AI CAMP

school of ai Algiers

Dear wizards

Welcome to NLP Track

What text data is?

Text Preprocessing steps

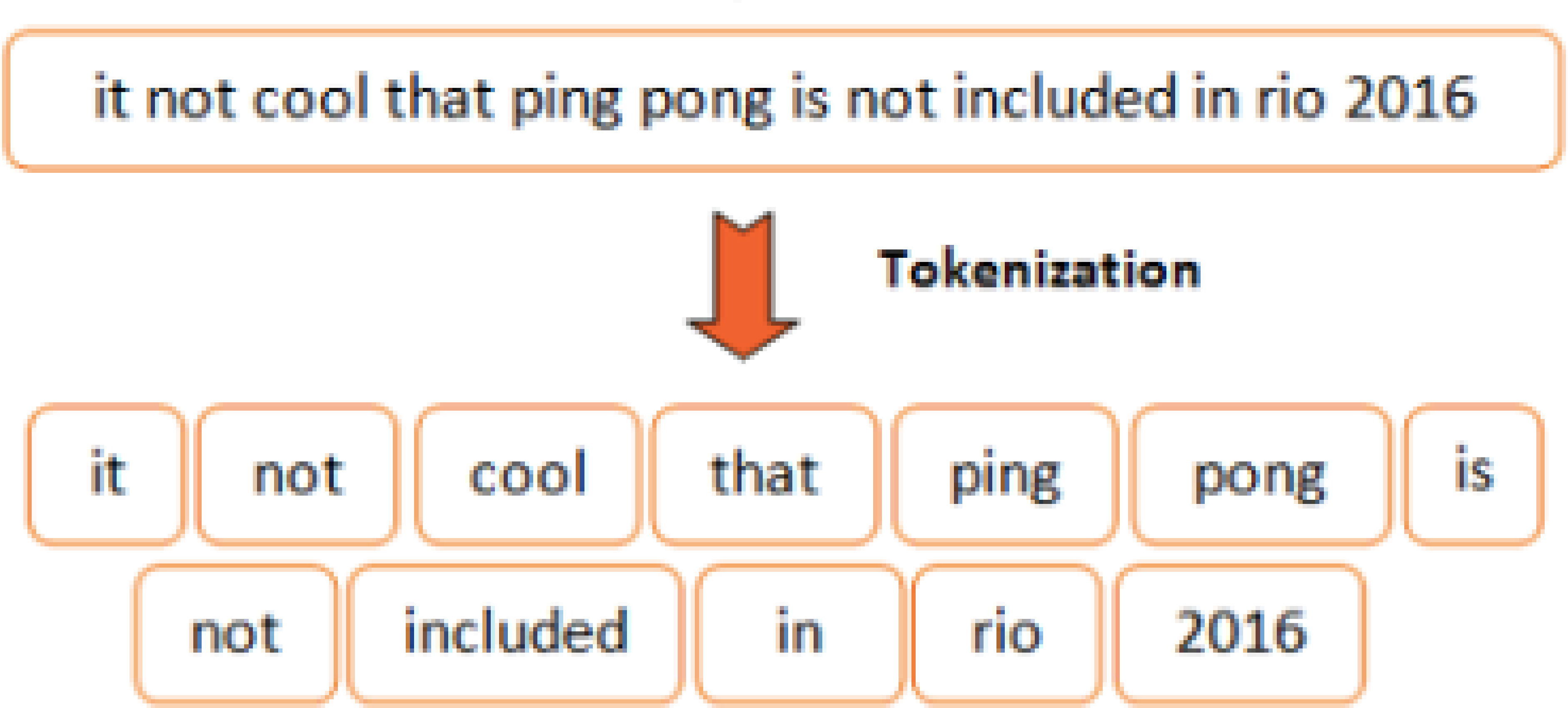
We start by:

- Removing punctuations: ‘!”#$%&'()*+,-./:;?@[\]^_`{|}~’
- Removing URLs
- Removing numbers
- Lower casing: Text → text

Punctuation marks can be considered as noise in some contexts

Tokenization:

the text is splited into smaller meaningful units and tokens



it not cool that ping pong is not included in rio 2016

Tokenization

it | not | cool | that | ping | pong | is

not | included | in | rio | 2016

Issues in tokenization:

Finland’s capital → Finland ? Finlands ? Finland’s

San Francisco → San Francesco or San | Francesco

For exemple japanese and german the sentence have a lot of prefix and suffix as お元気ですか (How are you?)

Max match principle:

it identifies the longest known word in the vocabulary and splits that word off the front

Thecatinthehat → The cat in the hat

Thetabledownthere → Theta bled own there

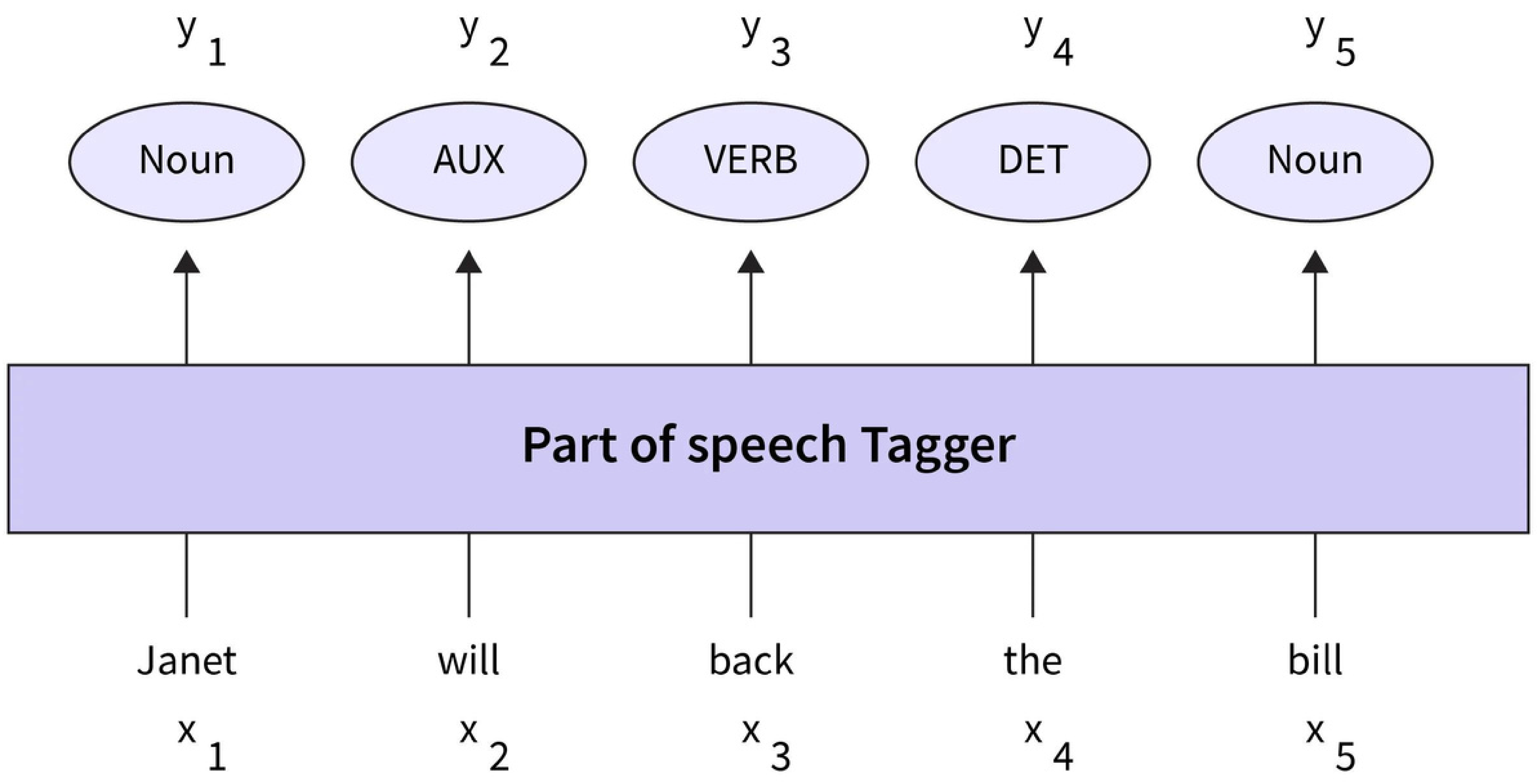
Part of speach (POS):

assigning a grammatical category or part-of-speech label such as noun, verb, adjective, pronoun, etc

The back door → adj

On my back → noun

Promissed to back the bill → verb



AI CAMP

school of ai Algiers

It serves several purposes as a preprocessing step:

POS tagging helps in understanding the grammatical structure of a sentence. It provides information about the roles of words in forming phrases and sentences

Feature Extraction: machine translation, ner, text classification

Lemmatization and Stemming

stemming & lemmatization:

Text Normalization techniques, where we return each word to the root word from wich it is derived

am, are, is → be

the boy’s cars are different colors →

the boy car be different color

Stemming is the process of removing the last few characters of a given word, to obtain a shorter form, even if that form doesn't have any meaning.

automate, automatic, automation → automat

stem

Lemmatization is a text normalization technique in natural language processing (NLP) that **involves reducing words to their base or root form** based on the word meaning and the POS consideration

Word: "meeting" (verb)
Lemmatized form: "meet"
Word: "meeting" (noun)
Lemmatized form: "meeting"

Stemming and lemmatization in Iformation Retrieval.

Grouping words with common stem together.

For exemple, a search on reads, also finds read, reading, and readable

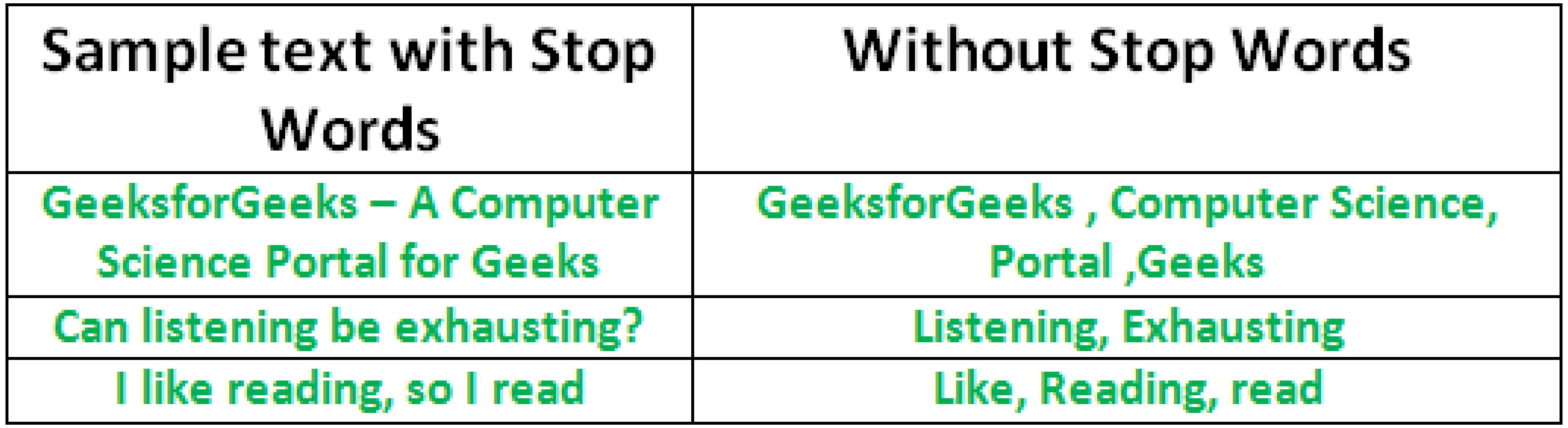
Stop Words

the process of eliminating words that are so widely used that they carry very little useful information

[”this”, “is”, “a”, “test”, “sentence”]

✅ X X ✅ ✅

Stop Words

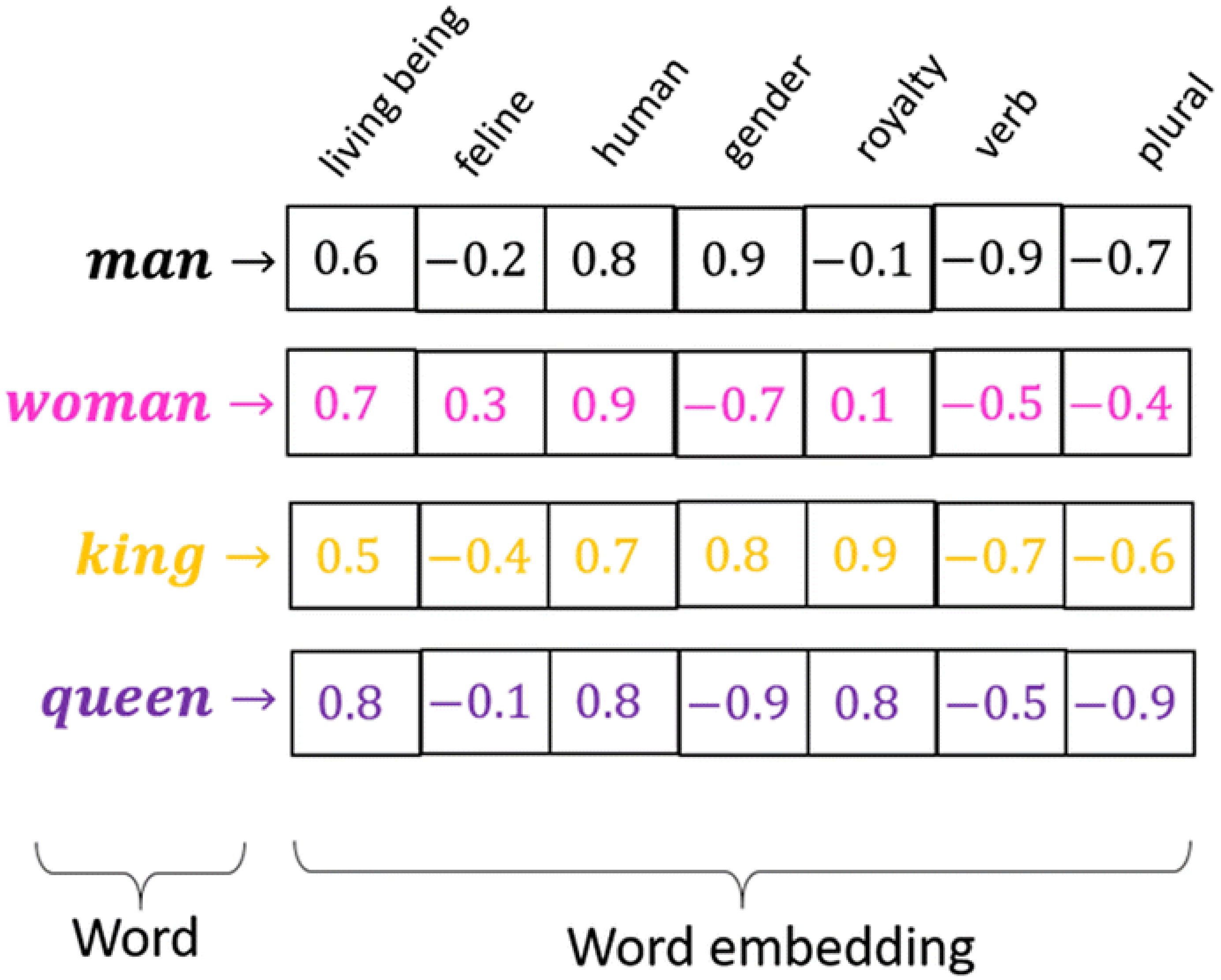


Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

removing the stop words depends on the task it self

Word Embedding:

word embedding is the step where we give a numerical represation to every single word or token, so that we could know its meaning and relationships between each other



	living being	feline	human	gender	royalty	verb	plural
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

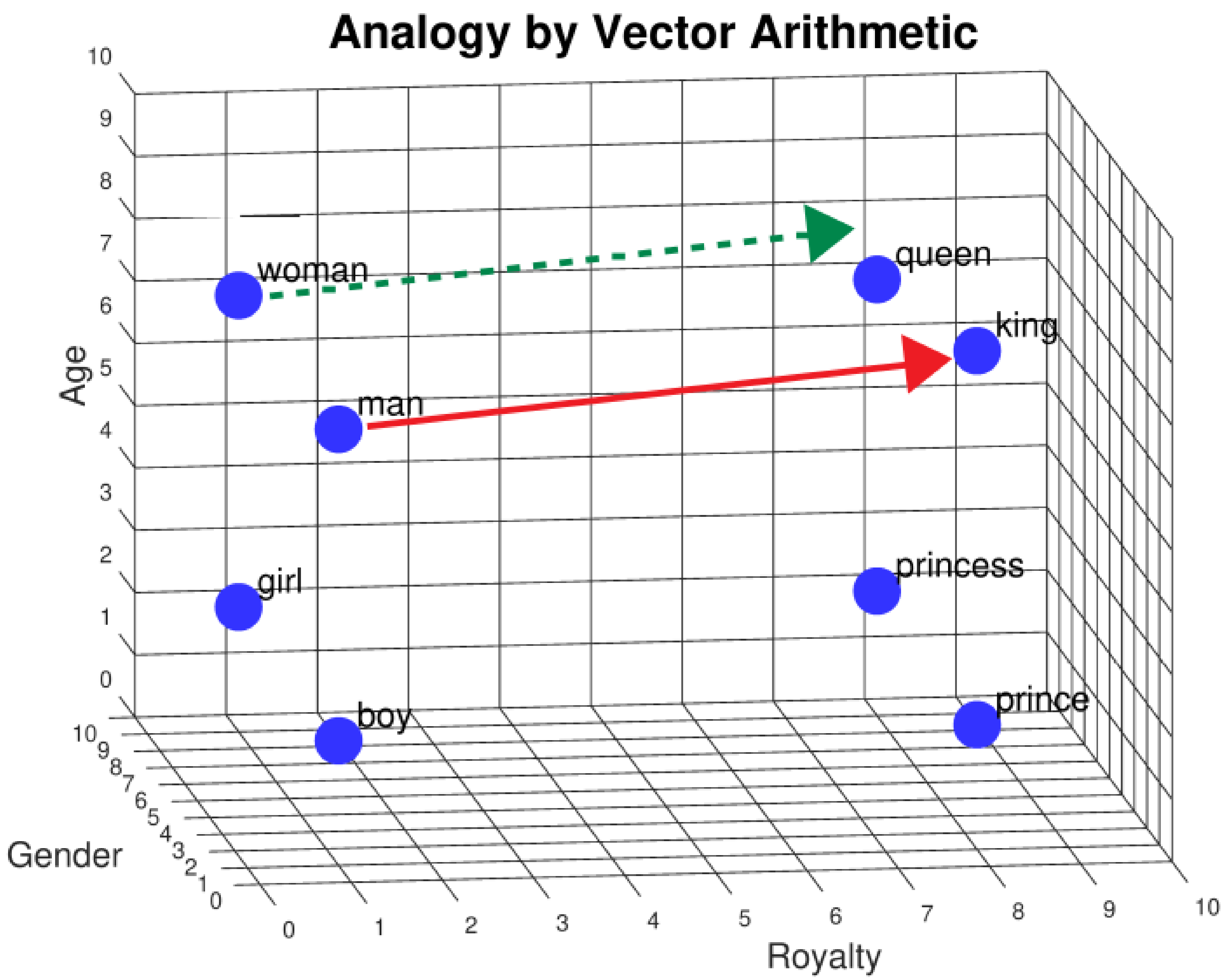
Word

Word embedding

Word Embedding:

$E_{man} \rightarrow E_{woman}$

$E_{king} \rightarrow$?



visualazing word’s embedding in 3D plot using dim reduction

AI CAMP

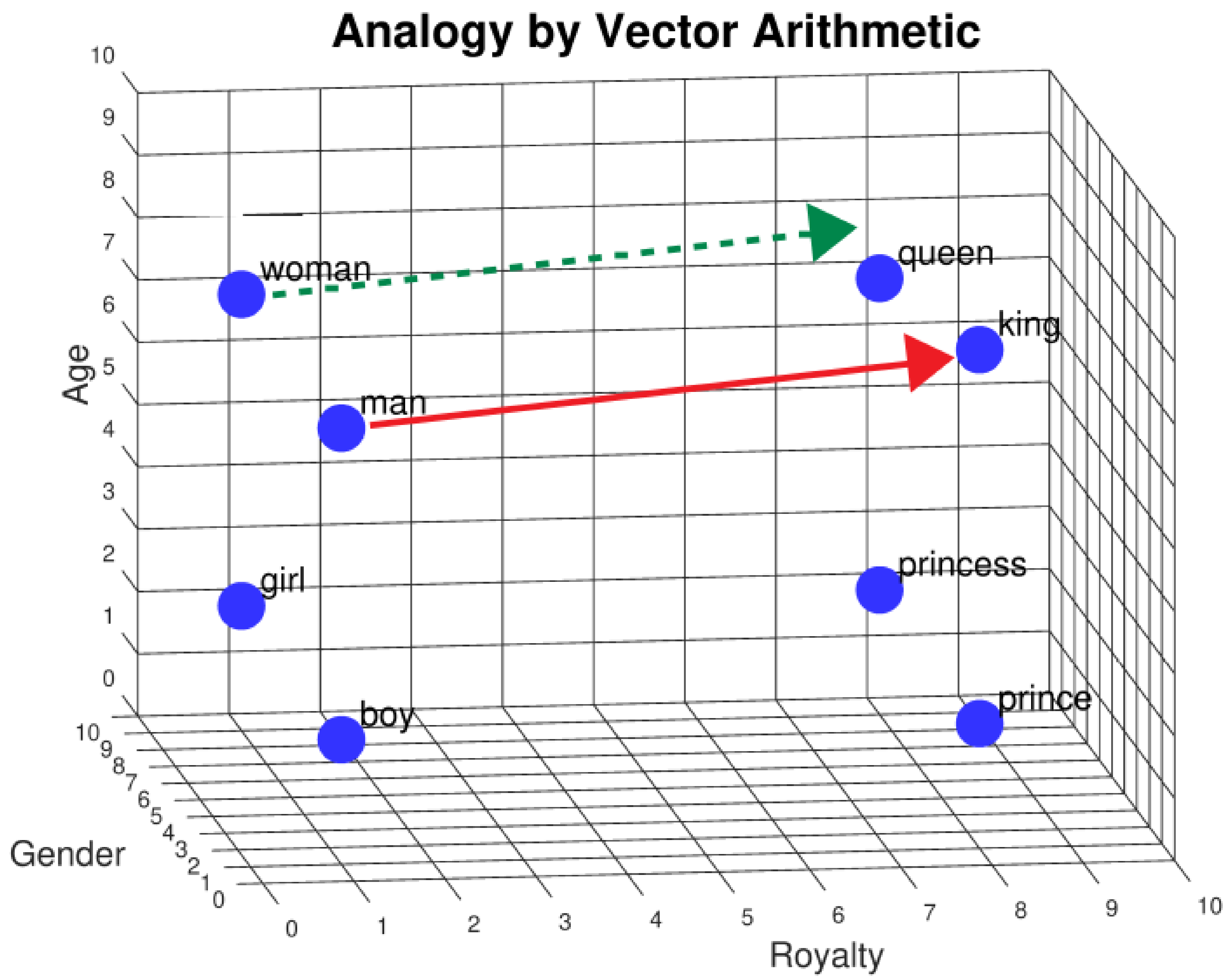
school of ai Algiers

Word Embedding:

Eking - Eman + Ewoman

=

Equeen

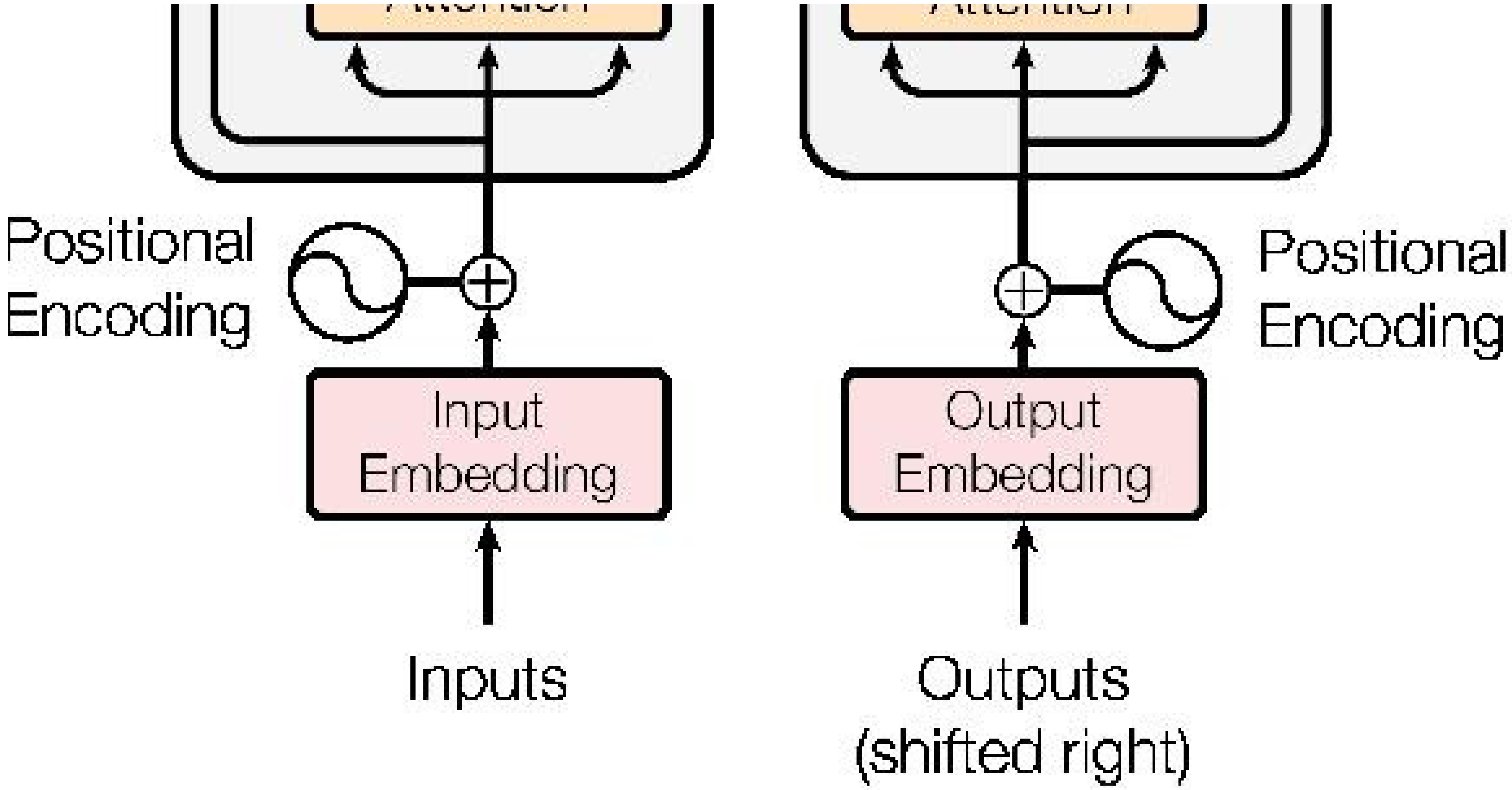


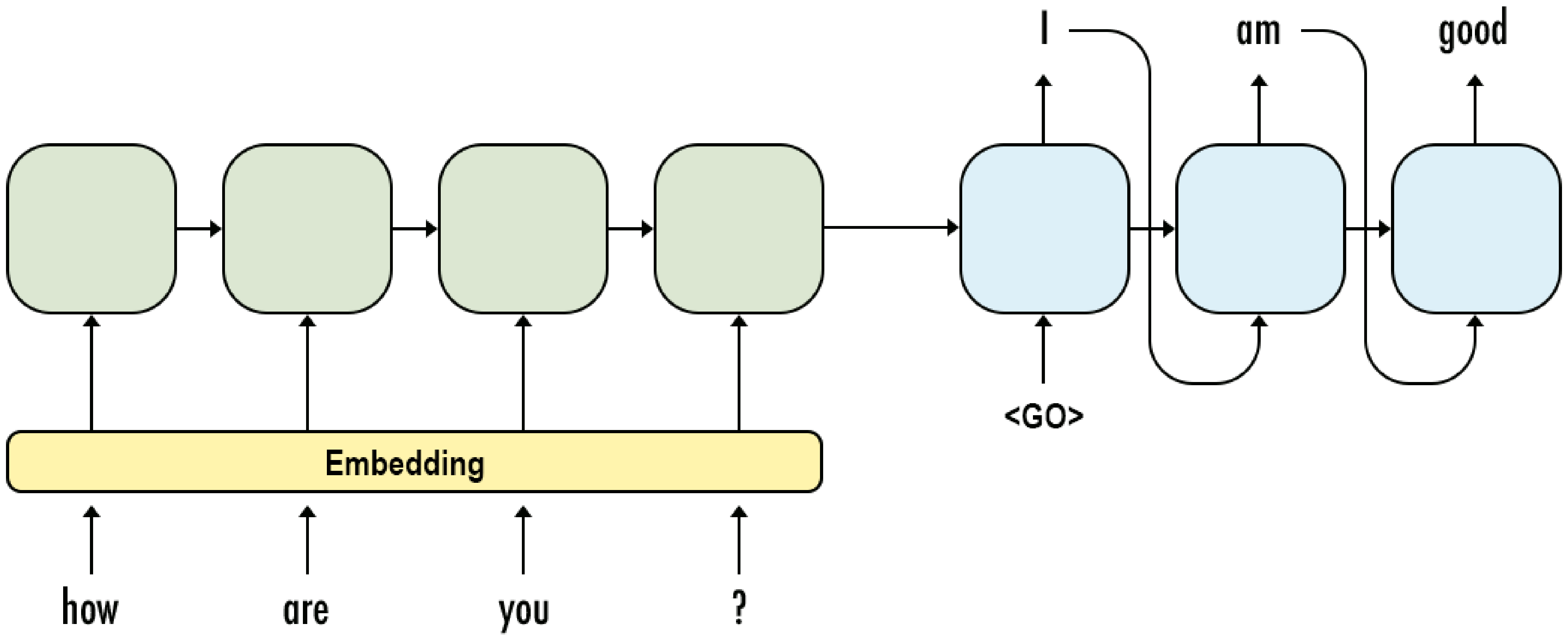
AI CAMP

school of ai Algiers

Word Embedding:

word ebedding is done by training a model or by using a pre-trained word embedding models like Flair, fastText, SpaCy





Bag Of Word (BOW):

The bag-of-words model is a way of feature extraction and representing text data when modeling text with machine learning algorithms

It involves two things:

A vocabulary of known words.

A measure of the presence of known words.

Bag Of Word (BOW):

It was the best of times,it was the worst of times,
it was the age of wisdom,it was the age of foolishness,

designing the vocabuary:

[“it”,“was”,“the”,“best”,“of”,“times”,“worst”,“age”,“wisdom”, “foolishness”]

Bag Of Word (BOW):

creating docs vectors:

it was the worst of times	1	1	1	0	1	1	1	0	0	0
"it was the age of wisdom"	1	1	1	0	1	0	0	1	1	0
it was the age of foolishness	1	1	1	0	1	0	0	1	0	1

AI CAMP

school of ai Algiers

Bag Of Word (BOW):

some use cases of the **Bow :**

Bow is widely used for text classification tasks, such as spam detection, sentiment analysis, and topic categorization.

Bow allows measuring the similarity between documents using metrics like cosine similarity

AI CAMP

school of ai Algiers

Thank you for your attention!