CHAPTER II

Text preprocessing & Data representation

Before a model processes text for a specific task, the text often needs to be **preprocessed** to improve model **performance** or to turn words and characters into a **format** the model can understand

Preprocessing

- Character Encoding
- Text Segmentation
- Text Cleaning
- Tokenization
- Text Encoding (Feature extraction)
- Corpora & Datasets
- Exploratory Data Analysis

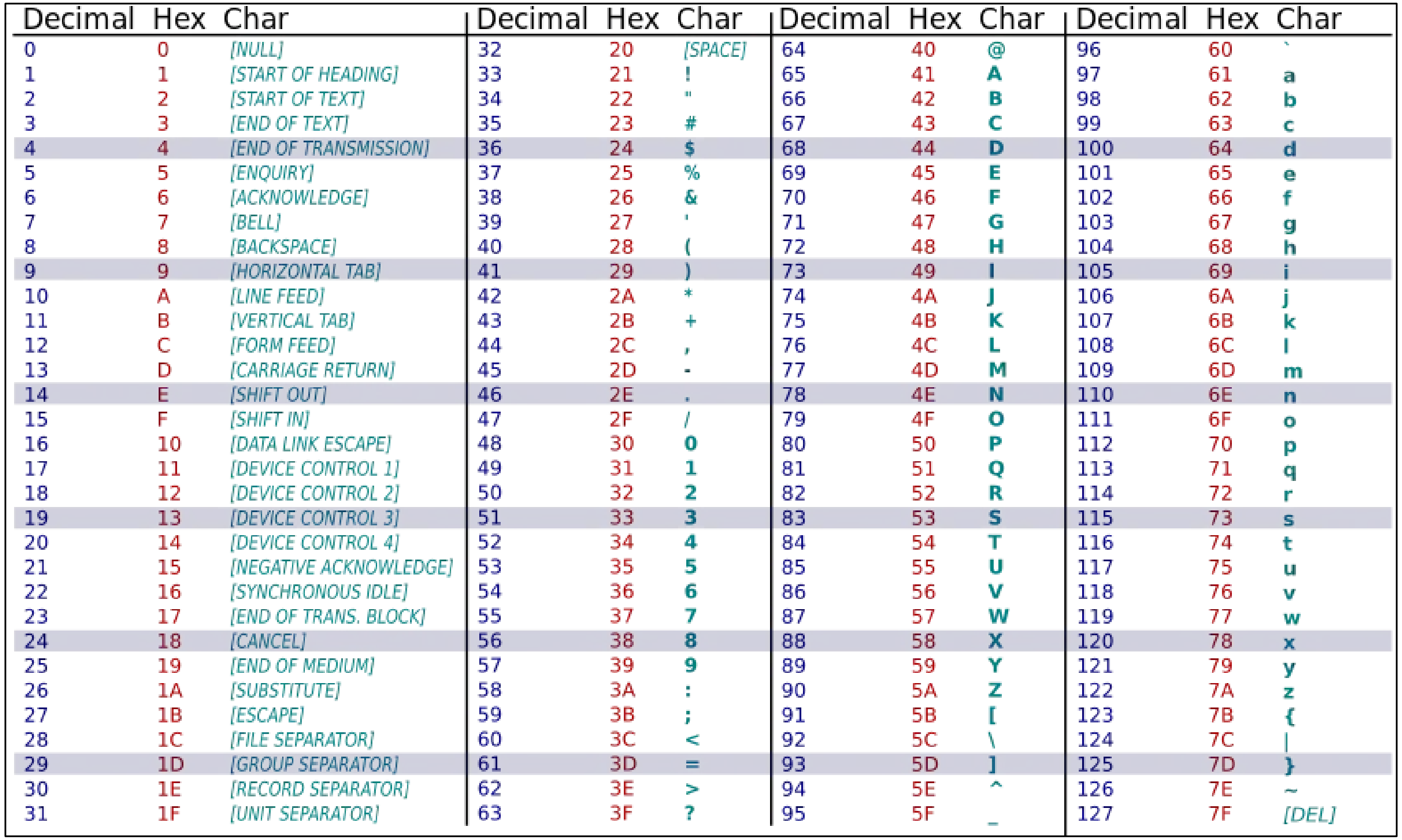
Charater encoding

Character encoding is a system for representing characters as numerical values, known as **code points**. These code points allow computers to store and manipulate text, which can then be displayed or used in other ways

- **Two main encoding standards:**
  - **ASCII:** assigns unique numbers to each symbol (128 code points).
  - **Unicode:** universal character encoding standard that could represent all the world's languages.

ASCII (American Standard Code for Information Interchange)

ASCII has **128 code points**, which means that it can represent 128 characters and symbols. Some of these code points represent instructions for the computer, while others represent printable characters such as letters and digits.



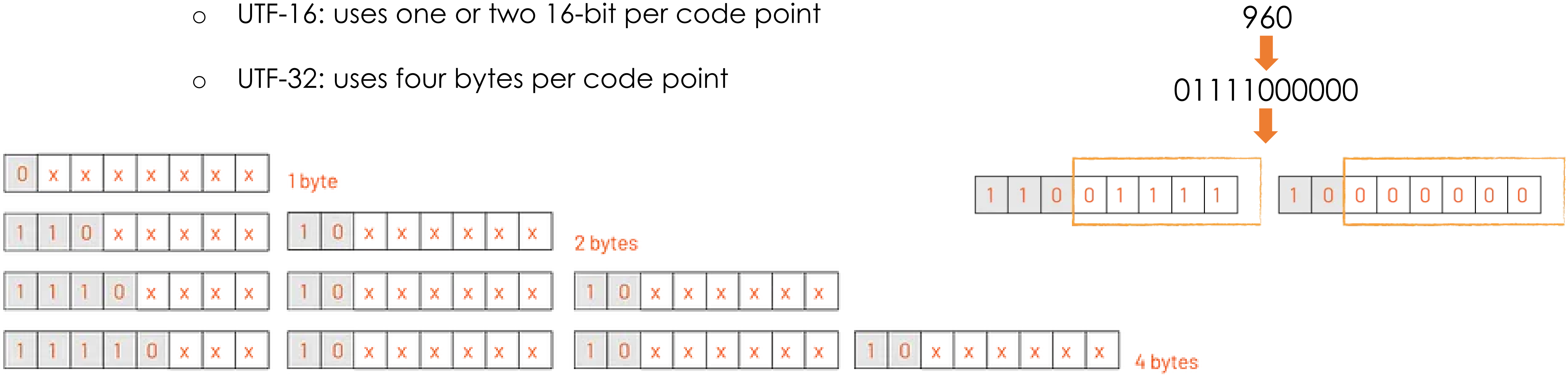
Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[END OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

- **ASCII Limitations:**
  - 7-bit encoding
  - Only 94 printable characters
  - Not suited for any language
  - Extended ASCII (8-bit encoding)

Unicode

A single, universal character encoding standard that could represent all the world’s languages

- 16-bit encoding scheme
- 65635 characters (0 to 10FFFF) noted (U+0000 to U+10FFFF)
- Suited for any language
  - UTF-8: uses one to four bytes per code point(compatible with ASCII)
  - UTF-16: uses one or two 16-bit per code point
  - UTF-32: uses four bytes per code point



960

01111000000

1	1	0	0	1	1	1	1

1	0	0	0	0	0	0	0

```
0xxxxxxx                                        1 byte
110xxxxx  10xxxxxx                              2 bytes
1110xxxx  10xxxxxx  10xxxxxx
11110xxx  10xxxxxx  10xxxxxx  10xxxxxx          4 bytes
```

UTF-8 Format having leading bits for 1 byte, 2 bytes, 3 bytes, 4 bytes

Unicode

Hello world!

⬇

ASCII/UTF-8: 48 65 6c 6c 6f 20 77 6f 72 6c 64 21

UTF-16: 0048 0065 006c 006c 006f 0020 0077 006f 0072 006c 0064 0021

UTF-32: 00000048 00000065 0000006c 0000006c 0000006f 00000020 00000077 0000006f 00000072 0000006c 00000064 00000021

UTF-8

100 → 64
233 → C3 A9
2357 → E0 A4 B5
128077 → F0 9F 91 8D

Hello world👍!

⬇

UTF-8: 48 65 6c 6c 6f 20 77 6f 72 6c 64 F0 9F 91 8D 21

Text segmentation

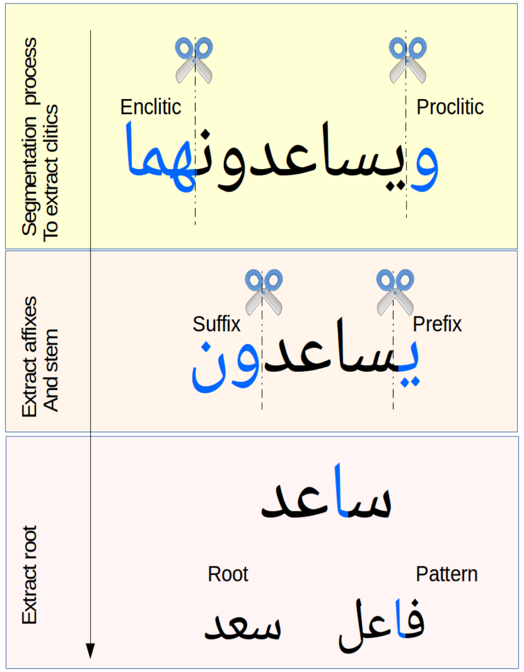
Text segmentation is the process of dividing written text into meaningful units, such as words, sentences or topics, using boundary markers

1- Word segmentation: dividing a string of written language into its component words (e,g: using word space)

- **Problem:**
  - Compounds:

    [ice box = ice-box = icebox], [cordon bleu], [ورد، وفر]
  - No word delimiter in some written scripts (e,g: Chinese)

    [美国会不同意。: The US will not agree **vs** The US Congress does not agree]
  - Detecting morphemes

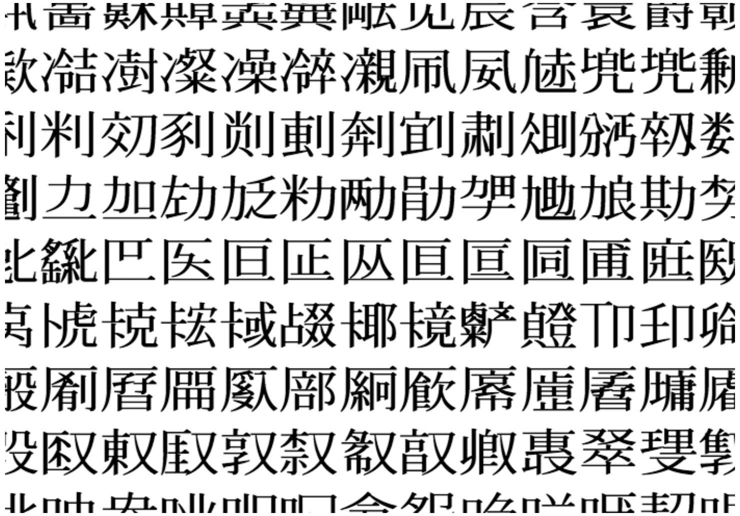


Text segmentation

Text segmentation is the process of dividing written text into meaningful units, such as words, sentences or topics, using boundary markers

2- Sentence segmentation: dividing a string of written language (text) into sentences (e,g: using punctuation). The period (.) is a reasonable approximation

- **Problem:**
  - Use of period in abbreviations (Mr. Smith went to..)
  - Not all written scripts contain regular punctuation
  - Sentence vs Paragraph detection

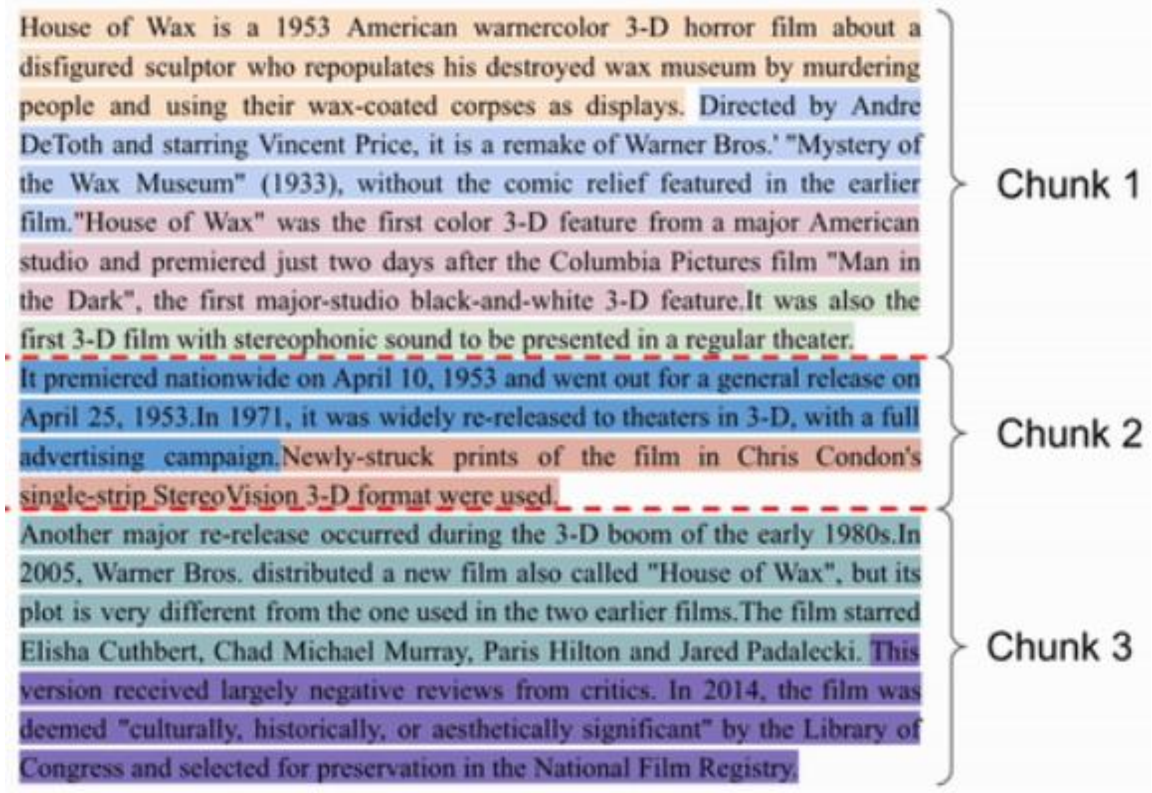


Text segmentation

Text segmentation is the process of dividing written text into meaningful units, such as words, sentences or topics, using boundary markers

3- Topic segmentation: dividing a text into topics or discourse

- **Subtasks:**
  - Topic identification
  - Text segmentation



House of Wax is a 1953 American warnercolor 3-D horror film about a disfigured sculptor who repopulates his destroyed wax museum by murdering people and using their wax-coated corpses as displays. Directed by Andre DeToth and starring Vincent Price, it is a remake of Warner Bros.' "Mystery of the Wax Museum" (1933), without the comic relief featured in the earlier film."House of Wax" was the first color 3-D feature from a major American studio and premiered just two days after the Columbia Pictures film "Man in the Dark", the first major-studio black-and-white 3-D feature.It was also the first 3-D film with stereophonic sound to be presented in a regular theater.

Chunk 1

It premiered nationwide on April 10, 1953 and went out for a general release on April 25, 1953.In 1971, it was widely re-released to theaters in 3-D, with a full advertising campaign.Newly-struck prints of the film in Chris Condon's single-strip StereoVision 3-D format were used.

Chunk 2

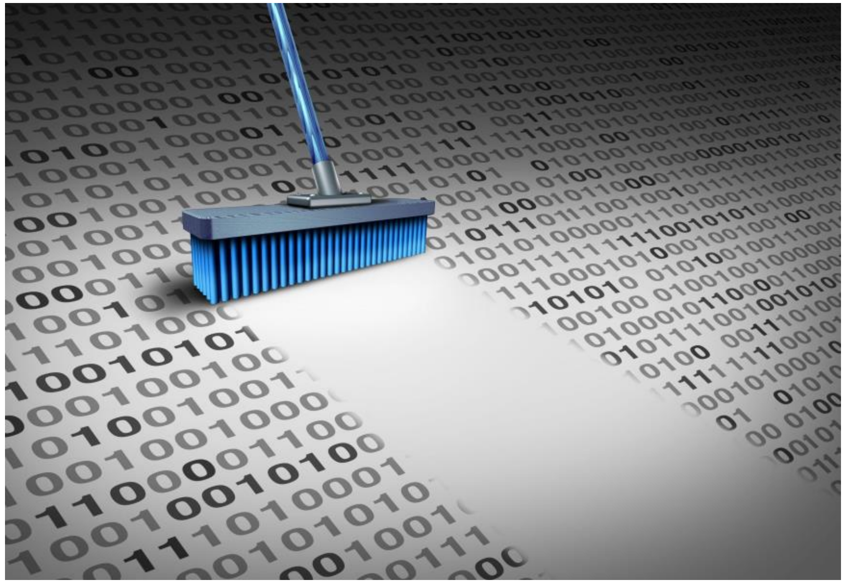
Another major re-release occurred during the 3-D boom of the early 1980s.In 2005, Warner Bros. distributed a new film also called "House of Wax", but its plot is very different from the one used in the two earlier films.The film starred Elisha Cuthbert, Chad Michael Murray, Paris Hilton and Jared Padalecki. This version received largely negative reviews from critics. In 2014, the film was deemed "culturally, historically, or aesthetically significant" by the Library of Congress and selected for preservation in the National Film Registry.

Chunk 3

Text cleaning

Text cleaning is the process of removing unnecessary data from text in order to get more consistent and standardized format. Cleaning depends on the type of task and data.

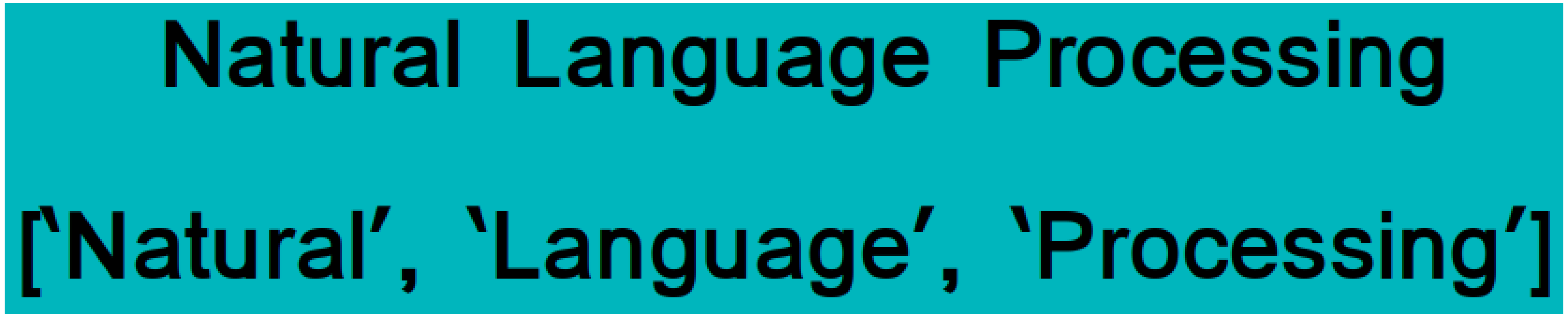
- Lowercasing Text
- Removing Punctuations
- Removing Numbers
- Removing Extra space
- Replacing the repetitions of punctations
- Removing Emojis and emoticons
- Removing Stop words
- Removing Diacritics



Tokenization

Tokenization is **splitting** a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words, terms, numbers or punctuation marks. Each of these smaller units are called **tokens.**

- **Word Tokenization:** Splitting a sentence into individual words.
- **Sentence Tokenization:** Breaking a paragraph into separate sentences.

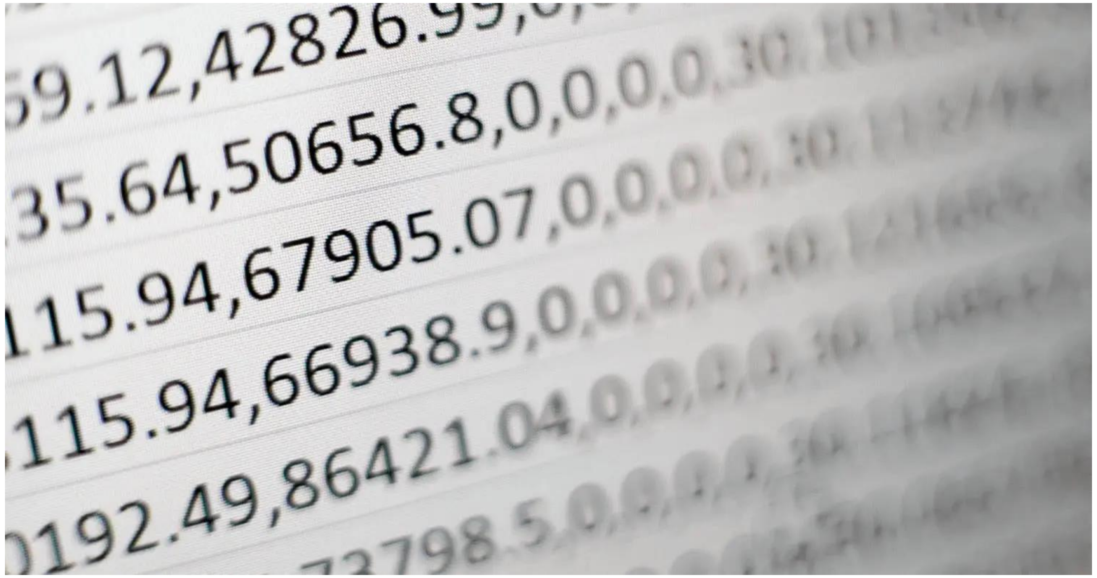


Natural Language Processing

[‘Natural’, ‘Language’, ‘Processing’]

Text encoding

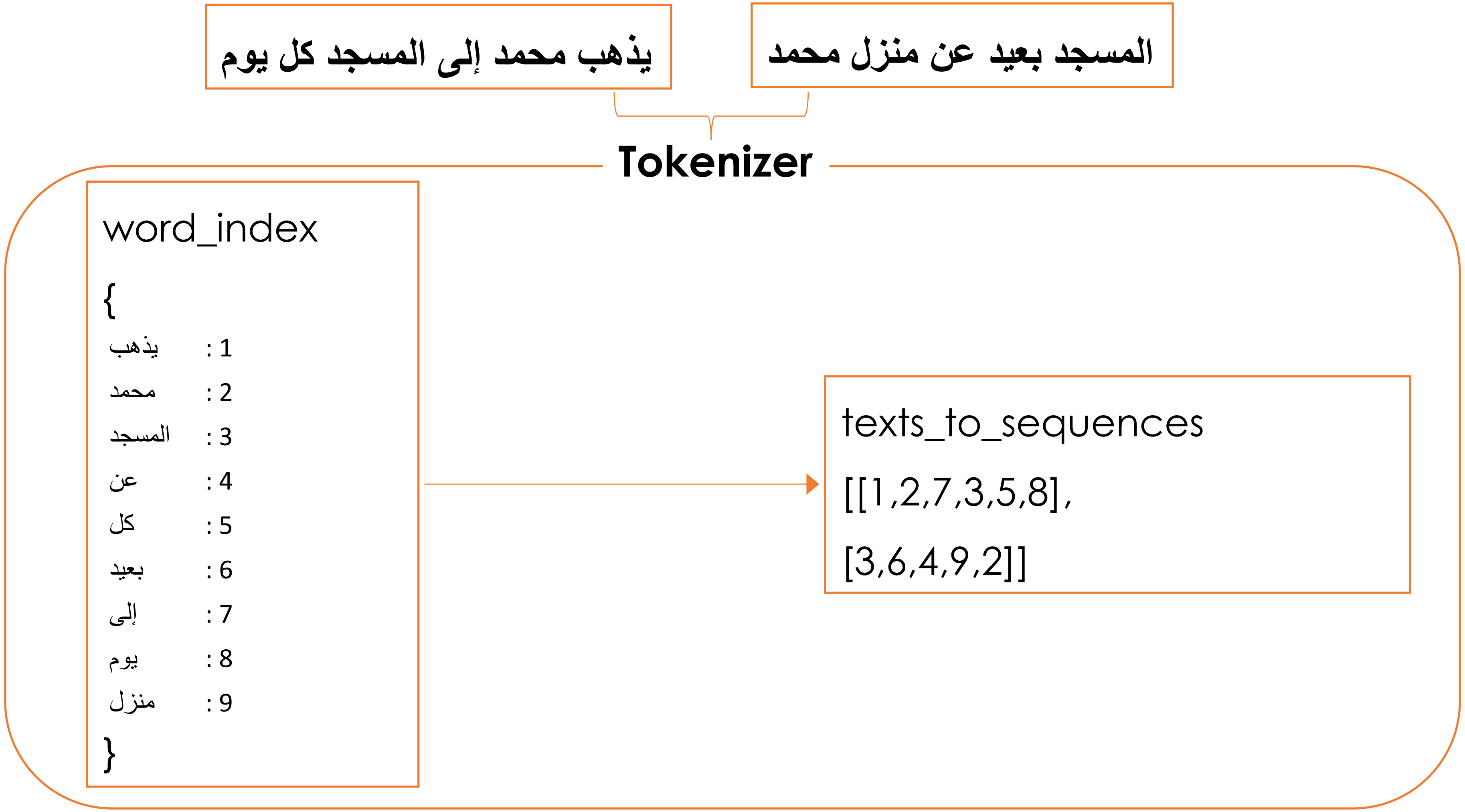
Text encoding is a process to **convert** meaningful **text** into **number/vector** representation so as to preserve the **context** and relationship between words and sentences, such that a machine can understand the pattern associated in any text and can make out the context of sentences



Text encoding techniques

- **Index-based encoding**
- **Bag-of-Words**
- **N-Grams**
- **TF-IDF**
- **One-Hot-Encoding**
- **Word Embeddings**
  - **Word2Vec (CBoW, Skip-Gram)**
  - **GLoVE**

Index-based encoding



يذهب محمد إلى المسجد كل يوم

المسجد بعيد عن منزل محمد

Tokenizer

word_index

{

يذهب : 1

محمد : 2

المسجد : 3

عن : 4

كل : 5

بعيد : 6

إلى : 7

يوم : 8

منزل : 9

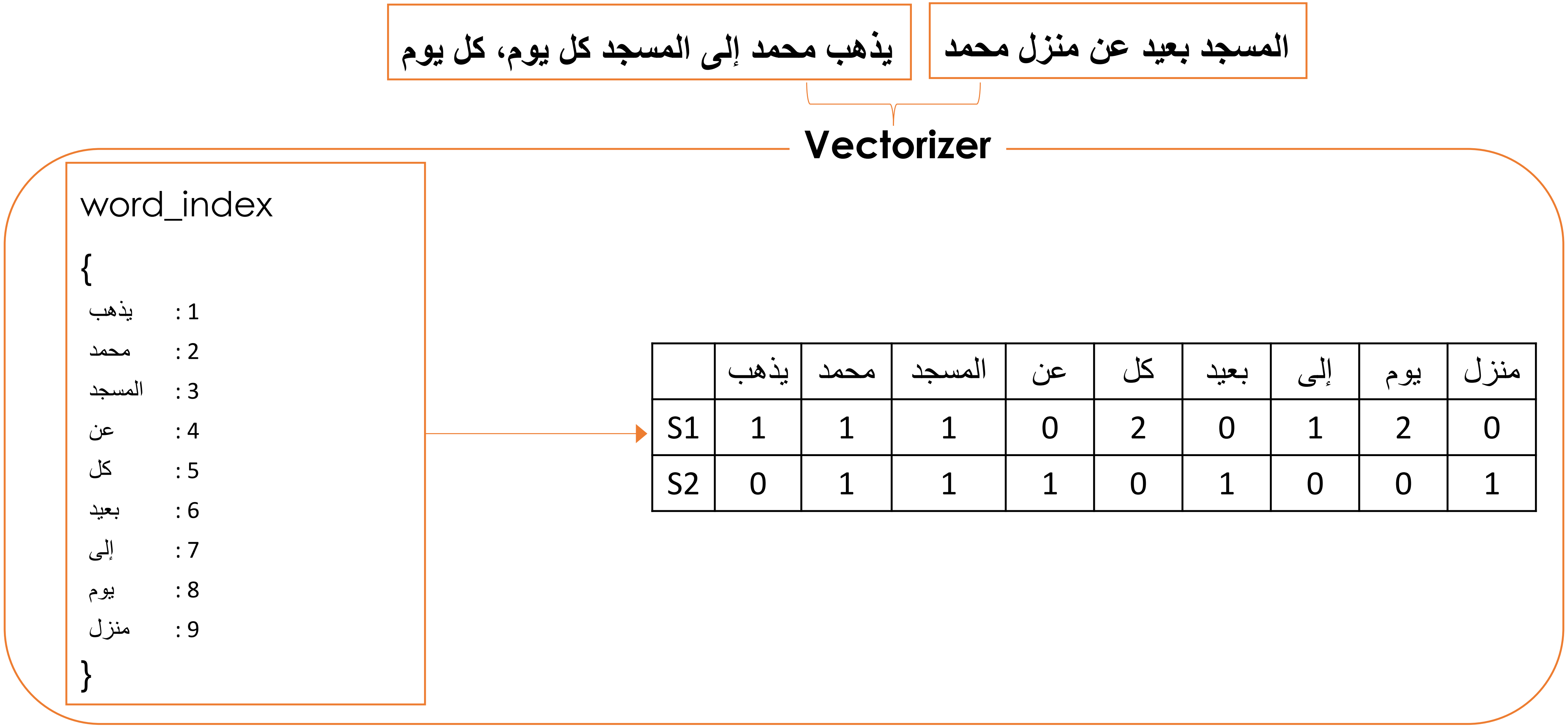
}

texts_to_sequences

[[1,2,7,3,5,8],

[3,6,4,9,2]]

Bag-of-Words (BoW)



يذهب محمد إلى المسجد كل يوم، كل يوم

المسجد بعيد عن منزل محمد

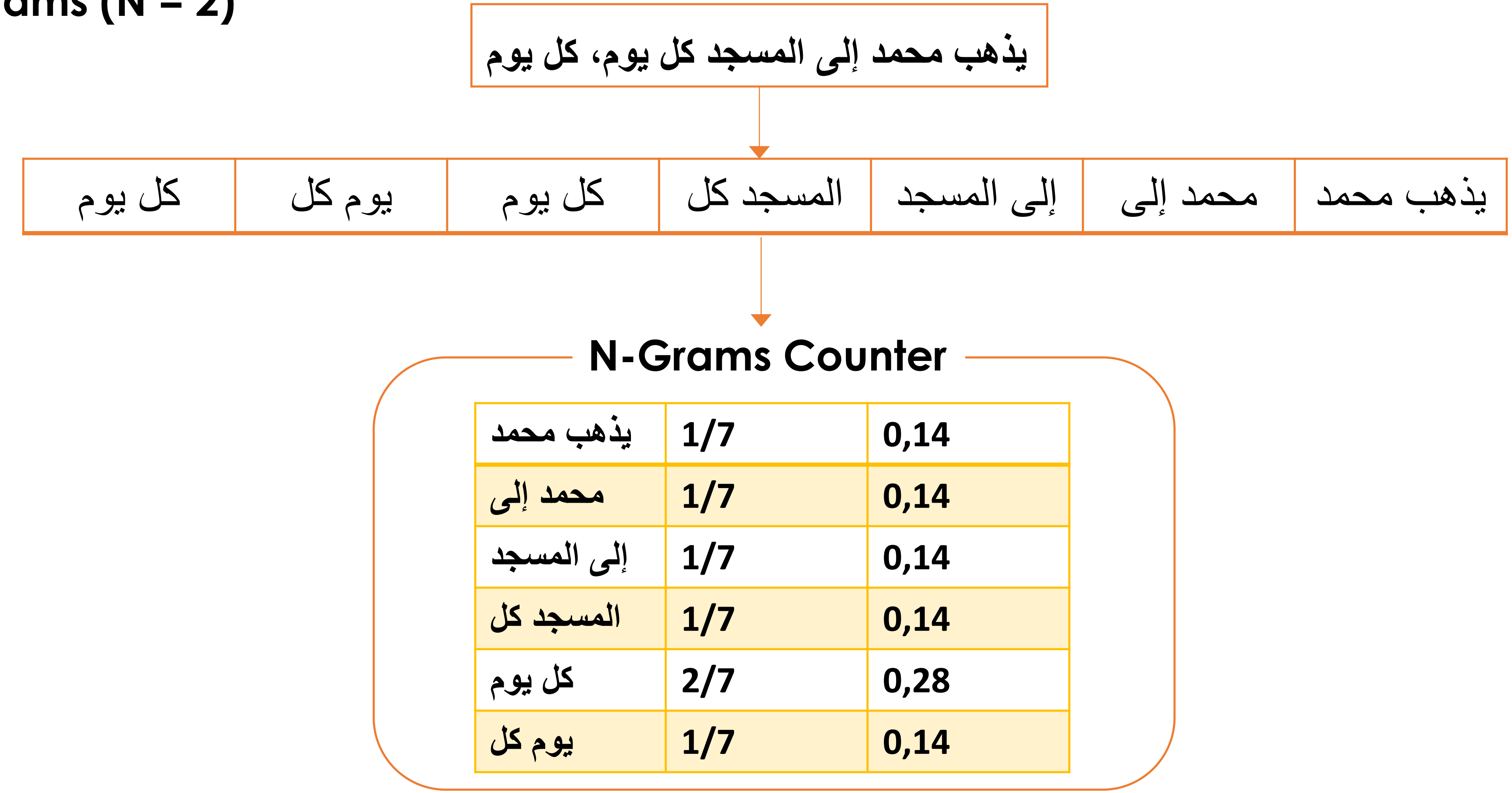
Vectorizer

```
word_index
{
 يذهب     : 1
 محمد     : 2
 المسجد   : 3
 عن       : 4
 كل       : 5
 بعيد     : 6
 إلى      : 7
 يوم      : 8
 منزل     : 9
}
```

	يذهب	محمد	المسجد	عن	كل	بعيد	إلى	يوم	منزل
S1	1	1	1	0	2	0	1	2	0
S2	0	1	1	1	0	1	0	0	1

N-Grams

- **N-Grams (N = 2)**



يذهب محمد إلى المسجد كل يوم، كل يوم

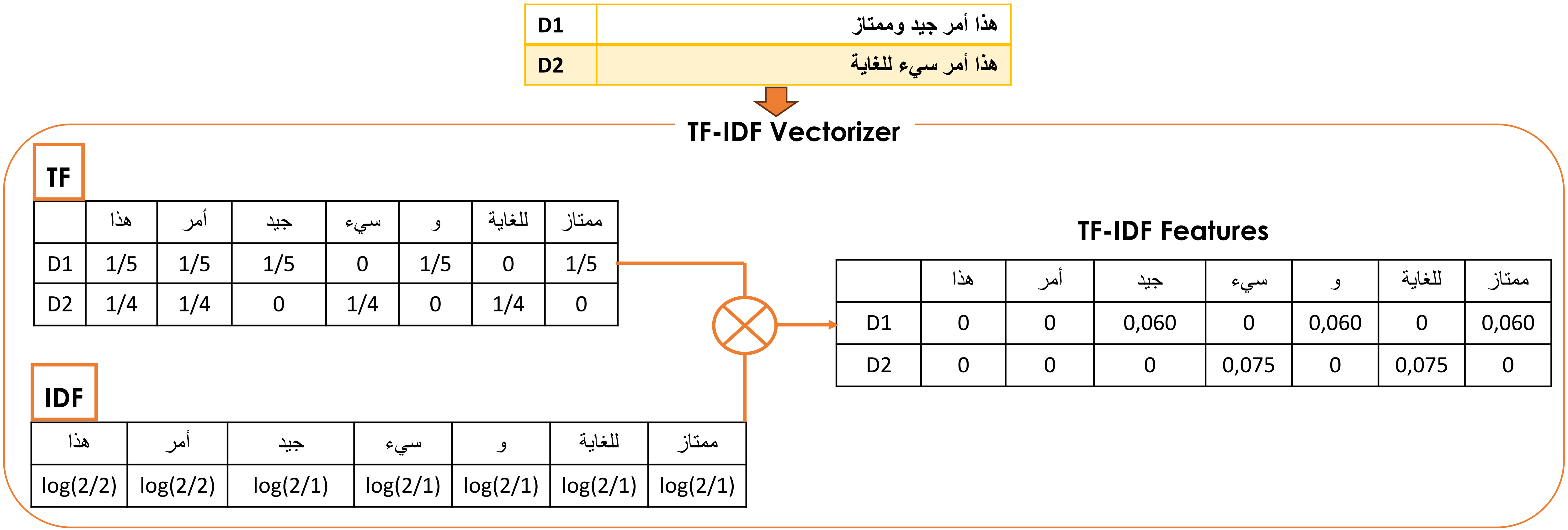
يذهب محمد	محمد إلى	إلى المسجد	المسجد كل	كل يوم	يوم كل	كل يوم

N-Grams Counter

يذهب محمد	1/7	0,14
محمد إلى	1/7	0,14
إلى المسجد	1/7	0,14
المسجد كل	1/7	0,14
كل يوم	2/7	0,28
يوم كل	1/7	0,14

TF-IDF

- Weights each word by its importance
- TF (Term Frequency) = *Number of occurrences of the word in document / Number of words in document*
- IDF (Inverse Docment Frequency) = *log(number of documents in the corpus / number of documents that include the word)*



D1	**هذا أمر جيد وممتاز**
D2	**هذا أمر سيء للغاية**

TF-IDF Vectorizer

TF

	هذا	أمر	جيد	سيء	و	للغاية	ممتاز
D1	1/5	1/5	1/5	0	1/5	0	1/5
D2	1/4	1/4	0	1/4	0	1/4	0

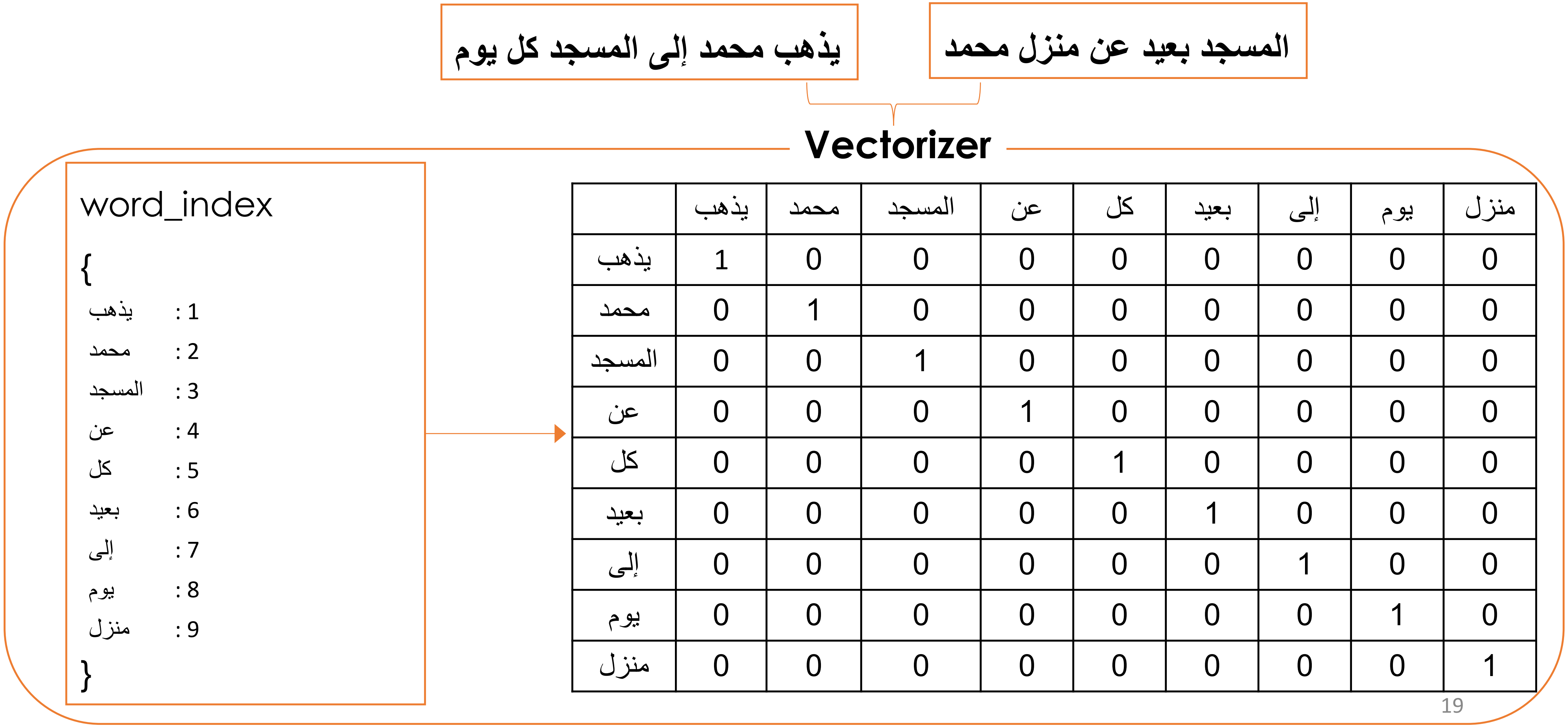
IDF

هذا	أمر	جيد	سيء	و	للغاية	ممتاز
log(2/2)	log(2/2)	log(2/1)	log(2/1)	log(2/1)	log(2/1)	log(2/1)

TF-IDF Features

	هذا	أمر	جيد	سيء	و	للغاية	ممتاز
D1	0	0	0,060	0	0,060	0	0,060
D2	0	0	0	0,075	0	0,075	0

One-Hot-Encoding



يذهب محمد إلى المسجد كل يوم

المسجد بعيد عن منزل محمد

Vectorizer

word_index

{

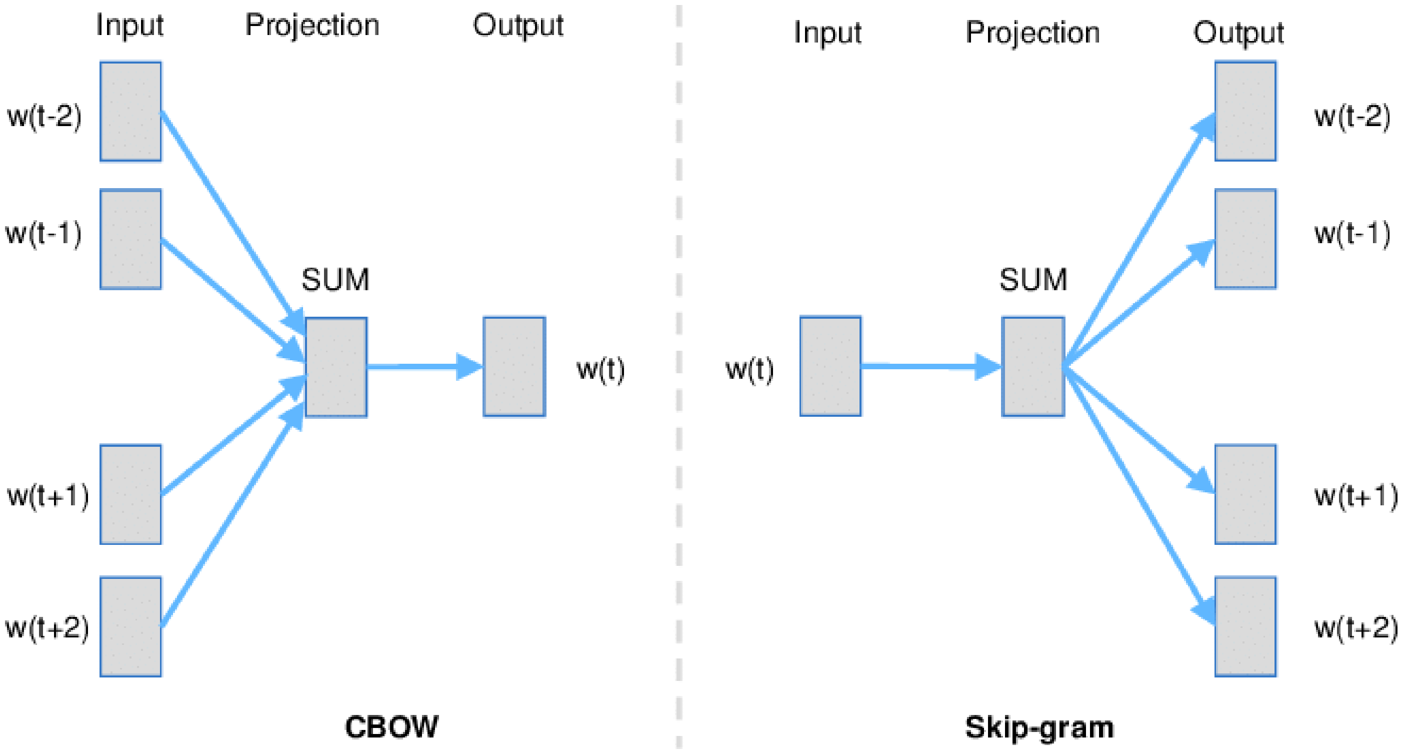
يذهب	: 1
محمد	: 2
المسجد	: 3
عن	: 4
كل	: 5
بعيد	: 6
إلى	: 7
يوم	: 8
منزل	: 9

}

	يذهب	محمد	المسجد	عن	كل	بعيد	إلى	يوم	منزل
يذهب	1	0	0	0	0	0	0	0	0
محمد	0	1	0	0	0	0	0	0	0
المسجد	0	0	1	0	0	0	0	0	0
عن	0	0	0	1	0	0	0	0	0
كل	0	0	0	0	1	0	0	0	0
بعيد	0	0	0	0	0	1	0	0	0
إلى	0	0	0	0	0	0	1	0	0
يوم	0	0	0	0	0	0	0	1	0
منزل	0	0	0	0	0	0	0	0	1

Word2Vec

A word2vec model is a simple neural network model with a single hidden layer. The task of this model is to predict the nearby words for each and every word in a sentence.



يذهب محمد إلى المسجد كل يوم

Word2Vec

يذهب	[0.2, 0.3, -0.1, 0.5, ...]
محمد	[0.1, -0.4, 0.6, -0.2, ...]
إلى	[0.3, -0.2, 0.4, 0.1, ...]
المسجد	[-0.5, 0.2, 0.3, -0.1, ...]
كل	[0.4, 0.1, -0.3, 0.2, ...]
يوم	[0.6, -0.3, 0.2, 0.4, ...]

Word2Vec (Training data)

We need a labeled dataset to train a neural network model. This means the dataset should have a set of inputs and an output for every input.

يذهب محمد إلى المسجد كل يوم

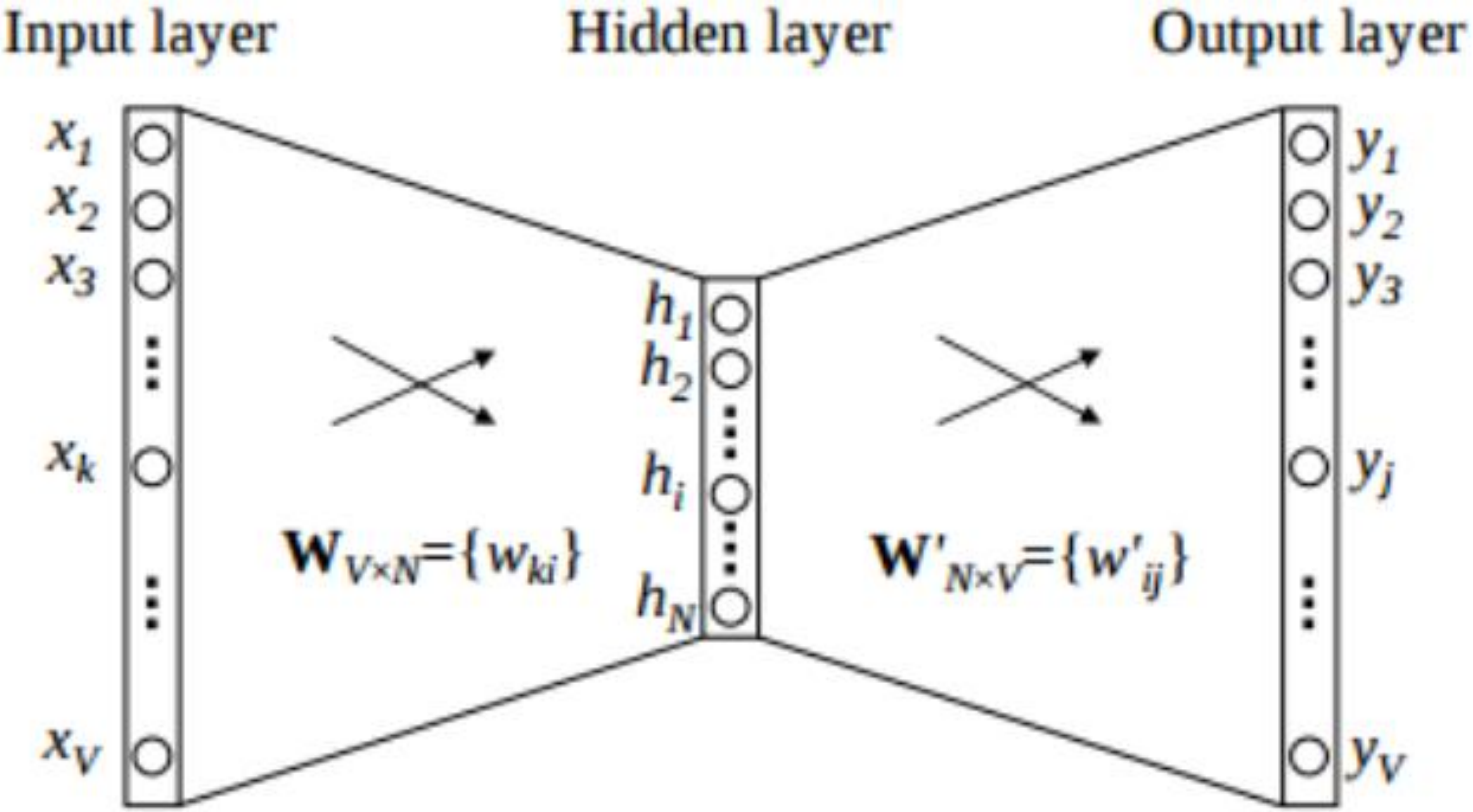
W = 2

Input	Output
يذهب	محمد
يذهب	إلى
محمد	يذهب
محمد	إلى
محمد	المسجد
إلى	يذهب
إلى	محمد
إلى	المسجد
إلى	كل
…	…

Word2Vec (Embedding)

Suppose the number of unique words in this **dataset** is **5,000** and we wish to create word **vectors** of size **100** each

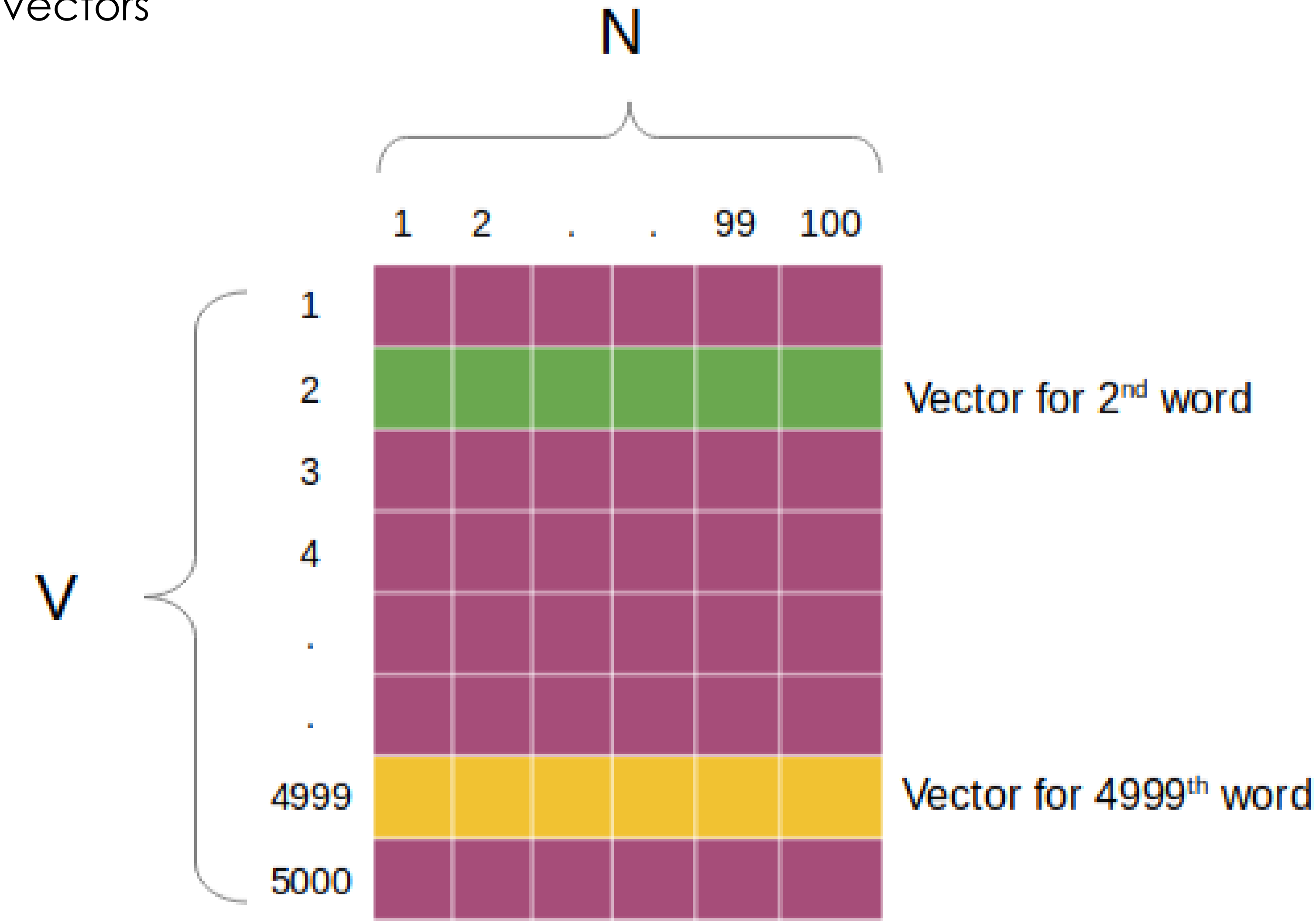
V = 5,000
N = 100



Word2Vec Model Architecture

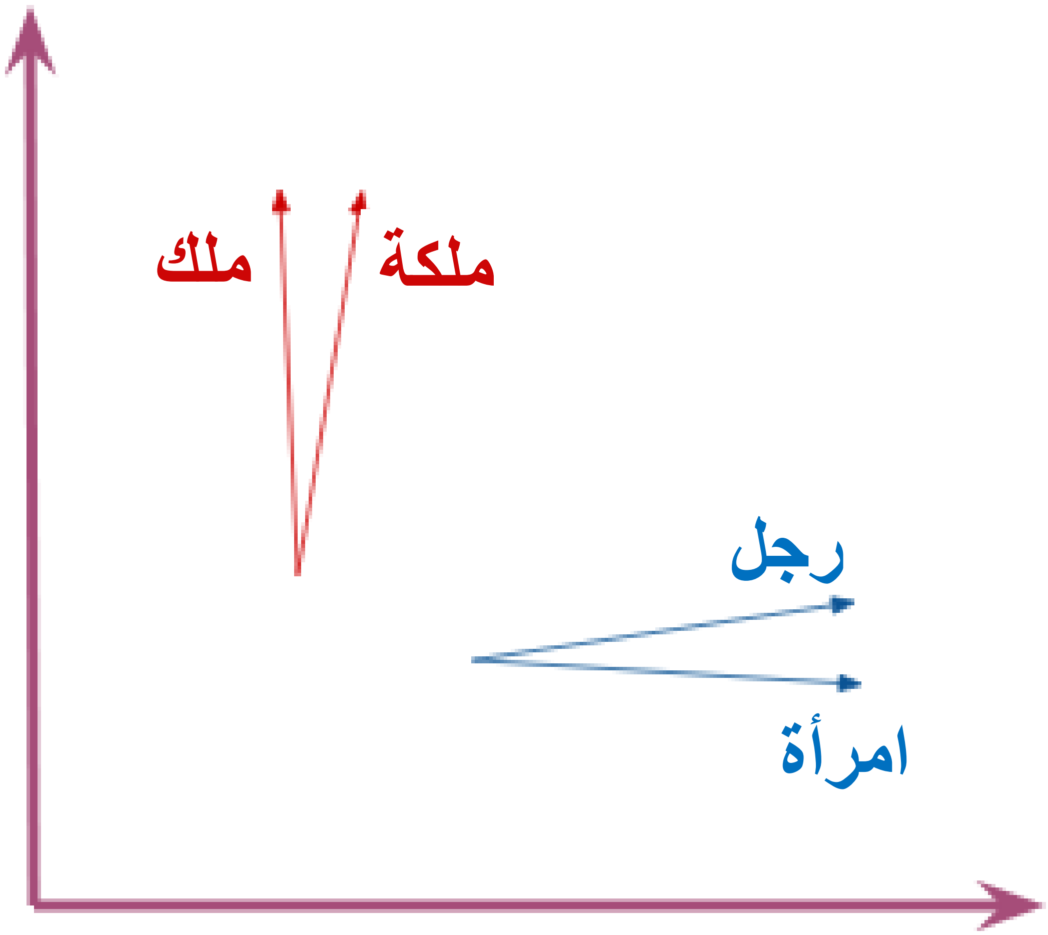
Word2Vec (After training)

Once this model is trained, we can easily extract the learned weight matrix $\mathbf{W}_{\mathbf{V \times N}}$ and use it to extract the word vectors



W_{VxN} weight matrix

Word2Vec (2D Visualization)



Corpora

A corpus is a significant collection of texts written in everyday language that computers can read.

Sources: Digital text, Audio transcripts, Scanned documents

Importance of corpora in NLP:

- Understand languages
- Text structure
- Discover word relationships
- Learning process

Corpora use cases in NLP

- **Training Machine Learning Models**: Corpora are used to teach (train and refine) machine learning models.
- **Language Understanding**: Learn how words and phrases are used in context. Help to generate new languages
- **Rule-Based Systems**: Used by linguists and NLP experts to develop and test linguistic rules and patterns.
- **Lexicon and Semantics**: Lexicons (dictionaries) are created and expanded with the help of corpora.
- **Statistical Analysis**: Corpora give information that is necessary for probabilistic NLP approaches to examine word frequency distributions, co-occurrence patterns, and other statistical features.
- **Domain-Specific Knowledge**: Specific to particular topics or fields. (legal documents, medical records,..)

Types of Corpora

Text Corpora
Multimodal Corpora
Parallel Corpora
Time-Series Corpora
Annotated Corpora

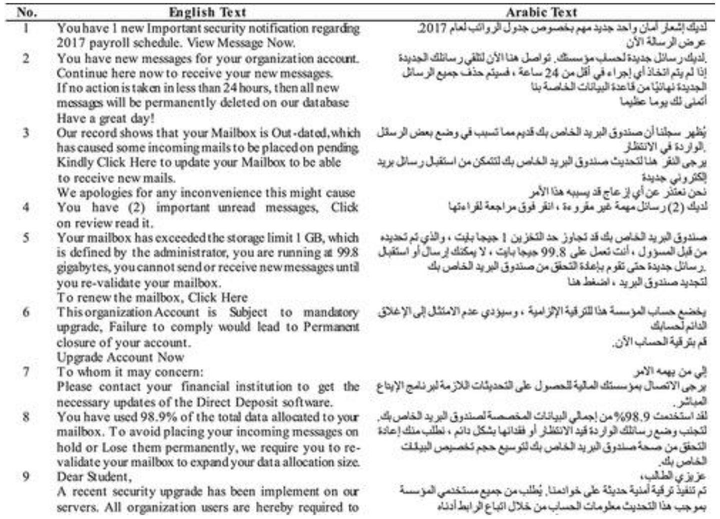
Text Corpora
- General-Purpose
- Specialized
- Comparable

Multimodal Corpora
- Text-Image
- Text-Speech
- Text-Video

Parallel Corpora
- Bilingual
- Comparable

Time-Series Corpora
- Historical
- Temporal

Annotated Corpora
- Linguistically
- Sentiment



No.	English Text	Arabic Text
1	You have 1 new Important security notification regarding 2017 payroll schedule. View Message Now.	لديك إشعار أمان واحد جديد مهم بخصوص جدول الرواتب لعام 2017. عرض الرسالة الآن
2	You have new messages for your organization account. Continue here now to receive your new messages. If no action is taken in less than 24 hours, then all new messages will be permanently deleted on our database Have a great day!	.لديك رسائل جديدة لحساب مؤسستك. تواصل هنا الآن لتلقي رسائلك الجديدة إذا لم يتم اتخاذ أي إجراء في أقل من 24 ساعة ، فسيتم حذف جميع الرسائل الجديدة نهائيًا من قاعدة البيانات الخاصة بنا أتمنى لك يوما عظيما
3	Our record shows that your Mailbox is Out-dated, which has caused some incoming mails to be placed on pending. Kindly Click Here to update your Mailbox to be able to receive new mails. We apologies for any inconvenience this might cause	يُظهر سجلنا أن صندوق البريد الخاص بك قديم مما تسبب في وضع بعض الرسائل .الواردة في الانتظار يرجى النقر هنا لتحديث صندوق البريد الخاص بك لتتمكن من استقبال رسائل بريد إلكتروني جديدة نحن نعتذر عن أي إزعاج قد يسببه هذا الأمر
4	You have (2) important unread messages, Click on review read it.	لديك (2) رسائل مهمة غير مقروءة ، انقر فوق مراجعة لقراءتها
5	Your mailbox has exceeded the storage limit 1 GB, which is defined by the administrator, you are running at 99.8 gigabytes, you cannot send or receive new messages until you re-validate your mailbox. To renew the mailbox, Click Here	صندوق البريد الخاص بك قد تجاوز حد التخزين 1 جيجا بايت ، والذي تم تحديده من قبل المسؤول ، أنت تعمل على 99.8 جيجا بايت ، لا يمكنك إرسال أو استقبال .رسائل جديدة حتى تقوم بإعادة التحقق من صندوق البريد الخاص بك لتجديد صندوق البريد ، اضغط هنا
6	This organization Account is Subject to mandatory upgrade, Failure to comply would lead to Permanent closure of your account. Upgrade Account Now	يخضع حساب المؤسسة هذا للترقية الإلزامية ، وسيؤدي عدم الامتثال إلى الإغلاق الدائم لحسابك قم بترقية الحساب الآن.
7	To whom it may concern: Please contact your financial institution to get the necessary updates of the Direct Deposit software.	إلى من يهمه الامر يرجى الاتصال بمؤسستك المالية للحصول على التحديثات اللازمة لبرنامج الإيداع المباشر.
8	You have used 98.9% of the total data allocated to your mailbox. To avoid placing your incoming messages on hold or Lose them permanently, we require you to re-validate your mailbox to expand your data allocation size.	لقد استخدمت 98.9% من إجمالي البيانات المخصصة لصندوق البريد الخاص بك. لتجنب وضع رسائلك الواردة قيد الانتظار أو فقدانها بشكل دائم ، نطلب منك إعادة التحقق من صحة صندوق البريد الخاص بك لتوسيع حجم تخصيص البيانات الخاص بك.
9	Dear Student, A recent security upgrade has been implement on our servers. All organization users are hereby required to	عزيزي الطالب، تم تنفيذ ترقية أمنية حديثة على خوادمنا. يُطلب من جميع مستخدمي المؤسسة بموجب هذا التحديث معلومات الحساب من خلال اتباع الرابط أدناه

Corpus features

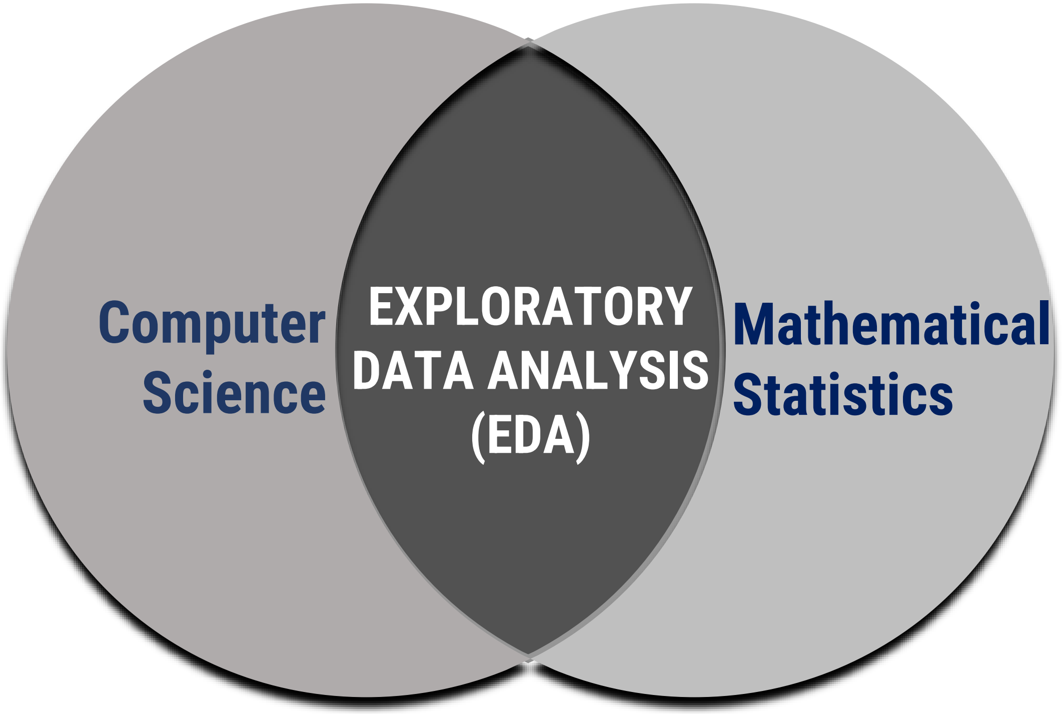
Large Corpus Size
High-Quality Data
Clean Data
Diversity
Annotation
Metadata

- As large as possible
- Influence on output quality
- Eliminate usefulness data
- Representative data
- Help supervise learning
- Provide context and origin

Exploratory Data Analysis (EDA)

DEFINITION

EDA involves examining and visualizing data sets to summarize their main characteristics, often with the help of statistical graphics and other data visualization techniques.



Exploratory Data Analysis (EDA)

KEY STEPS & TECHNIQUES

Understand The Data

- Examine the dataset's structure
- Check for missing values
- Explore basic summary statistics (mean, median,..)

Univariate Analysis

- Examine the distribution of each variable individually
- Use histograms, box plots, and summary statistics
- Identify outliers and potential errors in the data

Bivariate Analysis

- Explore relationships between pairs of variables
- Use scatter plots, correlation matrices, and cross-tabulations
- Identify potential patterns or trends

Multivariate Analysis

- Extend the analysis to multiple variables simultaneously
- Use techniques like heatmaps and pair plots to visualize relationships
- Identify potential clusters or groups in the data

Visualization

- Utilize various data visualization techniques such as bar charts, pie charts, line plots,…
- Consider using tools like matplotlib, seaborn, or Plotly in Python for creating interactive visualizations

Feature Engineering

- Create new variables or transform existing ones to extract more information from the data
- Handle categorical variables through encoding or creating dummy variables

Dimensionality Reduction

- Use techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality and visualize high-dimensional datasets

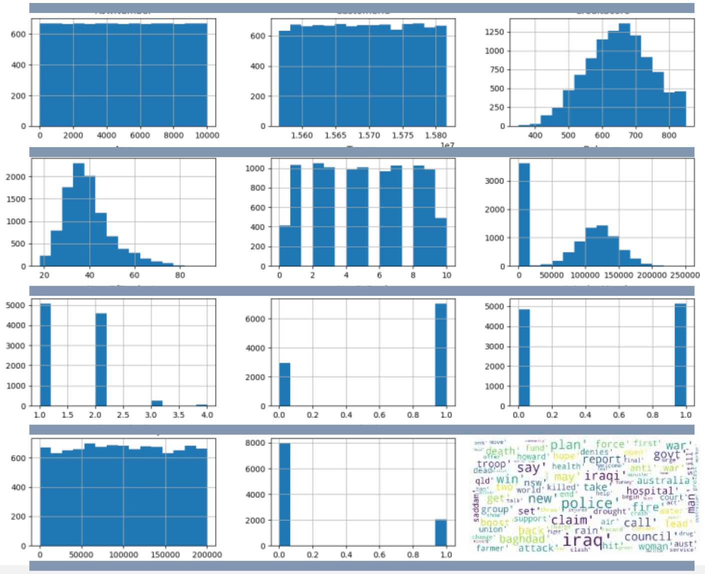
Statistical Testing

- Conduct hypothesis testing to validate assumptions or test for significant differences between groups

Exploratory Data Analysis (EDA)

EDA ON TEXTUAL DATA

- Word frequency analysis
  - Number of words in the text/corpus
  - Most frequently used words
  - Stop words frequency
  - N-gram exploration
  - Wordcloud
- Sentence/text length analysis
- Average word length analysis
- Topic modeling
- Sentiment analysis
- POS analysis


1.560 1.565 1.570 1.575 1.580 1e7
plan force war death report govt say iraqi australia police fire claim call council iraq baghdad attack farmer hospital new win troop