Text Parsing Basics

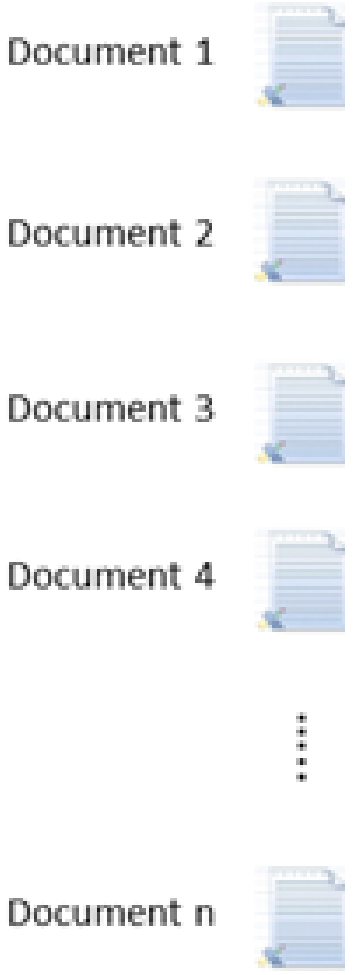
Dr. Goutam Chakraborty

Objectives

- Explain how textual data are converted to numeric data for analysis purposes.
- Explain typical terms used in text parsing such as *tokenization*, *lemmatization*, and *parts-of-speech tags*.
- Explain differences between phrase, entities, and concepts.

Goal of Text Parsing

- Converting unstructured text to spreadsheet (structured) type format for ease of analy

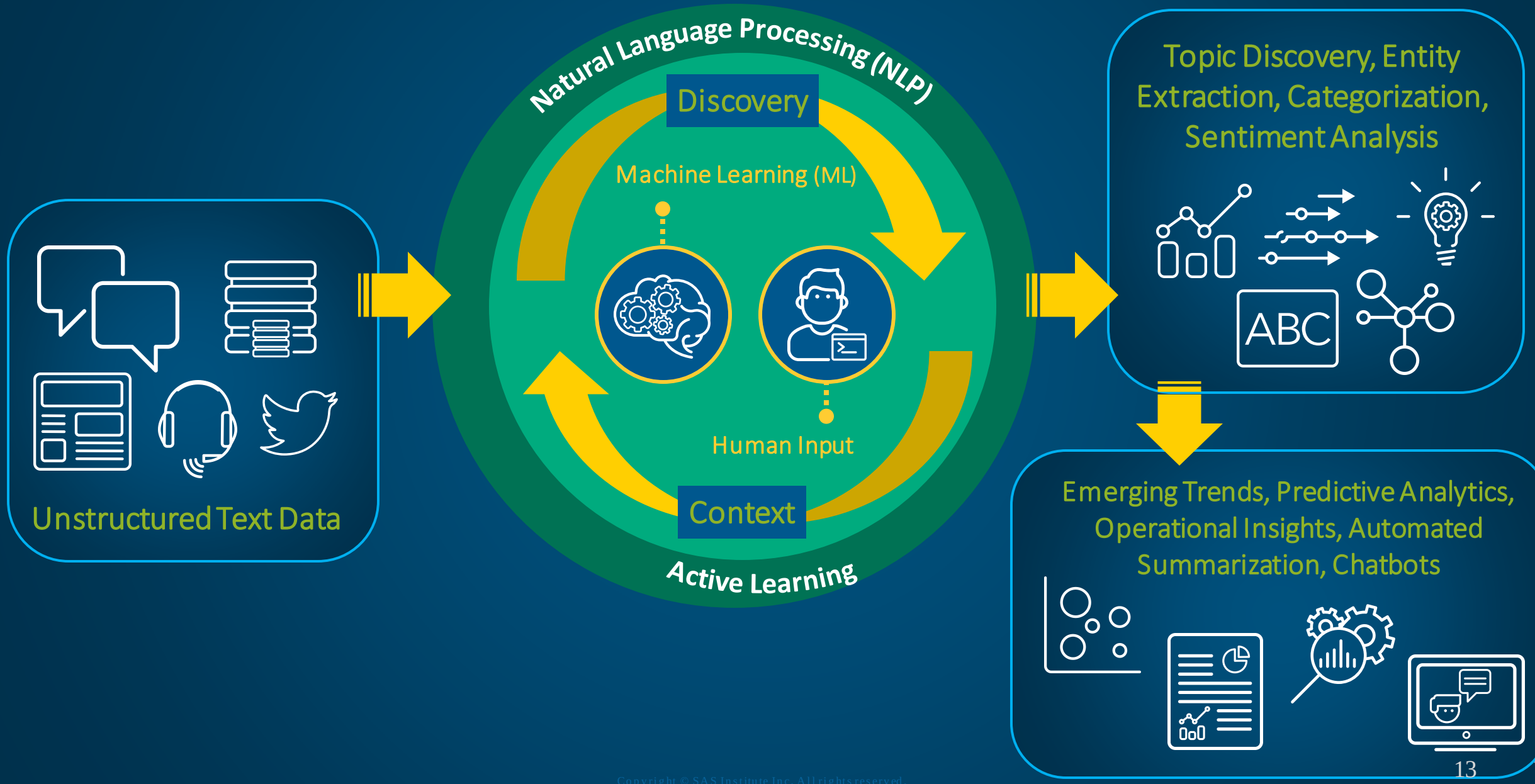


	X1	X2	X3	.....	Xp
Document 1	1	0	1	.....	1
Document 2	0	1	0	.....	0
Document 3	1	1	0	.....	1
Document 4	0	0	1	.....	1
.....	.....	.....	.....	.....	.....
Document n	1	0	0	.....	0

Approaches of Analyzing Text

- Bag-of-words approach
  - Counting of the words in a text to summarize and classify text documents
- Linguistic (natural language processing or NLP) approach
  - Taking into account syntax (structure) as well as semantics (meaning of words) for understanding and classifying text
- Hybrid approach
  - Modern text analytics applications represent a hybrid of the above two approaches plus brings in the concept of machine learning (ML) with human input (active learning)

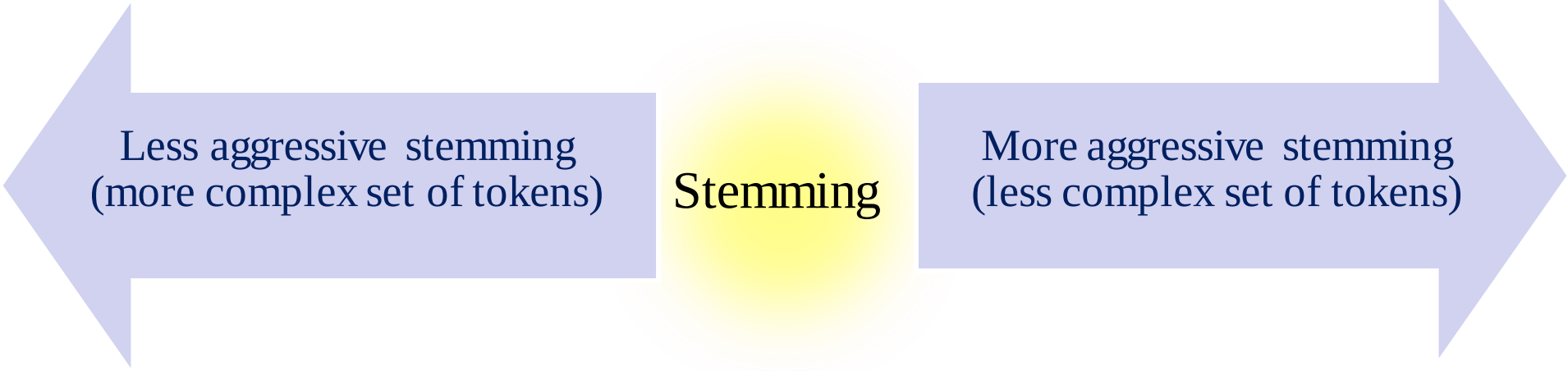
Modern Text Analytics



Copyright © SAS Institute Inc. All rights reserved.

Lemmatization

- *Lemmatization* is the process of applying normalization to tokens to reduce complexity.
- Inflectional stemming (grammatical variants) or morphological analysis
  - Singular/plural form (such as car/cars)
  - Present/past tense (such as run/ran/running)
- Stemming to a root (synonyms)
  - Car and automobile



Tokenization

- Converting a stream of characters (such as a collection of sentences in a text document) and breaking it down to tokens (units that are either a word, a number, or a punctuation mark).
- I love eating chocolate but I worry about calories in each.
  - How many tokens?
  - Frequency of tokens?

	I	Love	Eating	Chocolate	But	Worry	About	Calories	In	Each
Text 1	2	1	1	1	1	1	1	1	1	1

Parts-of-Speech (POS) Tags

- The goal of POS tags is to extract more sophisticated features from text beyond the token. It involves labeling each token (word) in a sentence with its appropriate part-of-speech such as the following:
  - Noun
  - Verb
  - Adjective
  - Adverb
  - Preposition
  - Conjunction
  - And many more

An Example of POS Tags

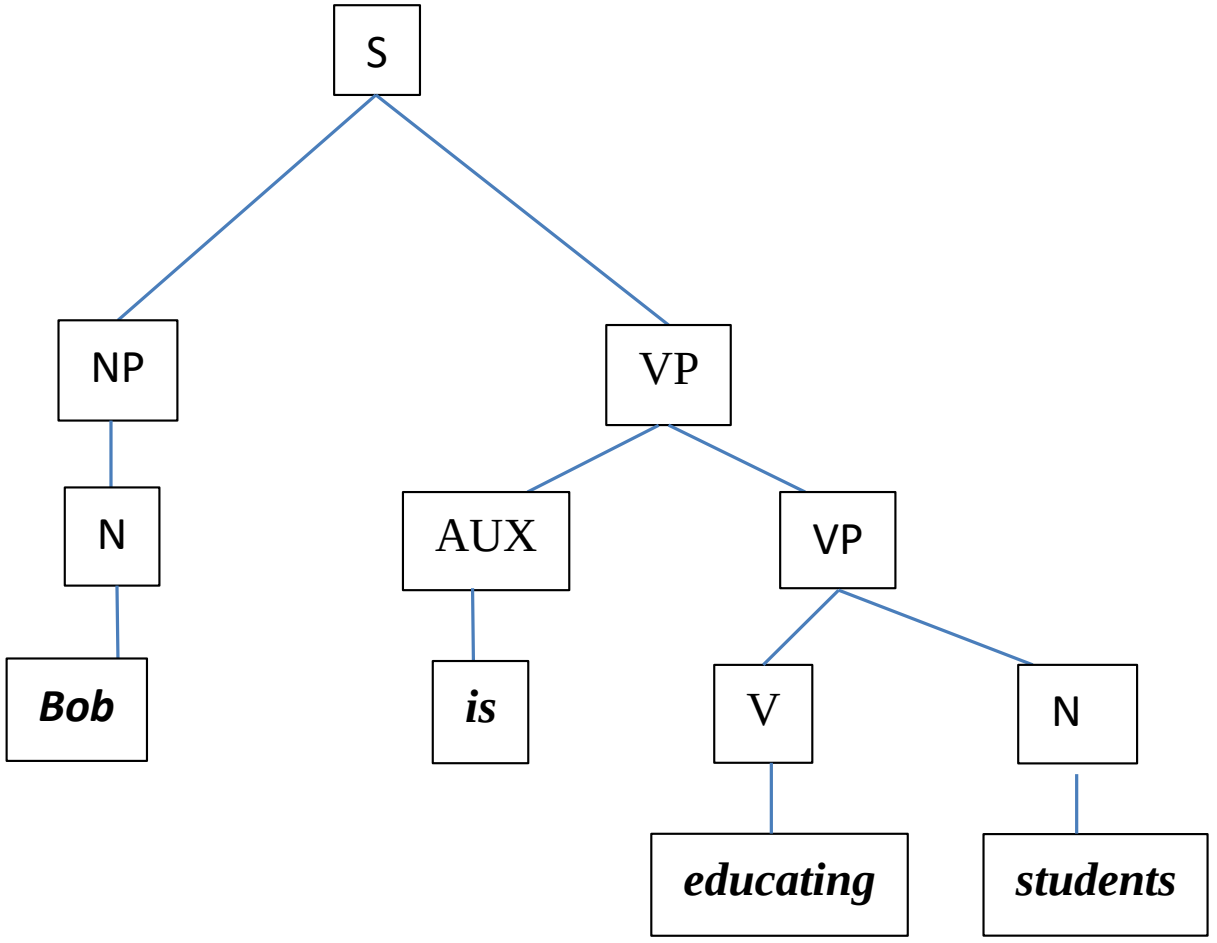
- "The waiter placed 4 glasses on the table"

Number	Tag	Description
1.	AT	Article
2.	CD	Cardinal number
3.	IN	Preposition
4.	JJ	Adjective
5.	JJR	Adjective, comparative
6.	NN	Noun, singular, or mass
7.	NNS	Noun, plural
8.	RB	Adverb
9.	RBR	Adverb, comparative
10.	VB	Verb, base form
11.	VBD	Verb, past tense
12.	VBG	Verb, gerund, or present participle
13.	VBN	Verb, past participle

The (**AT**) waiter (**NN**) placed (**VBD**) 4 (**CD**) glasses (**NNS**) on (**IN**) the (**AT**) table (**NN**)

Parsing Tree (Full Parse)

- Bob is educating students



https://medium.com/greyatom/learning-pos-tagging-chunking-in-nlp-85f7f811a8cb

continued...

Phrase and Entity Recognition

- The goal is to recognize particular types of noun phrases such as persons, organizations, and locations along with numeric entities such as money, date, and time.
- Features for entity recognition include
  - Lexicons
  - Word shape
  - POS tags.

Phrase and Entity Recognition

- Approaches in phrase and entity recognition include
  - **Rule-based** models
    - Use a combination of lexicons, word shapes, and grammar rules to identify named entities
    - Example " a capitalized noun following Ms. is likely to be a named entity"
  - **Statistical** models
    - Typically, these are supervised classification models that predict the probability that a word corresponds to a particular class label (such as person, location, and organization)
    - Models are trained on training data (that contains true labels) and then the probabilities are applied to new text.

Statistical Models for Entity Recognition

- Hidden Markov Model (HMM):
  - HMM attempts to model the most likely sequence of labels given a sequence of terms by maximizing the joint probability of terms and labels.
- Conditional Random Fields (CRF):
  - CRF are designed to model the conditional probability of a label sequence given a sequence of words.
  - Typically, CRFs can be estimated more efficiently by using less training data than HMMs.

Entities in SAS Text Miner

Standard Entity	Description	Example
ADDRESS	Postal address or a street name	2526 Carywood Drive
COMPANY	Name of a company	SAS Institute Inc.
CURRENCY	Currency or Currency Expression	$450,000 or 6000 USD
DATE	Full Date, Year, Month, or Day	1st Oct 1949 or January or 1983
INTERNET	URL of a website or e-mail address	http://www.sas.com or support@sas.com
LOCATION	Name or a city, state, country, or any other geographical place or region	Istanbul or Malaysia or Utah
MEASURE	Measurement or measurement expression	100 miles or 25 sec
ORGANIZATION	Name of a government or service agency	FBI – Federal Bureau of Investigation or SEC – Securities Exchange Commission
PERCENT	Percentage or percent expression	45% or 20 PERCENT
PERSON	Name of a person	Abraham Lincoln or John F. Kennedy
PHONE	Phone number in a standard format	1-XXX-XXX-XXXX or (XXX)XXX-XXXX
PROP_MISC	Proper noun with an ambiguous classification	U.S. President George Bush
SSN	Social Security number in a standard format	ZZZ-ZZ-ZZZZ
TIME	Time or time expression	2:45 pm 14:45
TIME_PERIOD	An expression of measured time	10 days or 5 years
TITLE	Title of a person or a position	Mr. or Ms. or Dr.
VEHICLE	Make, Name, Model, Year, and Color of a motor vehicle	Honda Accord EX 2005 Black

Concept Extraction

- A simple concept can be defined as an entity such as the name of a place, person, organization, or location.
  - An example of a concept of a place is New York, or an organization is SAS Institute Inc.
- A relational concept is defined as two or more entities in relation with each other.
  - For example, “CEO of SAS Dr. Jim Goodnight” is a relational concept.
- Concept extraction is best handled in SAS by using SAS Content Categorization Studio.