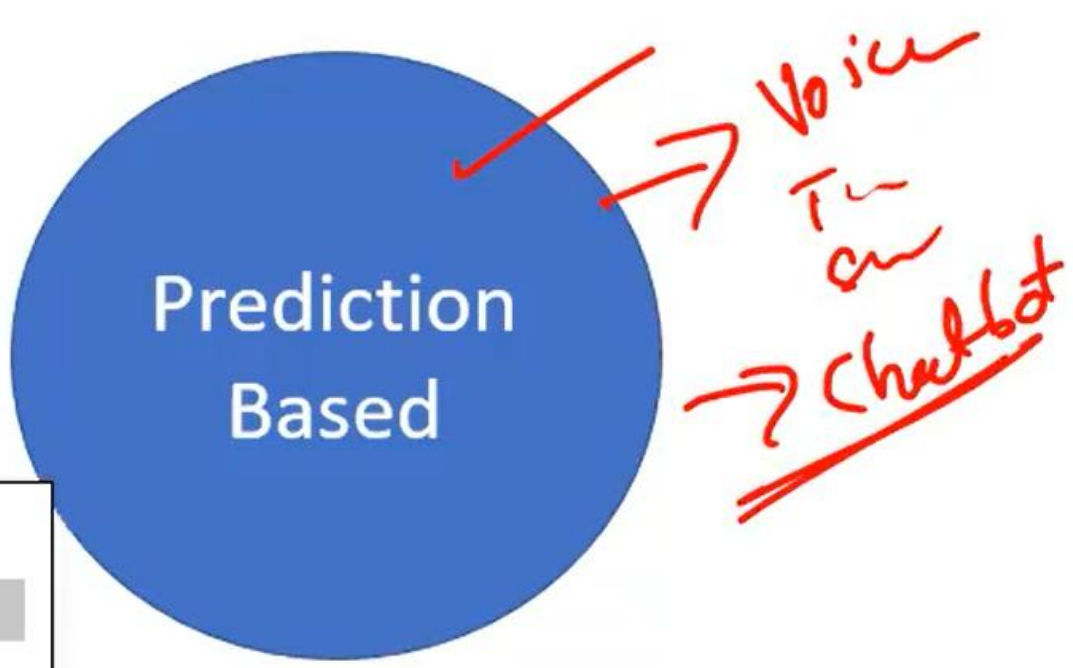
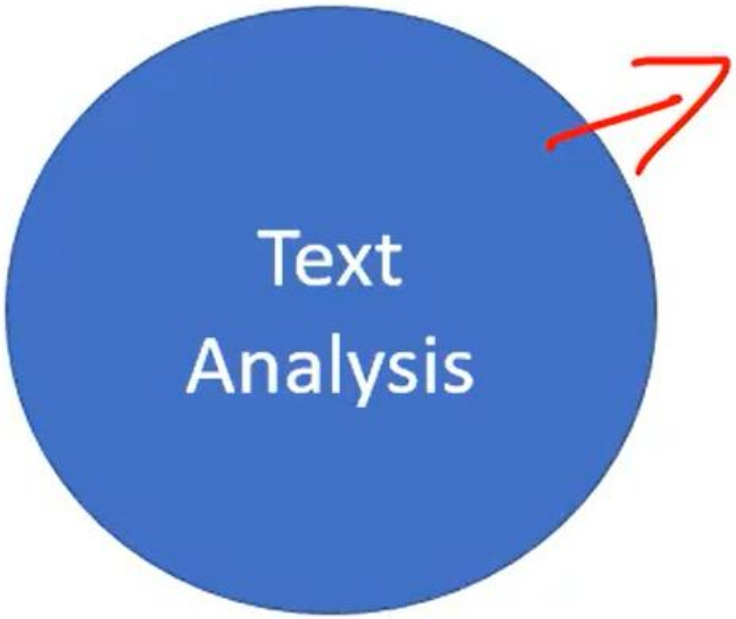
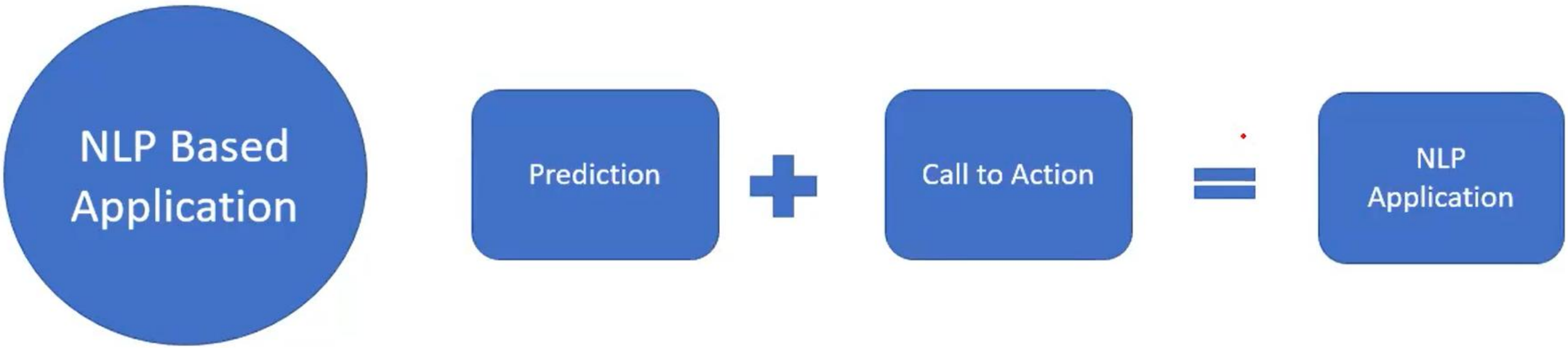
Natural Language Processing

Two Type of process



Voice
To
Text
Chatbot

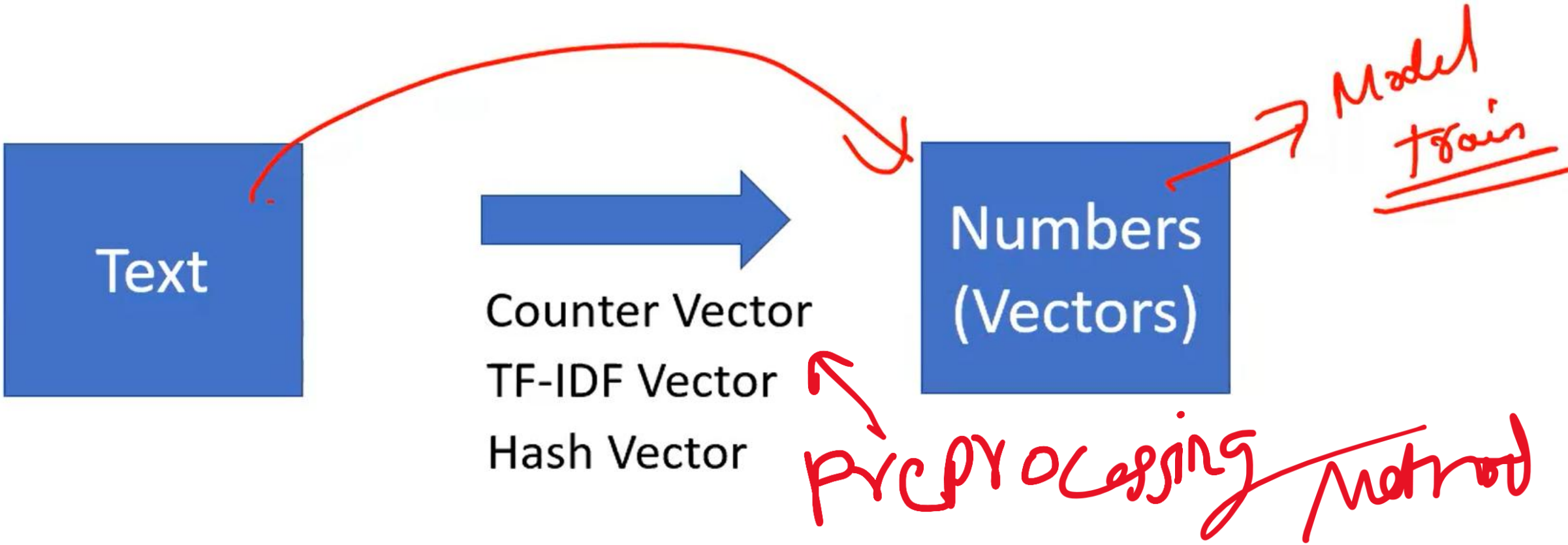



NLP Based Application
Prediction
Call to Action
NLP Application

Natural Language Processing

HOPES
Your Future Starts Here

How Computer Understands a Text?



Method	Meaning?	Context?	Size	Modern?
BoW	❌ No	❌ No	Large	Old
TF-IDF	❌ No	❌ No	Large	Old
One-Hot	❌ No	❌ No	Huge	Bad
Word2Vec	✓ Yes	❌ No	Small	Good
FastText	✓ Yes	❌ No	Small	Good
Sentence-BERT	✓✓	✓✓ Full context	Medium	Excellent
BERT/Transformers	✓✓	✓✓✓ Best	Medium/High	Best

⭐ 1 Bag of Words (BoW)

Text → Count of words

Example:

Sentence: "I love AI"

Word	Count
I	1
love	1
AI	1

Simple, but **no meaning** and no context.

⭐ 2 TF-IDF (Term Frequency – Inverse Document Frequency)

Better than BoW.

Gives **importance** of words.

Common words (the, is, a) have LOW weight.

Rare but meaningful words (cancer, rocket, diabetes) have HIGH weight.

Used in:

✓ Document classification

✓ Search ranking

✓ Keyword extraction

⭐ 3 One-Hot Encoding

Each word becomes a binary vector.

Example vocabulary:

```
["cat", "dog", "lion"]
```

Vector for “dog” →

```
[0, 1, 0]
```

Problem:

❌ Very large vectors

❌ No meaning relationship between words

⭐ 3 One-Hot Encoding

Each word becomes a binary vector.

Example vocabulary:

```
["cat", "dog", "lion"]
```

Vector for “dog” →

```
[0, 1, 0]
```

Problem:

❌ Very large vectors

❌ No meaning relationship between words

⭐ 4️⃣ Word Embeddings (Dense Vectors)

Words become **dense numeric vectors** that capture meaning.

Example:

"king" – "man" + "woman" ≈ "queen"

This means the model understands relationships!

Popular embedding models:

✓ **Word2Vec**

✓ **GloVe**

✓ **FastText**

Example vector (shortened):

"king" → `[0.12, -0.24, 0.58, 0.11, ...]`

⭐ 5️⃣ Sentence Embeddings

Entire sentence → 1 vector

Much more powerful than word vectors.

Models:

✓ **Sentence-BERT (SBERT)**

✓ **USE (Universal Sentence Encoder)**

✓ **MPNET**

✓ **E5 models**

Use cases:

✓ Search

✓ Chatbots

✓ Semantic similarity

✓ Clustering

Example (short):

"I love AI" → `[0.42, -0.21, 0.88, ...]`

⭐ 6️⃣ Transformer Embeddings (BERT, GPT, etc.)

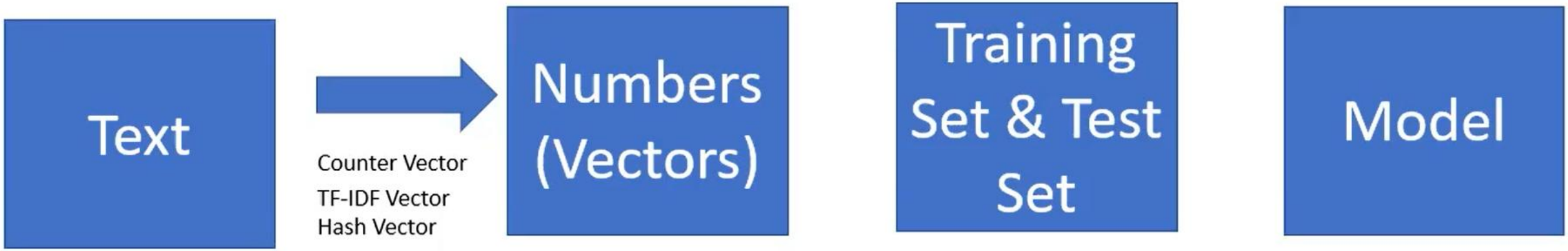
Most powerful method today.
Model like BERT converts each token → vector
AND gives full context understanding.
Example:
"In bank of river"
"bank" → embedding about water
"In bank deposit money"
"bank" → embedding about finance
Context-aware = 🔥
Used in:
✓ ChatGPT
✓ Google Search
✓ NLP research
✓ Classification
✓ Summarization
✓ Q&A

⭐ 7️⃣ Token IDs (for deep learning models)

Text → tokens → numbers
Example (subword tokenization):
Sentence: `"unbelievable"`
Tokens: `"un"`, `"believe"`, `"able"`
IDs: `[1872, 4512, 907]`
Used in LLMs, GRU/LSTM, Transformers.

NLP Model Creation

Text → ML → NLP predi



Counter Vector

	text
0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.

	king	was	the	not	But	him	one	north	kings	is	in	he	Eddard	everywhere	different	could	change	but	are	Stark	North	Hodor	without
0	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
1	2	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0
2	1	2	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0
3	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1

AutoSave Off

A1

	A	B	C	D	E
		title	text	label	
2	8476	You Can Smell Hillaryâ€™s Fear	Daniel Greenfield, a Shillman Journalism Fellow at the Freedom	FAKE	
3	10294	Watch The Exact Moment Paul Ryan Committed Polit	Google Pinterest Digg Linkedin Reddit Stumbleupon Print	FAKE	
4	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Monday that he will stop	REAL	
5	10142	Bernie supporters on Twitter erupt in anger against t	â€" Kaydee King (@KaydeeKing) November 9, 2016 The lesson	FAKE	
6	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners Hillary Clinton and	REAL	
7	6903	Tehran, USA		FAKE	
8	7341	Girl Horrified At What She Watches Boyfriend Do Aft	Share This Baylee Luciani (left), Screenshot of what Baylee caught	FAKE	
9	95	â€˜Britainâ€™s Schindlerâ€™ Dies at 106	A Czech stockbroker who saved more than 650 Jewish children fror	REAL	
10	4869	Fact check: Trump and Clinton at the 'commander-in	Hillary Clinton and Donald Trump made some inaccurate claims	REAL	
11	2909	Iran reportedly makes new push for uranium conces	Iranian negotiators reportedly have made a last-ditch push for	REAL	
12	1357	With all three Clintons in Iowa, a glimpse at the fire	CEDAR RAPIDS, Iowa â€" â€œI had one of the most wonderful	REAL	
13	988	Donald Trumpâ€™s Shockingly Weak Delegate Game	Donald Trumpâ€™s organizational problems have gone from bad	REAL	
14	7041	Strong Solar Storm, Tech Risks Today | S0 News Oct.2	Click Here To Learn More About Alexandra's Personalized	FAKE	
15	7623	10 Ways America Is Preparing for World War 3	October 31, 2016 at 4:52 am	FAKE	
16	1571	Trump takes on Cruz, but lightly	Killing Obama administration rules, dismantling Obamacare and pu	REAL	
17	4739	How women lead differently	As more women move into high offices,Â they often bring a style	REAL	
18	7737	Shocking! Michele Obama & Hillary Caught Glamoriz	Shocking! Michele Obama & Hillary Caught Glamorizing Date Rape	FAKE	

news

```
import pandas as pd
import numpy as np
import itertools
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, HashingVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
import matplotlib.pyplot as plt
```

Natural Language Processing

HOPES
Your Future Starts Here

TF-IDF Vector

Term Frequency – Inverse Doc

	text
0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.

	king	was	the	not	a		he	one	north	kings	is	in	him	everywhere	A	different	could	change	but	are	Stark	North	Hodor	Eddard
0	0.333333	0.0	0.5	0.0	0.5	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1	0.666667	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
2	0.333333	2.0	0.0	0.5	0.5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
3	0.000000	0.0	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0

King ⇒ 4

Inve

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents