Lesson 2

Text Mining Process Phases

Text Document Components

- Syntactically, characters that form words, which can be further combined to generate phrases or sentences

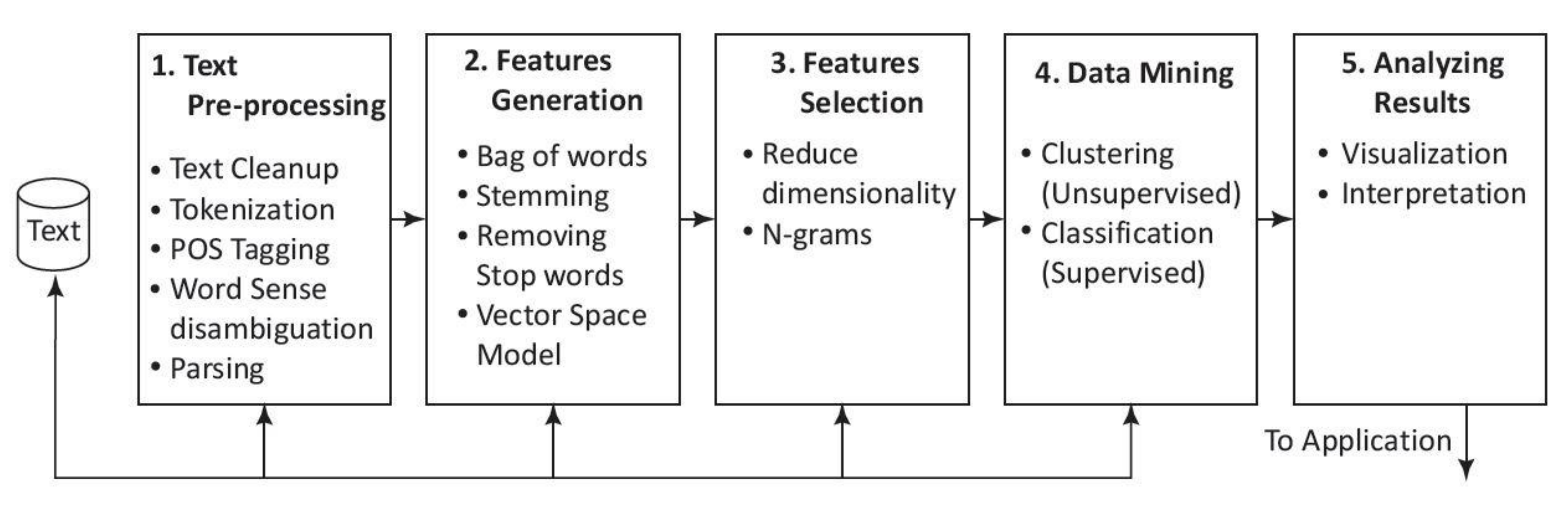
Text Mining Steps

- Recognizing, extracting and using the information present in words
- Along with searching of words, mining involves search for semantic patterns

Text Mining Process

- Consists of a process-pipeline executing in several phases
- Mining uses the iterative and interactive processes
- The processing in pipeline does text mining efficiently and mines the new information.

Figure 9.2 Five phases in a process pipeline



Phase 2: Preprocessing

- Clean-up
- Tokenization
- Part of Speech (POS) tagging
- Word sense disambiguation
- Parsing

Phase 2: Feature Generation

1. Bag of words—Order of words is not that important for certain applications.

Text document represented by the words it contains (and their occurrences) and for finding occurrence (frequency) of each word as a feature

Feature Generation

2. Stemming—identifies a word by its root

- Reduces the word to its most basic element. (impure → pure)

  Normalizes or unifies variations of the same concept

  Removes plurals, normalizes verb tenses and remove affixes

Feature Generation

3. Removing stop-words from the feature space—unlikely to help text mining, the search program tries to ignore stop-words

Ignores a, at, for, it, in, are, as, such, so, ….

Vector Space Model (VSM)

- An algebraic model for representing text documents as vector of identifiers, word frequencies or terms in the document index
- Term frequency-inverse document frequency (TF-IDF) for evaluating how important is a word in a document

Weight of a Word

- TF-IDF Weight
- May assign higher weights to keywords and Titles

Use of Vectors and Matrices

- Represent a collection of web documents as vectors
- Represent by a matrix with |D| × F shape, where |D| is the cardinality of the document space (total number of documents) and the F is the number of features. F represents the vocabulary size.

Example 9.2

- Shows that the matrices representing term frequencies tend to be very sparse (with majority of terms zeroed)
- A common representation of such matrix is thus the sparse matrices

Phase 3: Features Selection

- Process that selects a subset of features by rejecting irrelevant and/ or redundant features (variables, predictors or dimension) according to defined criteria

1. Feature Selection

- 1. Dimensionality reduction—Feature selection is one of the methods of division and therefore, dimension reduction. The basic objective is to eliminate irrelevant and redundant data. Redundant features are those, which provide no extra information

Feature Selection

- Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) for dimension reduction methods
- Discrimination ability of a feature measures relevancy of features. Correlation helps in finding the redundancy of the feature.

2. Feature Selection

2. N-gram evaluation—finding the number of consecutive words of interest and extract them

For example, 2-gram is a two words sequence, [“tasty food”, “Good one”]. 3-gram is a three words sequence, [“Crime Investigation Department”].

Feature Selection

- Discrimination ability of a feature measures relevancy of features. Correlation helps in finding the redundancy of the feature
- Two features are redundant to each other if their values correlate with each other.
- .

2. Feature Selection

2. N-gram evaluation—finding the number of consecutive words of interest and extract them. For example, 2-gram is a two words sequence, [“tasty food”, “Good one”]. 3-gram is a three words sequence, [“Crime Investigation Department”].

3. Feature Selection

- 3. Noise detection and evaluation of outliers methods do the identification of unusual or suspicious items, events or observations from the data set
- Step helps in cleaning the data from irrelevant words/information

Phase 4: Data Mining Techniques

- Unsupervised learning (for example, clustering)
- (i) The class labels (categories) of training data are unknown
- (ii) Establish the existence of groups or clusters in the data

Clustering

- Good clustering methods use high intra-cluster similarity and low inter-cluster similarity
- Examples of uses – blogs, patterns and trends

Supervised learning (for example, classification)

- (i) The training data is labeled indicating the class
- (ii) New data is classified based on the training set

Identifying evolutionary patterns in temporal text streams

- Useful in a wide range of applications, such as summarizing of events in news articles and extracting the research trends in the scientific literature

Phase 5: Analysing results

(i) Evaluate the outcome of the complete process.

(ii) Interpretation of Result– If acceptable then results obtained can be used as an input for next set of sequences. Else, the result can be discarded, and try to understand what and why the process failed.

Phase 5: Analysing results

(iii) Visualization – Prepare visuals from data, and build a prototype.

(iv) Use the results for further improvement in activities at the enterprise, industry or institution.

Summary

We learnt:

- Text Mining Process Pipeline
- Preprocessing Phase
- Feature Generation Phase: Bag of Words. TF-IDF, Weights to the words and terms
- VSM: Use of vectors and matrices

Summary

We learnt:

- Feature Selection Phase
- Data Mining Phase
- Supervised and unsupervised methods
- Analyzing the Results Phase

End of Lesson 2 on
Text Mining Process Phases