# Text Analysis Pipeline

*Visit the Dome*

Ruta Bakhda(119464)

# What is text analysis pipeline?

- *It is a text mining approach which makes information search more intelligent by interpreting relevant information in text with optimal run time efficiency.*

- Text mining combines approaches from :

# Text Mining

Large number of unstructured text

Automatic or semi automatic discovery of high quality unknown information

Collection of Texts
(Text Corpus)
Streams of text

Tokens
Part-of-speech tags
Concrete type of entities and relations
Classification tasks

# 1. Information Retrieval

- Gathers input texts that are potentially relevant for the give task
- Generally in the form of query

- *Input* : Large collection of unstructured text
- *Output* : Needed information
- *Goal* : Search and obtain relevant information

- *How* :
  - All texts in the given collection are indexed
  - Using *Vector Space Model* that maps all texts and and queries into vectors and similarity is measured using *Cosine Similarity*.

# 2. Natural Language Processing

- *Goal :*Analyze the input texts in order to identify and structure relevant information and produce annotations

- Algorithms derive
  - Lexical information about the words in a text
  - Syntactic information about the structure between words
  - Semantic information about the meaning of words

- Problem:
  - Ambiguity : ‘She is an apple fan’

# Different Linguistic Levels

- Lexical and Syntactic analyses:
  - The segmentation of a text into single units (Word token, sentence splitting, paragraph splitting)
  - The tagging of units (Categorizing tokens, lemmatization)
  - The parsing of syntactic structure / Chunking (Identify different types of phrases or to infer dependency tree)

- Approaches for information extraction
  - *Rule based approach* : Based on regular expression or lexicon
  - *Statistical approach* : Based on machine learning

# What is Text Classification?

- Assigning one of the predefined class to the to the each text in a collection

- Examples:
  - Topic detection
  - Identification of genre of a text in terms of the form, purpose and / or intended audience of the text
  - Authorship attribution
  - Automatic essay grading
  - Stance recognition

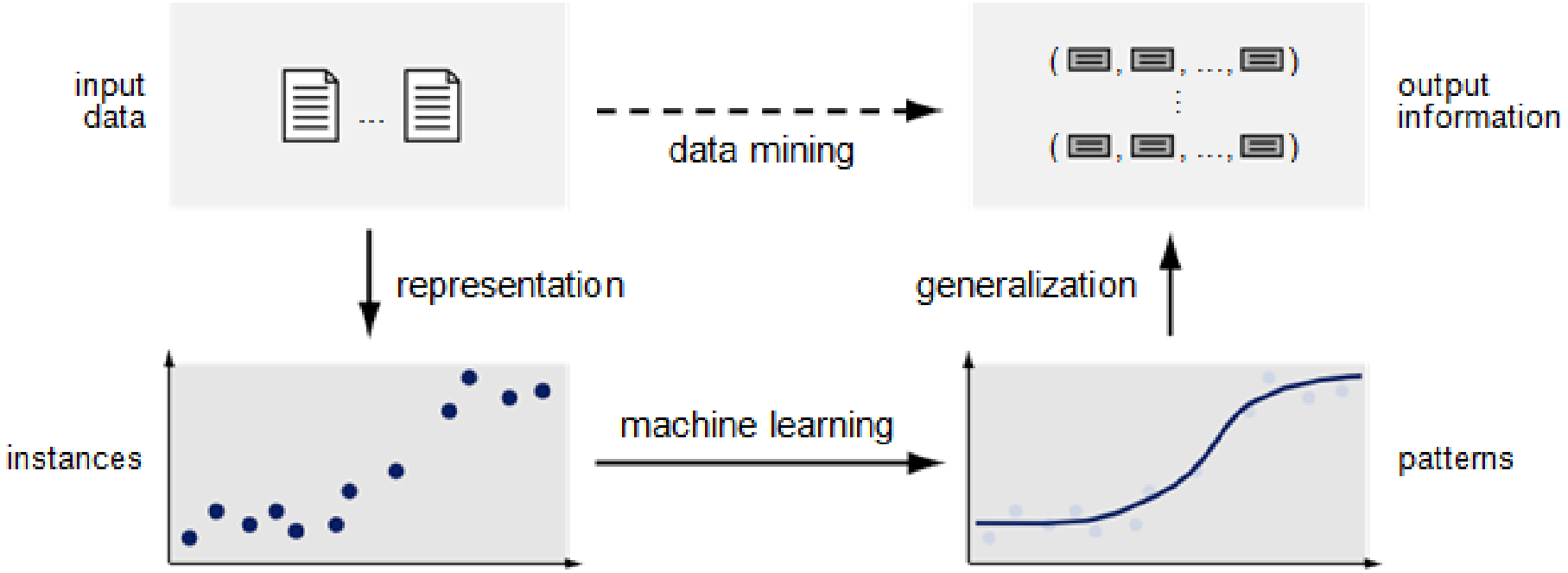
# Opinion Mining and Sentiment Analysis

- Opinion Mining :
  - Classifying subjectivity of a single course of units
  - Units can be seen as facts and subjective units as opinions

- Sentiment Analysis:
  - Sentiment polarity of text being negative or positive
  - Sentiment Scoring - Can also be accessed on numeric scales

# Data Mining

- Deriving new information of specified type from typically huge amounts of input data



# Machine learning

- *Goal* : An algorithm that learns without being explicitly programmed and learn target function
- *Target function* : It maps input space to an output space
- *Representation* : Y: X->C
- *Input* : Set of feature vectors
- *Output* : Prediction classes
- Types:
  - *Supervised learning*: Classification of text into various groups
  - *Unsupervised learning* : Finding common structures in data

# Text Corpora

- Corpora is the sample of real world texts
- A text corpus is a principled collection of texts that has been compiled to analyze a problem related to language or text analysis
- Often contains annotations

# Evaluation

- Quality is measured by
  - Effectiveness
    - Accuracy = (|TP|+|TN|) / (|TP|+|TN|+|FP|+|FN|)
    - Precision = |TP| / (|TP|+|FP|)
    - Recall = |TP| / (|TP|+|FN|)
    - F1-score = 2 * precision * recall / (precision + recall)
  - Efficiency
    - In terms of consumption of time and memory
    - Absolute over all run time on an input text
    - Average run time per instance of a text

# Validation

- Text corpus in divided into
  - Training set
  - Validation set
  - Test set
- Performance is measure using N-fold cross-validation.
- The measured efficiency and effectiveness are compared to alternatives ways of addressing tasks such as human-annotated ground truth or compare to some baseline.