# Introduction to Text Mining

Instructor: Junghye Lee

February 26th, 2018

# Outline

# What is Text Mining (TM)?

- The use of computational methods and techniques to extract high quality information from text
- A computational approach to the discovery of new, previously unknown information and/or knowledge through automated extraction of information from often large amounts of unstructured text

# Why is TM relevant/useful?

- Unstructured text is present in various forms, and in huge and ever increasing quantities:
  - books
  - financial and other business reports
  - various kinds of business and administrative documents
  - news articles
  - blog posts
  - wiki
  - messages/posts on social networking and social media sites
  - ...
- It is estimated that about 80% of all the available data are unstructured data

# Why is TM relevant/useful?

- To enable effective and efficient use of such huge quantities of textual content, we need computational methods for
  - automated extraction of information from unstructured text
  - analysis and summarization of extracted information
- TM research and practice are focused on the development, continual improvement and application of such methods

# TM Application Domains

- Document classification
- Clustering/organizing documents
- In particular, difficulties with automated text comprehension are caused by the fact that the human/natural language:
  - is full of ambiguous terms and phrases
  - often strongly relies on the context and background knowledge for defining and conveying meaning
  - is full of fuzzy and probabilistic terms and phrases
  - strongly based on commonsense knowledge and reasoning
  - is influenced by and is influencing peoples mutual interactions
- Visualization of document space (often aimed at facilitating document search)
- Making predictions (e.g., predicting stock market prices based on the analysis of news articles and financial reports)
- Content-based recommender systems (for news articles, movies, books, articles, ...)

# The Complexity of Unstructured Text

- In general, interpretation/comprehension of unstructured content (text, images, videos) is (often) easy for people, but very complex for computer program
  - Document refers to any kind of unstructured piece of text: blog post, news article, tweet, status update, business document, ...
- Clustering/organizing documents
- Document summarization
- Visualization of document space (often aimed at facilitating document search)
- Making predictions (e.g., predicting stock market prices based on the analysis of news articles and financial reports)
- Content-based recommender systems (for news articles, movies, books, articles, ...)

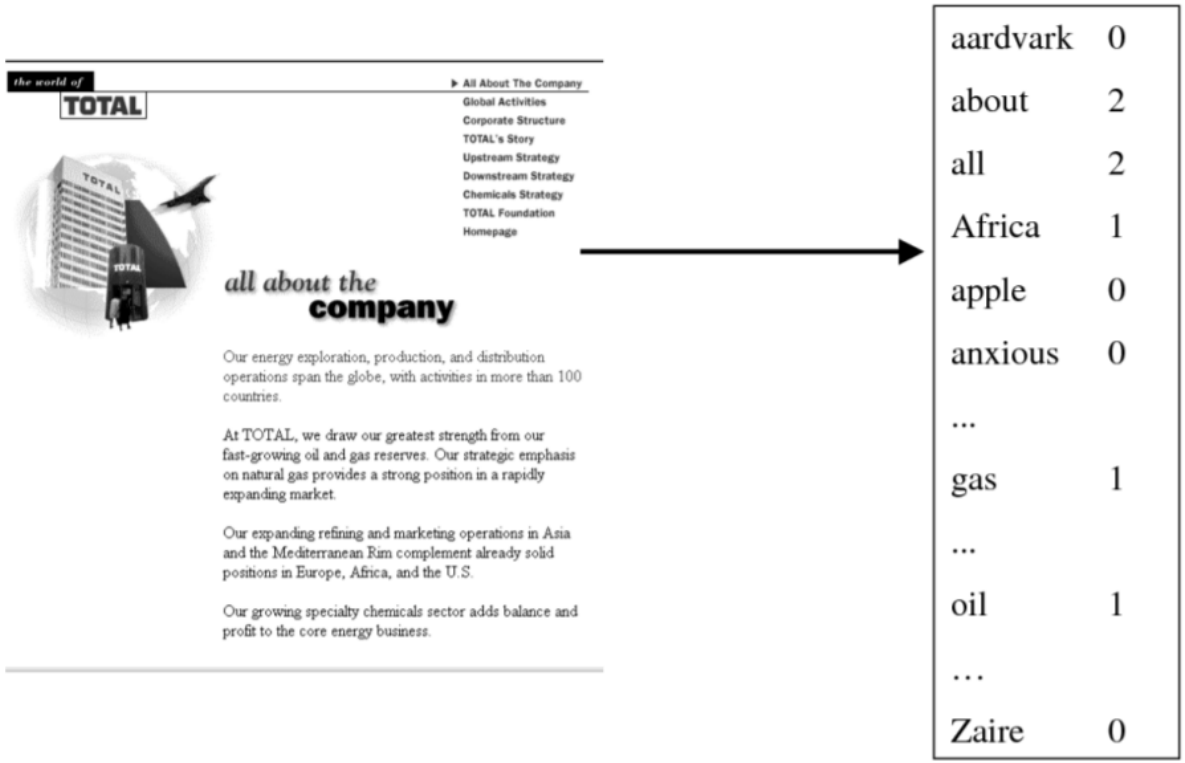
# The Complexity of Unstructured Text

- The use of supervised machine learning (ML) methods for TM is often very expensive
  - This is caused by the need to prepare high number of annotated documents to be used as the training dataset
  - Such a training set is essential for, e.g., document classification or extraction of entities, relations and events from text
- High-dimension of the attribute space:
  - Documents are often described with numerous attributes, which further impedes the application of ML methods
  - Most often, attributes are either all terms or a selection of terms and/or phrases from the collection of documents to be analyzed

# Bag-of-Words Representation of Text

- Considers text a simple set/bag of words
- Based on the following (unrealistic) assumptions:
  - words are mutually independent
  - word order in text is irrelevant
- Despite its unrealistic assumptions and simplicity, this approach to text modeling proved to be highly effective, and is often used in TM

# Bag-of-Words Representation of Text




Unique words from the corpus are used to create the corpus **dictionary**; then, each document from the corpus is represented as a vector of (dictionary) word frequencies

# Vector Space Model

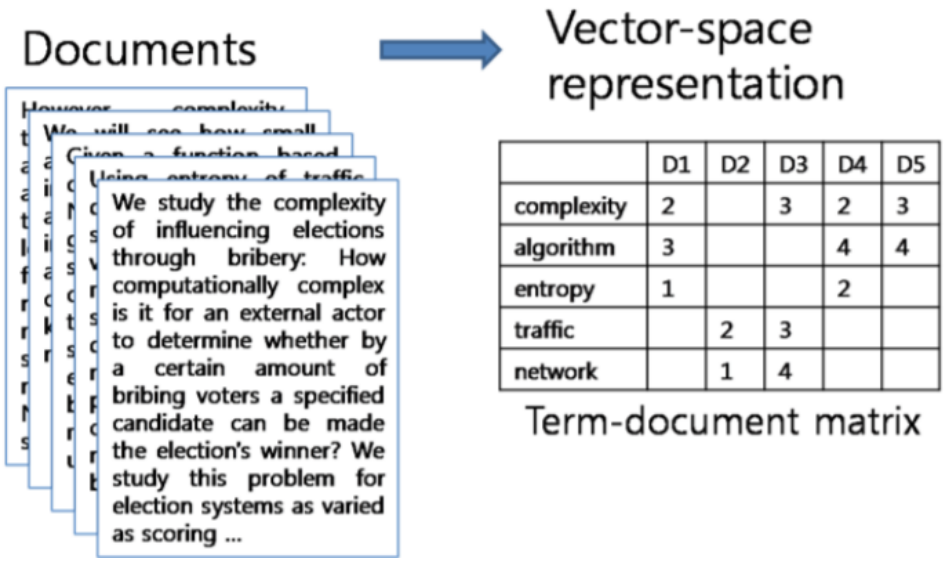
- Generalization of the Bag of Words model
- Each document from the corpus is represented as a multidimensional vector
  - Corpus refers to a collection of documents to be processed/analyzed
- Generalization of the Bag of Words model
- Each unique term from the corpus represents one dimension of the vector space
- Term can be a single word or a sequence of words (phrase)
- The number of unique terms in the corpus determines the dimension of the vector space

# Vector Space Model

- Vector elements are weights associated with individual terms; these weights reflect the relevancy of the corresponding terms in the given corpus
- If a corpus consists of $n$ terms $(t_i, i = 1, ..., n)$, document $d$ from that corpus would be represented with the vector: $d = w_1, w_2, ..., w_n$, where $w_i$ are weights associated with terms $t_i$

# Vector Space Model

- In VSM, corpus is represented in the form of Term Document Matrix (TDM), i.e., an *mxn* matrix with following features:
  - Rows ($i = 1, ..., m$) represent terms from the corpus
  - Columns ($j = 1, ..., n$) represent documents from the corpus
  - Cell *ij* stores the weight of the term *i* in the context of the document *j*




| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |

Term-document matrix

# Outline

# VSM: Text Pre-processing

- Before creating the TDM matrix, documents from the corpus need to be pre-processed
- Rationale/objective: to reduce the set of words to those that are expected to be the most relevant for the given corpus
- Pre-processing (often) includes:
  - Normalizing the text
  - Removing terms with very small/high frequency in the given corpus
  - Removing the so-called stop-words
  - Reducing words to their root form through stemming or lemmatization

# Normalization of Text

- Before creating the TDM matrix, documents from the corpus need to be pre-processed
- Rationale/objective: to reduce the set of words to those that are expected to be the most relevant for the given corpus
- Objective: transform various forms of the same term into a common, normalized form
  - e.g.1: Apple, apple, APPLE $\rightarrow$ apple
  - e.g.2: Intelligent Systems, Intelligent systems, Intelligent-systems $\rightarrow$ intelligent systems
- How it is done:
  - Using simple rules:
    - Remove all punctuation marks (dots, dashes, commas, ...)
    - Transform all words to lower case
  - Using a dictionary, such as **WordNet**, to replace synonyms with a common, often more general, concept
    - e.g., automobile, car $\rightarrow$ vehicle

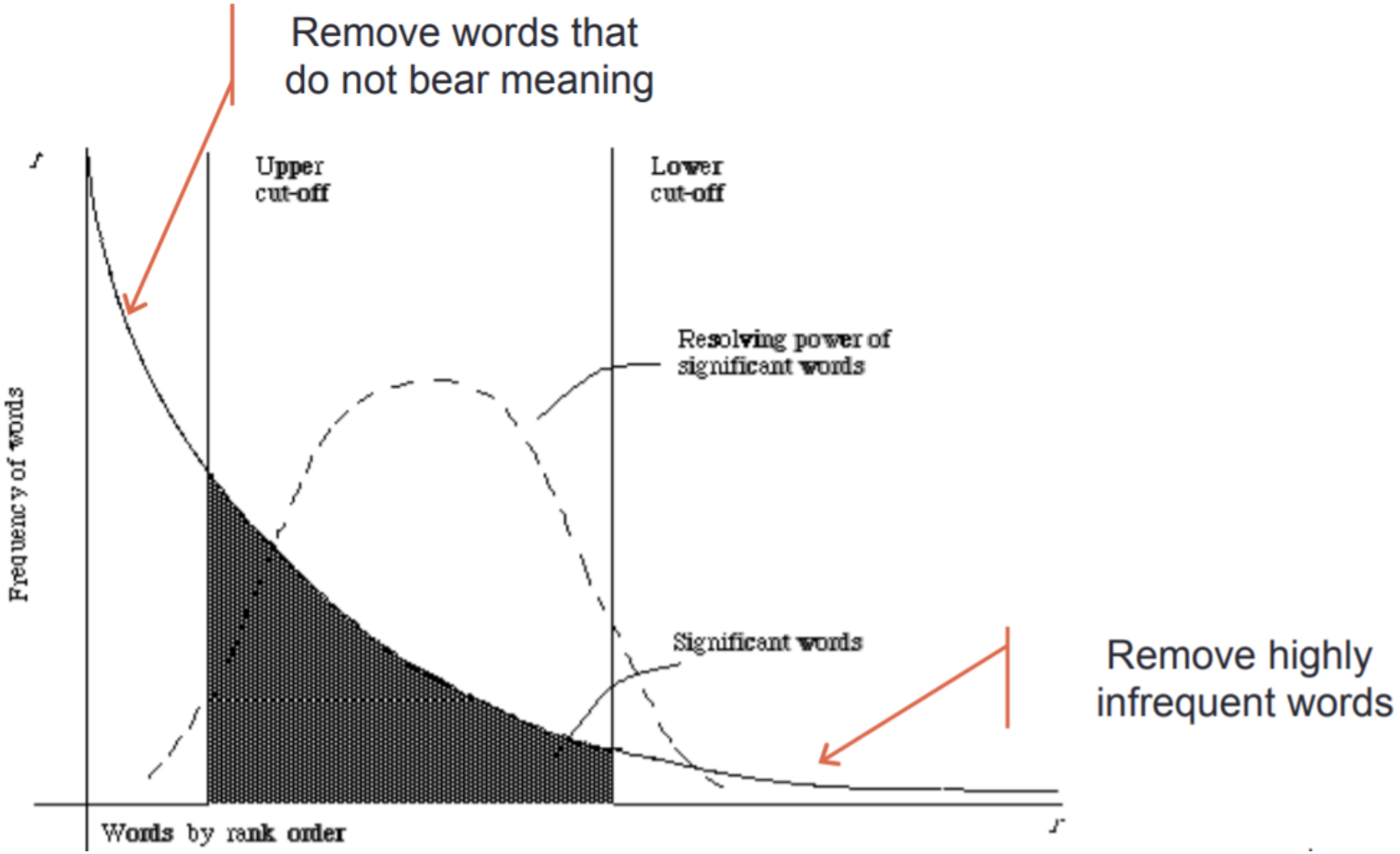
# Removing High and Low Frequency Terms

- Empirical observations (in numerous corpora):
  - Many low frequency words
  - Only a few words with high frequency
- Formalized in the **Zipf's rule**: the frequency of a word in a given corpus is inversely proportional to its rank in the frequency table (for that corpus)

# Implications of the Zipf's Rule

- Words in the upper part of the frequency table comprise a significant proportion of all the words in the corpus, but are semantically almost useless
  - Examples: the, a, an, we, do, to
- On the other hand, words towards the bottom of the frequency table are semantically rich, but are of very low frequency
  - Example: dextrosinistral
- The rest of the words are those that represent the corpus the best and thus should be included in the VSM model

# Implications of the Zipf's Rule



# Stop-Words

- An alternative or a complementary way to eliminate words that are (most probably) irrelevant for corpus analysis
- Stop-words are those words that (on their own) do not bear any information / meaning
- It is estimated that they represent 20-30% of words in any corpus
- There is no unique stop-words list
  - Frequently used lists are available at: http://www.ranks.nl/stopwords
- Potential problems with stop-words removal:
  - the loss of original meaning and structure of text
  - examples: "this is not a good option" $\rightarrow$ "option"
    "to be or not to be" $\rightarrow$ null

# Stemming and Lemmatization

- Two approaches to decreasing variability of words by reducing different forms of words to their basic/root form
- Stemming is a crude heuristic process that chops off the ends of words without considering linguistic features of the words
  - e.g., argue, argued, argues, arguing $\rightarrow$ argu
- Lemmatization refers to the use of a vocabulary and morphological analysis of words, aiming to return the base or dictionary form of a word, which is known as the lemma

  e.g., argue, argued, argues, arguing $\rightarrow$ argue

# Outline

# Outline