Natural Language Processing (NLP)

Discourse and Pragmatic processing

- Relationships may hold between phrases and parts of their discourse contexts, including:
- **Identical entities.** Consider the text:
  - Bill had a red balloon.
  - John wanted it.
  - The word **“it”** should be identified as referring to **red balloon.** This type of references are called **anaphora.**

Discourse and Pragmatic processing

- **Parts of entities.** Consider the text:
  - Rahul opened the book he just bought.
  - The **title page** was torn.
  - The phrase “title page” should be recognized as part of the book that was just bought.

Discourse and Pragmatic processing

- **Parts of actions.** Consider the text:
  - Sunil went on a business trip to New York.
  - He left on an early morning flight.
  - **Taking a flight** should be recognized as part of going on a trip.

Discourse and Pragmatic processing

- **Entities involved in actions**. Consider the text:
  - **They** took the TV and the stereo.
  - My house was broken into last week.

  The pronoun “they” should be recognized as referring to the burglars who broke into the house.

Discourse and Pragmatic processing

- **Elements of sets.** Consider the text:
  - The **designs** on the shirts, we have in stock are stars, the **moon**, item and a flag.
  - I'll take two moons.
  - Moons means shirts having moon design.

Discourse and Pragmatic processing

- Names of individuals:
  - Dave went to the movies.
- Causal chains
  - There was a big snow storm yesterday.
  - The schools were closed today.

Discourse and Pragmatic processing

- Planning sequences:
  - Sally wanted a new car
  - She decided to get a job.
- Implicit presuppositions (take for granted):
  - Did Mack fail IT980?

Discourse and Pragmatic processing

- We focus on using following kinds of knowledge:
  - The current focus of the dialogue
  - A model of each participant's current beliefs
  - The goal-driven character of dialogue
  - The rules of conversation shared by all participants.

Text Mining

- **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 ~80% of all the available data are unstructured data**

Text Mining (TM)

- **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 (hinders) 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

Text Mining (TM)

Bag Of Words Representation Of Text

- Considers text a simple set/bag of words
- Based on the following (unrealistic) assumptions:

?

Text Mining (TM)

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
- But, **highly effective**, and is often used in TM

Bag of Words Model (BOW)

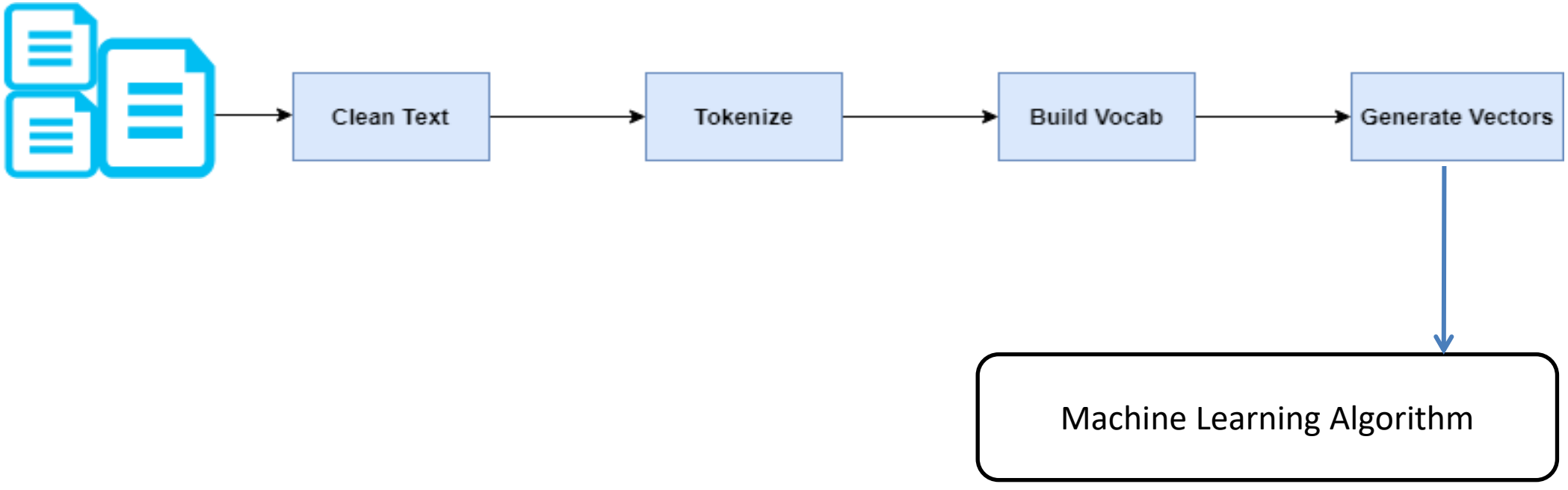
- BOW extracts features from text documents.
- These features can be used for training machine learning algorithms.
- It creates a vocabulary of unique words occurring in all the documents in the training set.
- it’s a collection of words to represent a sentence with word count and mostly disregarding the order in which they appear.

Bag of Words Model (BOW)

BOW is an approach widely used with:

- Natural language processing
- Information retrieval from documents
- Document classifications

Bag of Words Model (BOW)



Bag of Words Model (BOW)

Consider the below two sentences.

1. "John likes to watch movies. Mary likes movies too.“

2. "John also likes to watch football games."

Bag of Words Model (BOW)

These two sentences can be also represented with a collection of words.

1. ['John', 'likes', 'to', 'watch', 'movies.', 'Mary', 'likes', 'movies', 'too.']

2. ['John', 'also', 'likes', 'to', 'watch', 'football', 'games']

Bag of Words Model (BOW)

Remove multiple occurrences and use the word count.

1. {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1}

2. {"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1, "games":1}

Bag of Words Model (BOW)

- Tabular form for all documents

Words	Frequencies
John	2
Likes	3
To	2
Watch	2
Movies	2
Mary	1
Too	1
also	1
Football	1
games	1

Bag of Words Model (BOW)

Create a vector whose length is equals to total length of vocabulary

“John likes to watch movies. Mary likes movies too”
[1, 2, 1, 1, 2, 1, 1, 0, 0, 0]

“John also likes to watch football games”
[1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

t1	t2	t3	t4	t5	t6	….	…..	tn
w1	w2	w3	w4	w5	w6	…..	……	wn

Words	Frequencies
John	2
Likes	3
To	2
Watch	2
Movies	2
Mary	1
Too	1
also	1
Football	1
games	1

Which is incorrect entry ? For given sentences

"The black cat is happy"

"The black cat and dog are friends"

A. Rows 1, 2, 7

B. Rows 1,2,3,4

C. Rows1,2,5,6

D. All of these

Words	Frequencies
The	2
black	3
To	1
cat	2
is	1
happy	1
dog	1
are	1
friends	1
games	1

Which is incorrect entry ? For given sentences

"The black cat is happy"

"The black cat and dog are friends"

A. Rows 1, 2, 7

B. Rows 1,2,3,4

C. Rows1,2,5,6

D. All of these

Words	Frequencies
The	2
black	3
To	1
cat	2
is	1
happy	1
dog	1
are	1
friends	1
games	1