insAnalytics

Consulting. Training. Research & Development.

Insights.. Through Analytics…

Data Mining

Introduction to Text Mining

Key Objectives

After successful completion of the topic, participants will be able to:

- ❑ Develop awareness of **Text Mining** as a technique to analyze **free text data**

- ❑ Articulate the **key concepts** of **Text Mining process**
- ❑ Understand the different text processing methods like, **removal of special characters**, **abbreviations**, **stop words**, **parts-of-speech tagging**, **stemming** etc.
- ❑ Articulate the concepts like "**bag of words**", "**term-document matrix**", "**frequency of a word**", "**occurrence of a word**", "**association between two words**", etc.
- ❑ **Gather a basic idea** of different analysis done under text mining like **Taxonomy, Link Analysis, Sentiment Analysis** with example
- ❑ Get a ground level hands on activity on solving text mining using R like
  - ❑ Creation of Word Cloud
  - ❑ Finding some basic sentiment measure of a feedback, etc.

**To learn & understand the subject matter better,
participants need to be aware of the following areas:**

- ❑ Descriptive Statistics
- ❑ Statistical Inference
- ❑ Data Mining – CRISP DM Methodology

- ❑ Logistic Regression
- ❑ Classification Techniques
- ❑ Cluster Analysis

Introduction to Text Mining

What is Text Mining

Text Mining Process

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery

Introduction to Text Mining

What is Text Mining

Text Mining Process

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery

insAnalytics

Introduction to Text Mining

What is Text Mining

Text Mining Process

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery

What is Text Mining – An Illustration

What is Text Mining – Definition

Search vs. Discover

Comparison with Other Analytical Methods

Few Applications

Popular Text Mining Tools

What is Text Mining – An Illustration

SmartCommunicate, a leading mobile manufacturer, has recently launched a new smart phone in the market.

After 3 months from launch Mark, the CMO of SmartCommunicate, wants to know what people are thinking about the phone. The easiest way to learn this is to log on to different social media like Facebook, or Twitter, etc. and look at the comments.

But, Mark found there are so many different type of comments made about the product in the social media. It was too huge to read through and find any conclusion out of them.

He then asks Charlie, the analyst in his team, if she can analytically find some inference out of these comments.

How Charlie can help?

What is Text Mining – An Illustration

Charlie used some software to extract all the comments to a file in her machine.

She found more that 1,000 comments of various length. Different kinds of words are used.

Hence, she has to use the data mining technique called **Text Mining**

How Experts Define

> **"** The objective of Text Mining is to exploit information contained in textual documents in various ways, including …discovery of patterns and trends in data, associations among entities, predictive rules, etc.**"**

(Grobelnik et al., 2001)

> **"**Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known.**"**

(Hearst, 1999)
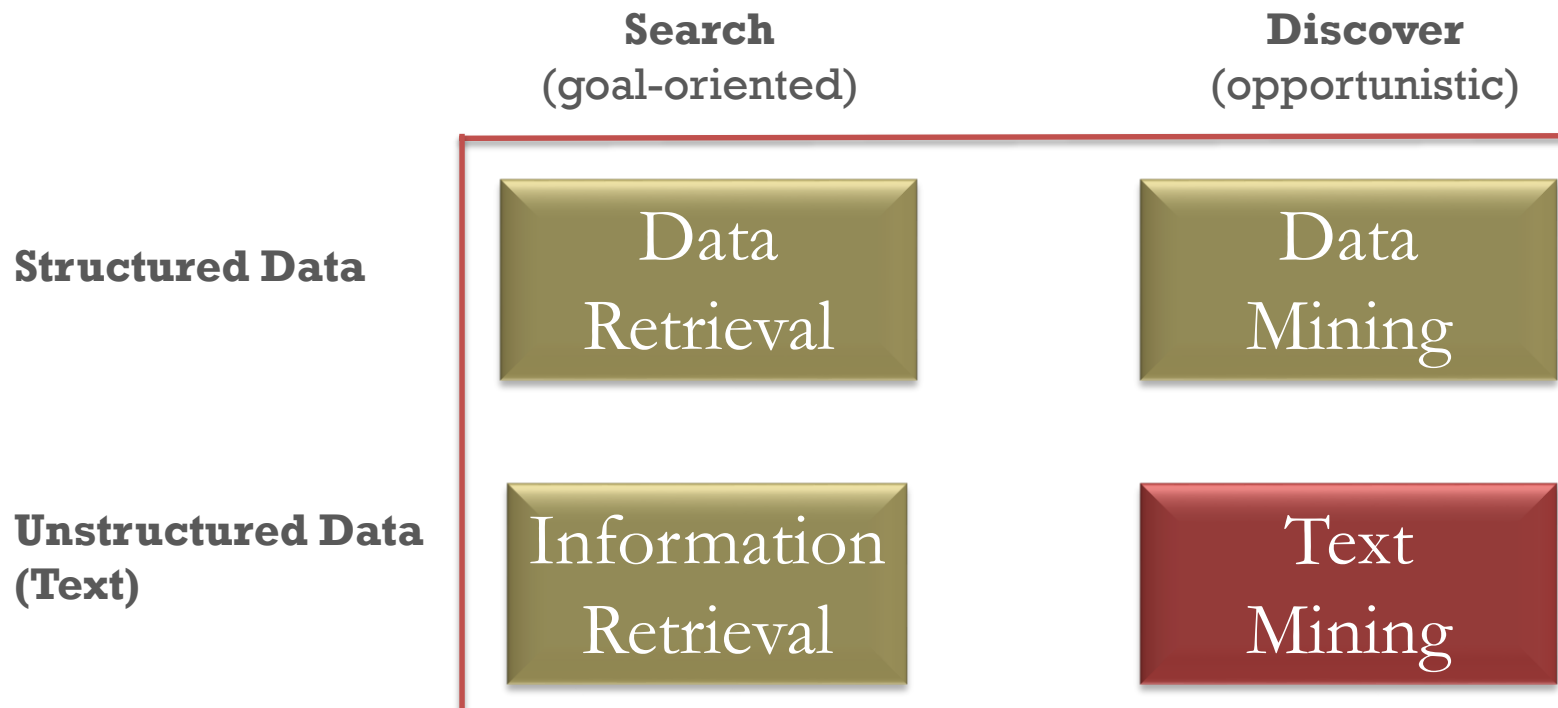
As We Define Text Mining

Text mining is the process of **extracting implicit, previously unknown potentially useful information from free-text data**.
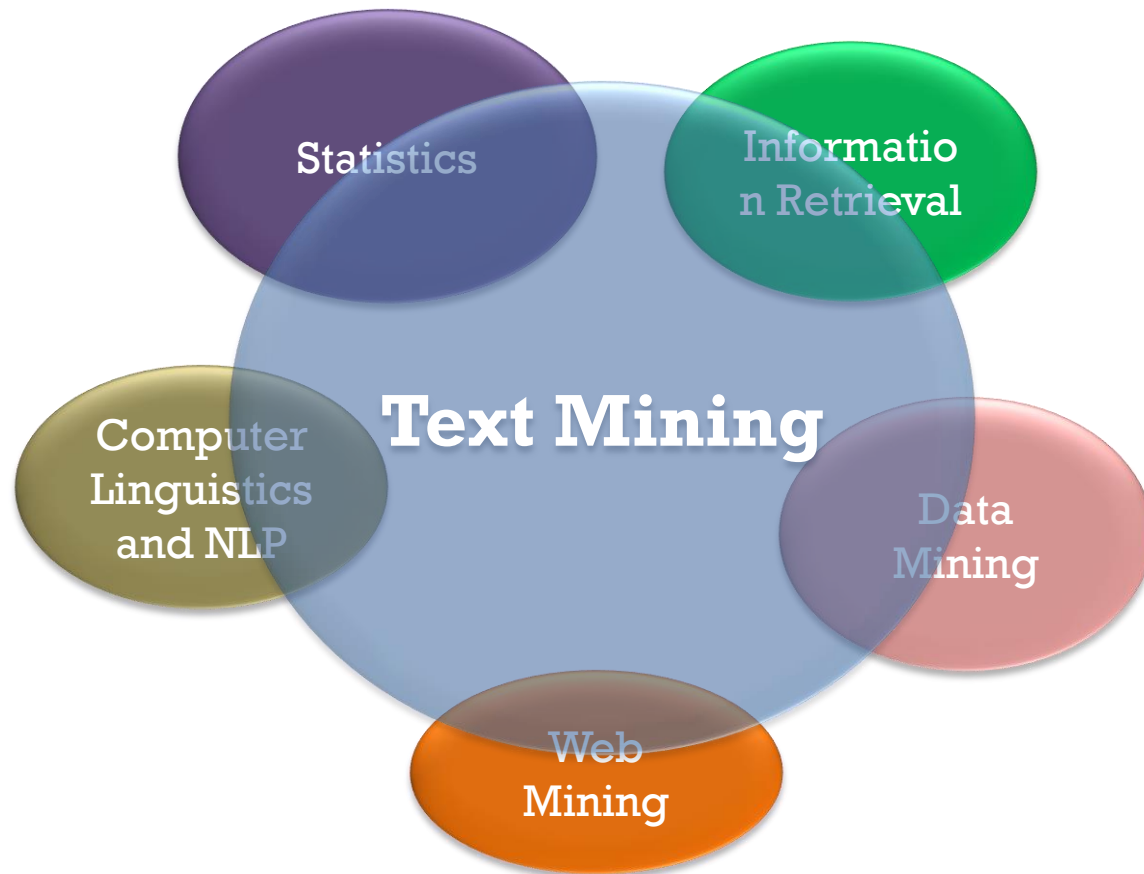
The information extracted has to be:

- ❑ **Implicit**
- ❑ **Previously unknown**
- ❑ **Potentially useful**

Search vs. Discover

	Search (goal-oriented)	**Discover** (opportunistic)
Structured Data	Data Retrieval	Data Mining
Unstructured Data (Text)	Information Retrieval	Text Mining

Text Mining – Comparison with Other Analytical Methods

Text Mining – Few Applications

Marketing	**Discover distinct groups of potential buyers according to various users text based profile**
Customer Feedback	**Understand the customer sentiment based on the feedback received**
Social Media	**Understand the highly discussed topics in social media like Facebook, Twitter etc.**
Spam Filtering	**Identify if a mail is spam mail or not**
Document Categorization	**Identify the topic of a document**

Popular Text Mining Tools

PolyAnalyst **6**

abc

IBM SPSS Modeler 15.0

SAS® Text Miner

CLARABRIDGE

LEXALYTICS

R

ATTENSITY

an attensity group company

Introduction to Text Mining

What is Text Mining ✓

Text Mining Process

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery

Data Mining

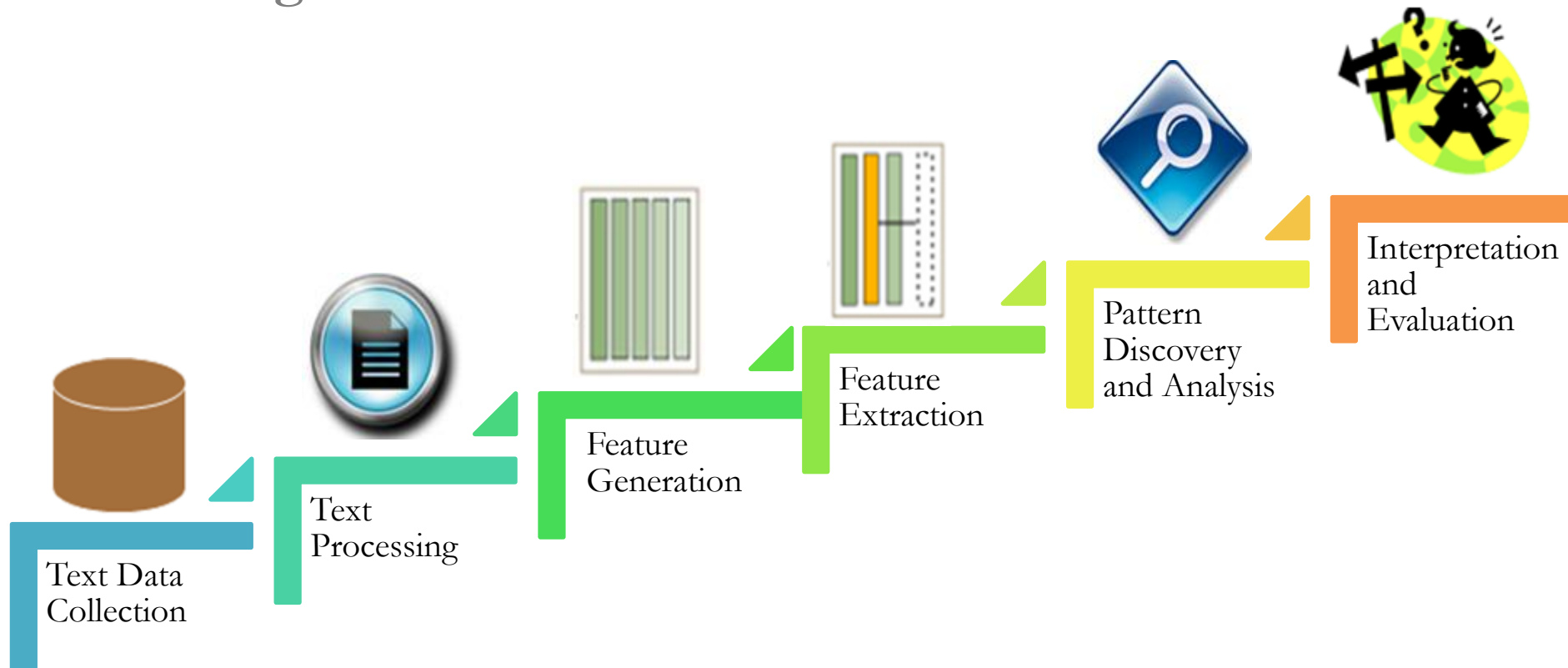

Introduction to Text Mining

What is Text Mining ✓

Text Mining Process

Text Processing

Feature Generation

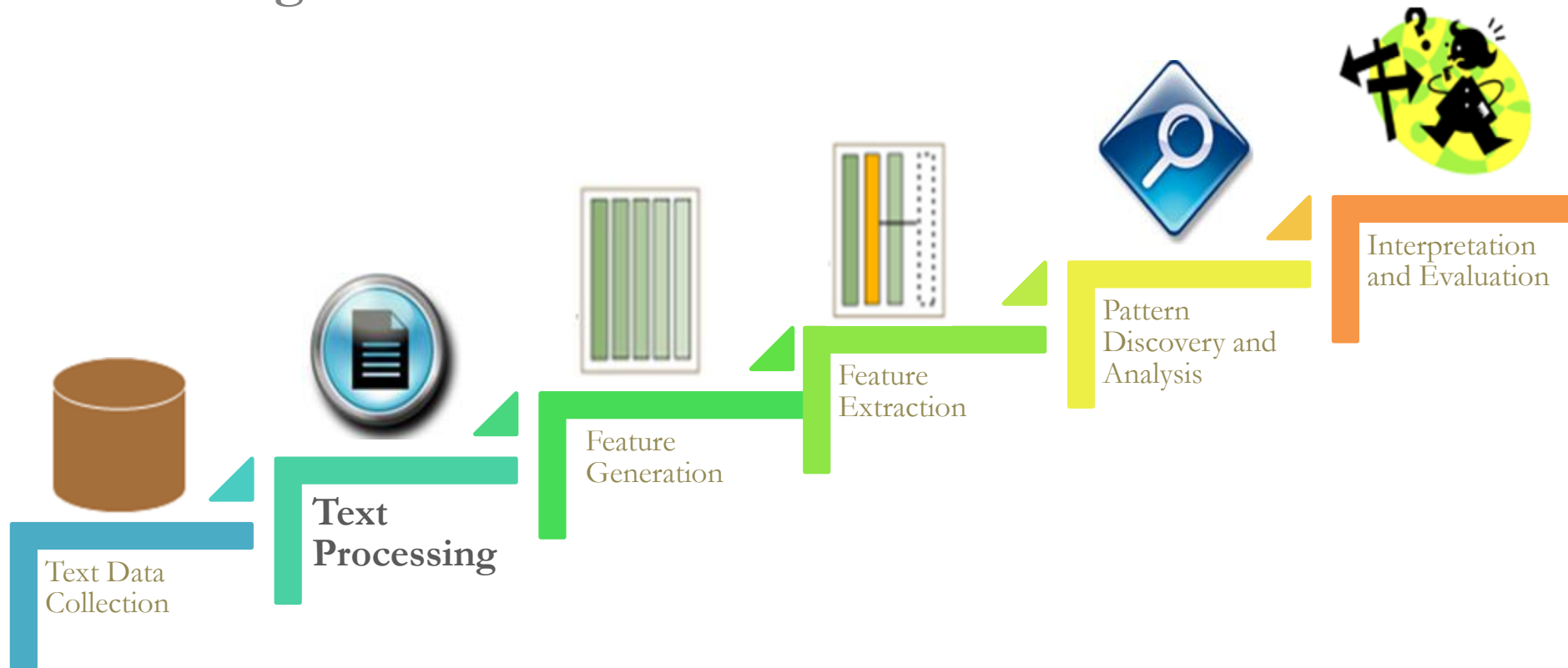Feature Extraction

Pattern Discovery

Text Mining Process Flow

Text Data
Collection

Text
Processing

Feature
Generation

Feature
Extraction

Pattern
Discovery
and Analysis

Interpretation
and
Evaluation

Text Mining Process Flow

Text Data Collection

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery and Analysis

Interpretation and Evaluation
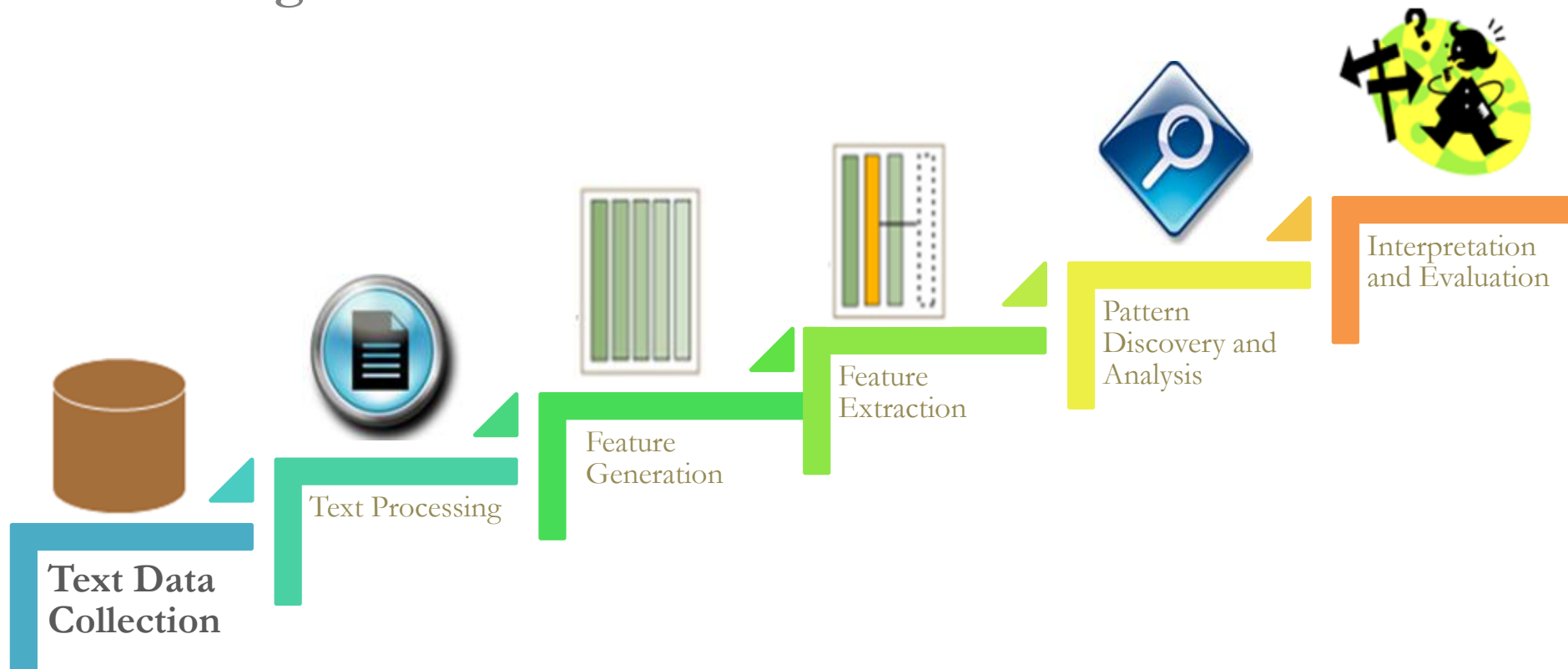
The free text data is the collected from the source.

The source can be **customer complain log** or **twitter feeds** or **social media comments** or **blogs** or **documents** or **feedback to a survey**, etc.

Text Mining Process Flow

Text Data Collection

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery and Analysis

Interpretation and Evaluation

After collection of data, the text is processed or cleaned

Text Mining Process Flow

Text Data Collection

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery and Analysis

Interpretation and Evaluation
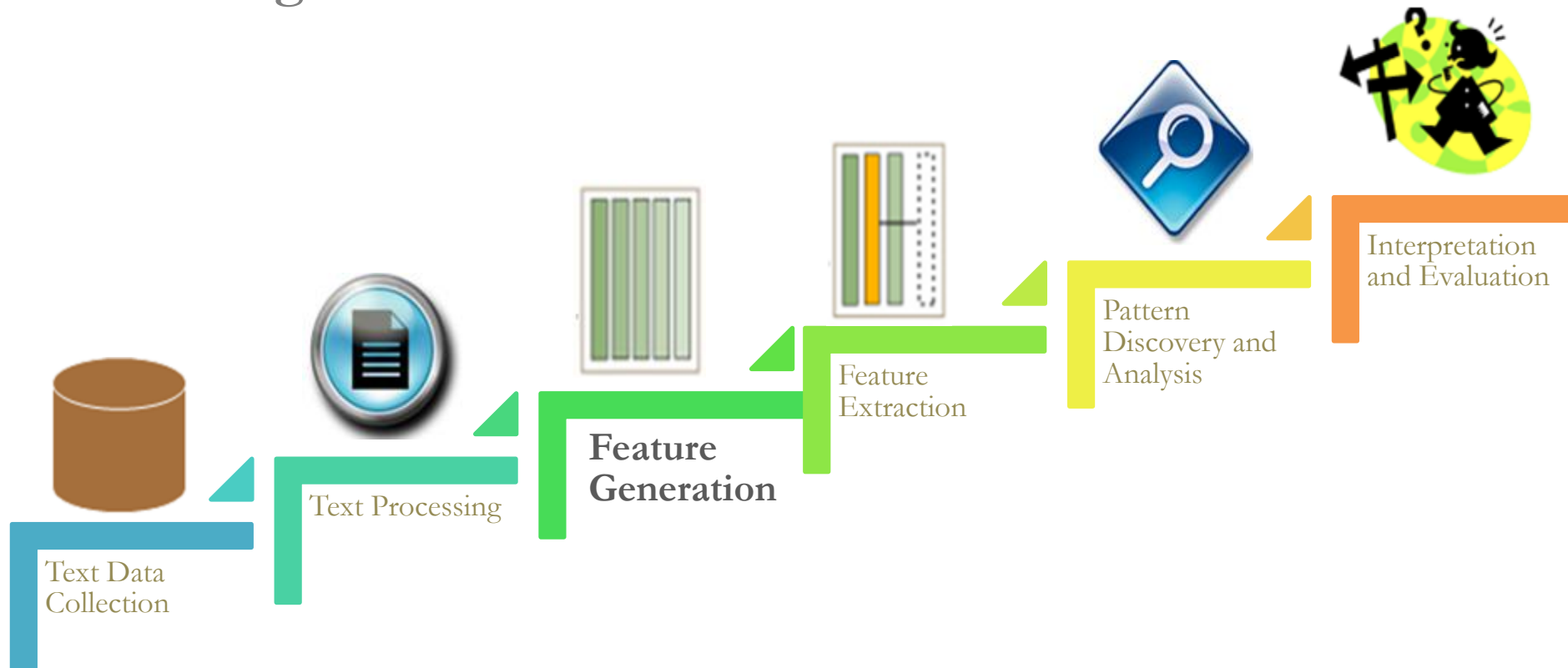
This is the first stage of analysis, where we find the words or phrases used in the texts.

Text Mining Process Flow

Interpretation and Evaluation

Pattern Discovery and Analysis

Feature Extraction
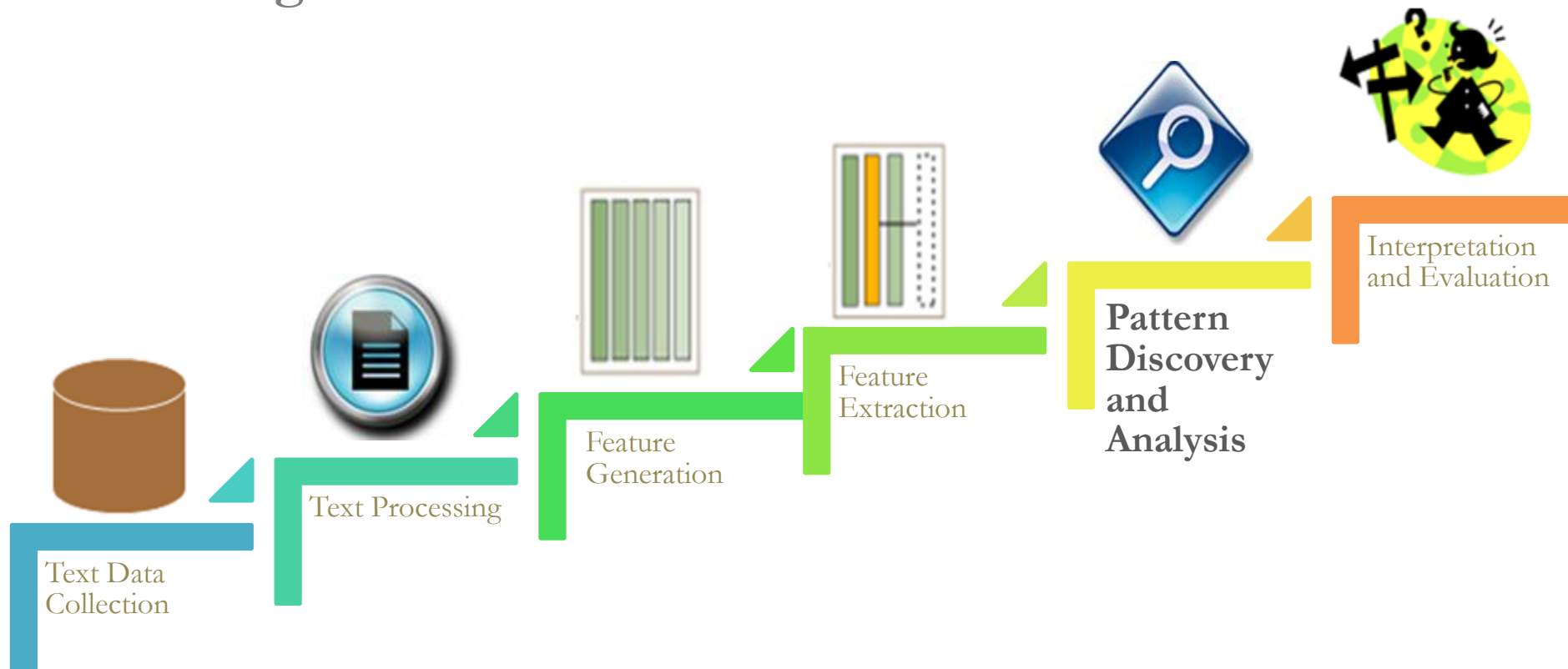
Feature Generation

Text Processing

Text Data Collection
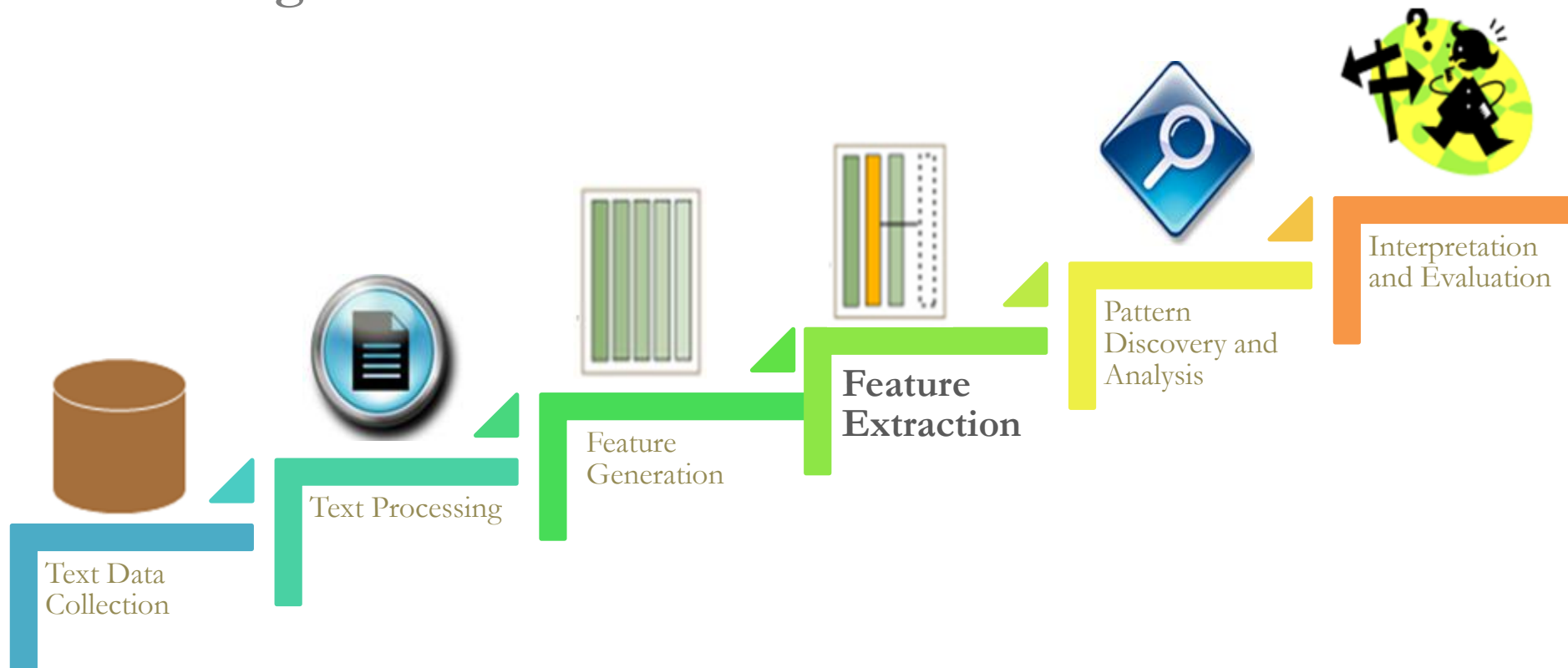
Here, we find what is being discussed. We find the frequently used proper nouns here.

For example, if the data is about the feedback of a tourist place, what tourist spots are being discussed. If it is a feedback on a car, what model is being discussed

Text Mining Process Flow

Text Data Collection → Text Processing → Feature Generation → Feature Extraction → **Pattern Discovery and Analysis** → Interpretation and Evaluation
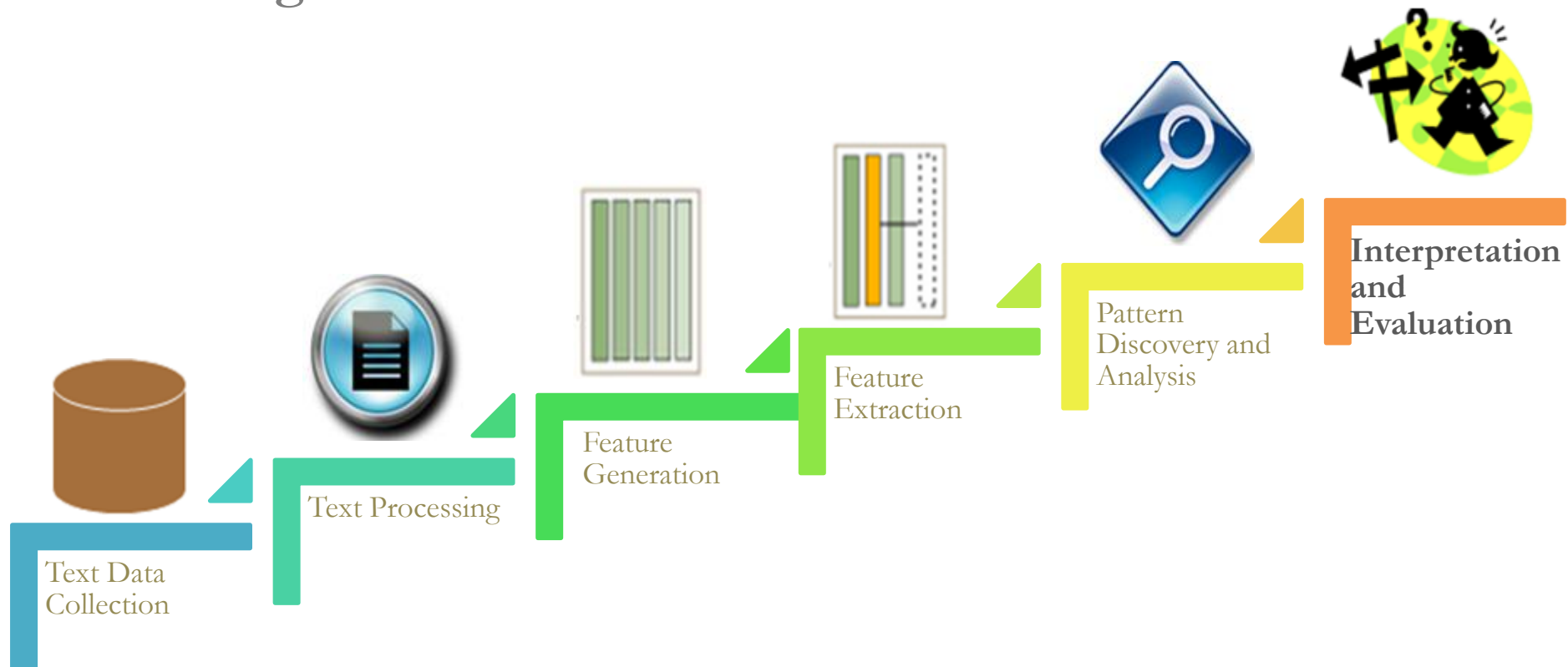
Here we use different analytical techniques to understand the hidden pattern.

For example, the link or association between two words/ phrases, sentiment of the comments, etc.

Based on the need and scenario we can employ supervised or unsupervised learning techniques

Text Mining Process Flow

Text Data Collection

Text Processing

Feature Generation

Feature Extraction

Pattern Discovery and Analysis

Interpretation and Evaluation

Finally we find conclusions from the analysis and evaluate the findings.

Introduction to Text Mining

What is Text Mining	✓
Text Mining Process	✓
Text Processing	
Feature Generation	
Feature Extraction	
Pattern Discovery	

Data Mining
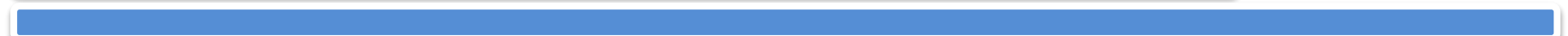
Introduction to Text Mining

What is Text Mining	✓
Text Mining Process	✓

Text Processing

Feature Generation	Text Processing – What is Done
	Text Processing – Removal of Special Characters
Feature Extraction	Text Processing – Abbreviations
	Text Processing – Synonyms and Stemming
Pattern Discovery	Text Processing – Parts of Speech Tagging

Text Processing – What is Done Here

As text mining is all about working with free text data,

it in very important to process the data and make it free from any kind of undesirable text

The text processing includes:

✓ **Removal of any special characters**

✓ **Replace any abbreviation by the actual word**

✓ **Identify the synonymous words used**

✓ **Identify the "parts of speech" of each of the words used in each sentence**

insAnalytics

Text Processing – Removal of Special Characters

As the data is a free text data, we often get some special characters in the text.

The first step to data processing is removal of the special characters like @, #, $, % etc.

Before Processing	After Processing
"What a nice place!!!!.... We were amazed__to watch setting sun @ sun set point!!"	"What a nice place! We were amazed to watch setting sun sun set point!"
"The food was horrible! # %@%#$$#!!!??"	"The food was horrible!"

Text Processing – Abbreviations

Often the data contains about abbreviations of different words.

For example, "account" written as "acct" or "A/C" or "acnt" or "accnt" or in many other ways. We need to change them to the correct format.

Before Processing	After Processing
"Just came across this web <u>thru</u> a friend of mine."	"Just came across this web <u>through</u> a friend of mine."
"<u>r u</u> available?"	"<u>Are</u> <u>you</u> available?"
"As I typed <u>pwd</u> wrong for 3 times, my <u>acctn</u> got locked!"	"As I typed <u>password</u> wrong for 3 times, my <u>account</u> got locked!"

Text Processing – Synonyms

Often different words to mean the same thing. The synonymous words needs to be identified and marked as synonyms.

Before Processing	After Processing
"I returned the product as it was not working properly. But, my money has not been refunded yet!! "	"I _returned_ the product as it was not working properly. But, my money has not been _refunded_ yet!! "
"My contribution to the teams success was overlooked. Also the extra efforts I gave during project closure, was ignored."	"My contribution to the teams success was _overlooked_. Also the extra efforts I gave during project closure, was _ignored_."

Text Processing – Parts of Speech (POS) Tagging

Often the same word can be used as noun as well as verb.

Hence, we need to tag them.

This is done through in-built process of processing the language in the software we are using. This kind of processing is called **natural language processing** (**NLP**).

Before Processing	After Processing	Accuracy
"The sporty touch given to the exteriors similar to new model makes it adorable compared to the old model."		

Text Processing – Parts of Speech (POS) Tagging

Often the same word can be used as noun as well as verb.

Hence, we need to tag them.

This is done through in-built process of processing the language in the software we are using. This kind of processing is called **natural language processing** (**NLP**).

Before Processing	After Processing	Accuracy
"The sporty touch given to the exteriors similar to new model makes it adorable compared to the old model."	"The (*article*) sporty (*adjective*) touch (*verb*) given (*verb*) to (*preposition*) the (*article*) exteriors (*noun*) similar (*adjective*) to (*preposition*) new (*adjective*) model (*noun*) makes (*verb*) it (*preposition*) adorable (*adjective*) compared (*verb*) to (*preposition*) the (*article*) old (*adjective*) model (*noun*)."	

Text Processing – Parts of Speech (POS) Tagging

Often the same word can be used as noun as well as verb.

Hence, we need to tag them.

This is done through in-built process of processing the language in the software we are using. This kind of processing is called **natural language processing** (**NLP**).

Before Processing	After Processing	Accuracy
"The sporty touch given to the exteriors similar to new model makes it adorable compared to the old model."	"The (*article*) sporty (*adjective*) **touch (*verb*)** given (*verb*) to (*preposition*) the (*article*) exteriors (*noun*) similar (*adjective*) to (*preposition*) new (*adjective*) model (*noun*) makes (*verb*) it (*preposition*) adorable (*adjective*) **compared (*verb*)** to (*preposition*) the (*article*) old (*adjective*) model (*noun*)."	touch : should be noun, not verb compared: should be adjective not verb So, 17 out of 19 tagging were correct. Accuracy is 89.5%

Introduction to Text Mining

What is Text Mining	✓
Text Mining Process	✓
Text Processing	✓
Feature Generation	
Feature Extraction	
Pattern Discovery	

Introduction to Text Mining

What is Text Mining ✓

Text Mining Process ✓

Text Processing ✓

Feature Generation

Feature Extraction

Pattern Discovery

Bag of Words

Spell Correction

Phrases

Stemming of Words

Removal of Stop Words

Bag Of Words

The words used in the texts are now separated from each other. This is the first step to identify what the comment is talking about.

We list the distinct words used in the texts. The list is called "**Bag of Words**"

Remember, often tools consider a series of uninterrupted (i.e. without any space or special characters) alphabets or numbers as words. **It does not look at the meaning of the words**.

Bag Of Words – Illustrations

Text	Distinct Words
"Recently I started with, again after almost 10 years, for second time, The God of Small things. This book is & will always be very close to my heart."	{"recently", "I", "started", "with", "again", "after", "almost", "10", "years", "for", "second", "time", "The", "God", "of", "small", "things", "this", "book", "is", "will", "always", "be", "very", "close", "my", "heart"}

Text	Distinct Words
"The sporty touch given to the exteriors similar to new model makes it adorable compared to the old model."	{"the", "sporty", "touch", "given", "to", "exteriors", "similar", "new", "model", "makes", " it", "adorable", "compared", "old"}

Spell Correction

As the text we are using is free text. Often people make spelling mistakes. Hence we need to identify the spelling errors and correct them.

For this we need a pre-loaded dictionary in the tool or we need to upload a similar library externally.

Text	Before Spell Corrections	Before Spell Corrections
"i have read it aftr watching 3 idiots. must say chetan had done a grt job. much interestng and preety good story."	{"i", "have", "read", "it", **"aftr"**, "watching", "3", "idiots", "must", "say", " chetan", "had", "done", "a", **"grt"**, "job", "much", **"interestng"**, "and", **"preety"**, "good", "story"}	{"i", "have", "read", "it", **"after"**, "watching", "3", "idiots", "must", "say", " chetan", "had", "done", "a", **"great"**, "job", "much", **"interesting"**, "and", **"pretty"**, "good", "story"}

Phrases

Often people use phrases to explain their thoughts rather than a single word. Hence, phrases makes more sense in the text than the individual words

We extract phrases used in the text. For the tool **any two consecutive words makes a phrase.**

Text	Distinct Words
"The sporty touch given to the exteriors similar to new model makes it adorable compared to the old model"	{"the sporty", "sporty touch", "touch given", "given to", "to the", "the exteriors", "exteriors similar", "similar to", "to new", "new model", "model makes", "makes it", "it adorable", "adorable compared", "compared to", "the old", "old model"}

Stemming of Words

Often in language one word can have different forms.

- Nouns can be singular or plural
- Verbs can have different tenses
- Verbs can be transformed to noun or adjectives

We need to identify the different format of a word being used in the text

Text	Distinct Words
""If you **_visiting_** Chennai for tour, you should probably not miss Marina beach." "I have **_visited_** Chennai for interview"	"visit", "visited", "visiting", "visitor" → "visit"

Removal of Stop Words

There are some words which are used in the text just to construct the sentences. But they occur so frequently in different text, that they do not help us to do any analysis or sometime may mislead our analysis.

Prepositions, **articles**, **punctuations**, **compositions**, and **interjections** are removed.

These are together called **stop words**.

Text	After Processing
"Recently I started _with_, again _after_ almost 10 years, _for_ second time, _The_ God _of_ Small things_. This_ book is _&_ will always be very close _to_ my heart."	"Recently I started, again almost 10 years, second time God Small things book is will always be very close my heart"

Introduction to Text Mining

What is Text Mining ✓

Text Mining Process ✓

Text Processing ✓

Feature Generation ✓

Feature Extraction

Pattern Discovery

© insAnalytics

Introduction to Text Mining

What is Text Mining ✓

Text Mining Process ✓

Text Processing ✓

Feature Generation ✓

Feature Extraction

Pattern Discovery

Entity Extraction

Term Document Matrix

Entity Extraction

Under "feature generation" we primarily clean the text data and make them ready for analysis.

In "Feature Extraction" we first try to understand the data. For example, we try to understand "what the texts are talking about"

As the first step towards that, we try to find the followings from the texts
- Names of individuals
- Price
- Names of places
- Names of products
- Names of entities

These words are different from other English words and should be treated differently.

These words are called "entity"

Entity Extraction– Example

Text	Distinct Entities
"What a phone god it is worth the money I got it for 45.589 rupees but what an investment! It is a phone in which you can do anything but some people refer iphone but the note 3 is better. I work in a very big company in Delhi so I check my email to but I use it more for doing work. The cheapest place to buy it is India because in UK and New Zealand it costs more than 60000."	• <u>Name of Places:</u> Delhi, India, UK, New Zealand • <u>Products:</u> iPhone, Note 3 • <u>Number:</u> 60000, 45.589
"Unlike The Namesake, the book has stories that are just not about a bunch of displaced people coping with life in an alien land, dealing with the emotional conundrum of cultural values. "Stories of Boston, Bengal and beyond "the maiden book of Jhumpa Lahiri, winner of the most coveted Pulitzer Prize has a lot more life to it. Her tales are vivid, sumptuous slices of human experiences put together in stories set across different context of time and places."	• <u>Name of Person:</u> Jhumpa Lahiri • <u>Name of Place:</u> Boston, Bengal • <u>Name of Prize:</u> Pulitzer Prize • <u>Name of Book:</u> Namesake

Term Document Matrix

We have the list of English words used in the texts.

Now we need to find which words are used in which documents.

For that we create the "**Term Document Matrix**" – a matrix where

      number of rows is equal to number of texts and

      number of columns is the number of distinct words used in all texts combined

      the value cell (i,j) indicates how many times the word 'j' has occurred in the ith text

Term Document Matrix – Example

Let us take the following 4 feedbacks on some book.

Text	Distinct Words	Distinct Words (after removal of stop words)
This is an excellent story. I recommend this to all."	"this", "is", "an", "excellent", "story", "I" "recommend", "this", "to", "all", "superb", "book", "read", "twice", "a", "must", "awesome"	"excellent", "story", "I" "recommend", "all", "superb", "book", "read", "twice", "must", "awesome"
"Superb book. I read this book twice."		
"A must read book!!"		
"Awesome story!!"		

How the term-document matrix will look like?

Term Document Matrix – Example

We created the following term document matrix.

Sl No	Text
1	This is an excellent story. I recommend this to all."
2	"Superb book. I read this book twice."
3	"A must read book!!"
4	"Awesome story!!"

Sl No	all	awsome	book	excellent	i	must	read	recommend	story	superb	twice
1	1	0	0	1	1	0	0	1	1	0	0
2	0	0	2	0	1	0	1	0	0	1	1
3	0	0	1	0	0	1	1	0	0	0	0
4	0	1	0	0	0	0	0	0	1	0	0

Introduction to Text Mining

What is Text Mining	✓
Text Mining Process	✓
Text Processing	✓
Feature Generation	✓
Feature Extraction	✓
Pattern Discovery	

Introduction to Text Mining

What is Text Mining	✓
Text Mining Process	✓
Text Processing	✓
Feature Generation	✓
Feature Extraction	✓

Pattern Discovery

Few Measures

Pattern Discovery – Some Use Cases

Few Measures – Frequency and Occurrence

Based on the term-document matrix we define some measures.

Frequency of a word is defined by the **number of times a word has appeared** in the entire dataset

Occurrence of a word is defined by the **number of documents/ comments / rows a word has appeared**

Few Measures – Frequency and Occurrence

Based on the term-document matrix we define some measures.

Frequency of a word is defined by the number of times a word has appeared in the entire dataset

Occurrence of a word is defined by the number of documents/ comments / rows a word has appeared

In the above example,

Sl No	all	awsome	book	excellent	i	must	read	recommend	story	superb	twice
1	1	0	0	1	1	0	0	1	1	0	0
2	0	0	2	0	1	0	1	0	0	1	1
3	0	0	1	0	0	1	1	0	0	0	0
4	0	1	0	0	0	0	0	0	1	0	0

The word "book" has frequency 3 but occurrence 2

Few Measures – Support

Based on the term-document matrix we define some measures.

Frequency of a word is defined by the number of times a word has appeared in the entire dataset

Occurrence of a word is defined by the number of documents/ comments / rows a word has appeared

We also define, **support** as the **percentage of documents the word has occurred**, i.e. occurrence divided by the number of records.

In the example, support of "book" is 2/4= 50%

Few Measures – Association of Two Words

"Association" measures how often two words (say, A and B) occur together. The measure is given as below:

$$Association\ between\ A\ \&\ B = \frac{\#\ of\ texts\ where\ both\ A\ \&\ B\ occured}{\#\ of\ texts\ where\ either\ A\ or\ B\ occured}$$

Pattern Discovery – Association of Two Words

"Association" measures how often two words (say, A and B) occur together. The measure is given as below:

$$Association\ between\ A\ \&\ B = \frac{\#\ of\ texts\ where\ both\ A\ \&\ B\ occured}{\#\ of\ texts\ where\ either\ A\ or\ B\ occured}$$

Sl No	all	awsome	book	excellent	i	must	read	recommend	story	superb	twice
1	1	0	0	1	1	0	0	1	1	0	0
2	0	0	2	0	1	0	1	0	0	1	1
3	0	0	1	0	0	1	1	0	0	0	0
4	0	1	0	0	0	0	0	0	1	0	0

In the above example,

Association between "book" & "superb" is ½ , i.e. 50%

Association between "book" & "read" is 2/2 , i.e. 100%

Association between "book" & "I" is 1/3 , i.e. 33%

Pattern Discovery – Some Use Cases

Illustration1: Catalogue Creation

BUY2SELL is an online marketplace, where buyers meets the sellers. Sellers list their products and buyers purchases the products.

While listing, often the sellers given very detailed description about their product. BUY2SELL wants to read the description and create a catalogue of products to make searching a product easier for the buyer.

Here is an illustration for "smart-phones"
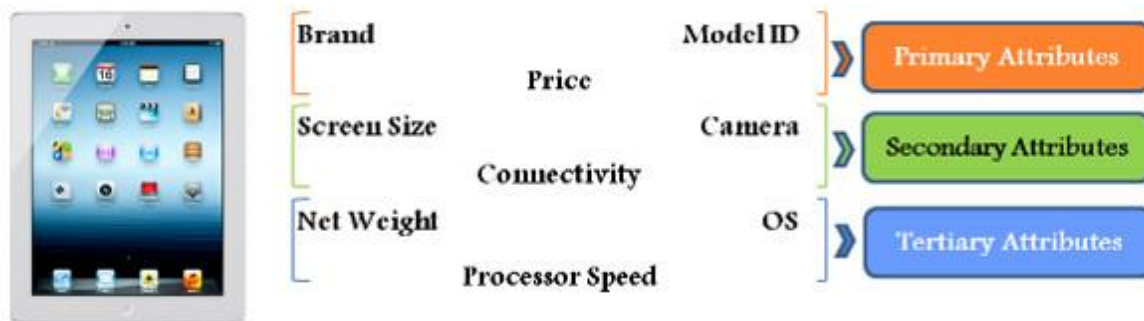
Primarily this is done through "**Feature/ Entity Extraction**"

Pattern Discovery – Some Use Cases

Illustration2: Word Cloud Creation

E2EIT is one major IT organization having offices across the globe. The organization has recently has done a employee satisfaction survey. During that survey there was a question where they asked "How the organization can be made a better place to work". Employees has given different comments on that.

Now, E2EIT wants to visually see what are the topics the employees are talking about?

An word cloud (like given) can be created to get a visual understanding of the words used.

Primarily this is done from "**Term Document Matrix**"

Pattern Discovery – Some Use Cases

Illustration3: Text Classification

VIBGYOR telecom in a leading telecom provider. Daily a huge number of complains comes to them trough internet, where people write their queries or complains on a free text field.

The company wants to build an analytics driven rule engine based on the words used in the feedback text to classify the queries or complains into pre-decided categories and directly divert them to the respective departments.

A rule engine, similar to decision trees, is developed based on the words used in the texts. This is often called "Taxonomy"

Pattern Discovery – Some Use Cases

Illustration4: Link Analysis

ROADKING is the leading automobile manufacturing company. It has recently launched a new car in hatch-back segment.

After 6 months, they want to look into different social media and see how people are comparing with the new model with a other models (both from their company and their competitors).
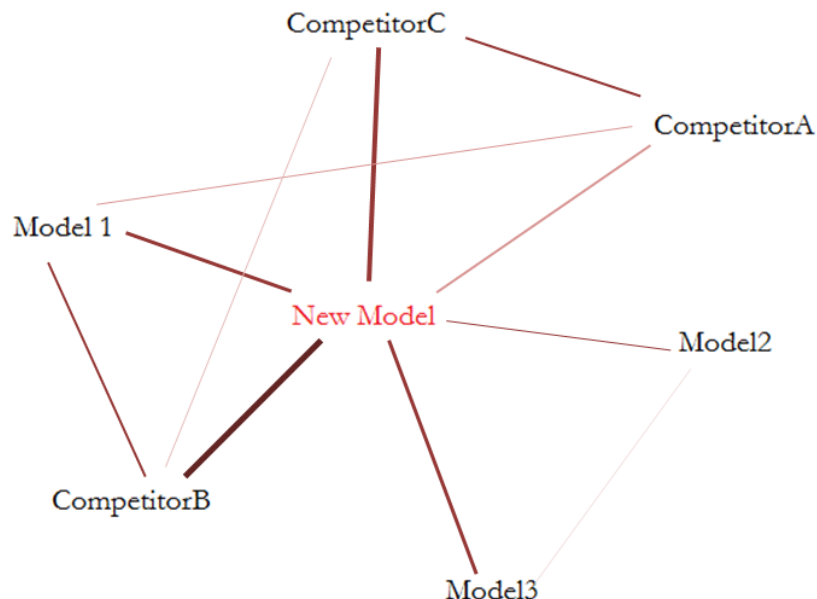
Link analysis using the association between different words and entities can be done here.
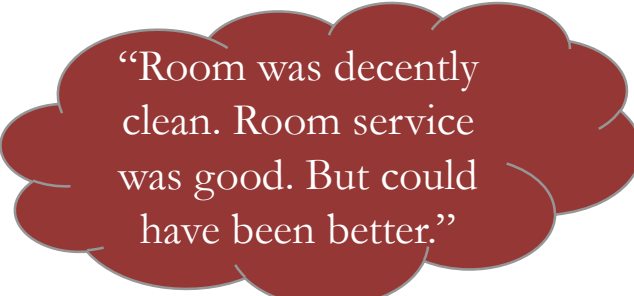
Pattern Discovery – Some Use Cases

Illustration5: Sentiment Analysis
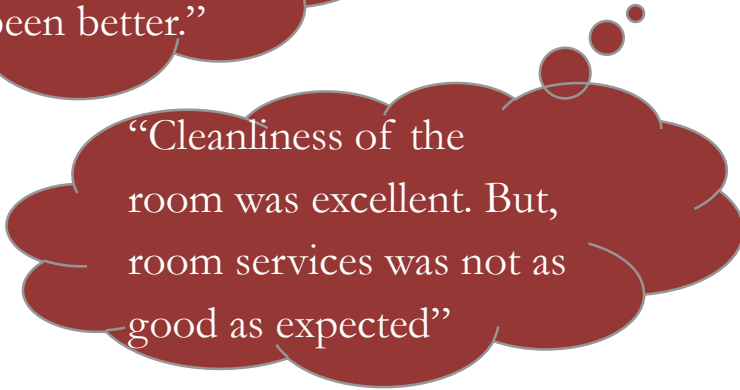
FEELHOME is the leading hotel chain with their outlets spread over the world.

They wanted to know how their delegates are thinking about their service after visiting their hotel. Hence, they looked into the social media and found what their past visitors are speaking.

Below are 4 comments from 4 different people.

"Room was decently clean. Room service was good. But could have been better."

"Room was very clean. Room service was bad."

"Cleanliness of the room was excellent. But, room services was not as good as expected"

"Room was fine. But, room service was pathetic"

Pattern Discovery

Pattern Discovery – Some Use Cases

Illustration5: Sentiment Analysis

Note that, here all 4 people are satisfied with room cleanliness, but dis-satisfied with room services. But, the moods are different. As, "pathetic", "bad" does not carry same weight.

Sentiment Analysis algorithm goes through the comments grammatically and word occurrence wise to score a comment in a -5 to +5 score on sentiments, where -5 means strong negative sentiment and +5 means strong positive sentiment.

"Room was decently clean. Room service was good. But could have been better."

"Room was very clean. Room service was bad."

"Cleanliness of the room was excellent. But, room services was not as good as expected"

"Room was fine. But, room service was pathetic"

Introduction to Text Mining

What is Text Mining	✓
Text Mining Process	✓
Text Processing	✓
Feature Generation	✓
Feature Extraction	✓
Pattern Discovery	✓

Recapitulation & Key Takeaways

- **Text mining** is the process of **extracting implicit, previously unknown potentially useful information from free-text data**.
- Text Mining is **a combination** of Statistics, Data Mining, Natural Language Processing, Information Retrieval, and Web Mining
- There are several specialized software available in the market for Text Mining, as it is very much tool dependent technique

- The source can be **customer complain log** or **twitter feeds** or **social media comments** or **blogs** or **documents** or **feedback to a survey**, etc.

- Once the text data is received, the data is processed like **removal of special characters, managing the abbreviations, spell corrections, finding synonyms, etc.**

- From the processed data, we find the **distinct words and phrases** used. From there we remove the **stops words**
- **Term-document matrix** is created to see which word is used in which comment. **Frequency** is the number of times it is used and **occurrence** is number of comments where it is used.

- **Sentiment analysis** is used to find the sentiment of a comment. The words and their positions in the comment gives an idea about the sentiment. NLP is used here.

- **Taxonomy** is the method of classification of a comment into a pre-decided categories based on the occurrence of different words

- **Link analysis** is the analysis to find which words occur together (in the same comment)

Data Mining

Introduction to Text Mining