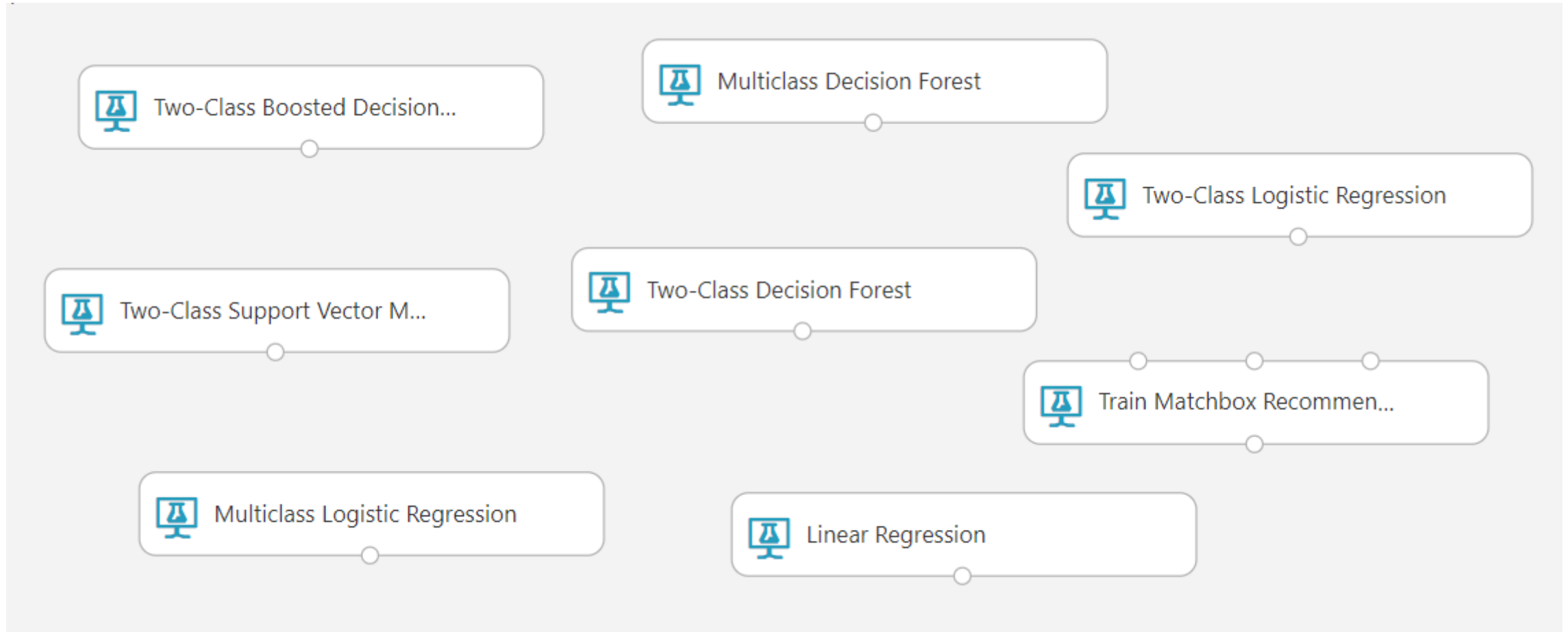


So far we learnt...



What is Text Analytics?

Introduction

- Interaction between computers and human

COBOL, FOTRAN, JAVA, C#, PYTHON

Natural Language Processing

- Computers Lack common sense and reasoning abilities
- Computers don't have Knowledge – Only processing and storage

Mala football khelne awadte.

J'adore Jouer au football

Ek is lief vir speel sokker

I Love Playing Football


Natural Language Processing

I want to **book** a flight from London to Paris.



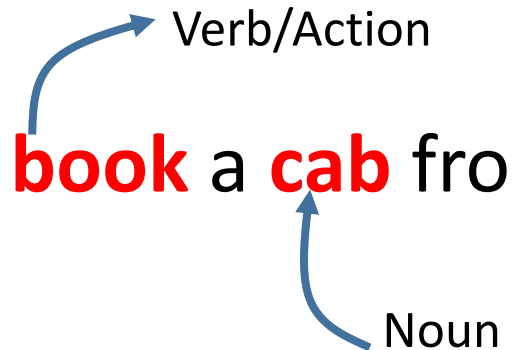
A blue curved arrow points from the word 'book' to the label 'Verb/Action'.

I read a **book** during the flight.



A blue curved arrow points from the word 'book' to the label 'Noun'.

I want to **book** a **cab** from Airport to Home.



Two blue curved arrows are present: one points from 'book' to 'Verb/Action' and another points from 'cab' to 'Noun'.

Structure of the Language

- Morphology – Structure of words
- Syntax – Order and combination of words, how sentences are related
- Phonology – The sound system of a language
- Semantics – Meaning of words and sentences
- Pragmatics – Functionally and Socially appropriate communication

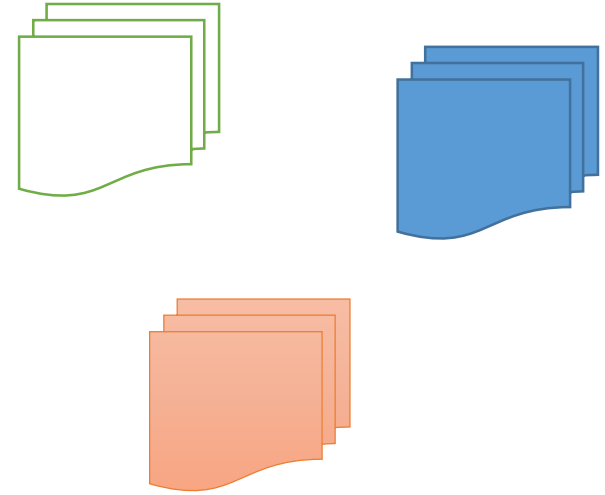
NLP Defined

- Interaction between computers and human or natural language
- Processes the naturally written or spoken text by analysing, processing and deriving information

Why Natural Language Processing is important?

Why Natural Language Processing?

- Text Classification
 - Document/Article Classification
 - News Stories Classification



Why Natural Language Processing?

- Text Classification
 - Document/Article Classification
 - News Stories Classification
 - Spam Filters



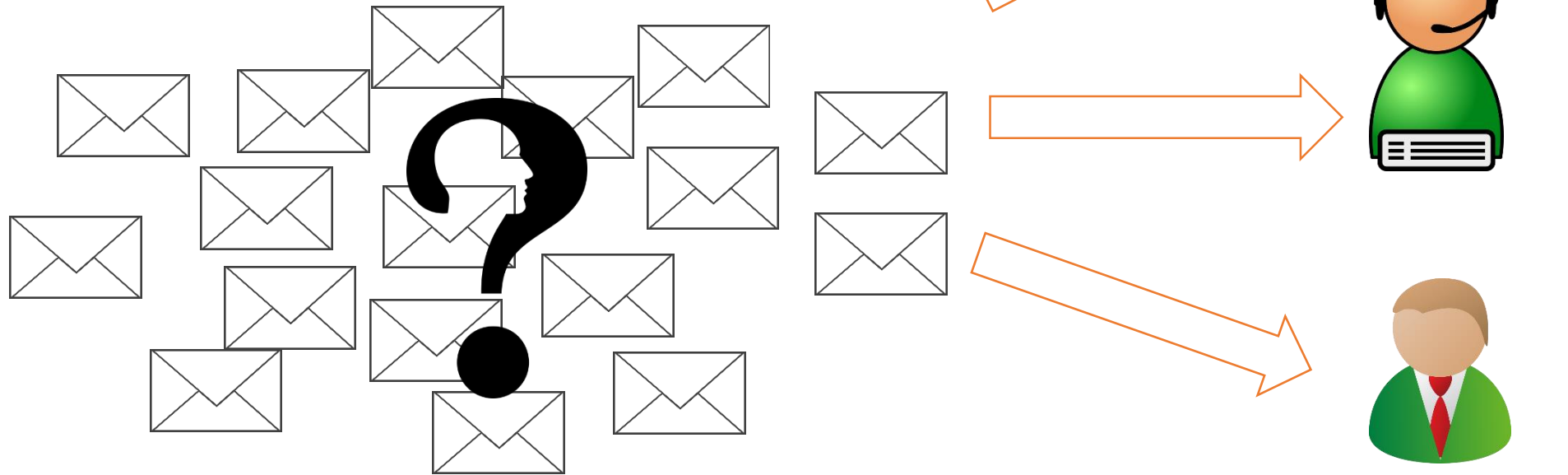
Why Natural Language Processing?

- Text Classification
 - Document/Article Classification
 - News Stories Classification
 - Spam Filters
 - Sentiment Analysis – Social as well as Text requests



Why Natural Language Processing?

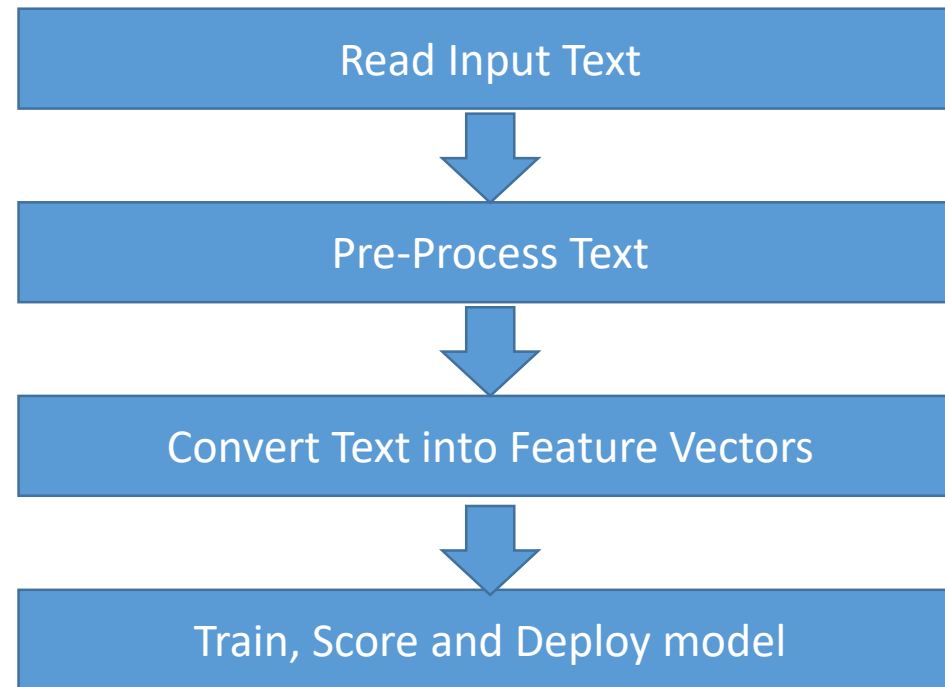
- Text Classification
 - Document/Article Classification
 - News Stories Classification
 - Spam Filters
 - Sentiment Analysis – Social as well as Text requests
 - Customer/Service Request Classification



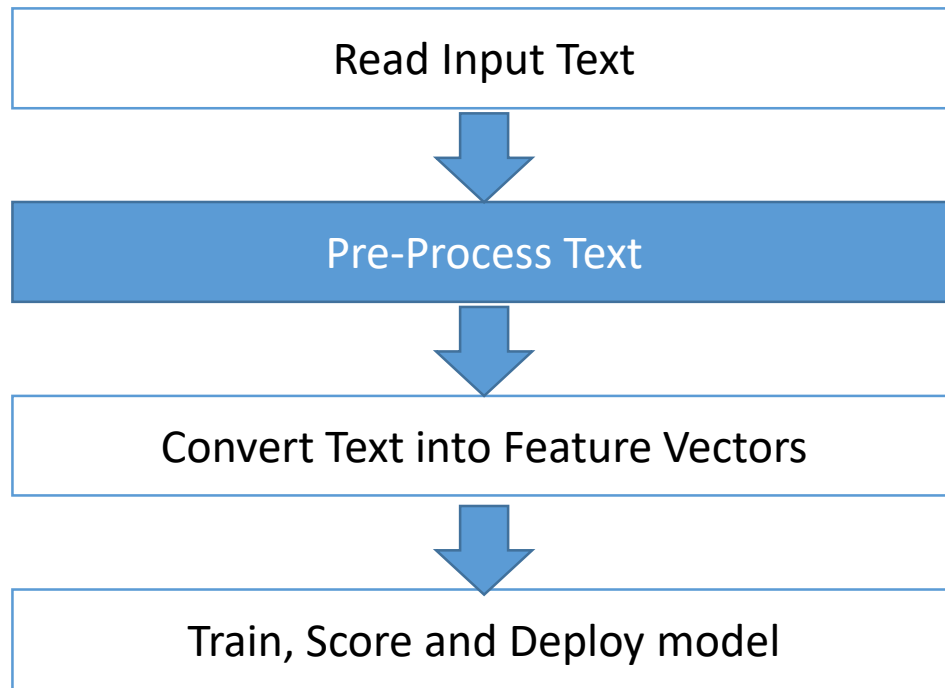
Why Natural Language Processing?

- Text Classification
 - Document/Article Classification
 - News Stories Classification
 - Spam Filters
 - Sentiment Analysis – Social as well as Text requests
 - Customer/Service Request Classification
- Automatic Translation
- Speech Recognition
 - Call centre applications
- Text Generations or Dialogs in AI

Text Analytics Steps



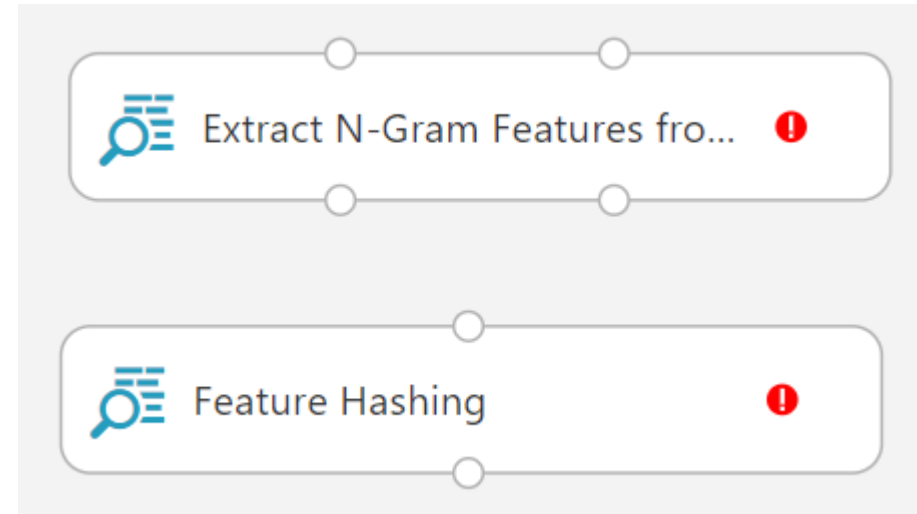
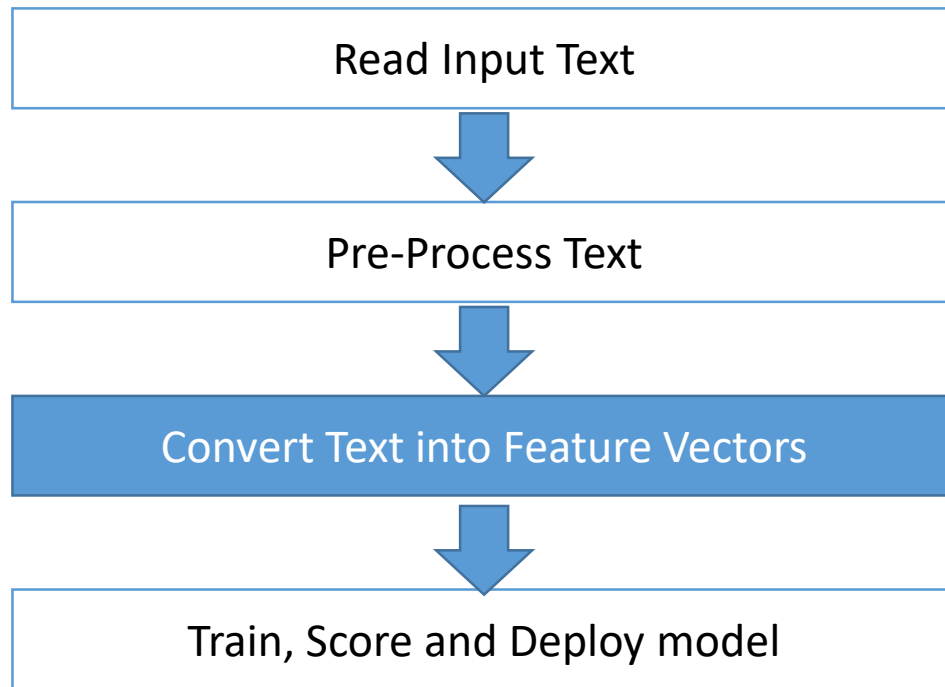
Text Analytics Steps



The screenshot shows the 'Preprocess Text' module in the Azure ML interface. The module is labeled '1' and has a red error icon. The 'Properties' pane on the right shows the following settings:

- Language:** English
- Remove by part of speech:** False
- Text column to clean:** Selected columns: Column type: String, Feature
- Launch column selector** button
- Checkboxes (all checked):**
 - Remove stop words
 - Lemmatization
 - Detect sentences
 - Normalize case to lowercase
 - Remove numbers
 - Remove special characters
 - Remove duplicate characters
 - Remove email addresses
 - Remove URLs
 - Expand verb contractions
 - Normalize backslashes to slashes
 - Split tokens on special characters

Text Analytics Steps



Text Pre-processing

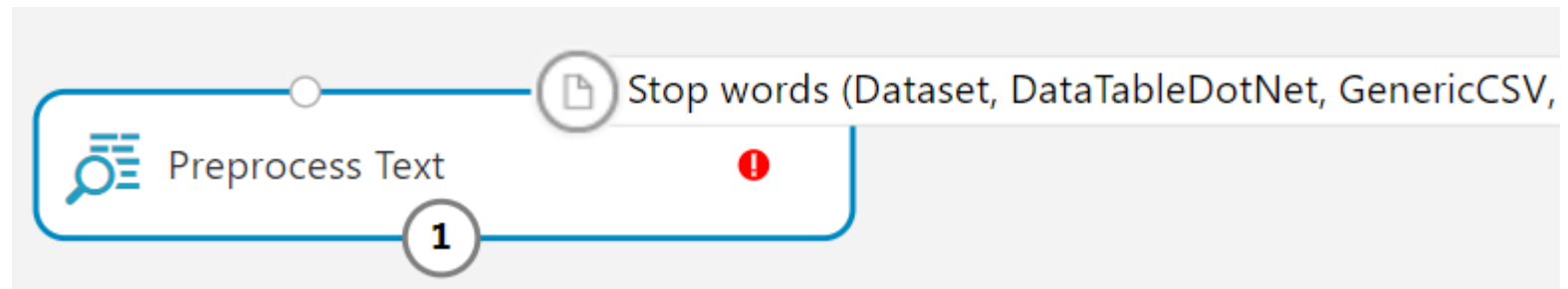
Noise Removal

- Stop words

a, an, the, etc, at, about

“I liked a car” vs “ I liked the car”

<http://az754797.vo.msecnd.net/docs/Stopwords.zip>



Noise Removal

- Stop words
- Normalize case

Car is parked in the bay.

car is parked in the bay.

Noise Removal

- Stop words
- Normalize case
- **Numbers**

I paid \$2 for a cup of coffee.
I paid \$10 for a cup of coffee.

I paid \$2 for a cup of coffee. Great value for money.
I paid \$10 for a cup of coffee. It was not worth it.

I paid \$2 for a cup of coffee. It's very expensive
compared to my home country.

Noise Removal

- Stop words
- Normalize case
- Numbers
- **Special Characters**

I paid \$2 for a cup of coffee. Great value for money.
I paid \$10 for a cup of coffee. It was not worth it.

Noise Removal

- Stop words
- Normalize case
- Numbers
- Special Characters
- Duplicate Characters

I paid \$2 for a cup of **coffee coffee**. Great value for money.
I paid \$10 for a cup of coffee. It was not worth it.

Noise Removal

- Stop words
- Normalize case
- Numbers
- Special Characters
- Duplicate Characters
- Email Addresses
- URLs
- Slashes

I was not happy to read an email from abc@xyz.com

Felt great after reading a carefully crafted email from abc@xyz.com.

Lemmatization

- Generating Dictionary Form
- Sentence Separation
- Tokenization
- Part-of-speech identification

Lemmatization

- A lemma is the canonical form, dictionary form, or citation form of a set of words – Wikipedia
- Lemmatization is the process of identifying a single canonical form to represent multiple word tokens - Microsoft

run, ran, running, runs



Lemma

run

Build, built, building, builds



Verb

Lemma

build

building



noun

building

Lemmatization

- A lemma is the canonical form, dictionary form, or citation form of a set of words – Wikipedia
- Lemmatization is the process of identifying a single canonical form to represent multiple word tokens - Microsoft

“Nothing better than having a cup of coffee after a hectic day.”

“Came home after a hectic day and the coffee made me feel the best.”

“It was good to have a cup of coffee after the hectic day.”

Lemmatization

- A lemma is the canonical form, dictionary form, or citation form of a set of words – Wikipedia
- Lemmatization is the process of identifying a single canonical form to represent multiple word tokens - Microsoft

“Nothing **better** than having a cup of coffee after a hectic day.”

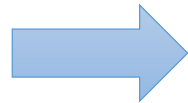
“Came home after a hectic day and the coffee made me feel the **best**.”

“It was **good** to have a cup of coffee after the hectic day.”

Lemmatization

- Lemmatization is the process of identifying a **single canonical form** to represent **multiple word tokens** – Microsoft

Good
Better
Best



Good

Is
Am
Are
Will be
Shall be



be

Lemmatization

- Generating Dictionary Form
- Sentence Separation
- Tokenization
- Part-of-speech identification



Sentence Separation

Fields are closed. No Trespassing. Violators will be prosecuted.

Fields are closed. No Trespassing Violators will be prosecuted.

Sentence Separation

What? I just did not know about this “news”.


What ! I just did not know about this “news”.

“What....I just didn’t know about this “news”....”

What....? I just did not know about this news...

What.... I just did not know about “this news...”

Lemmatization

- Generating Dictionary Form
- Sentence Separation
- Tokenization 
- Part-of-speech identification

Tokenization

Process of breaking sentences into tokens.

What? I just didn't know about this "news".

What ! I just did not know about this "news".

Tokenization

What? I just **didn't** know about this “news”.

What ! I just **did not** know about this “news”.

Lemmatization

- Generating Dictionary Form
- Sentence Separation
- Tokenization
- Part-of-speech identification



Part of Speech

- How word is used in a given sentence?
- “Grammatical Intelligence”

Part-Of-Speech

- Nouns – name of a person, thing, or place
- Pronouns – Noun phrase to avoid repetition
- Adjectives – Describes, modifies or gives more information about a noun or pronoun
- Verbs – Verb shows an action or state of being
- Adverbs – Describes or modifies a verb
- Prepositions – Shows relationship of a noun or pronoun with another word
- Conjunctions – Joins two words in a sentence
- Interjections – Expresses strong feeling or emotion

Part-Of-Speech

The Willis Tower, **built** as and still commonly referred to as the Sears Tower, is a 442.1 meter tall **building** in Chicago.

The heavens are **above**. (Adverb)

The moral code of conduct is **above** the civil code of conduct. (Preposition)

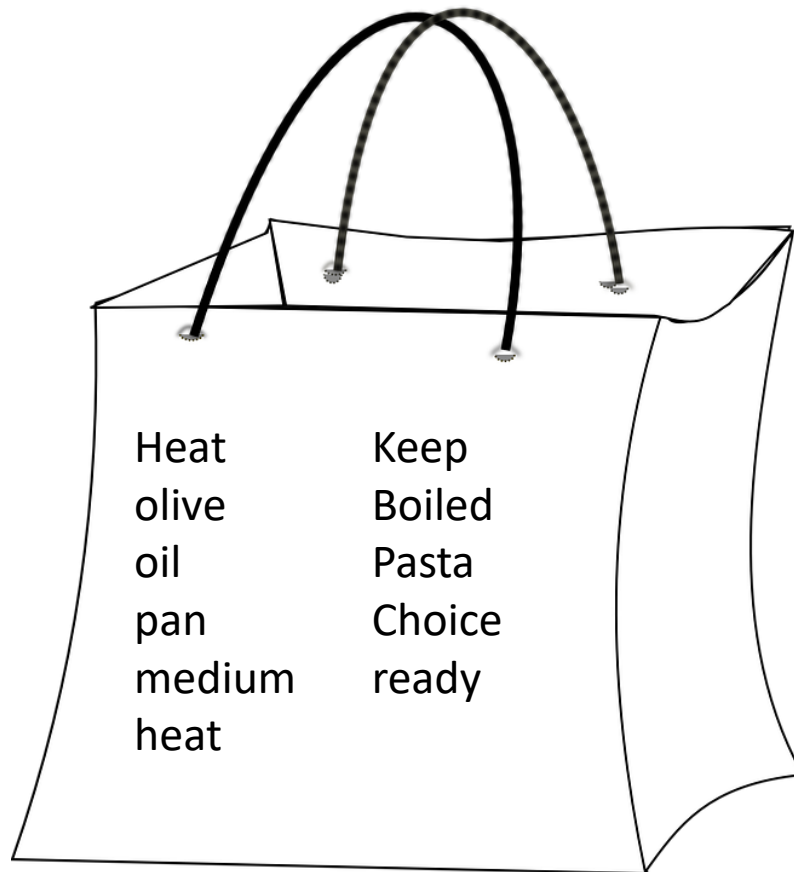
Read the sentence given **above**. (Adjective)

Our blessings come from **above**. (Noun)

<https://www.englishgrammar.org/word-part-speech/>

Bag Of Words

- Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.



N-gram Model

- Contiguous sequence of “n” items from a given sample of text or speech – Wikipedia

Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.

Bi-Gram

Heat olive
Olive oil
Oil in
in pan
Pan on
On medium
Medium heat

tri-Gram

Heat olive oil
Olive oil in
Oil in pan
in pan on
Pan on medium
On medium heat

Text Data Feature Engineering

Features

Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	CreditHistory	Property_Area	Status
LP001002	Male	No	0	No	\$5,849.00		60	1	Urban	Y
LP001003	Male	Yes	1	No	\$4,583.00	\$128.00	120	1	Rural	N
LP001005	Male	Yes	0	Yes	\$3,000.00	\$66.00	60	1	Urban	Y
LP001006	Male	Yes	2	No	\$2,583.00	\$120.00	60	1	Urban	Y

Text	Category
Heat olive oil in a pan on medium heat.	Food
Illinois is also known as Land of Lincoln	Politics
Red sauce pasta is among the most popular Italian dishes.	Food
Best time to visit Goa is December-January	Travel
You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.	Food
The Se Cathedral at Goa is an example of Portuguese architecture and one of the largest churches in Asia.	Travel
Beat the heat with this summer special treat prepared using orange juice and pineapple.	Food
Barrack Obama served as the 44 th President of the United States.	Politics

Features

Text	Category
Heat olive oil in a pan on medium heat.	Food
Illinois is also known as Land of Lincoln. Abraham Lincoln, Statesman and Lawyer, served as the 16 th President of the United States.	Politics
Red sauce pasta is among the most popular Italian dishes.	Food
Best time to visit Goa is December-January	Travel
You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.	Food
The Se Cathedral at Goa is an example of Portuguese architecture and one of the largest churches in Asia.	Travel
Beat the heat with this summer special treat prepared using orange juice and pineapple.	Food
Barrack Obama served as the 44 th President of the United Stated.	Politics

Features

Text	Category
Heat olive oil in a pan on medium heat.	Food
Illinois is also known as Land of Lincoln . Abraham Lincoln , Statesman and Lawyer, served as the 16 th President of the United States .	Politics
Red sauce pasta is among the most popular Italian dishes.	Food
Best time to visit Goa is December-January	Travel
You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.	Food
The Se Cathedral at Goa is an example of Portuguese architecture and one of the largest churches in Asia.	Travel
Beat the heat with this summer special treat prepared using orange juice and pineapple.	Food
Barrack Obama served as the 44 th President of the United States .	Politics

Features

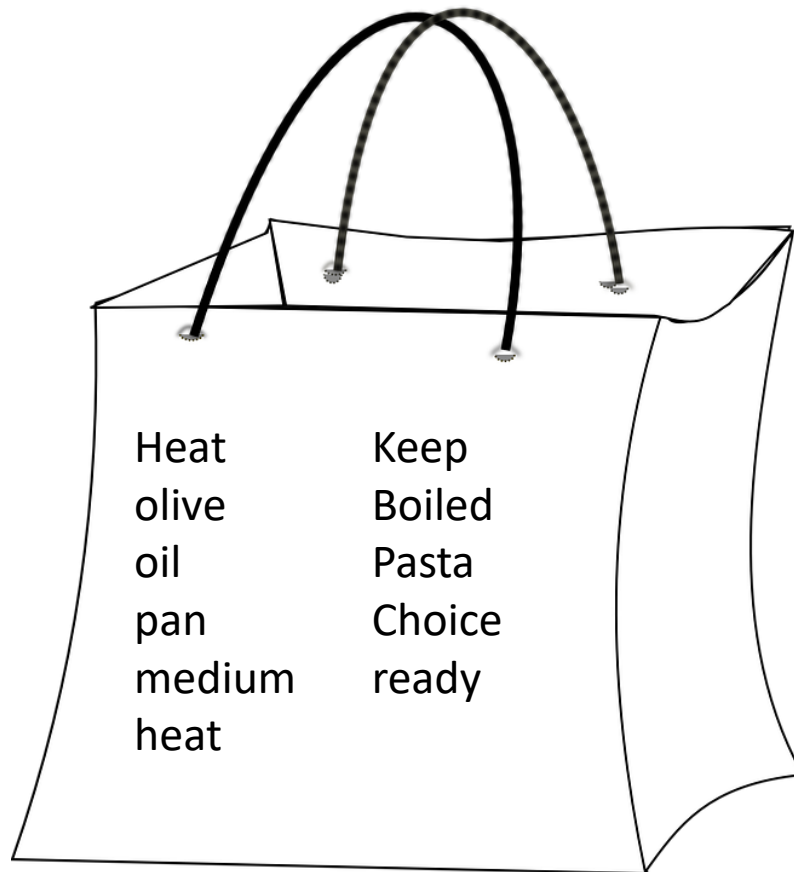
Text	Category
Heat olive oil in a pan on medium heat.	Food
Illinois is also known as Land of Lincoln. Abraham Lincoln, Statesman and Lawyer, served as the 16 th President of the United States.	Politics
Red sauce pasta is among the most popular Italian dishes.	Food
Best time to visit Goa is December-January	Travel
You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.	Food
The Se Cathedral, a must visit , at Goa is an example of Portuguese architecture and one of the largest churches in Asia.	Travel
Beat the heat with this summer special treat prepared using orange juice and pineapple.	Food
Barrack Obama served as the 44 th President of the United States.	Politics

Text Data Feature Engineering

- Bag-Of-Words
- N-gram model

Bag Of Words

- Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.



Bag Of Words

- a. Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- b. You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.

Heat	Olive	Oil	Pan	Medium	Heat	Keep	Boiled	Pasta	Choice	Ready	Need	Some	Penne	Extra	Virgin	parsley	garnish
------	-------	-----	-----	--------	------	------	--------	-------	--------	-------	------	------	-------	-------	--------	---------	---------

Bag Of Words

- a. Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- b. You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.

	Heat	Olive	Oil	Pan	Medium	Heat	Keep	Boiled	Pasta	Choice	Ready	Need	Some	Penne	Extra	Virgin	parsley	garnish
a	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
B	0	1	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1

Bag Of Words

- a. Heat **olive oil** in a pan on medium heat. Keep **boiled pasta** of your choice ready.
- b. You will need some **boiled pasta** penne, extra virgin **olive oil** and parsley for garnish.

	Heat	Olive	Oil	Pan	Medium	Heat	Keep	Boiled	Pasta	Choice	Ready	Need	Some	Penne	Extra	Virgin	parsley	garnish
a	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
B	0	1	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1

N-gram Model

N-gram Model

- Contiguous sequence of “n” items from a given sample of text or speech – Wikipedia

Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.

Bi-Gram

Heat olive
Olive oil
Oil in
in pan
Pan on
On medium
Medium heat

tri-Gram

Heat olive oil
Olive oil in
Oil in pan
in pan on
Pan on medium
On medium heat

N-gram Model

- Contiguous sequence of “n” items from a given sample of text or speech – Wikipedia
- Probabilistic model to predict the nth word from n-1 words.

Bi-grams

Machine _____

Machine **Learning**

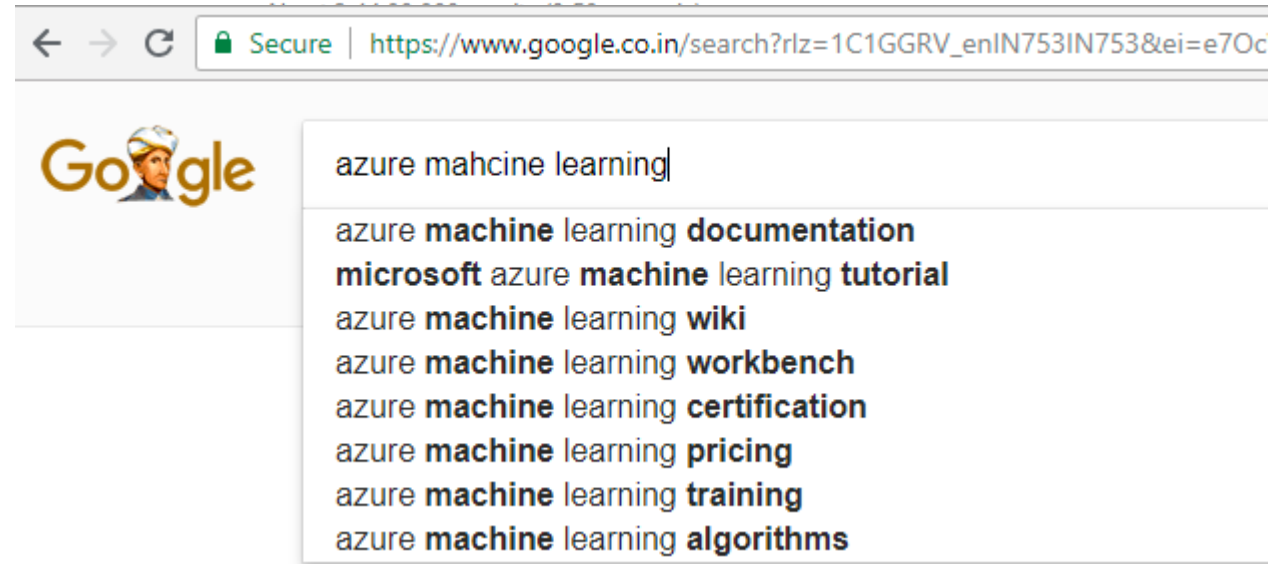
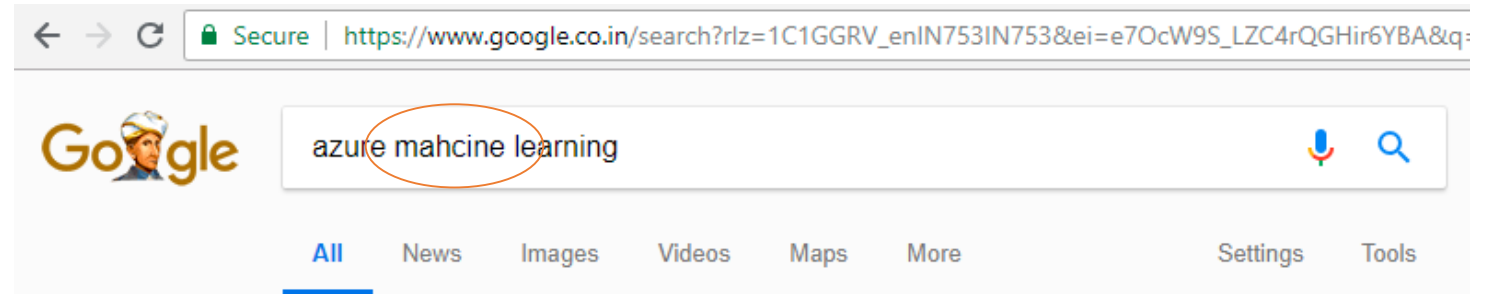
Tri-grams

Azure Machine _____

Azure Machine **Learning**

N-gram Model Applications

- Auto-Correct spelling of words
- Auto-suggestion of words
- Optical Character Recognition



N-gram model

- a. Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- b. You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.

Heat olive
Olive oil
Oil in
In a
A pan
Pan on
On medium
Medium heat
Heat keep
Keep boiled
Boiled pasta
.....

You will
Will need
Need some
Some boiled
Boiled pasta
Pasta penne
Penne extra
Extra virgin
Virgin olive
Olive oil
.....

N-gram model

- a. Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- b. You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.

Heat olive
Olive oil
Oil in
In a
A pan
Pan on
On medium
Medium heat
Heat keep
Keep boiled
Boiled pasta
.....

You will
Will need
Need some
Some boiled
Boiled pasta
Pasta penne
Penne extra
Extra virgin
Virgin olive
Olive oil
.....

Oilve _____

N-gram model

- Heat olive oil in a pan on medium heat. Keep boiled pasta of your choice ready.
- You will need some boiled pasta penne, extra virgin olive oil and parsley for garnish.

Heat olive
Olive oil
Oil in
In a
A pan
Pan on
On medium
Medium heat
Heat keep
Keep boiled
Boiled pasta
.....

You will
Will need
Need some
Some boiled
Boiled pasta
Pasta penne
Penne extra
Extra virgin
Virgin olive
Olive oil
.....

Oilve _____

Boiled _____

Problems with Bag-of-Words or n-gram

- Curse of Dimensionality

171,476 words in the Oxford dictionary

A tweet has 280 characters or 50+ words

Illinois is also known as Land of Lincoln. Abraham Lincoln, Statesman and Lawyer, served as the 16th President of the United States.

Problems with Bag-of-Words or n-gram

- Curse of Dimensionality

171,476 words in the Oxford dictionary

A tweet has 280 characters or 50+ words

Illinois is also known as Land of Lincoln. Abraham Lincoln, Statesman and Lawyer, served as the 16th President of the United States.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Illinois	is	also	known	as	land	of	Lincoln	Abraham	statesman	and	lawyer	served	the	16th	president	United	States
1	1	1	1	2	1	2	2	1	1	1	1	1	2	1	1	1	1

Problems with Bag-of-Words or n-gram

- Curse of Dimensionality

171,476 words in the Oxford dictionary

A tweet has 280 characters or 50+ words

Sentences	John	likes	movies	Liza	watching	Frans	loves	playing	football	merry	enjoys	action
John likes movies	1	1	1	0	0	0	0	0	0	0	0	0
Liza likes watching movies	0	1	1	1	1	0	0	0	0	0	0	0
Frans loves playing football	0	0	0	0	0	1	1	1	1	0	0	0
Merry enjoys action movies	0	0	1	0	0	0	0	0	0	1	1	1

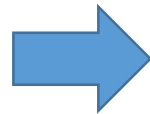
Problems with Bag-of-Words or n-gram

- Curse of Dimensionality
- Semantic relation

171,476 words in the Oxford dictionary

A tweet has 280 characters or 50+ words

Liza likes watching movies



Sentence	Liza likes	likes watching	watching movies
	1	1	1

4 word sentence

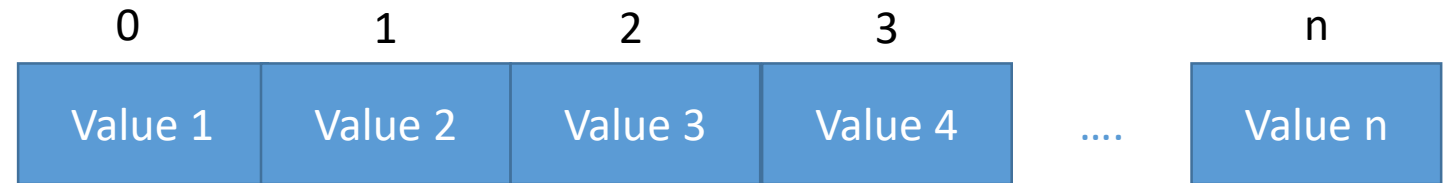
$N-1 = 3$ features

What is Feature Hashing?

What are Array, Linkedlist and Hashtables?

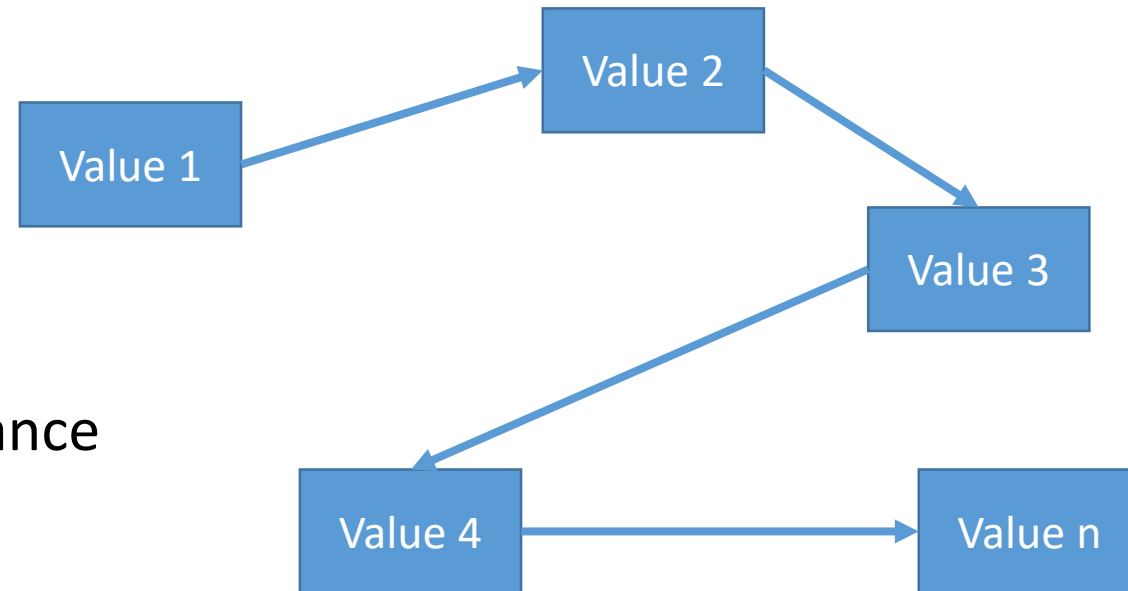
Arrays and Linked List

Array size is fixed.
Fixed memory allocation.

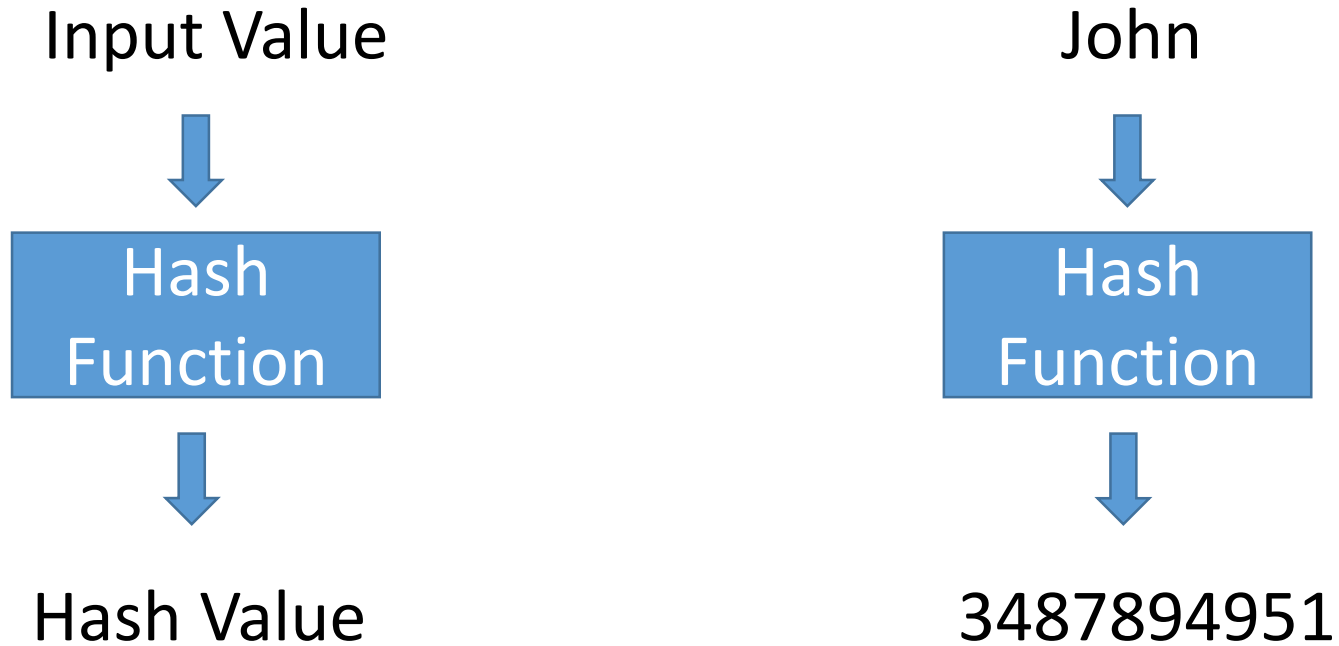


Array (3) = value 4

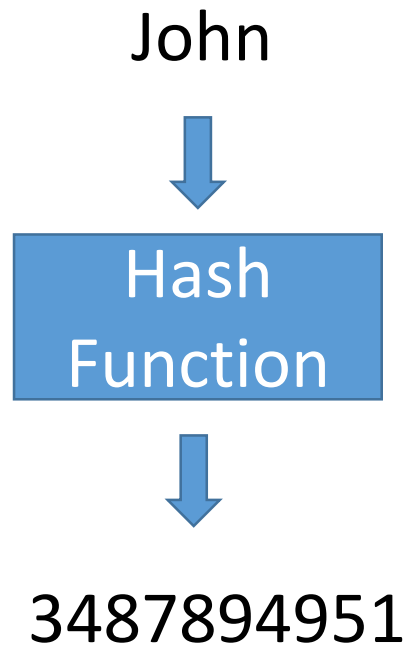
LinkedList can grow dynamically.
Sequential Read hence bad performance



Hashing

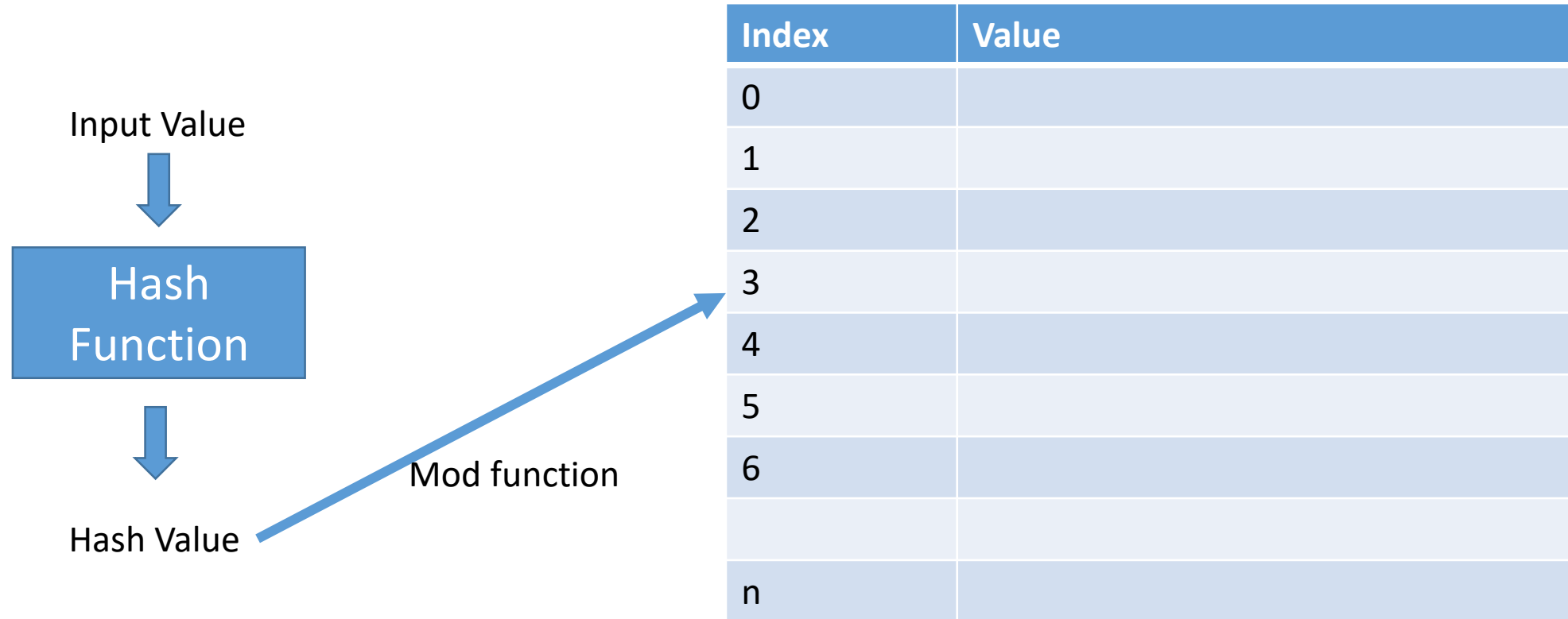


MOD Function



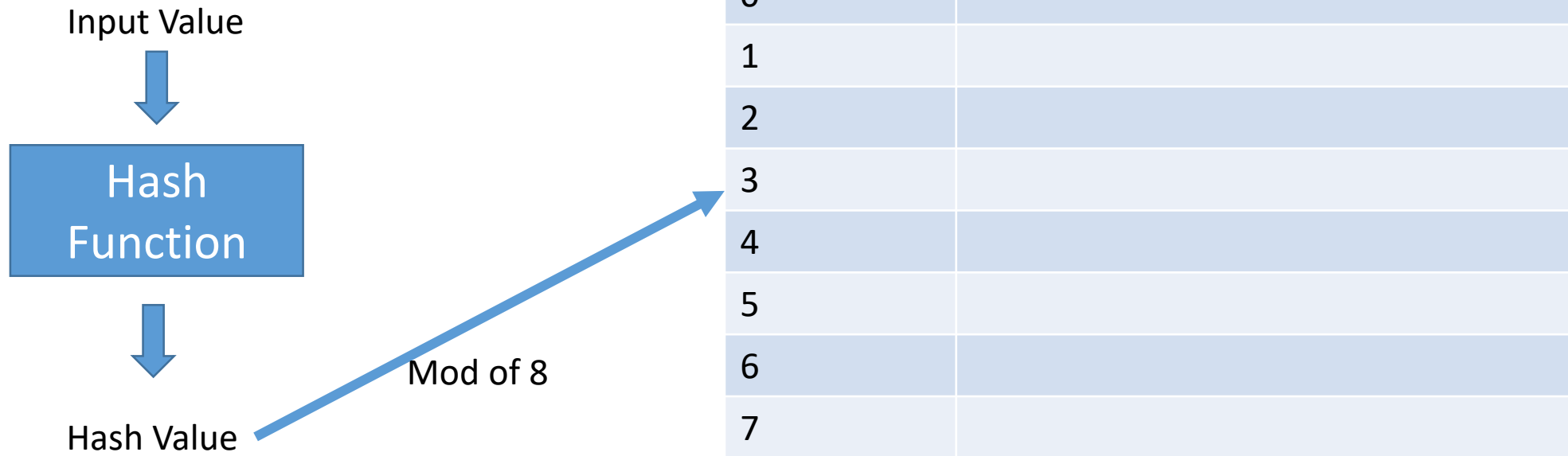
		Reminder
MOD 8	→	7
MOD 50	→	1
MOD 100	→	51

Hash Table

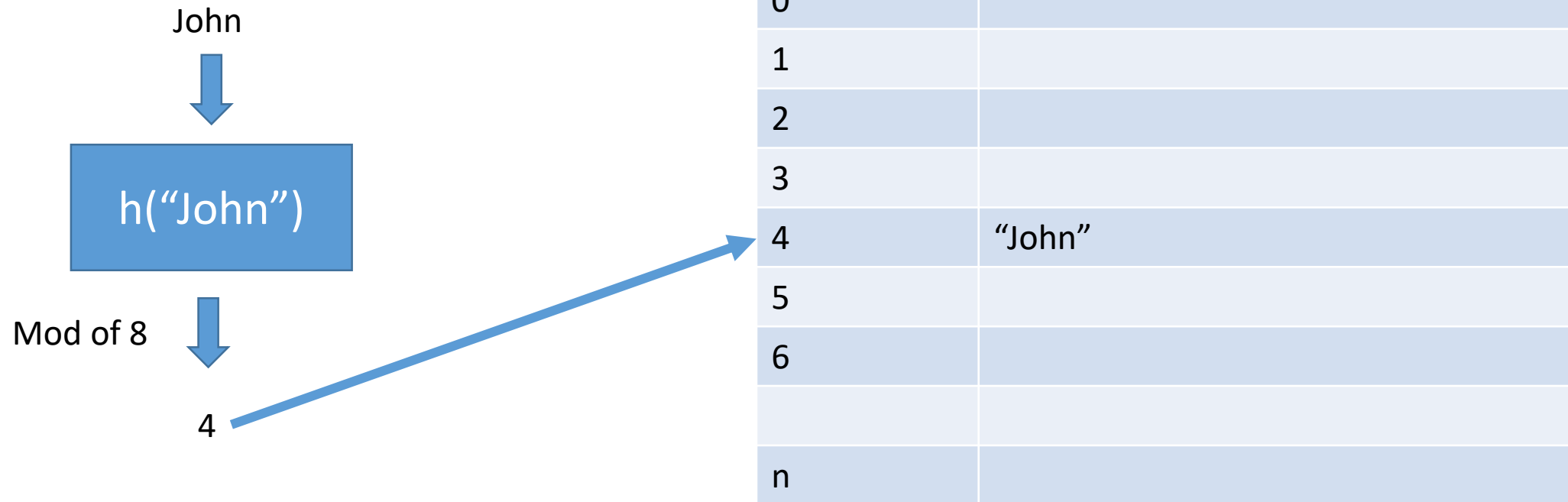


Hash Table

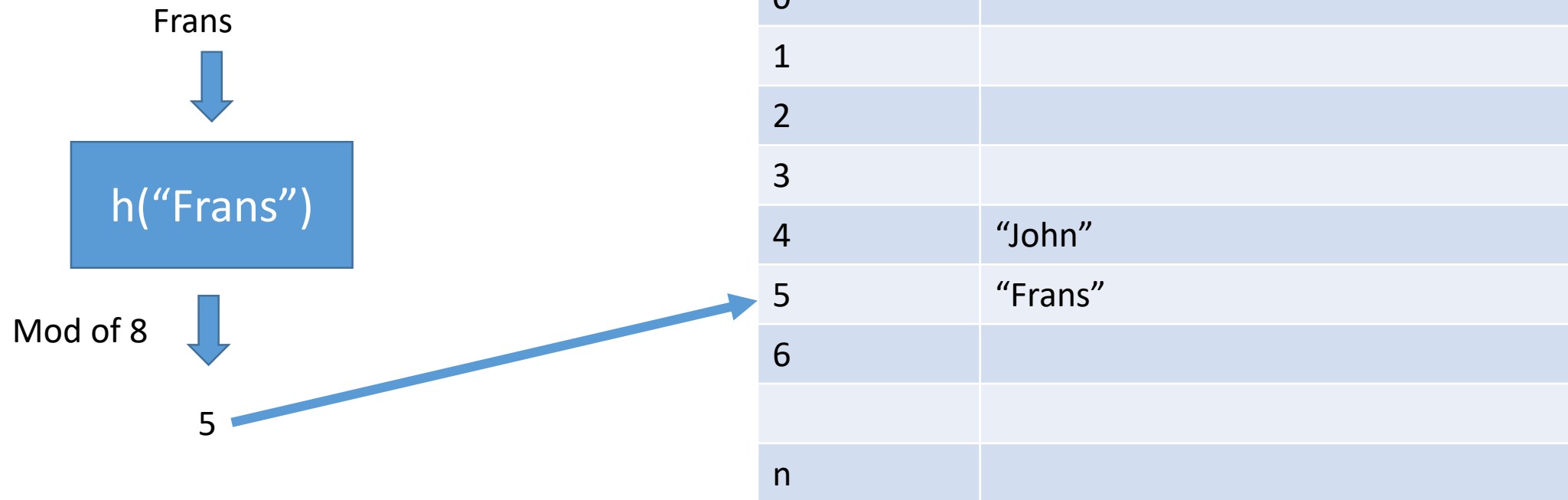
Index size = 8



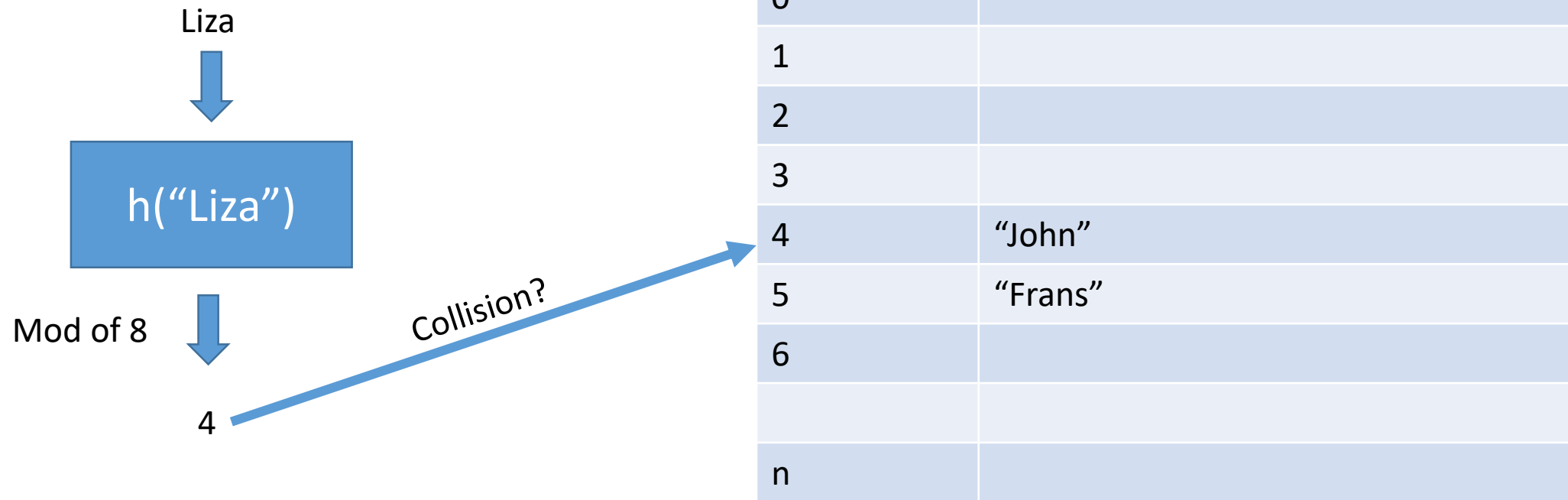
Hash Table



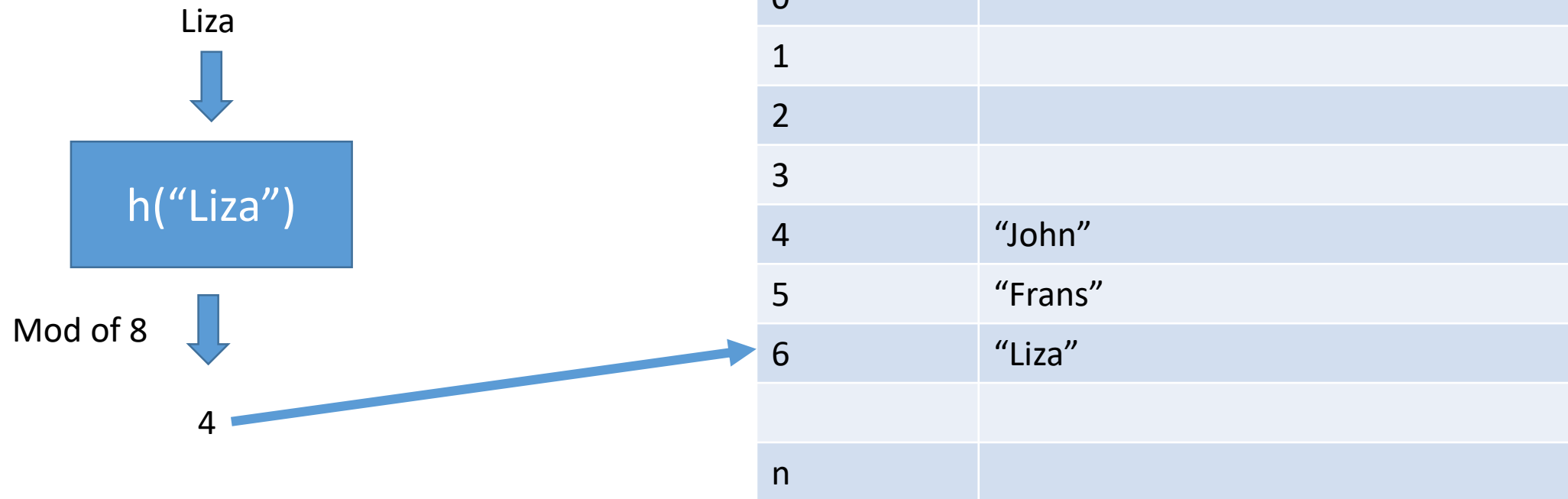
Hash Table



Hash Table

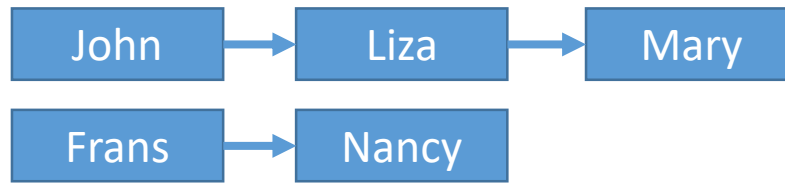


Hash Table



Hash Table with Separate Chaining

Index
0
1
2
3
4
5
6
n



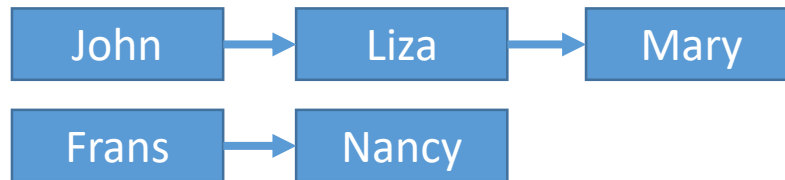
Hash Table with Separate Chaining

Combination of Arrays and Linkedlists

Faster search and insertion

Makes use of all the information provided by the key value

Index
0
1
2
3
4
5
6
n



Feature Hashing

Feature Hashing

Sentence	Murmurhash3	Divide by	Reminder
john	3487894951	8	7
likes	1103617568	8	0
movies	3188341541	8	5

Index	Value
0	likes
1	
2	
3	
4	
5	movies
6	
7	john

Feature Hashing

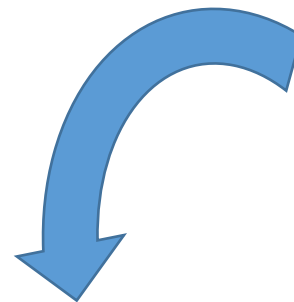
Sentence	Murmurhash3	Divide by	Reminder
john	3487894951	8	7
likes	1103617568	8	0
movies	3188341541	8	5

Index	Value
0	likes
1	
2	
3	
4	
5	movies
6	
7	john

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8

Feature Hashing

Sentence	Murmurhash3	Divide by	Reminder
john	3487894951	8	7
likes	1103617568	8	0
movies	3188341541	8	5



Index	Value
0	likes
1	
2	
3	
4	
5	movies
6	
7	john

Value	likes					movies		john
Index	0	1	2	3	4	5	6	7
Feature	1	2	3	4	5	6	7	8

Feature Hashing

Sentence	Murmurhash3	Divide by	Reminder
john	3487894951	8	7
likes	1103617568	8	0
movies	3188341541	8	5

Reminder -->	0	1	2	3	4	5	6	7
	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
	0	0	0	0	0	0	0	1
	1	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0

Bag Of Words

	John	likes	movies	Liza	watching	Frans	loves	playing	football	merry	enjoys	action
John likes movies	1	1	1	0	0	0	0	0	0	0	0	0
Liza likes watching movies	0	1	1	1	1	0	0	0	0	0	0	0
Frans loves playing football	0	0	0	0	0	1	1	1	1	0	0	0
Merry enjoys action movies	0	0	1	0	0	0	0	0	0	1	1	1

Feature Hashing

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
John likes movies	1	0	0	0	0	2	0	0
Liza likes watching movies	2	0	1	0	0	1	0	0
Frans loves playing football	0	1	1	0	1	0	0	1
Merry enjoys action movies	0	0	0	0	0	3	0	1

Thank You...!