IST 3420: Introduction to Data Science and Management

Langtao Chen, Fall 2017

10: Advanced Topics – Text Analytics

Learning Objectives

- Understand general methods of processing textual data
- Be able to use tm package to conduct text analytics

Textual Data Are Ubiquitous

- It's a truism that 80 percent of business-relevant information originates in unstructured form, primarily text.
  - Web pages
  - Social media (e.g., Twitter feed)
  - Customer reviews
  - News articles
  - ……

What is Text Mining?

Text mining is a process that employs a set of algorithms for **converting unstructured text into structured data objects** and then using **quantitative methods to analyze these data objects**.

Text mining derives quantitative representation from text.

Quantify Textual Data (to generate numbers)

- Text needs to be structured before we can analyze it

- Two major approaches
  - Bag of Words (BoW)
  - Natural Language Processing (NLP)

Bag-of-Words (BoW) Approach

- Considers a document simply as a collection of words
- Each word is an individual item for analysis
  - Meaning of the text, order of words, and syntax are ignored

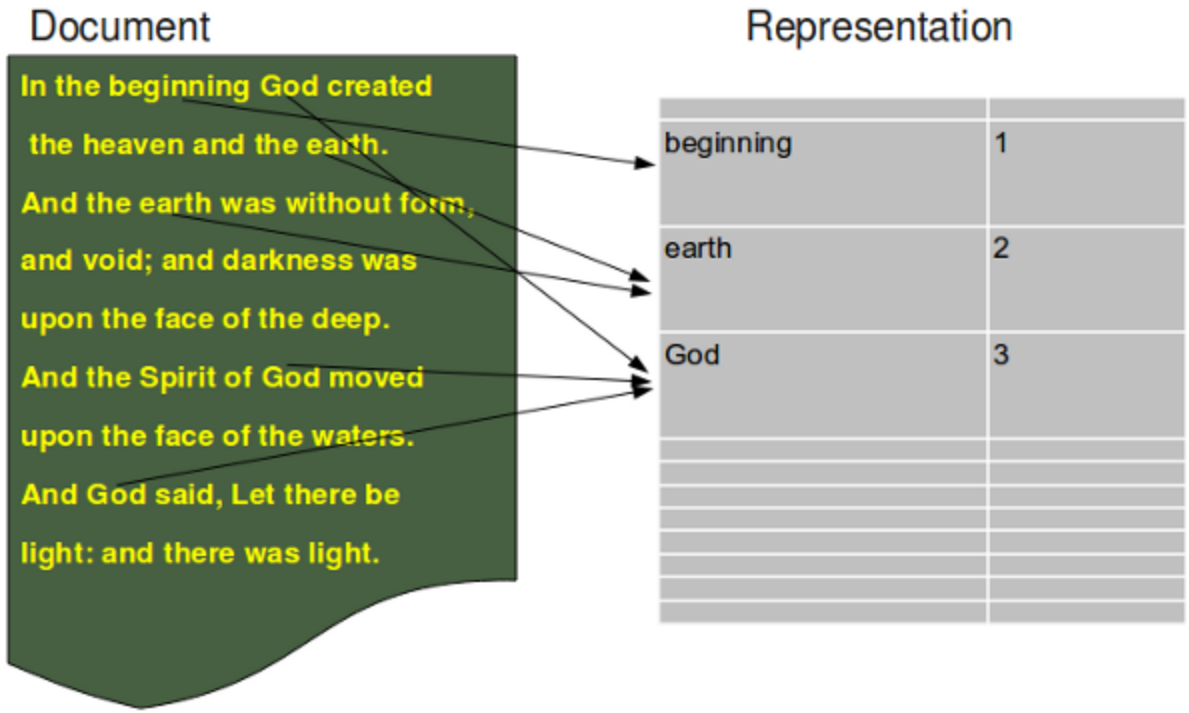


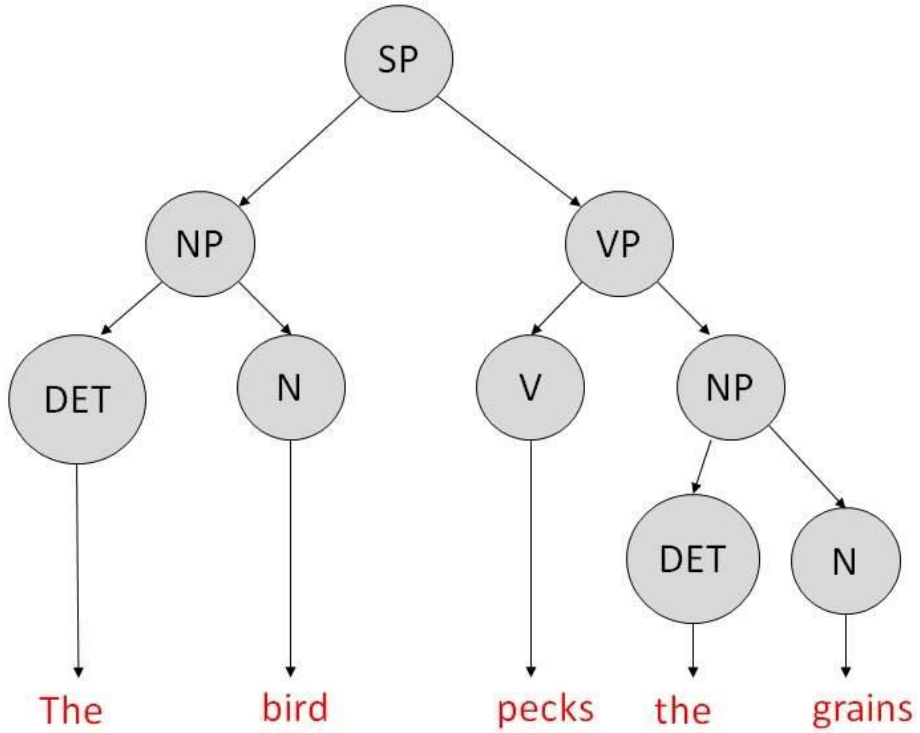
Image source: http://www.python-course.eu/text_classification_python.php

Natural Language Processing (NLP) Approach

- Natural Language Understanding (NLU):
  - Mapping the given input in natural language into useful representations.
  - Analyzing different aspects of the language.

- Natural Language Generation (NLG)
  - It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.
  - It involves −
    - **Text planning** − It includes retrieving the relevant content from knowledge base.
    - **Sentence planning** − It includes choosing required words, forming meaningful phrases, setting tone of the sentence.
    - **Text Realization** − It is mapping sentence plan into sentence structure.

- The NLU is harder than NLG.



A grammar parser

https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm

BoW vs. NLP

- "San Francisco beat Boston in last night's baseball game"
- "Boston beat San Francisco in last night's baseball game"

NLP tries to understand or deal with the meaning of the text.

Text Classification/Categorization

- To classify a set of documents
- Textual feature extraction + machine learning

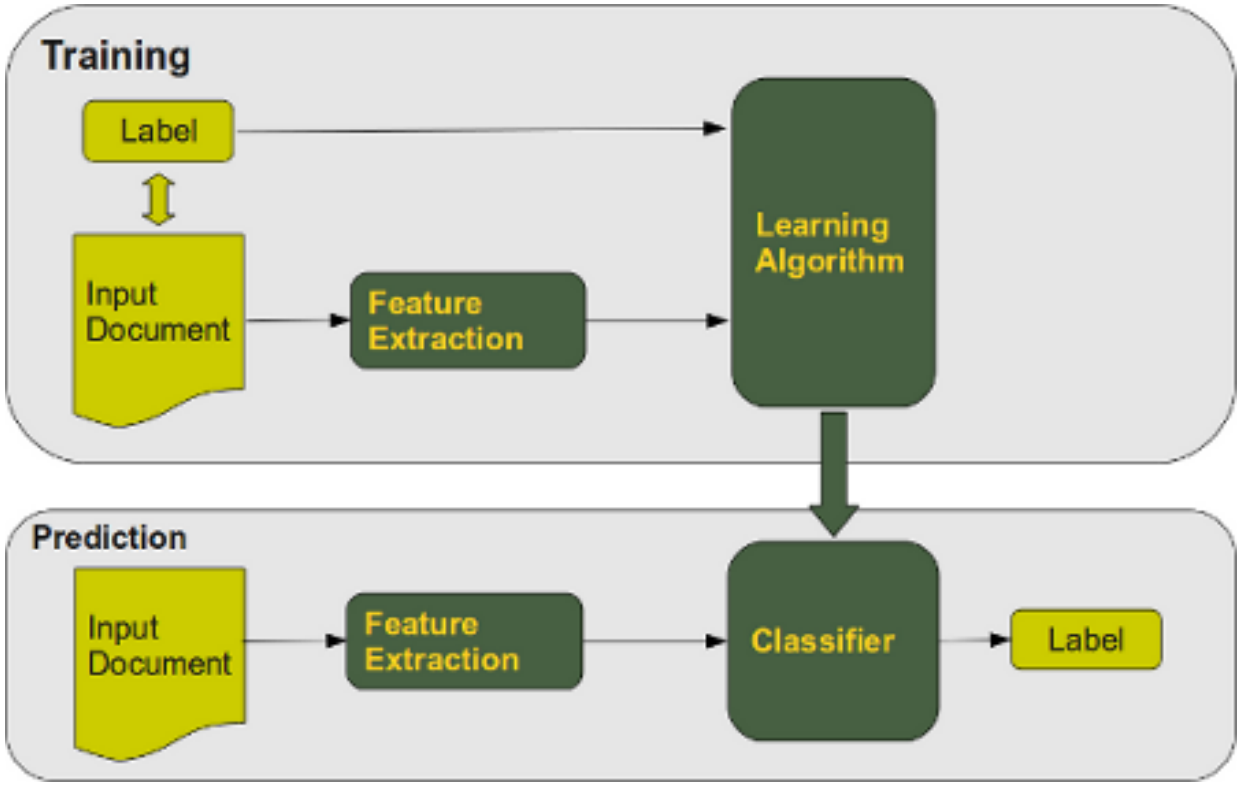
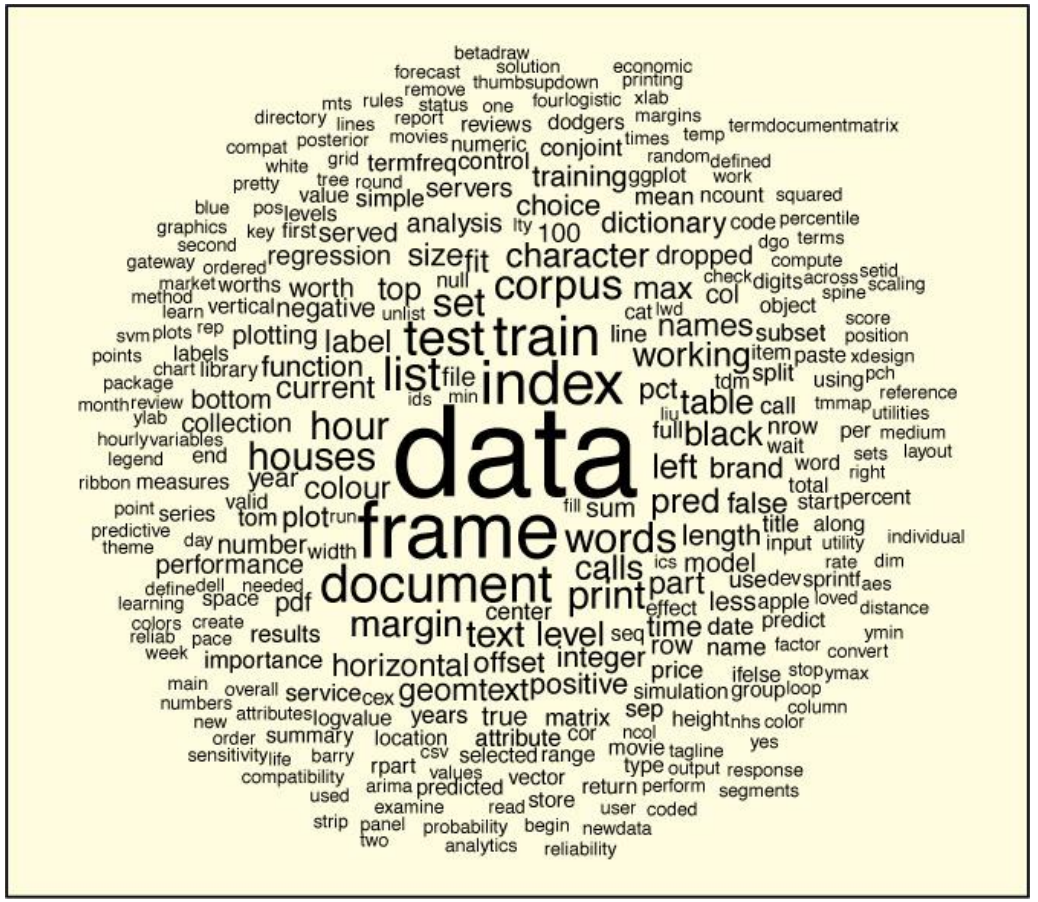


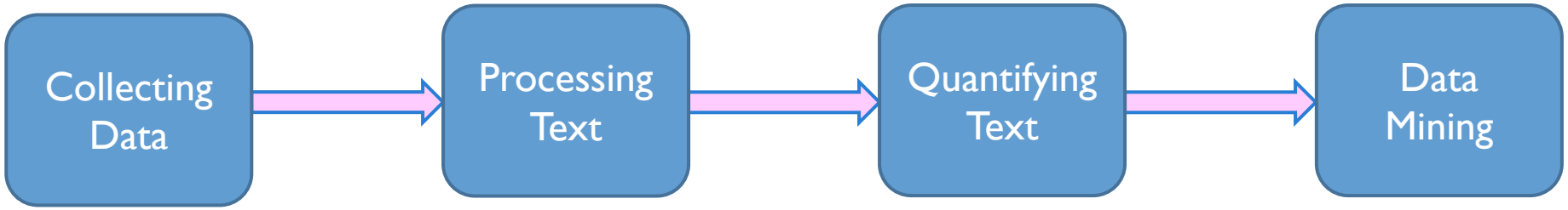
Image source: http://www.python-course.eu/text_classification_introduction.php

Word Cloud



An R Programmer's Word Cloud

A General Text Analytics Procedure



- Documents
- Web pages
- Social media
- …

- Tokenization
- Part-of-speech
- Stemming
- Synonyms
- Stop words
- Term filtering
- …

- Term count
- Term-by-document matrix
- Singular value decomposition (SVD)
- …

- Topic extraction (latent semantic analysis, latent Dirichlet allocation etc.)
- Text classification
- Sentiment analysis
- Predictive analytics
- …

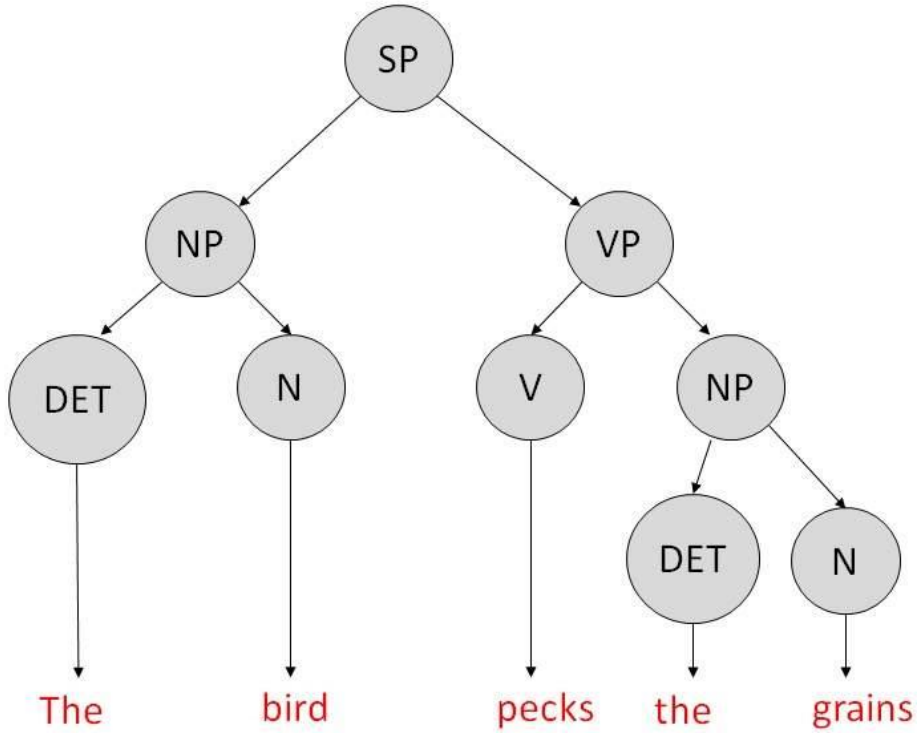
Tokenization

- The process of dividing text into separate “tokens” or terms.

Parts of Speech (POS) Tagging

- POS is a category of words which have similar grammatical properties.

- Commonly listed English parts of speech are:
  - noun
  - verb
  - adjective
  - adverb
  - pronoun
  - preposition
  - conjunction
  - interjection
  - sometimes numeral, article or determiner



A grammar parser

https://en.wikipedia.org/wiki/Part_of_speech

Penn Treebank POS Tags

- https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun

Number	Tag	Description
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Stop Words

- Many of the most frequently used words in English are worthless in retrieval and text mining – these words are called ***stop words***.
  - the, of, and, to, ….
  - Typically about 400 to 500 such words
  - For an application, an additional domain specific stop words list may be constructed
  - A list of English stop words
    - http://xpo6.com/list-of-english-stop-words/

- Why do we need to remove stop words?
  - Reduce indexing (or data) file size
    - Stopwords accounts 20-30% of total word counts.
  - Improve efficiency
    - Stop words are not useful for searching or text mining
    - Stop words always have a large number of hits

Stemming

- A linguistic method that reduces different variants of words to a common one (a.k.a. root/stem)
  - E.g.,
    - user
    - users
    - used
    - using
  - stem:
    - use

Benefits

- Improving effectiveness of retrieval and text mining
  - matching similar words
- Reducing indexing size
  - combing words with same roots may reduce indexing size as much as 40-50%.

Term-Document Matrix

- Most common form of representation in text mining is the *term - document* matrix
  - Term: typically a single word, but could be a word phrase like "data mining"
  - Document: a generic term meaning a source from which the text is to be retrieved
  - Can be large - terms are often 50k or larger, documents can be in the billions (www)
  - Can be binary, or use counts

Term-Document Matrix

- Example: 10 documents, 6 terms

The term "analytics" occurs 10 times in document D2

	Data	Analytics	R	Python	Business	Statistics
D1	24	21	9	0	0	3
D2	32	10	5	0	3	0
D3	12	16	5	0	0	0
D4	6	7	2	0	0	0
D5	43	31	20	0	3	0
D6	2	0	0	18	7	6
D7	0	0	1	32	12	0
D8	3	0	0	22	4	4
D9	1	0	0	34	27	25
D10	6	0	0	17	4	23

Each document row is just a vector of terms, sometimes Boolean (indicating presence/absence of terms)

TF-IDF Weighing Scheme

- The term frequency-inverse document frequency (TF-IDF) is a popular weighing scheme that identifies documents with frequent occurrences of rare terms.

$$TF - IDF = Term\ Frequency * Inverse\ Document\ Frequency$$

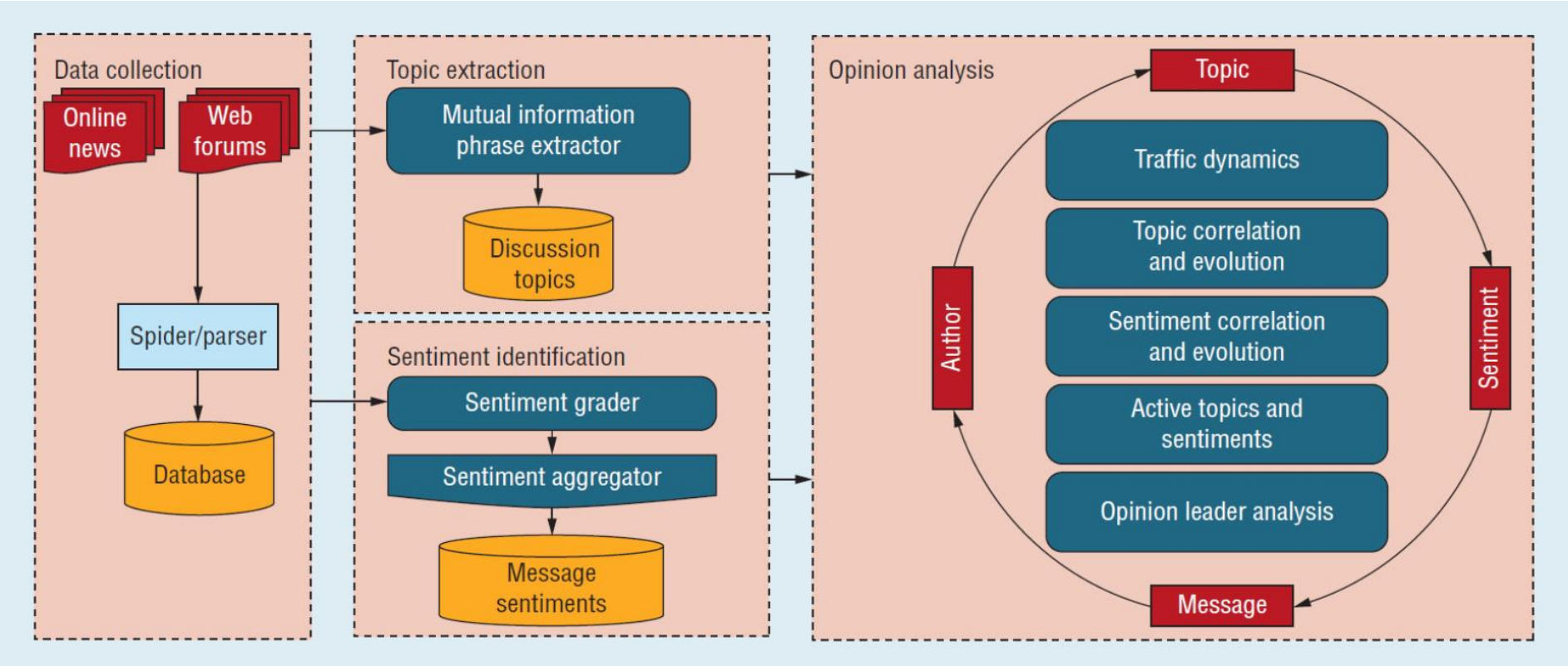
Feature Selection

- Performance of text classifiers can be optimized by choosing only a subset of term features

- Benefits of feature selection:
  - Improve efficiency of the algorithm by reducing the size of the effective vocabulary.
  - Improve classification accuracy by eliminating noise features

Case: Sentiment Analysis

Sentiment analysis (aka opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.



Text Classification Models for Sentiment Analysis

- We have a dataset with labels for training and testing
- Use supervised machine learning methods
  - Extract textual features (document-term matrix, TFIDF etc.)
    - List-based sentiment scores can be used as features
  - Predictive modeling

Text Mining Demo: Sentiment Analysis

- Refer to:
  - "Text_Classification-Sentiment_Analysis.Rmd"

Q & A