Workshop
Part 5 - Text Mining

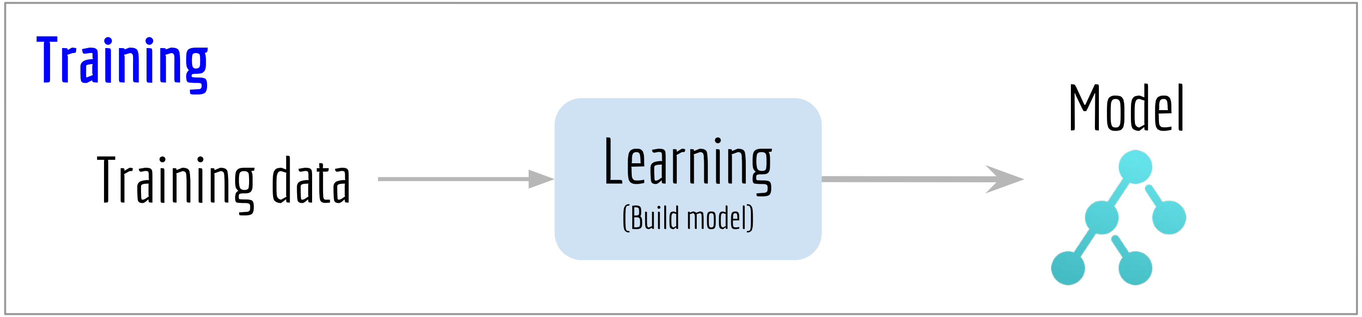
by Kanda Tiwatthanont

Overview

Part 5 : Sentiment Analysis

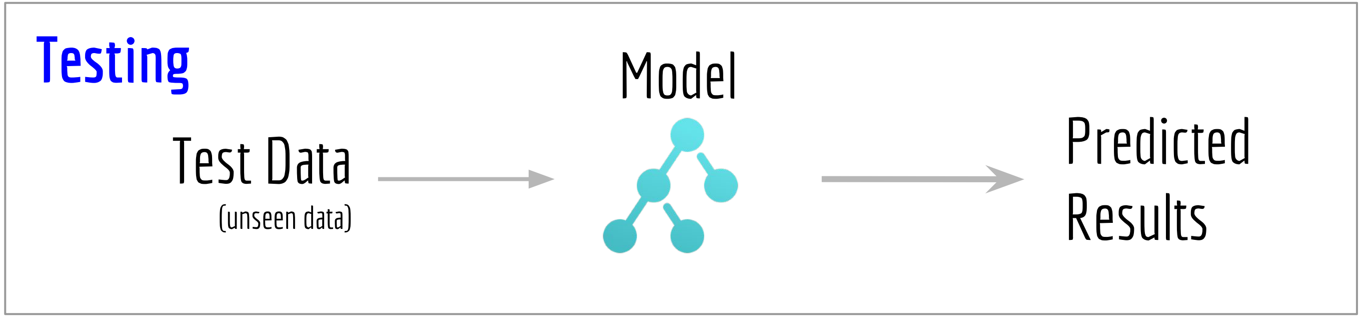
- Introduction Text Mining
- From Unstructured to Structured Data
- Text Mining Methodology
    - Data Preprocessing
    - Sentiment Analysis
- Hands-on -- Sentiment Analysis

Introduction Text Mining



Training

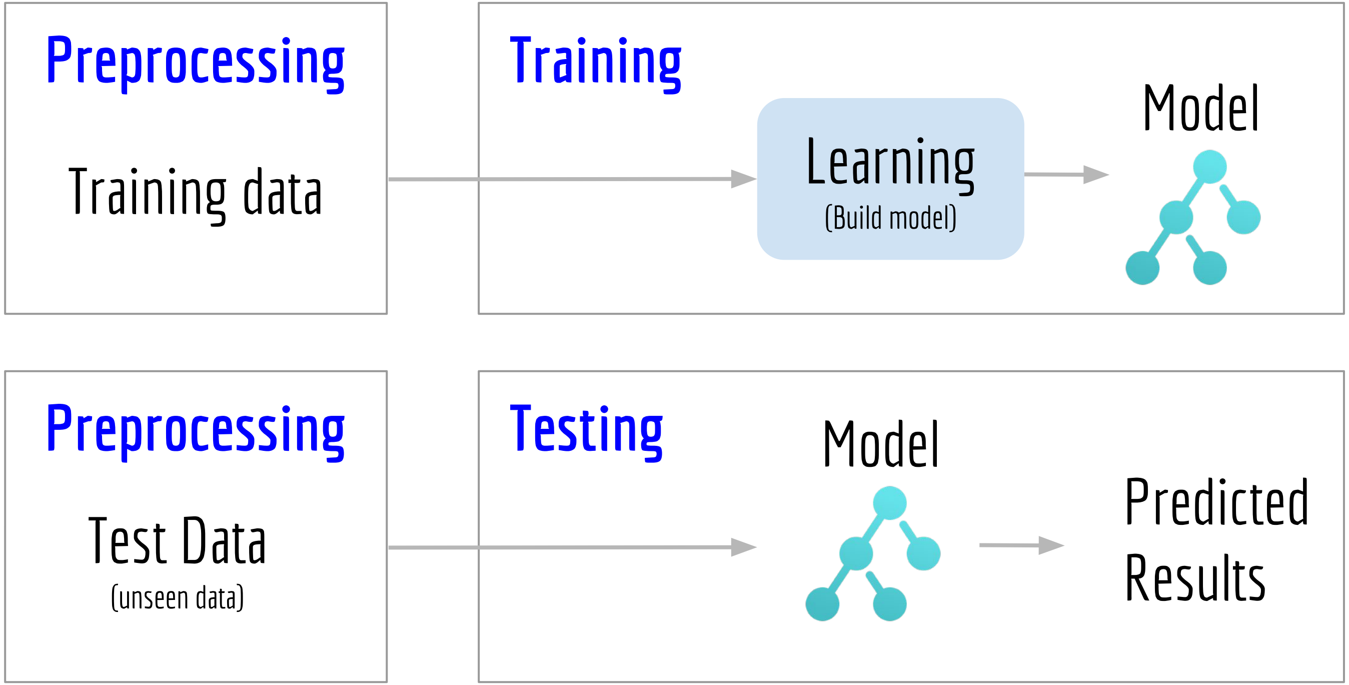
Training data → Learning (Build model) → Model



Testing

Test Data (unseen data) → Model → Predicted Results

Introduction Text Mining



Preprocessing

Training data

Training

Learning

(Build model)

Model

Preprocessing

Test Data

(unseen data)

Testing

Model

Predicted Results

From Unstructured to Structured data

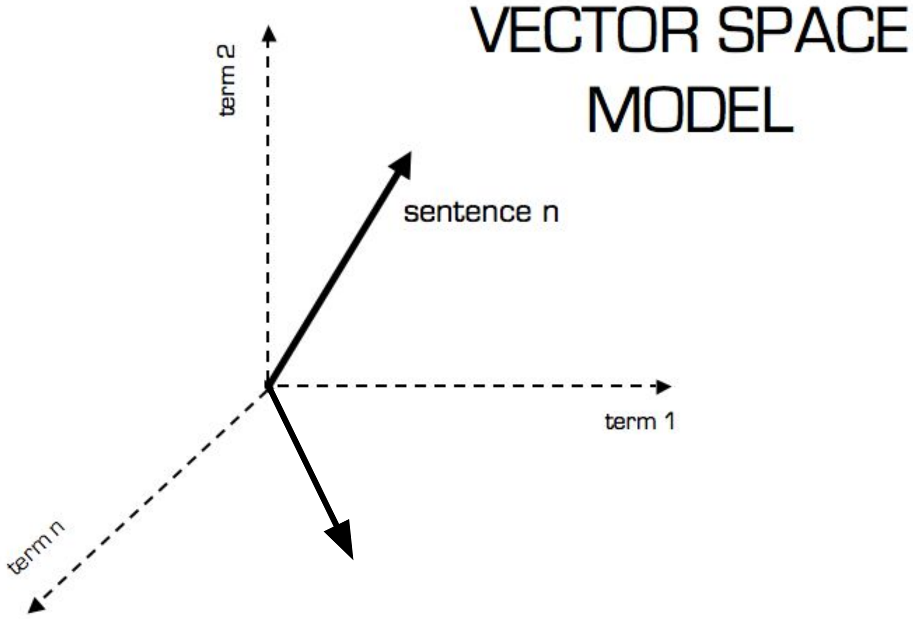
Sentence 1 : The sky is blue

the	sky	is	blue
1	1	1	1

From Unstructured to Structured data

vectorization

- A space where text is represented as a vector of numbers instead of its original string textual



Ref : http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/

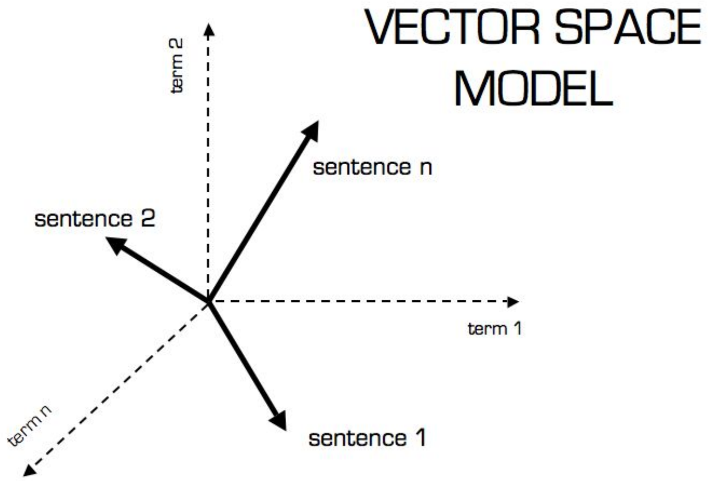
From Unstructured to Structured data

Train Document Set:

d1: The sky is blue.

d2: The sun is bright.

$$E(t) = \begin{cases} 1, \text{ if } t \text{ is ``the''} \\ 2, \text{ if } t \text{ is ``sky''} \\ 3, \text{ if } t \text{ is ``is''} \\ 4, \text{ if } t \text{ is ``blue''} \\ 5, \text{ if } t \text{ is ``sun''} \\ 6, \text{ if } t \text{ is ``bright''} \end{cases}$$



From Unstructured to Structured data

Train Document Set:

d1: The sky is blue. - category Earth

d2: The sun is bright. - category Universe

	the	**sky**	is	**blue**	**sun**	**bright**
d1	1	**1**	1	**1**	0	0
d2	1	0	1	0	**1**	**1**

From Unstructured to Structured data

Train Document Set:

d1: The sky is blue. - category Earth

d2: The sun is bright. - category Universe

Test Document Set:

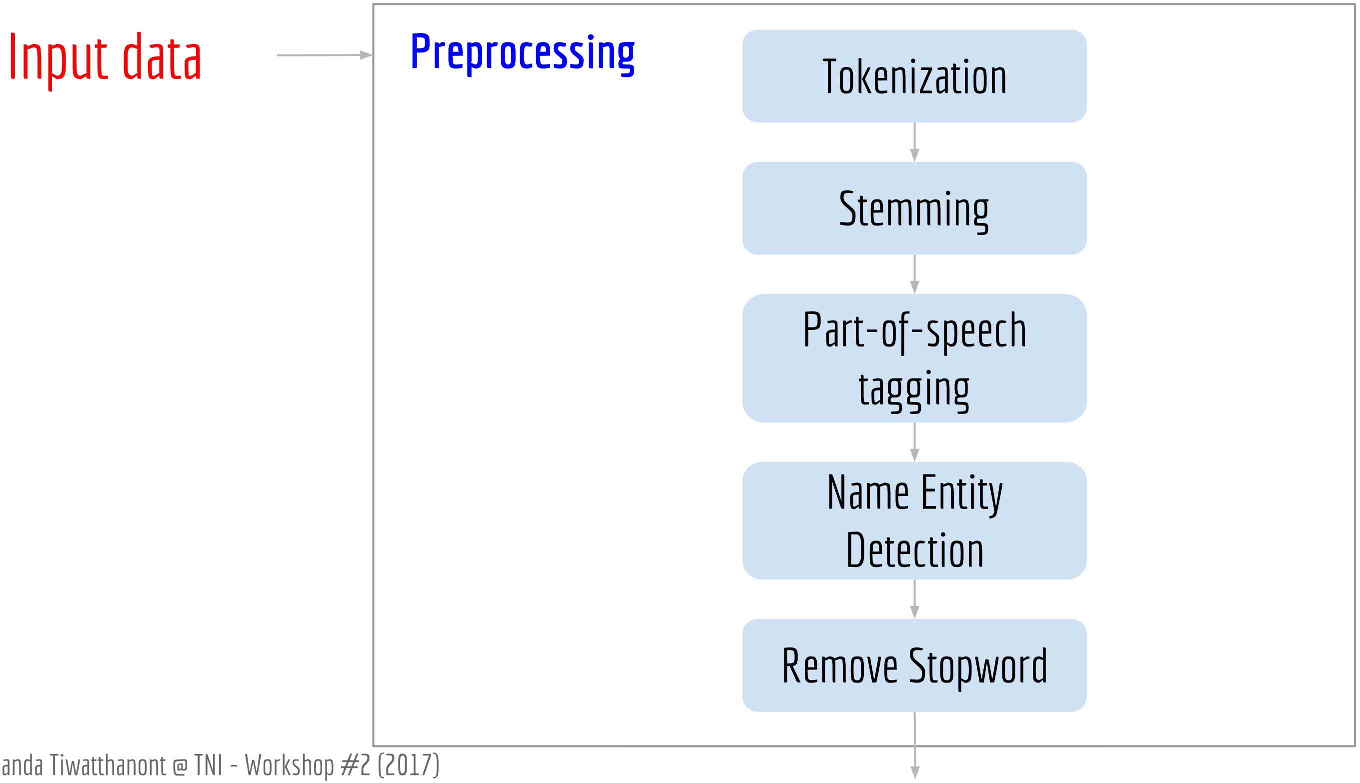
d3: The sun is bright on the sky.

	the	sky	is	blue	sun	bright
d1	1	1	1	1	0	0
d2	1	0	1	0	1	1
d3	1	1	1	0	1	1

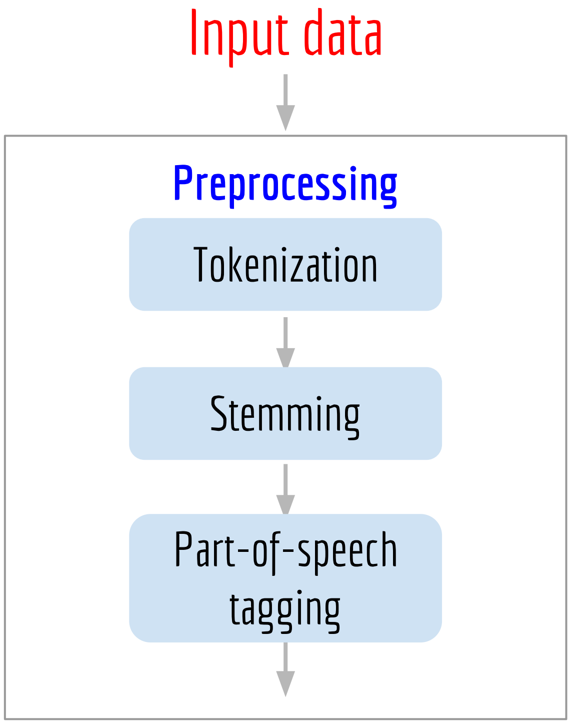
From Unstructured to Structured data

1. Term Occurrence
2. Term Frequency
3. Term Frequency - Inverse Document Frequency (TF-IDF)

Text Mining Methodology - Preprocessing



A Text Mining Methodology - Preprocessing


Input data

Preprocessing

Tokenization

Stemming

Part-of-speech tagging

“ The iPhone camera is amazing and easier to use. ”

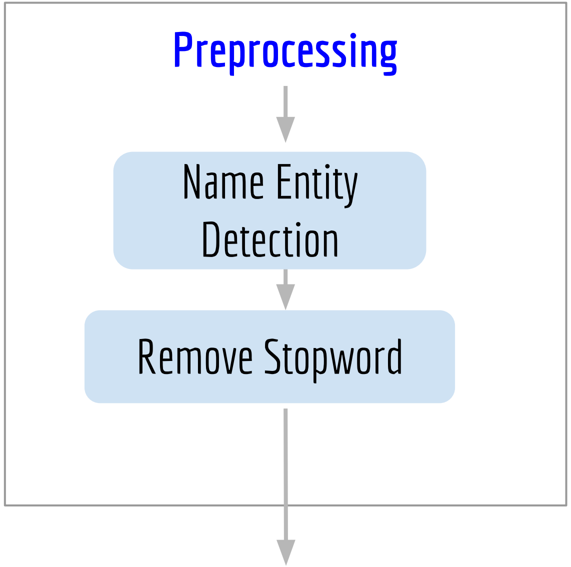
| The | iPhone | camera | is | amazing | and | easier | to | use | . |

| The | iPhone | camera | is | **amaze** | and | **easy** | to | use | . |

| The | iPhone | camera | is | amaze | and | easy | to | use | . |

Determiner Noun Noun Verb Adj Conj Adj Prep Verb

A Text Mining Methodology - Preprocessing



Preprocessing

Name Entity Detection

Remove Stopword

| The | iPhone | camera | is | amaze | and | easy | to | use | . |

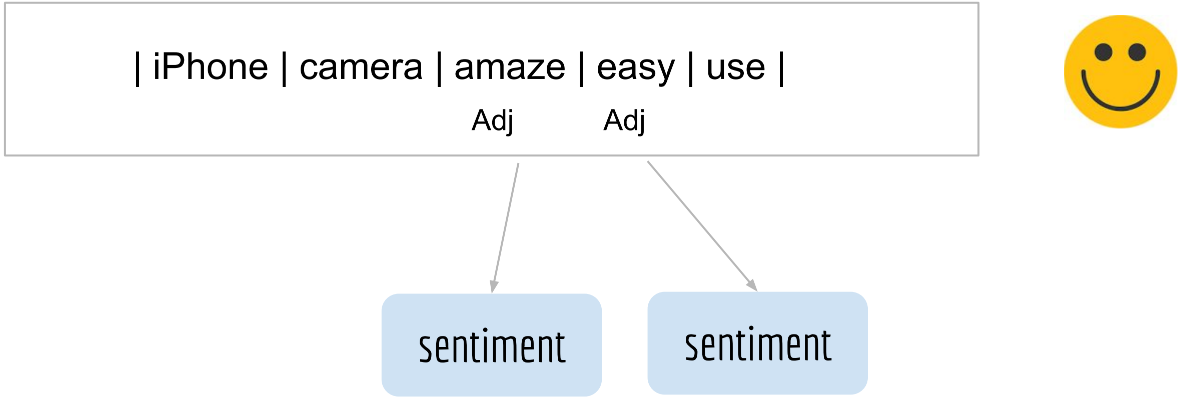
Determiner NE Noun Verb Adj Conj Adj Prep Verb

| iPhone | camera | amaze | easy | use |

NE Noun Adj Adj Verb

A Text Mining Methodology - Sentiment Analysis

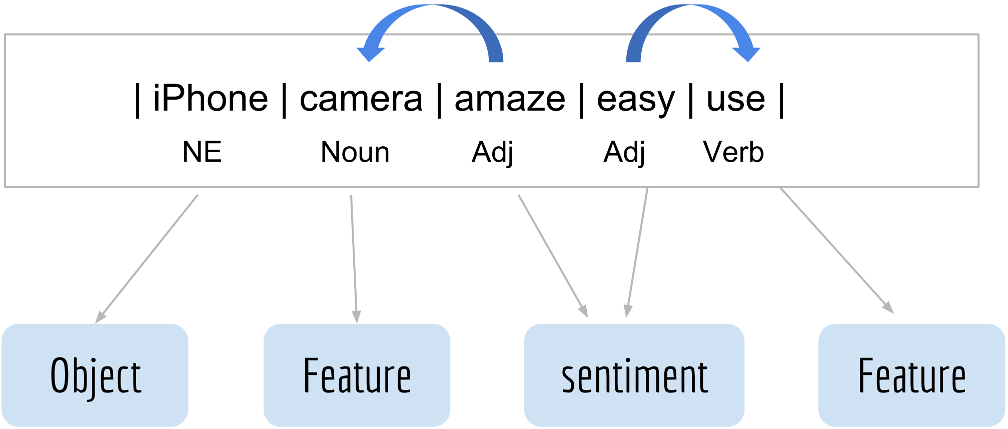
Document level



| iPhone | camera | amaze | easy | use |

Text Mining Methodology - Sentiment Analysis

Feature level (Aspect level)



| iPhone | camera | amaze | easy | use |

NE Noun Adj Adj Verb

Object

Feature

sentiment

Feature

Object : iPhone
Feature : Camera - Positive
Usability - Positive