

# Information Retrieval

- Introduction
- Text and Language

Development:  
Moshe Friedman

Credits:

Yoav Goldberg, Ido Dagan, Reut Tsarfaty , Moshe Koppel, Wei Song,  
David Bamman, Ed Grefenstette, Chris Manning, Tsvi Kuflik,  
Hinrich Schütze, Christina Lioma and more

# Information Retrieval - administration

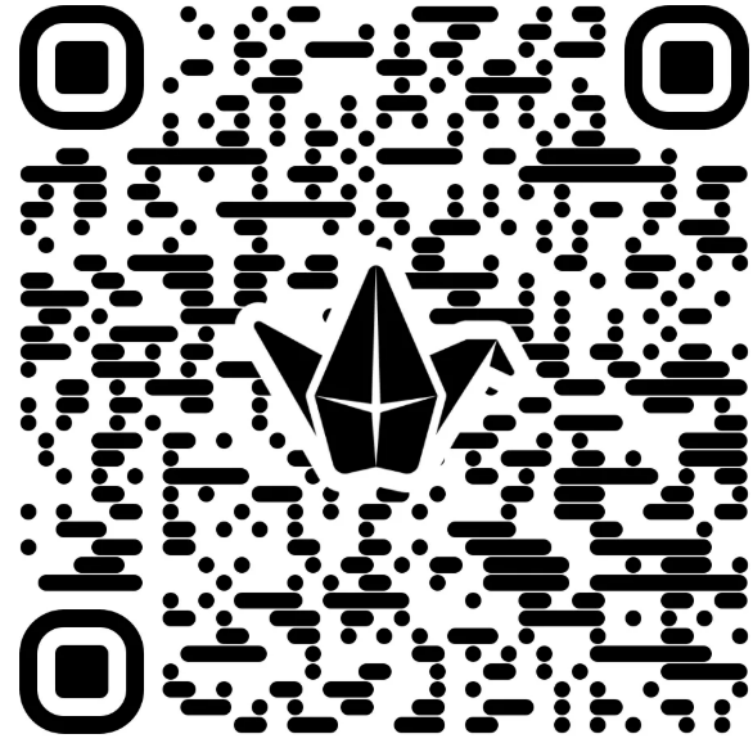
Moshe Friedman

Email: [moshefr.teach@gmail.com](mailto:moshefr.teach@gmail.com)

Reception time: before/after lesson/zoom with coordination

# What do you have in mind re: information retrieval (IR)

[https://padlet.com/moshe\\_cs/my-terrific-sandbox-8fe8d1ocox4uwqe1?frame\\_id=page%3A-laBA3DbY1Xf4ptwglVU4](https://padlet.com/moshe_cs/my-terrific-sandbox-8fe8d1ocox4uwqe1?frame_id=page%3A-laBA3DbY1Xf4ptwglVU4)



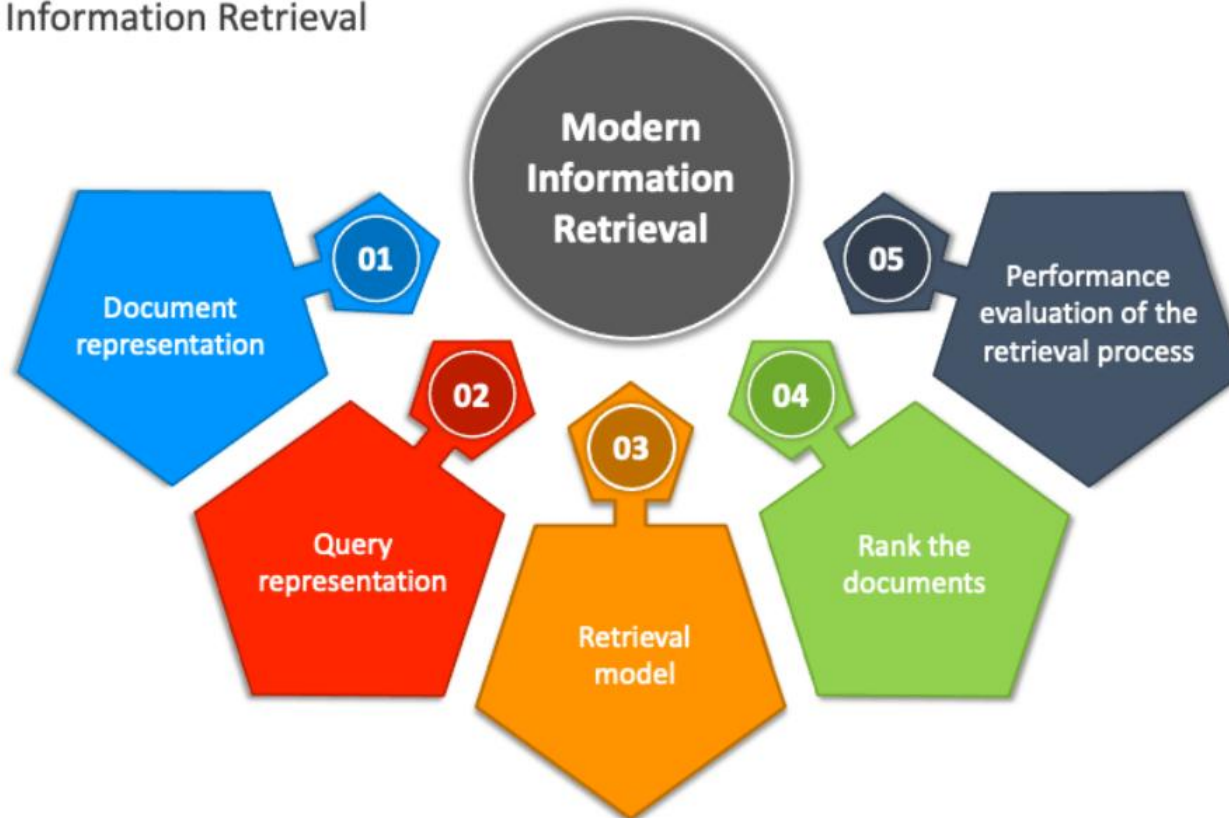
# Introduction to **Information Retrieval**

Introduction

# Classic Information Retrieval Components

## INFORMATION RETRIEVAL

Modern Information Retrieval

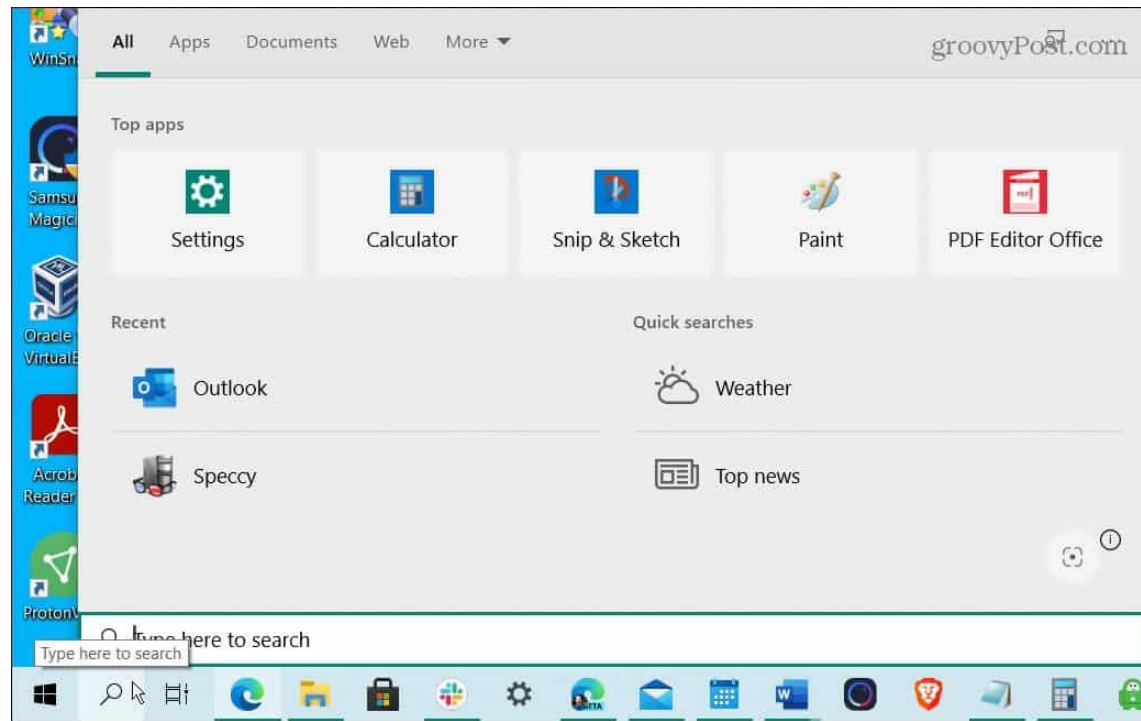


# Information Retrieval

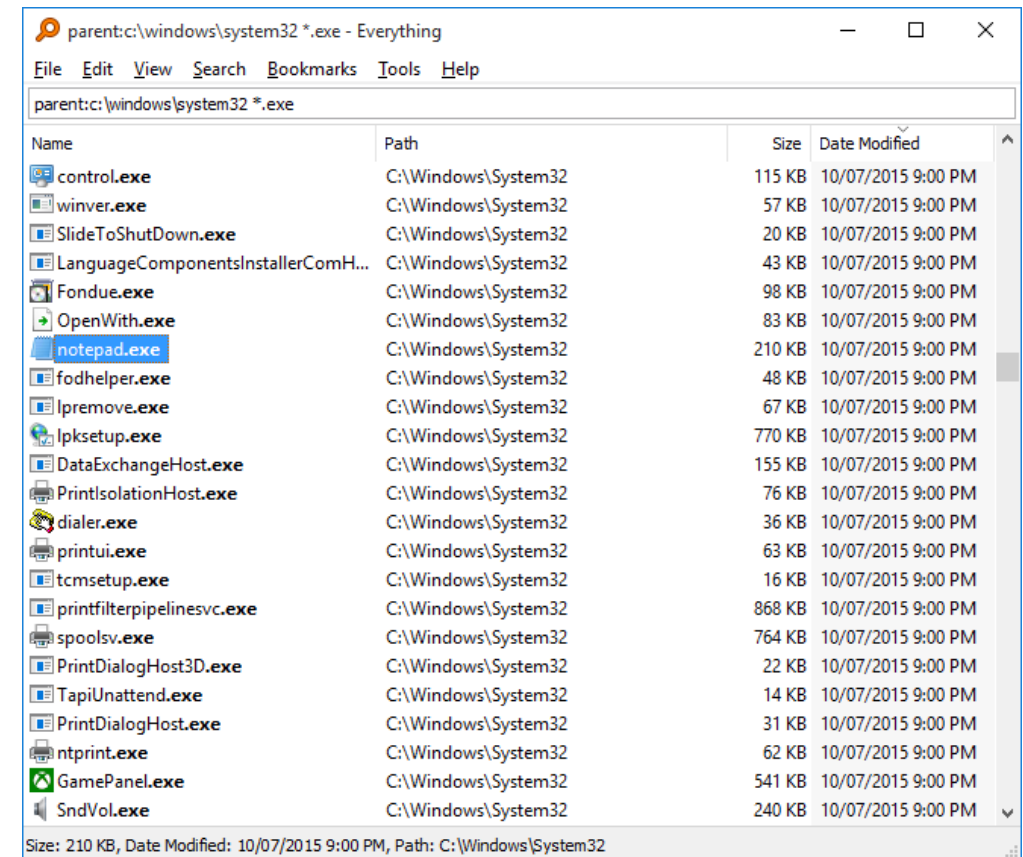
- Information Retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers).
- These days we frequently think first of **web search**, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval
  - Autocomplete

# Information Retrieval Use Cases: Desktop Search

## Window Search



## Everything



# Information Retrieval Use Cases

## Corporate KB Search

The screenshot displays a search interface for a corporate knowledge base. At the top, there is a search bar with the placeholder text "Search...". Below the search bar, the results are categorized under "Articles Search Results". The first result is "Article Links - Adding a link to an article in the knowledge base", dated Wed, Jan 30, 2013, with 3.43/5 (7 Votes) and 28210 views. The second result is "Article Summary", dated Fri, Aug 5, 2016, with 1/5 (1 Vote) and 1701 views. The third result is "Article Referrers", dated Sat, Sep 17, 2016, with 2677 views. The fourth result is "Article Contributors", dated Fri, Aug 12, 2016, with 1/5 (1 Vote) and 1514 views. The fifth result is "Article Settings", dated Thu, Aug 25, 2016, with 1904 views. The sixth result is "Article Interface in Front-end Area", dated Wed, Sep 14, 2016, with 1/5 (2 Votes) and 1560 views. The interface includes a sidebar with filters for "Search for" (Articles), "By ID" (Search by ID), "That contains" (article), "But not", "Search in" (All), "Categories" (Any), "Include sub-categories in search", "Search Filter" (Any Word), "Sort By" (Relevance), and "Sort Order" (Descending). Annotations highlight that individual search keywords are highlighted in the search results and that search results are formatted in a similar way to Google or Yahoo search results.

Search...

Home / Search

Search for

Articles

By ID

Search by ID

That contains

article

But not

Search in

All

Categories

Any

Include sub-categories in search

Search Filter

Any Word

Sort By

Relevance

Sort Order

Descending

Articles Search Results

Article Links - Adding a link to an article in the knowledge base

Wed, Jan 30, 2013

article linking, internal link

3.43/5 (7 Votes)

28210

1

Problem- When I try to enter a hyperlink in my article, the functionality works fine, but I have to know the URL to the article I want to link to, so I can manually enter the path in hyperlink. Is there any way I can browse existing articles using a... [Read More](#)

Article Summary

Fri, Aug 5, 2016

Article Summary Statistics, Statistics of Article summary

1/5 (1 Vote)

1701

0

Expand "Statistics" section in left navigation bar and click on "Articles" link to view reports/statistics related to articles in Articles Statistics page. By default, Summary is displayed which shows the summary of your Knowledge base articles... [Read More](#)

Article Referrers

Sat, Sep 17, 2016

Article Referrers, Referrers, KB Article referrers

2677

0

Article referrer's statistics show from where the request for article pages of your knowledge base has been originated. This feature lets you know which search engines are sending traffic to your knowledge base, for which keywords and which... [Read More](#)

Article Contributors

Fri, Aug 12, 2016

Articles Contributors, Users Contributions

1/5 (1 Vote)

1514

0

The report displays a list of contributors along with the number of articles they have posted in the knowledge base. A contributor can be a Superuser, an Editor, a Writer, or a Writer-trusted. Expand "Statistics" section in left navigation bar... [Read More](#)

Article Settings

Thu, Aug 25, 2016

Manage Article settings, Article settings, RSS feed settings, Article features settings

1904

0

3/5 (2 Votes)

on bar and click on "Manage Settings" link. Manage settings page will display, click on "Article" s. This setting page is categorized into various parts as follows: Add to... [Read More](#)

Article Interface in Front-end Area

Wed, Sep 14, 2016

1/5 (2 Votes)

1560

0

The report shows the list of newly added and recently updated articles in the knowledge base. Expand "Statistics" section in left navigation bar and click on "Articles" link. Article statistics page will display, click... [Read More](#)

Individual search keywords are highlighted in the search results. Non-word characters like punctuation are ignored.

Search results are formatted in a similar way to the results of Google or Yahoo search, so end users find searching easy.

End users can refine the search results using the advanced search form. The results are sortable based on relevance, popularity & rating.

## Email Search

The screenshot displays an email search interface. At the top, there is a search bar with the placeholder text "Search Current Mailbox (Ctrl+E)". Below the search bar, the results are categorized under "Inbox - KatieJ@OCreative.onmicrosoft.com". The first result is "Presentation for Friday", dated Tue 3:44 PM, with 24 KB. The second result is "Let's do lunch at the new cafe!", dated 5/10/2013, with 20 KB. The third result is "Need report", dated 5/10/2013, with 22 KB. The fourth result is "Sports statistics", dated 5/9/2013, with 9 KB. The fifth result is "Expense reports", dated 5/9/2013, with 45 KB. The interface includes a sidebar with filters for "Inbox", "Sent Items", "Deleted Items", "Drafts", "Junk Email", "Outbox", "RSS Feeds", "Search Folders", "Online Archi...", and "Project Falcon". The bottom status bar shows "ITEMS: 14 UNREAD: 9" and "ALL FOLDERS ARE UP TO DATE. CONNECTED TO: M".

Inbox - KatieJ@OCreative.onmicrosoft.com

FILE HOME SEND / RECEIVE FOLDER VIEW GROUPS

New New Clean Up Delete Reply Reply Forward More

New Email Items

New Delete Respond Quick Steps

Search Current Mailbox (Ctrl+E)

Current Mailbox

All Unread By Date (Conversations) Newest

Tuesday

Presentation for Friday

24 KB

Tue 3:44 PM

Meant to send this to you too. :)

Last Month

Let's do lunch at the new cafe!

20 KB

5/10/2013

Need report

22 KB

5/10/2013

Hi, Katie. Do you have the report? Best, <end>

Sports statistics

9 KB

5/9/2013

Do you LOVE sports? If so, read on... We are

Expense reports

45 KB

5/9/2013

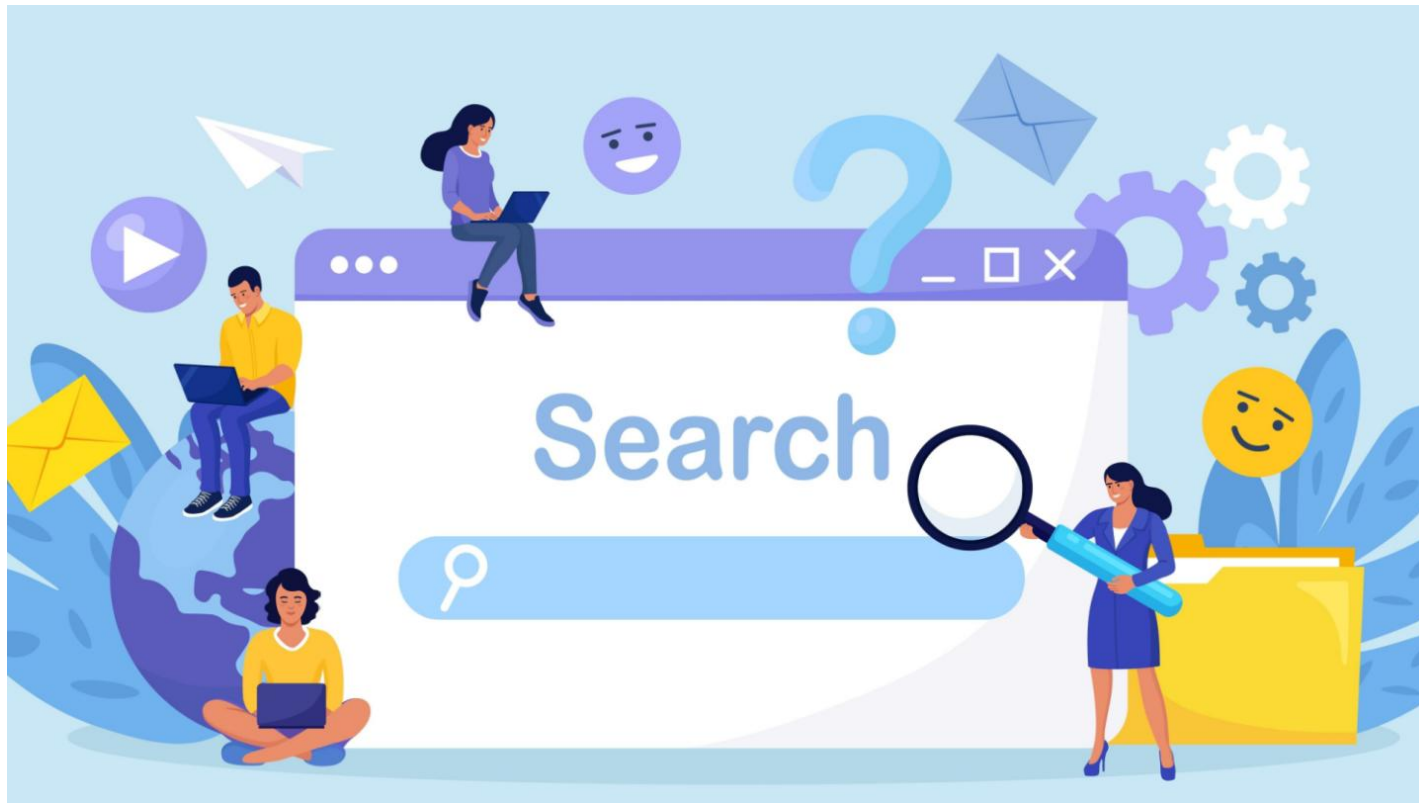
Mail Calendar People Tasks ...

ITEMS: 14 UNREAD: 9 ALL FOLDERS ARE UP TO DATE. CONNECTED TO: M



# Information Retrieval Use Cases

## Web Search



## Autocomplete

cross platform	Search
cross platform	
command prompt	
chkdsk	
control	
cpu	
caret	

ComputerHope.com

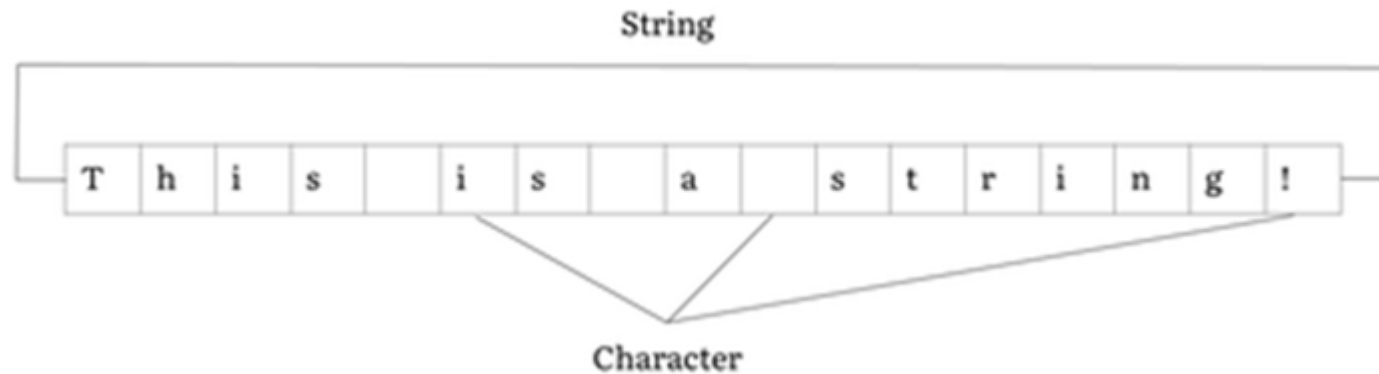
# Introduction to **Information Retrieval**

Text and Language



# What is a text?

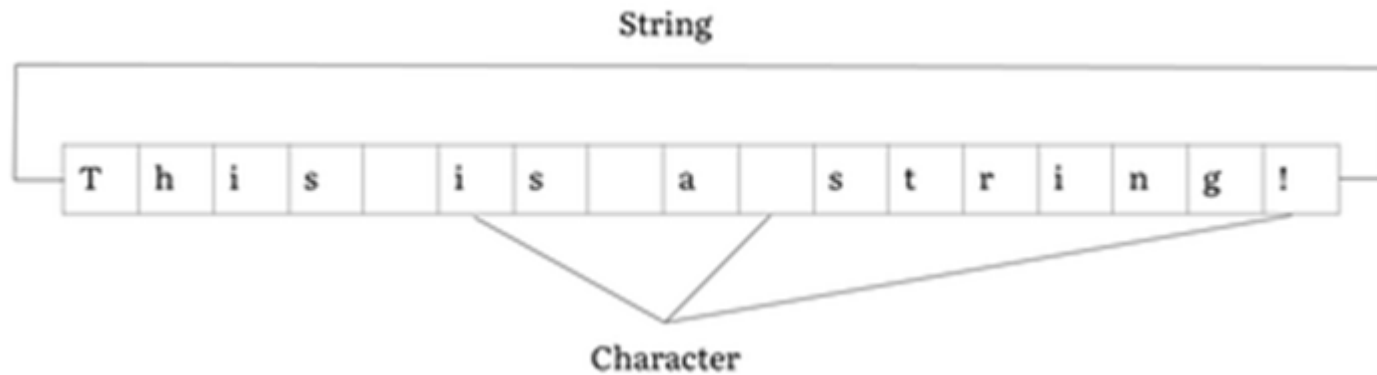
## A programmer's perspective



מחרוזת  
(וויקיפדיה) -  
בשפות תכנות  
מחרוזת היא טיפוס  
נתונים המכיל רצף  
של תווים.  
ולכן בראיה של  
מתכנת – טקסט  
היא מחרוזת תווים.

# What is a text?

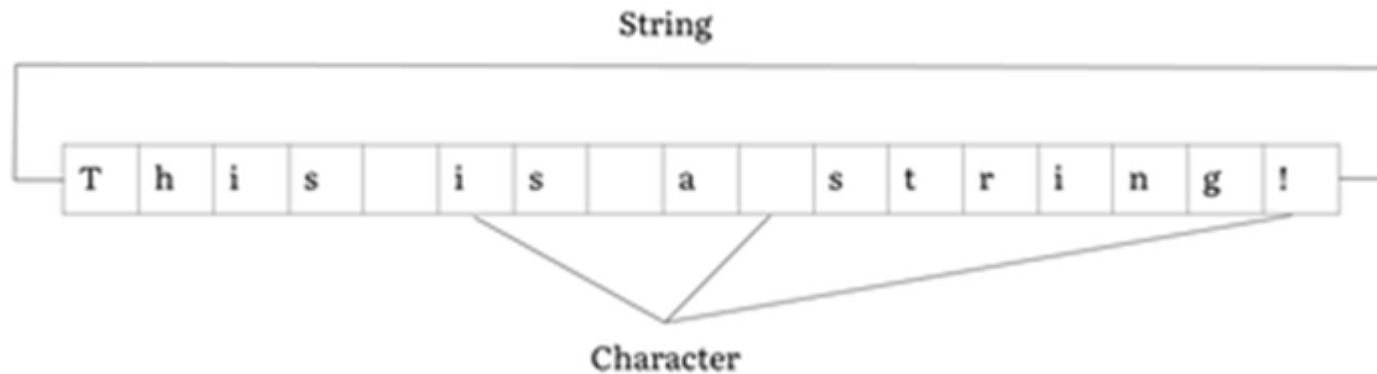
## A programmer's perspective



אך האם כל מחרוזת תווים משמעותה טקסט?

# What is a text?

## A programmer's perspective



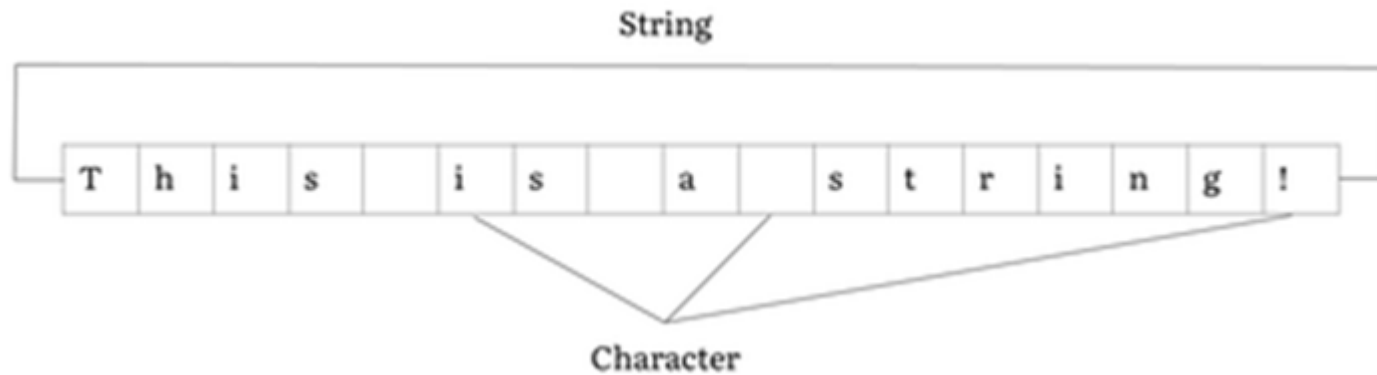
אך האם כל מחרוזת תווים משמעותה טקסט?

RSA encryption key

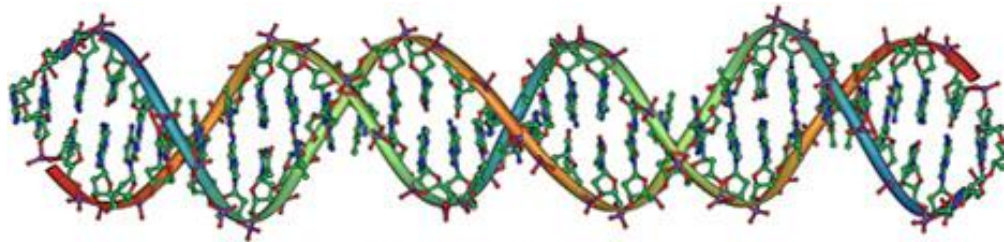
```
l1vBCavOBjD22a5RFZQDn+jpr8D/zPS+hhXnXHazZAKkOx1qKbZtIZWDt3Mv
8TFt6YgNZ9RSN19r7G7Vim9q3l35EXbG1Vb/UwLY4WtcNL9nHuvyclF8ebS
wn3gS+ArRdOa2KsSdg13LU1OsPsRmOlroPrS7jUM3jZS9g3FnukRRCYqJMZZ
nCf/2C330aeLHCLmNn6pegRqGb6hTy6717g6+xw/bGtLEIrd75AMcW+q72E/
x3UkMOx1wtRRougEfujI8YF/j2PM85YoZgW5HTc9+xrLxJ9Z8N13/TWujgC7
xmFO3VsP6IqfSZCvvatktwkXmvmf9S7e6uNzFs33Rw==
```

# What is a text?

## A programmer's perspective



אך האם כל מחרוזת תווים משמעותה טקסט?

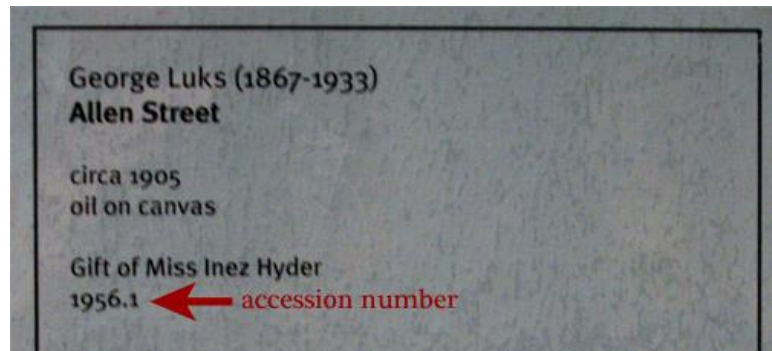
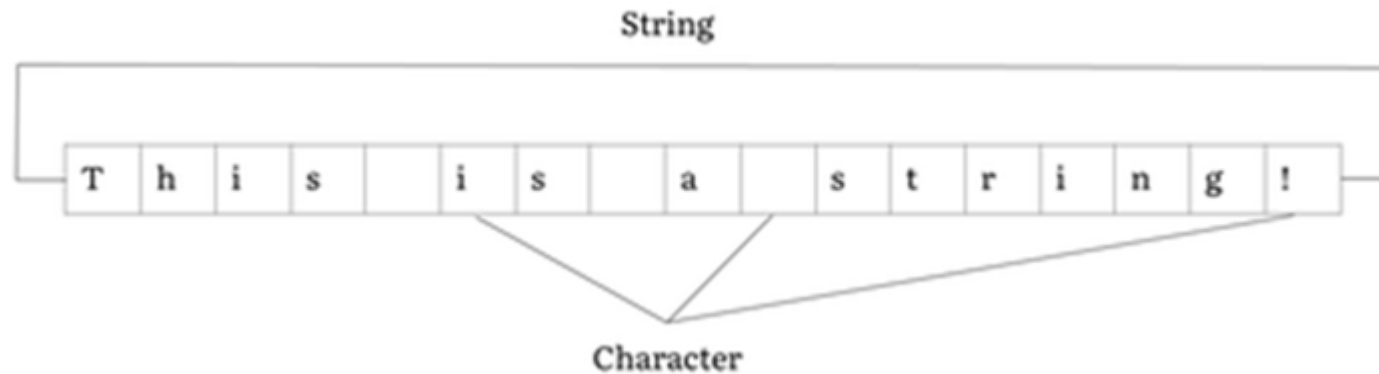


Genome Sequence

AGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGGTAGTAGA

# What is a text?

## A programmer's perspective



אך האם כל מחרוזת תווים משמעותה טקסט?

**Accession number (Wikipedia)**- In libraries, art galleries, museums and archives, an accession number is a unique identifier assigned to, and achieving initial control of, each acquisition



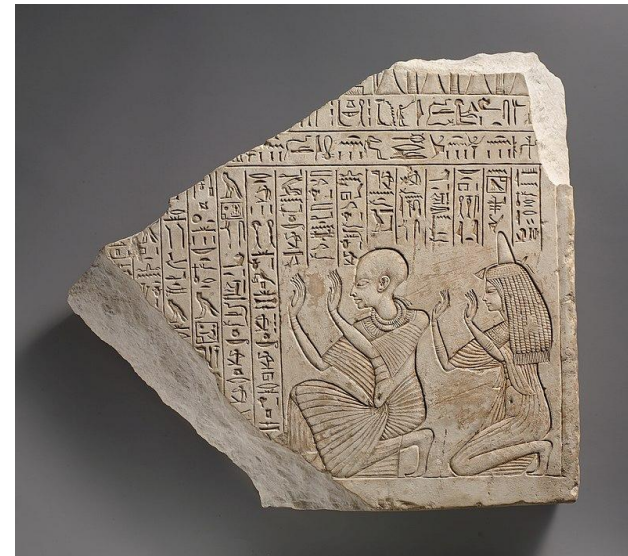
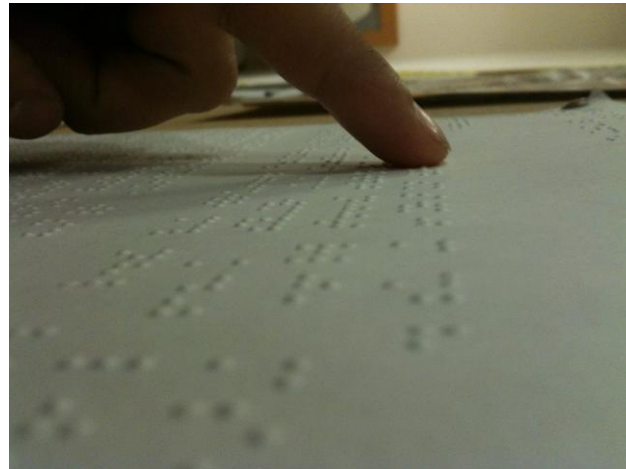
# What is a text?

מסקנה – לא כל מחרוזת תווים היא טקסט.

אז מהו טקסט?

**Text** (Wikipedia) - In literary theory, a text is any object that can be "read".

בחזרה לטקסט עוד מעט



# נבחן עוד כמה שאלות ...

## מה הוא ידע?



**Knowledge** (Wikipedia) - Knowledge is a familiarity, awareness, or understanding of someone or something, such as facts (descriptive knowledge).

# נבחן עוד כמה שאלות ...

מהי שפה?



**Language** is the use of a system of communication which consists of a set of sounds or written symbols.

הדרך בה אנחנו צורכים, מעבירים ומבינים ידע הוא דרך שפה

# נבחן עוד כמה שאלות ...

מהי שפה?

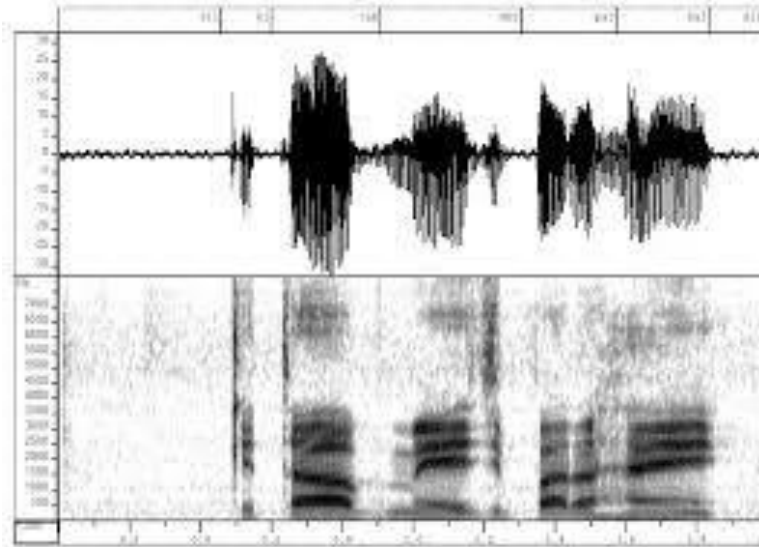


**Language** is the use of a system of communication which consists of a set of sounds or written symbols.

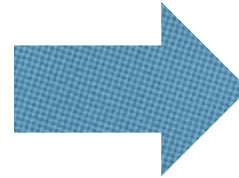
**שפה טבעית - להבדיל משפות אחרות (שפות תכנות למשל),**  
**שפה המבטאת תקשורת בין בני אדם נכנה שפה טבעית.**

# שפה טבעית ואחזור מידע

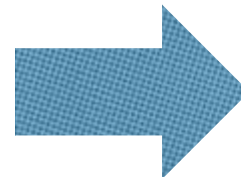
Voice/Speech recognition /phonology



TAVILIETAKADUR



text



תביא לי את  
הכדור

# שפה טבעית ואחזור מידע

Voice/Speech recognition /phonology

OCR – Optical character recognition

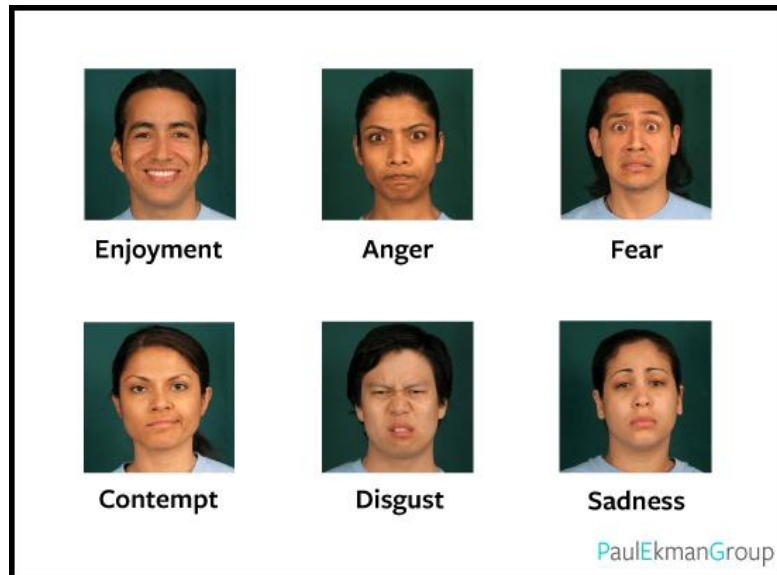
Optical		Character													
O	p	t	i	c	a	l	C	h	a	r	a	c	t	e	r
Recognition															
R	e	c	o	g	n	i	t	i	o	n					

# שפה טבעית ואחזור מידע

Voice/Speech recognition /phonology

OCR – Optical character recognition

Nonverbal / Gesture / Micro expression



1		Thumb up	7		"OK"
2		Index extension	8		"Victory"
3		Make fist	9		"Call"
4		Palm open	10		"Drag"
5		Wrist out	11		Wrist out (fist)
6		Wrist in	12		Wrist in (fist)

credit Wei Song, et al

# What do we mean by text?

**Text** (Wikipedia) - In literary theory, a text is any object that can be "read".

**String** (Wikipedia) - In computer programming, a string is traditionally a sequence of characters.

**Language** is the use of a system of communication which consists of a set of sounds or written symbols.

**A text string** - We refer to text of (verbal) language saved as a string.

## **We is not included:**

- Voice/Speech recognition/Phonology
- OCR – Optical character recognition
- Nonverbal Language / Gesture / Micro expression Recognition



# What do we mean by text?

**Text** (Wikipedia) - In literary theory, a text is any object that can be "read".

**String** (Wikipedia) - In computer programming, a string is traditionally a sequence of characters.

**Language** is the use of a system of communication which consists of a set of sounds or written symbols.

**A text string** - We refer to text of (verbal) language saved as a string.

## We is included – (written) non-spoken language:

- **Emoticons** - :-) :-(  
:-)
- **Emoji** - 😊 😞
- **Text messaging** – LOL ("laughing out loud"), "gr8" ("great")

And more

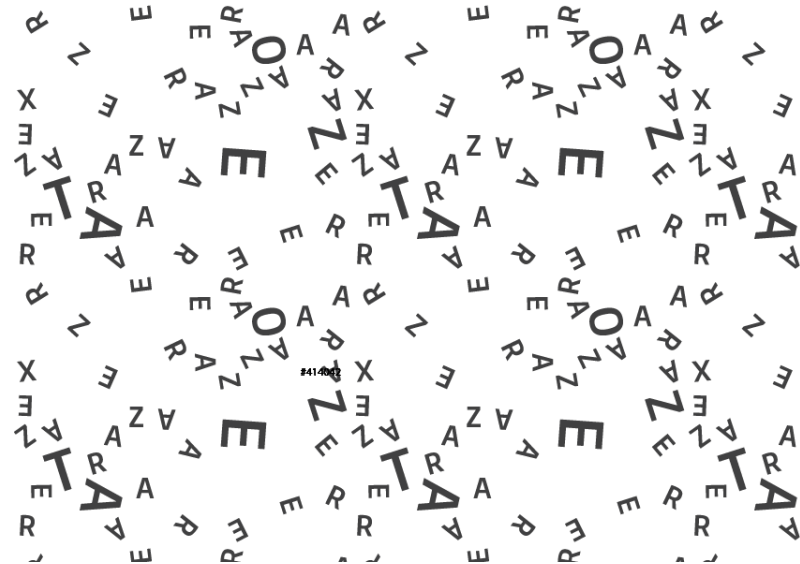
# What do we know so far

Terms: Natural language and Text (verbal and non-verbal)

We want to process text, which describes natural language.

What's next?

- ❖ Bytes and bits
- ❖ Why is language processing hard?



# From text to data

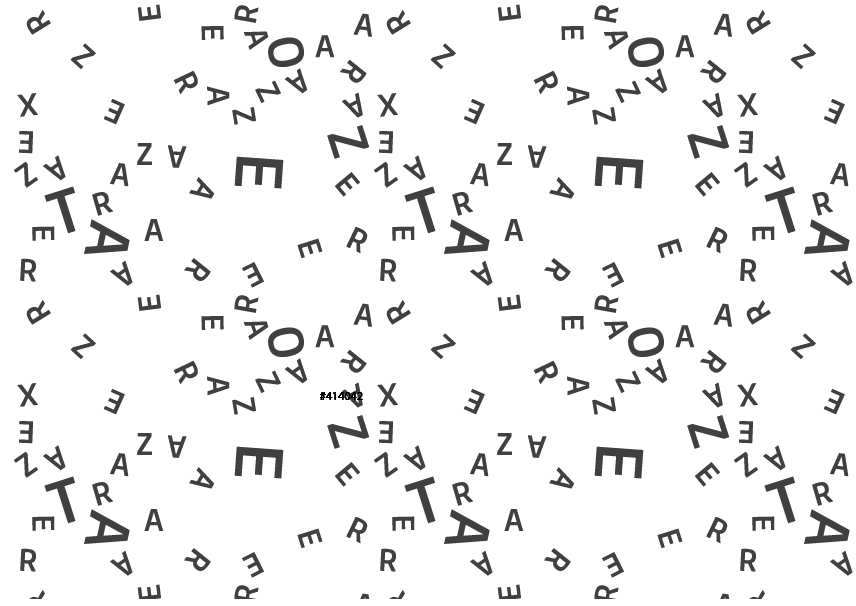
- **An introduction**
  - What is a text?
  - Natural Language and Information Retrieval
  - What do we mean by text?
- Some text analysis challenges
- Levels of Language Processing

# A lot of Buzz words

- Text analysis
- Text processing
- Text mining
- Natural language processing (aka NLP)
- Computational linguistics

What do these terms mean?

How do they relate to machine learning?



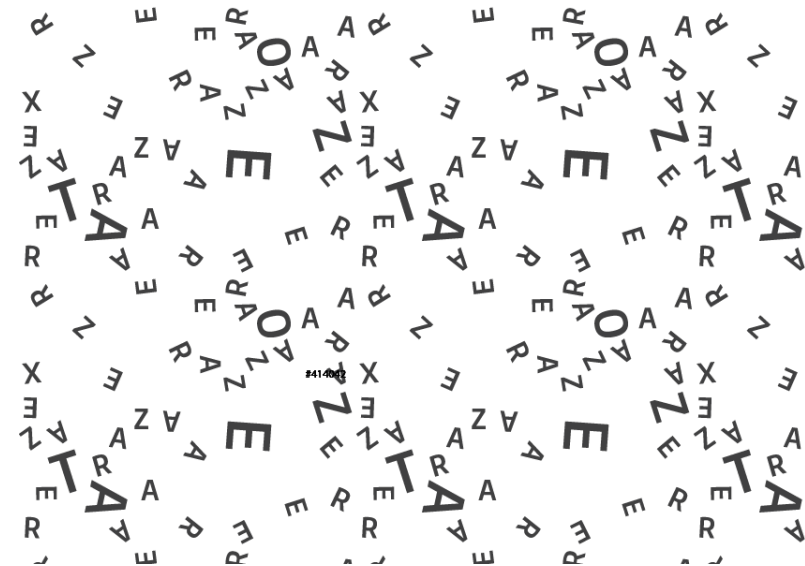
# What do we know so far

Terms: Natural language and Text (verbal and non-verbal)

We want to process text, which describes natural language.

# What's next?

- ❖ Bytes and bits
- ❖ Why is language processing hard?



# From text to data

- An introduction
- **Some text analysis challenges**
  - Variability - one meaning, many forms
  - Ambiguity - one form, many meanings
  - Zipf law
- Levels of Language Processing

# Text data – properties and challenges

## Variability - one meaning, many forms

2maro 2marrow 2mor 2mora

2moro 2morow 2morr 2morro 2morrow 2mr

2mro 2mrrw 2mrw 2mw tmmrw tmo tmoro tmorrow

tmoz tmr tmro tmrow tmrrow tmrrw tmrw tmrww tmw

tomaro tomarow tomarro tomarrow tomm

tommarow tommarrow tommoro tommorow

tommorrow tommorw tommrow tomo tomolo tomoro

tomorow tomorro tomorrw tomrw

# Text data – properties and challenges

## Variability - one meaning, many forms

he acquired it

he purchased it

he bought it

it was bought by him

it was sold to him

she sold it to him

she sold him that



# Text data – properties and challenges

## Ambiguity - one form, many meanings

Bank

noun



noun



# Text data – properties and challenges

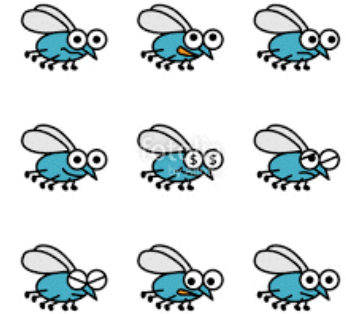
## Ambiguity - one form, many meanings

### Time Flies

If time is a  
**noun**



If time is a  
**verb**



#32292888

# Text data – properties and challenges

## Ambiguity - one form, many meanings

### Ambiguity – (extreme cases)

\* אם מישהו רכב על סוס ועזב, אז הוא פרש או לא?

\* להודות למישהו בפה מלא, זה מנומס או לא?

\* אם ארכיאולוג העלה חרס בידו, הוא הצליח או לא?

\* באיסטנבול לא סוגרים דלתות. הם טורקים.

\* שקלתי הרבה לפני שהחלטתי לעשות דיאטה

# Text data – properties and challenges

## Zipf law

Word frequencies follow a power-law distribution.  
--> Long tail - most words will occur only few times if they occur

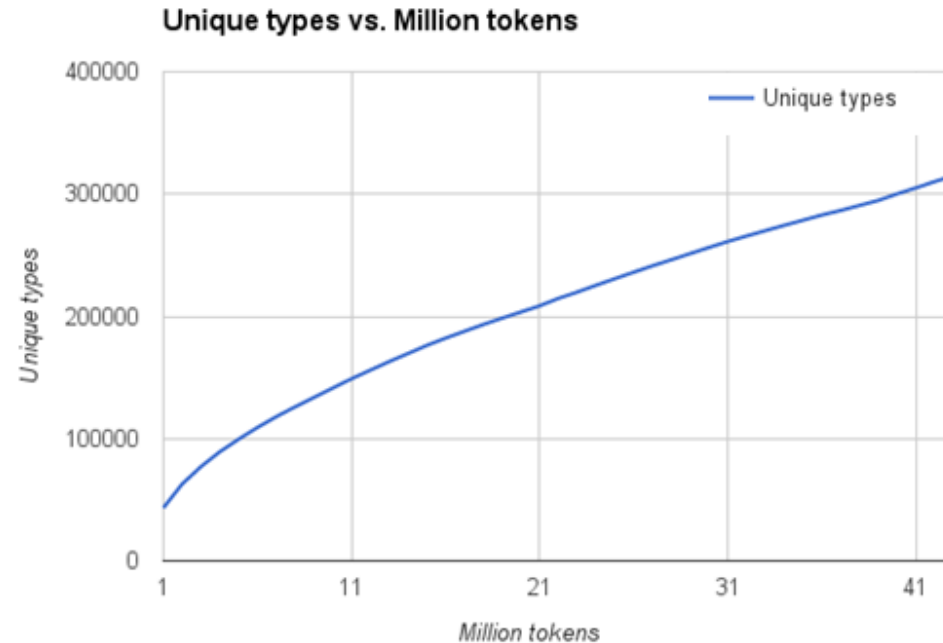
There are very likely to be word forms we did not see

In a 43M words text, there are:

- 316K unique words
- 144K words occur once
- 42K words occur twice

...

- 26K words occur >50 times



# From text to data

- An introduction
- Some text analysis challenges
- Levels of Language Processing
  - Phonology
  - Morphology
  - Syntax
  - Semantics

# Levels of Language Processing

## Phonology

TAVILITAKADUR



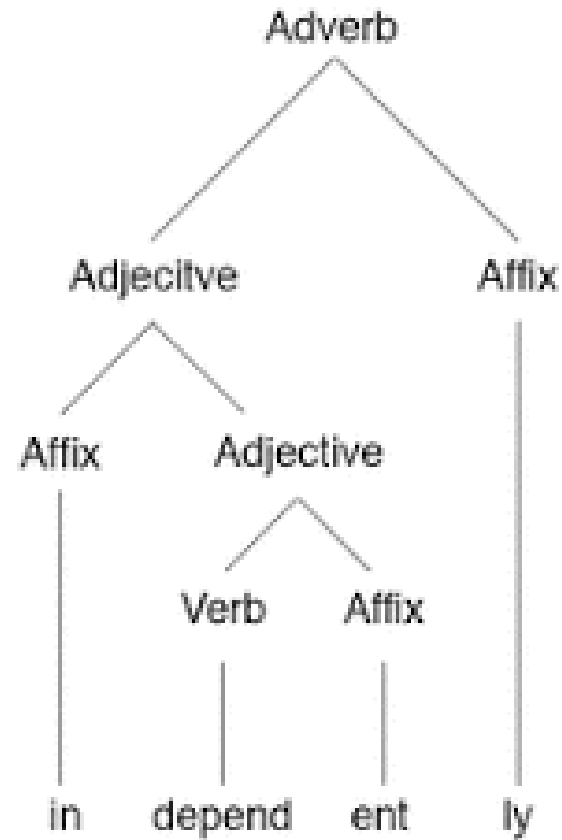
תביא לי את הכדור

# Levels of Language Processing

Phonology

**Morphology**

Independently



# Levels of Language Processing

Phonology

**Morphology**

בגרמנית – אפשר להרכיב מילים מורכבות  
מאוד בעלות תחיליות, אמצעיות וסופיות.



lebensversicherungsgesellschaftsangestellter

Life insurance company employee



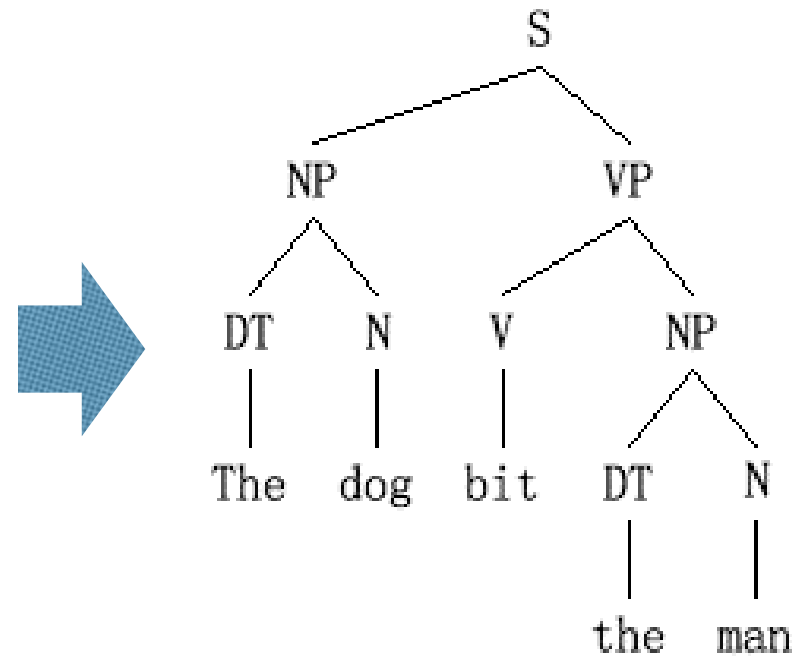
# Levels of Language Processing

Phonology

Morphology

**Syntax**

The dog bit the man



# Levels of Language Processing

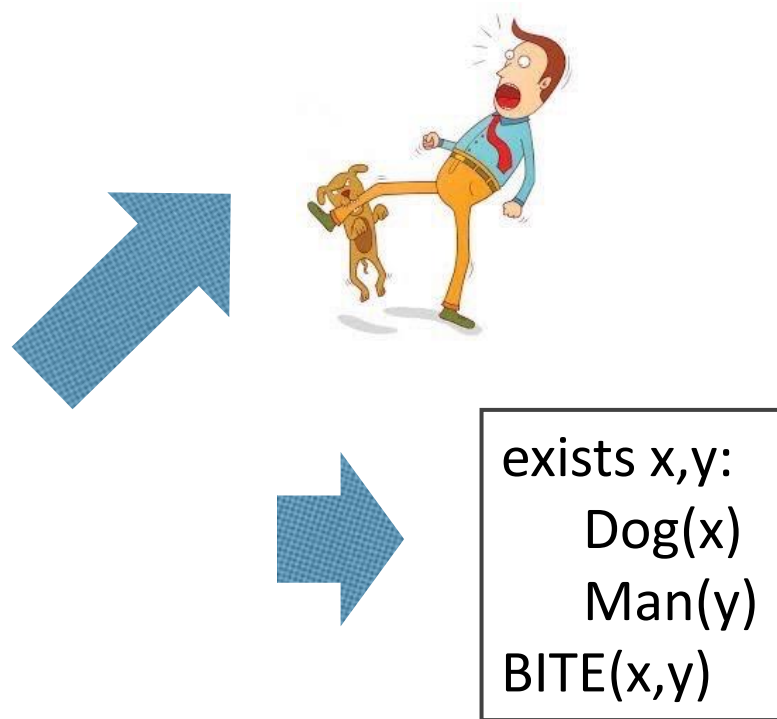
Phonology

Morphology

Syntax

**Semantics**

The dog bit the man



# Text documents

- Text data units
- Preprocessing text data
- Vectorization
- Text analysis flow and examples
- Text Analysis example

# Text data units – different granularity

## **The basic units of text processing:**

- Corpus- document collection
- Document
- Section
- Paragraph
- Sentence
- Phrase
- Word
- Character level

# Text data units – different granularity

## **Document --> label**

- Language classification
- Topic classification
- Author classification
- Sentiment classification

## **Sentence --> label**

- ❖ Usually the same as in "document --> label"  
shorter text --> harder task

Question: are each of these binary / multi-class / multi-label?

# Text data units – different granularity

## Word --> label

- ❖ Tokens vs. Types

### Type --> label

- Sad vs happy words
- Adjectives vs. Nouns

### Token --> label

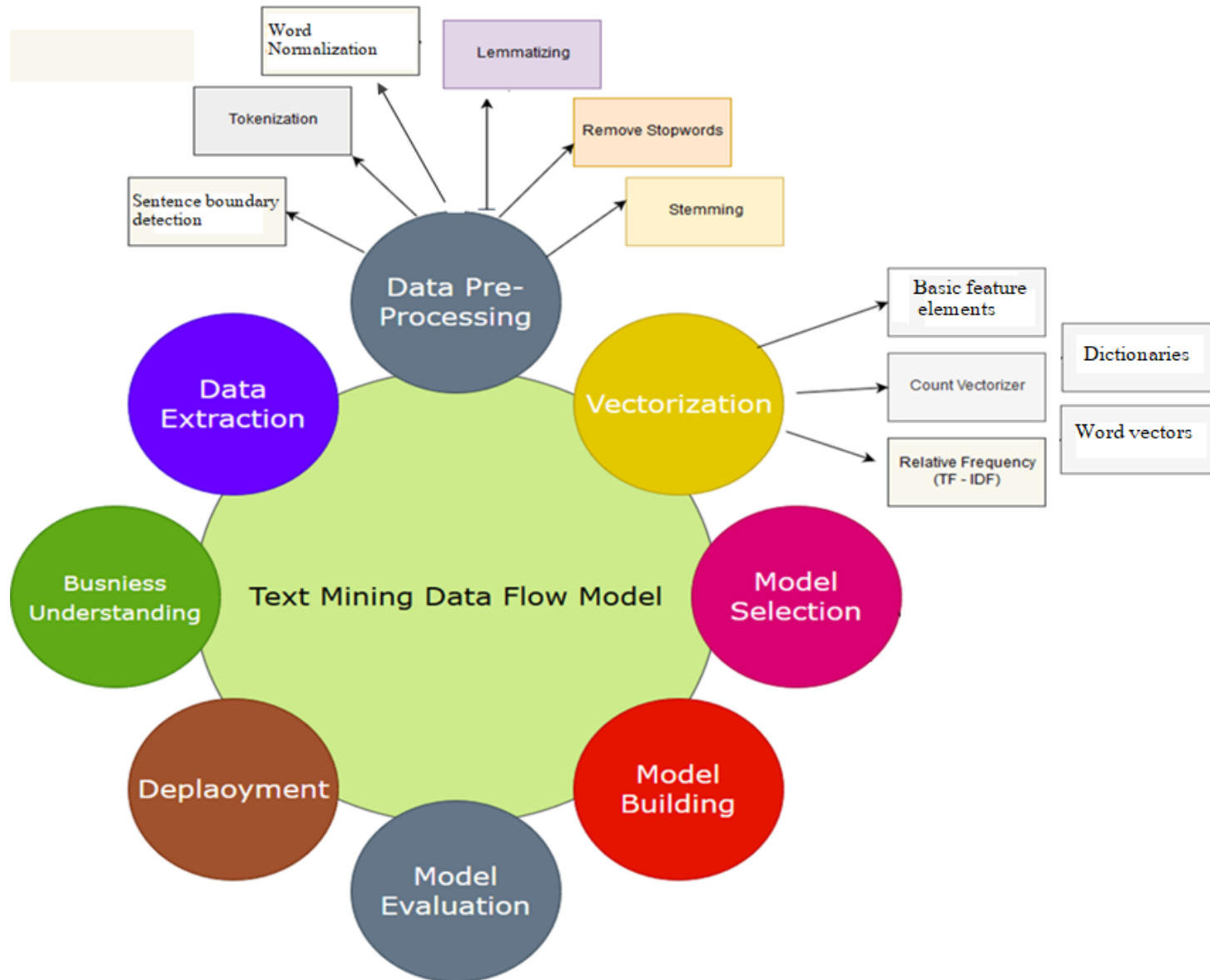
- Sentence boundary detection
- Common spelling mistakes ("then" vs "than")

# Retrieved Text Documents

What is the basic atom level of text we should relate to?

How do we retrieve documents?

# Basic Text Processing & Analysis Flow





# What is a word?

sequence of characters (perhaps letters)?

**white-space bounded?**

Input:

Friends, Romans, Countrymen, lend me your ears;

→ whitespace might not be enough

# Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens* (perhaps at the same time throwing away certain characters), such as punctuation.

Here is an example of tokenization:

Input:

Friends, Romans, Countrymen, lend me your ears;

Output:

•Friends	❖lend
•Romans	❖me
•Countrymen	❖your
	❖ears

**Tokenization:** dividing a text into tokens.

# What is a word?

## Twitter Text example

**sequence of characters (perhaps letters)?**

**@dbamman have you seen this :) <http://popvssoda.com>**

•@	❖this	❖/
•dbamman	❖:	❖/
•have	❖)	❖popvssoda
•you	❖http	❖.
•seen	❖:	❖com

# Words, Sentences and Punctuation

We typically don't want to just strip all punctuation, however.

- Punctuation signals boundaries (sentence, clausal boundaries, parentheticals, asides)
  - Some punctuation has illocutionary force, like exclamation points (!) and question marks (?)
  - Emoticons are strong signals of e.g., sentiment
- Most tokenization algorithms (for languages typically delimited by whitespace) use regular expressions to segment a string into discrete tokens.

How much can we rely on white spaces?

# Tokenization in languages without spaces

Many languages, such as Mandarin (Chinese), Japanese and Thai don't use spaces to separate words!

How do we decide where the token boundaries should be?

# Tokenization in Mandarin

Mandarin words are composed of characters called "**hanzi**" (or sometimes just "**zi**")

Each one represents a meaning unit called a morpheme.

Each word has on average 2.4 of them.

But deciding what counts as a word is complex and not agreed upon.

# How to do tokenization in Mandarin?

- 姚明进入总决赛 “Yao Ming reaches the finals”

- 3 words?

- 姚明 进入 总决赛

- YaoMing reaches finals

- 5 words?

- 姚 明 进入 总 决赛

- Yao Ming reaches overall finals

- 7 characters? (don't use words at all):

- 姚 明 进 入 总 决 赛

- Yao Ming enter enter overall decision game

# Tokenization / segmentation

So, in Chinese it's common to just treat each character (zi) as a token.

- So, the **segmentation** step is very simple

In other languages (like Thai and Japanese), more complex word segmentation is required.

- The standard algorithms are neural sequence models trained by supervised machine learning.



# Word / token normalization

## **Punctuation**

- U.S.A. → USA (acronym)
- Ltd. → Ltd

# Word / token normalization

## **Case folding**

- General Motors → general motors

Risk:

US → us

# Stemming

**Stem:** a "base form", based on heuristics.

create, created, creating, creator, creativity  
↓ ↓ ↓ ↓ ↓  
creat, creat, creat, creat, creat

**Stemming:** cutting the inflected words to their root form.

# How should we treat Hebrew and Arabic?

## Approach 1 - Transliteration Scheme

Each character is represented as a Latin character

א	ב	ג	ד	ה	ו	ז	ח	ט	י	כ	ל	מ	נ	ס	ע	פ	צ	ק	ר	ש	ת
a	b	g	d	h	w	z	x	v	i	k	l	m	n	s	y	p	c	q	r	e	t

Latin	Arabic	Latin	Arabic	Latin	Arabic	Latin	Arabic	Latin	Arabic
a	ا	O	و	b	ب	n	ن	N	ك
A	آ	U	ؤ	G	ع	t	ت	d	د
e	ي	x	ش	p	پ	q	ق	y	ي
i	ئ	s	س	c	ج	f	ف	w	و
o	و	m	م	H	ح	k	ك	z	ز
u	ؤ	l	ل	j	ج	g	گ	r	ر

For basic difficulty of the usage for these languages:

- Right-to-left
- Non-Unicode tools

Drawback:

- Not supported for most tools.

# How should we treat Hebrew & Arabic?

## Approach 2 – Adapting a general approach

- Tokenization for Hebrew & Arabic
- What is a word/token in Hebrew & Arabic
- Adding Morphology
- Word normalization for Hebrew & Arabic

# Tokenization – עכשיו בעברית

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens* (perhaps at the same time throwing away certain characters), such as punctuation.

Here is an example of tokenization:

Input:

אתמול, 8.6.2018, בשעה 17:00, הלכתי עם אמא למכולת.

Output:

• אתמול

• 8.6.2018

• בשעה

• 17:00

❖ הלכתי

❖ עם

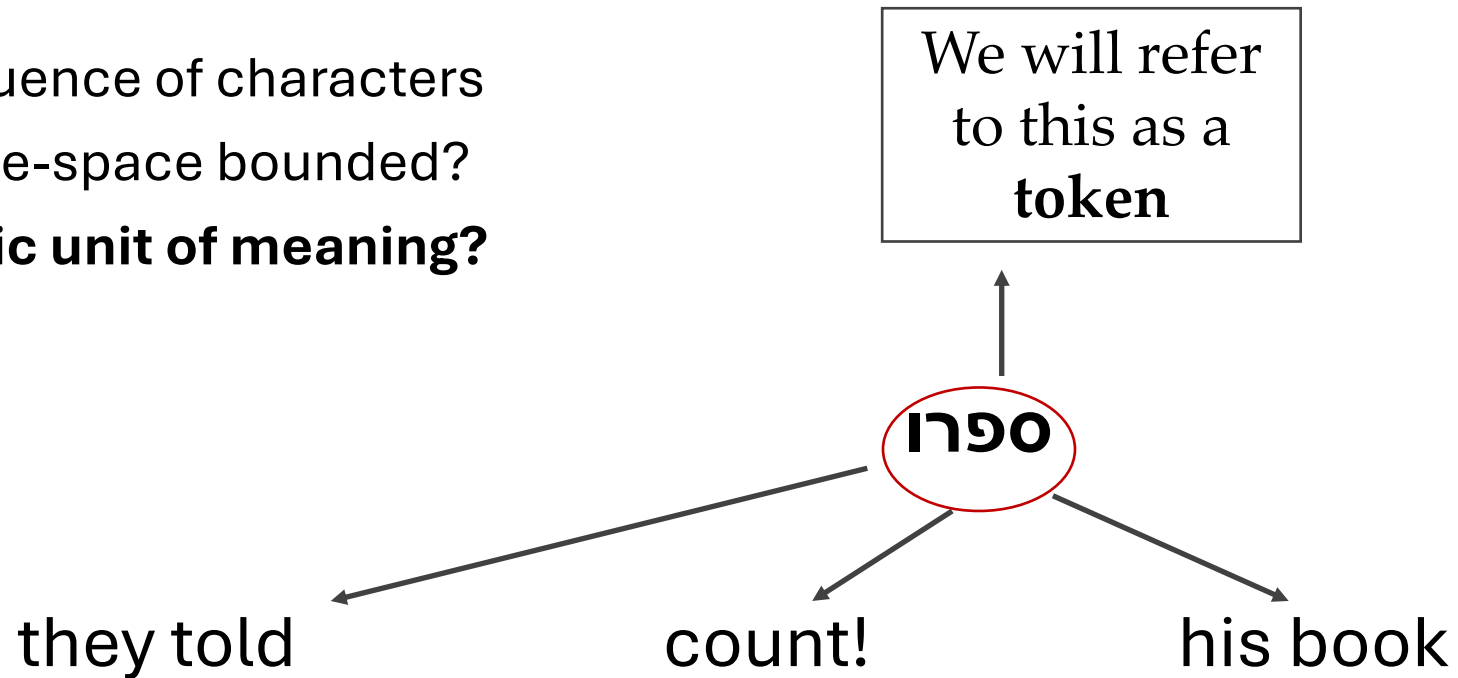
❖ אמא

❖ למכולת

Tokenization: dividing a text into tokens.

# עברית מקרה מסובך יותר – What is a word?

sequence of characters  
white-space bounded?  
**basic unit of meaning?**



→ whitespace might not be enough

# Morphological analysis

**Lemma:** the "dictionary entry" of a word.

create, created, creating, creator, creativity  
↓ ↓ ↓ ↓ ↓  
create, create, create, creator, creativity

**Stemming:** cutting the inflected words to their root form.

ספרו - ה-ספר-של-  
הוא

**Lemmatization:** reducing the inflected forms of a word into a single form for easy analysis.

**Stem:** a "base form", based on heuristics.

create, created, creating, creator, creativity  
↓ ↓ ↓ ↓ ↓  
creat, creat, creat, creat, creat



# Word / token normalization

## Morphology

Morphemes - The small meaningful units that make up words

- Lemma: The core meaning--bearing units
- Affixes: Bits and pieces that adhere to stems
  - Often with grammatical functions
  - יותר משמעותי בעברית וערבית (שפות בהם יחסית יש הרבה מורפולוגיה)

## Word / token normalization - נרמול בעברית (פיסוק)

### פיסוק – קיצורים, ראשי תיבות

- ב.ע.מ. <-- בעמ
- דוא"ל <-- דואל
- וכו'
- די. וי. איי. (לועזית)

### סיכון:

- גב' <-- גב
- ד"ר <-- דר

## Word / token normalization - נרמול בעברית (ניקוד)

### פיסוק – קיצורים, ראשי תיבות

- פְּרִי --> פרי

- צוּרָה --> צורה

### סיכון:

- דֶּרֶךְ --> דרך (איבוד מידע, דרך --> דֶּרֶךְ או דִּבְרֶה)

# Basic useful preprocessing operation - summary

**Tokenization (sometimes segmentation):** dividing a text into tokens.

**Input:** "I love playing soccer with my friends, mostly on the weekends!"

**Output:** ["I", "love", "playing", "soccer", "with", "my", "friends", "mostly", "on", "the", "weekends", "!"]

**Stemming:** cutting the inflected words to their root form.

**Input:** "The mice in the fields were running and jumping around."

**Output:** " The mice in the field were run and jump around ."

**Lemmatization:** reducing the inflected forms of a word into a single form for easy analysis.

**Input:** "The mice in the fields were running and jumping around."

**Output:** "The mouse in the field be run and jump around"

Other Useful operations:

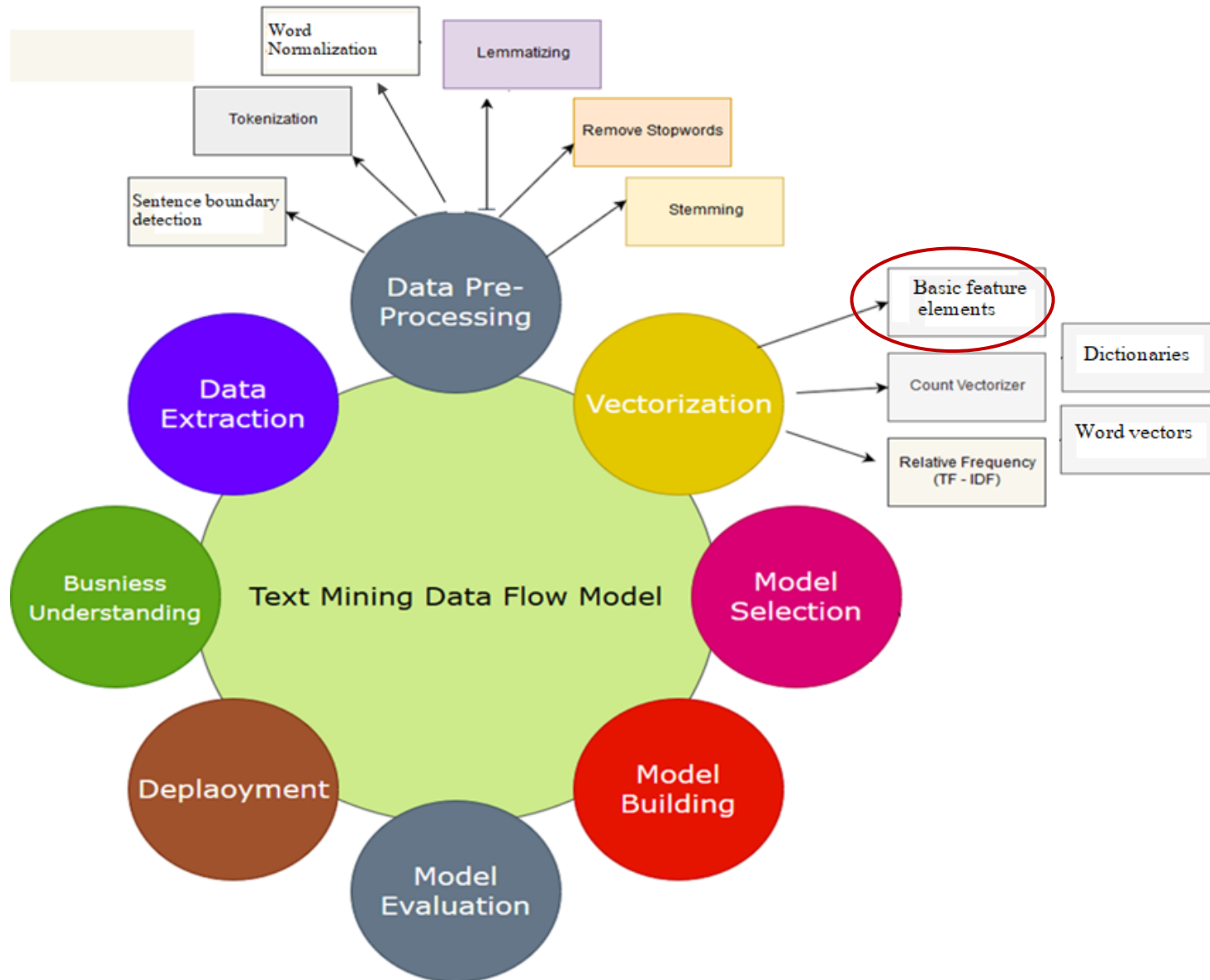
**Sentence breaking:** placing sentence boundaries on a text.

**Word Normalizations and cleaning**

**Stop word removal:** removing words, which are considered as such.

**Part-of-speech tagging:** identifying the part of speech for every word.

# Basic Text Processing & Analysis Flow



# Text as data – properties and challenges

## **Sparse data**

- Zipf's law, variability

## **Symbolic**

- abstract symbol to meaning mapping, variability, ambiguity

## **Many levels of granularity**

- document, paragraph, sentence, word, characters

How would these affect a classifier over text data?

Vectorization:  
from document to feature vector

**The bag of words (BOW) model:**

Each **word** is treated as a feature in a unit called **document**.

Each such word will become a feature

How do we measure their strength?

- Word Count
  - What about zipf law?

# Vectorization: extracting basic feature units

## **The bag of words (BOW) model:**

Each **word** is treated as a feature in a unit called **document**.

Each such word will become a feature

- Alternative: **original tokens – no processing**
- Alternative: **normalized words – e.g., lemmas, stems**
- Alternative: **partial word (prefix, suffix)**
- Alternative: **ngrams – unigram, bigram, trigram**
- Alternative: **characters – we will usually use with ngrams**
- More complex alternatives ...



# Vectorization: extracting basic feature units

## The bag of words (BOW) model:

Each **word** is treated as a feature in a unit called **document**.

Each such word will become a feature

- Alternative: **normalized words – e.g., lemmas, stems**

**Lemma:** the "dictionary entry" of a word.

create, created, creating, creator, creativity  
↓ ↓ ↓ ↓ ↓  
create, create, create, creator, creativity

**Stem:** a "base form", based on heuristics.

create, created, creating, creator, creativity  
↓ ↓ ↓ ↓ ↓  
creat, creat, creat, creat, creat

# Vectorization: extracting basic feature units

## The bag of words (BOW) model:

Each **word** is treated as a feature in a unit called **document**.

Each such word will become a feature

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

### Feature Vectors

(1, 5, 0, 0, 2, 0, 0)

(1, 1, 1, 0, 1, 0, 1)

(0, 1, 0, 1, 2, 1, 0)

# Vectorization: extracting basic feature units

## The bag of words (BOW) model:

Each **word** is treated as a feature in a unit called **document**.

Each such word will become a feature

- Alternative: **ngrams** – **unigram**, **bigram**, **trigram**

**ngrams:**

**unigrams**

**bigrams**

**trigrams**

```
['the', 'special', 'onion', 'soup', 'was', 'not', 'very', 'bad',  
'the special', 'special onion', 'onion soup', 'soup was',  
'was not', 'not very', 'very bad', 'the special onion', 'special onion soup',  
'onion soup was', 'soup was not', 'was not very', 'not very bad']
```

Until the next time 😊

