Many options for MATLAB and non-MATLAB users

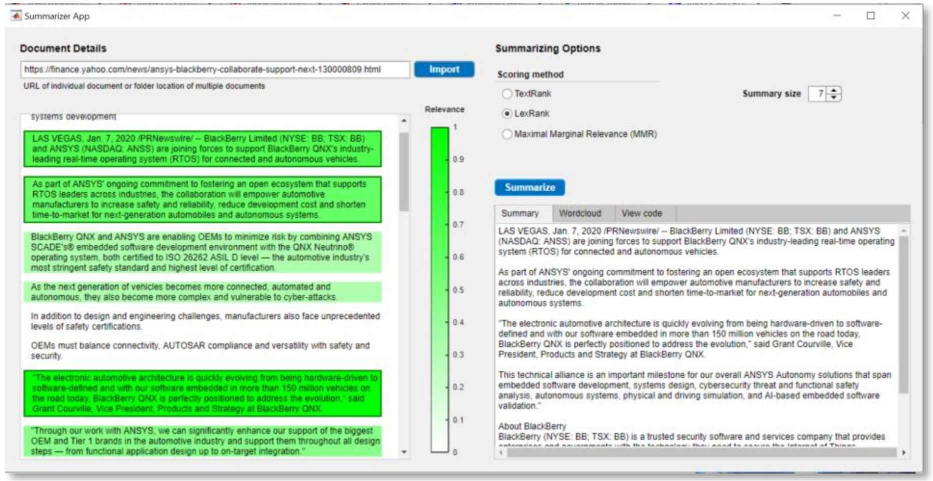
- Share interactive Live Editor Notebook
- Publish scripts as HTML, PDF, LaTeX, or Microsoft Word
- Generate formatted reports using MATLAB Report Generator
- Create standalone or web apps using MATLAB App Designer
- Host the application on MATLAB Production Server or MATLAB Web App Server

Document Summarization

Extract a summary from one or more documents automatically

Extractive Summarization





Document Details

https://finance.yahoo.com/news/ansys-blackberry-collaborate-support-next-130000809.html

URL of individual document or folder location of multiple documents

systems development

LAS VEGAS, Jan. 7, 2020 /PRNewswire/ -- BlackBerry Limited (NYSE: BB; TSX: BB) and ANSYS (NASDAQ: ANSS) are joining forces to support BlackBerry QNX's industry-leading real-time operating system (RTOS) for connected and autonomous vehicles.

As part of ANSYS' ongoing commitment to fostering an open ecosystem that supports RTOS leaders across industries, the collaboration will empower automotive manufacturers to increase safety and reliability, reduce development cost and shorten time-to-market for next-generation automobiles and autonomous systems.

BlackBerry QNX and ANSYS are enabling OEMs to minimize risk by combining ANSYS SCADE's® embedded software development environment with the QNX Neutrino® operating system, both certified to ISO 26262 ASIL D level — the automotive industry's most stringent safety standard and highest level of certification.

As the next generation of vehicles becomes more connected, automated and autonomous, they also become more complex and vulnerable to cyber-attacks.

In addition to design and engineering challenges, manufacturers also face unprecedented levels of safety certifications.

OEMs must balance connectivity, AUTOSAR compliance and versatility with safety and security.

"The electronic automotive architecture is quickly evolving from being hardware-driven to software-defined and with our software embedded in more than 150 million vehicles on the road today, BlackBerry QNX is perfectly positioned to address the evolution," said Grant Courville, Vice President, Products and Strategy at BlackBerry QNX.

"Through our work with ANSYS, we can significantly enhance our support of the biggest OEM and Tier 1 brands in the automotive industry and support them throughout all design steps — from functional application design up to on-target integration."

Summarizing Options

Scoring method

○ TextRank
● LexRank
○ Maximal Marginal Relevance (MMR)

Summary size 7

LAS VEGAS, Jan. 7, 2020 /PRNewswire/ -- BlackBerry Limited (NYSE: BB; TSX: BB) and ANSYS (NASDAQ: ANSS) are joining forces to support BlackBerry QNX's industry-leading real-time operating system (RTOS) for connected and autonomous vehicles.

As part of ANSYS' ongoing commitment to fostering an open ecosystem that supports RTOS leaders across industries, the collaboration will empower automotive manufacturers to increase safety and reliability, reduce development cost and shorten time-to-market for next-generation automobiles and autonomous systems.

"The electronic automotive architecture is quickly evolving from being hardware-driven to software-defined and with our software embedded in more than 150 million vehicles on the road today, BlackBerry QNX is perfectly positioned to address the evolution," said Grant Courville, Vice President, Products and Strategy at BlackBerry QNX.

This technical alliance is an important milestone for our overall ANSYS Autonomy solutions that span embedded software development, systems design, cybersecurity threat and functional safety analysis, autonomous systems, physical and driving simulation, and AI-based embedded software validation."

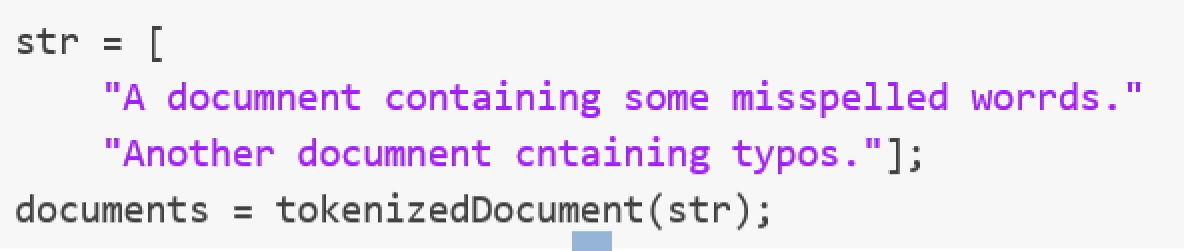
About BlackBerry
BlackBerry (NYSE: BB; TSX: BB) is a trusted security software and services company that provides

Relevant Applications

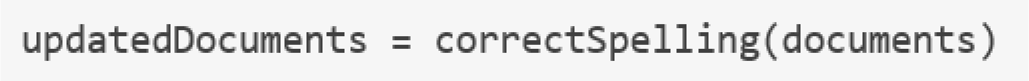
- Identify opportunities and gaps in scientific research by summarizing technical articles.
- Highlight and understand relevant information faster by summarizing internal reports.

Spelling Correction

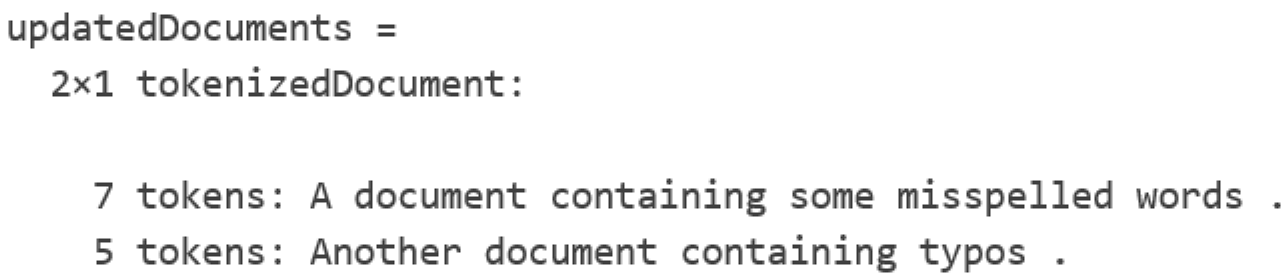
correctSpelling



```
str = [
    "A documnent containing some misspelled worrds."
    "Another documnent cntaining typos."];
documents = tokenizedDocument(str);
```



```
updatedDocuments = correctSpelling(documents)
```

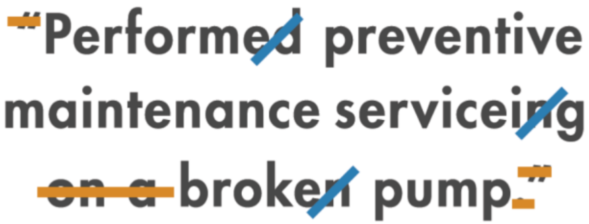


```
updatedDocuments = 
  2×1 tokenizedDocument:

    7 tokens: A document containing some misspelled words .
    5 tokens: Another document containing typos .
```

NOTE: Supports English, German, and Korean text.

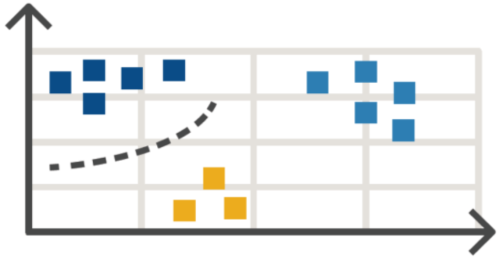
Text Analytics Toolbox

Import


~~"~~Performe~~d~~ preventive
maintenance servicei~~n~~g
~~on a~~ broke~~n~~ pump~~."~~

Preprocess

Visualize



Model and Predict

- Extract text from Microsoft Word files, PDFs, text files and spreadsheets.
- Remove less helpful artifacts and apply text normalization.
- Use word clouds and text scatter plots to summarize and validate results.
- Convert text into numeric representations and apply specialized machine learning algorithms for prediction and topic modeling.

More Resources

Text Analytics Toolbox

텍스트 데이터 분석 및 모델링

- Text Analytics Toolbox
- Text Analytics Toolbox – Documentation
- Getting Started with Text Analytics in MATLAB (White Paper)
- Text Analytics in MATLAB (23:36) – Video
- 8 MATLAB Cheat Sheets for Data Science - Cheat Sheets

Visit MathWorks at