Royal University of Phnom Penh

Faculty of Engineering

Thesis Topic:

Comparative Study on Khmer Text Retrieval

Advisor: Professor Rina Bouy, Phd

Student Name: Loeurm Vanda

Degree: Bachelor of Data Science & Engineering (**First Generation**)

Date: 21 Jun, 2025

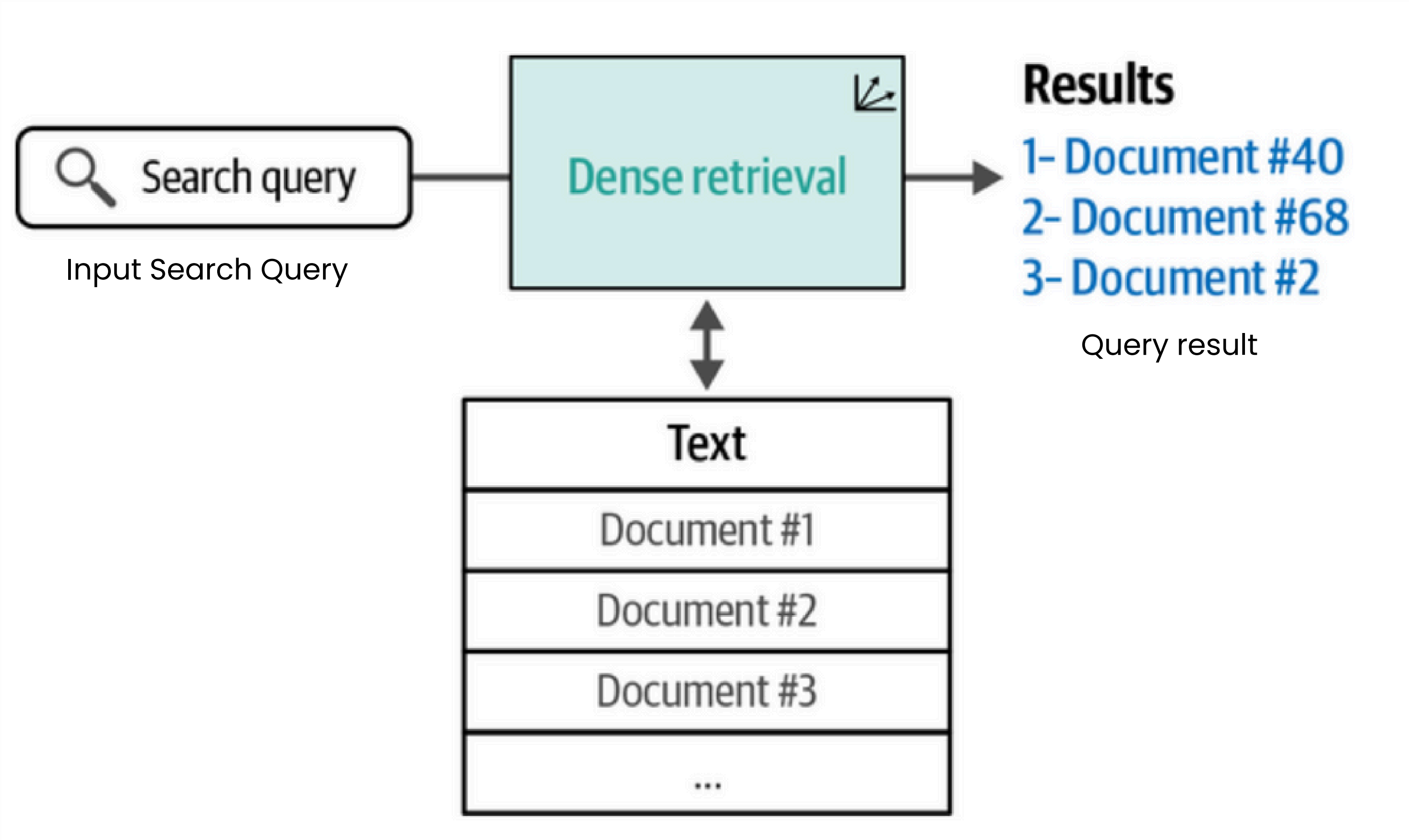
CONTENT

1. **Introduction**
2. **Problem**
3. **Research Objective**
4. **Literature Review**
5. **Methodology**
6. **Experiment Setup**
7. **Result**
8. **Limitations & Future Works**
9. **Conclusion**

- References
- Appendix

I. Introduction

AS Khmer digital content continue to growth, there are many available content such as document news, academic text or government documents in Khmer Language.



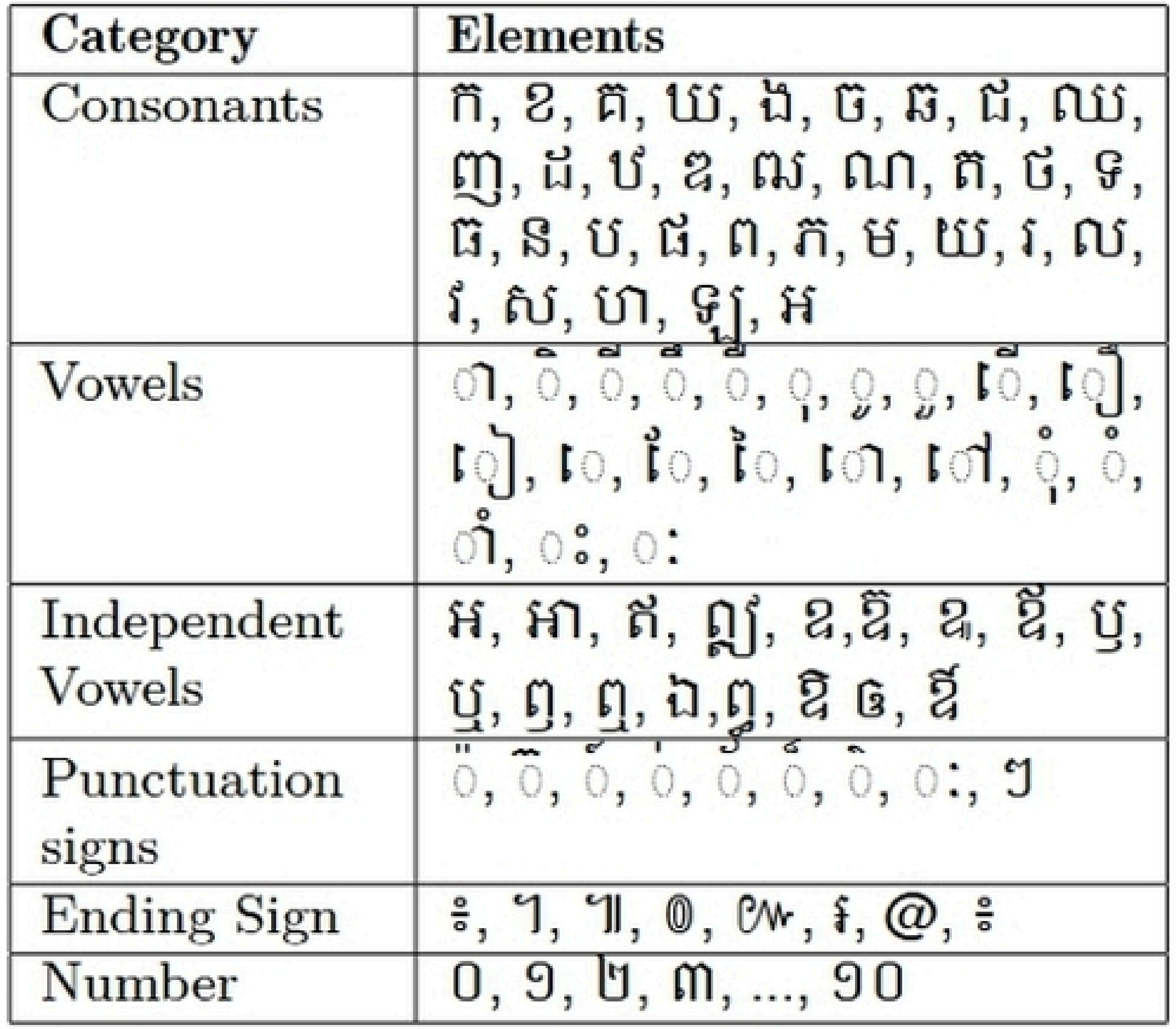
Document Embedding as Dense Vector

- Khmer text retrieval is the process of searching and retrieving the relevant documents from the large corpus of documents based on user query.
- Although there are many existing search systems that can perform with multi-language, the performance is still limited.

II. Problem

Searching for relevant information in the Khmer language is still challenging due to its complexity.

- Complex language characteristic and the absence of spaces.
- Lack of models or methods that can handle the Khmer Language.
- Fail to capture semantic meaning and focus only on lexical aspects.
- Lack of research and comparative studies on existing text retrieval.
- Low Khmer text data for NLP research.



Category	Elements
Consonants	ក, ខ, គ, ឃ, ង, ច, ឆ, ជ, ឈ, ញ, ដ, ឋ, ឌ, ឍ, ណ, ត, ថ, ទ, ធ, ន, ប, ផ, ព, ភ, ម, យ, រ, ល, វ, ស, ហ, ឡ, អ
Vowels	◌ា, ◌ិ, ◌ី, ◌ឹ, ◌ឺ, ◌ុ, ◌ូ, ◌ួ, ◌ើ, ◌ឿ, ◌ៀ, ◌េ, ◌ែ, ◌ៃ, ◌ោ, ◌ៅ, ◌ុំ, ◌ំ, ◌ាំ, ◌ះ, ◌ៈ
Independent Vowels	អ, អា, ឥ, ឦ, ឧ,ឩ, ឪ, ឳ, ឫ, ឬ, ឭ, ឮ, ឯ,ឰ, ឱ ឲ, ឳ
Punctuation signs	◌៉, ◌៊, ◌៍, ◌់, ◌៌, ◌៎, ◌័, ◌ៈ, ៗ
Ending Sign	៖, ។, ៕, ៙, ៚, ៛, @, ៖
Number	០, ១, ២, ៣, ..., ១០

III. Research Objectives

In this study, we aim to:

- Collect data from the internet website called **"Thmey Thmey News"**. Open datasets to public for future research in Hugging Face.
- Perform experiment and comparison on different methods:
  - **TF-IDF** (Term Frequency Inverse Document),
  - **BoW** (Bag of Words)
  - **Fasttext** embedding
  - **gte-multilingual-base** (General Text Embedding on multilingual).

IV. Literature Review

1. Challenges In Khmer Words

- Stacked VS Unstacked Syllable
  - **ចំការ** is writing in unstack syllable style.
  - **ចម្ការ** is writing in stack syllable style.

- **Semantic Meaning:**
  - **សប្បាយចិត្ត** and **រីករាយ** these two words meaning are semantically similar

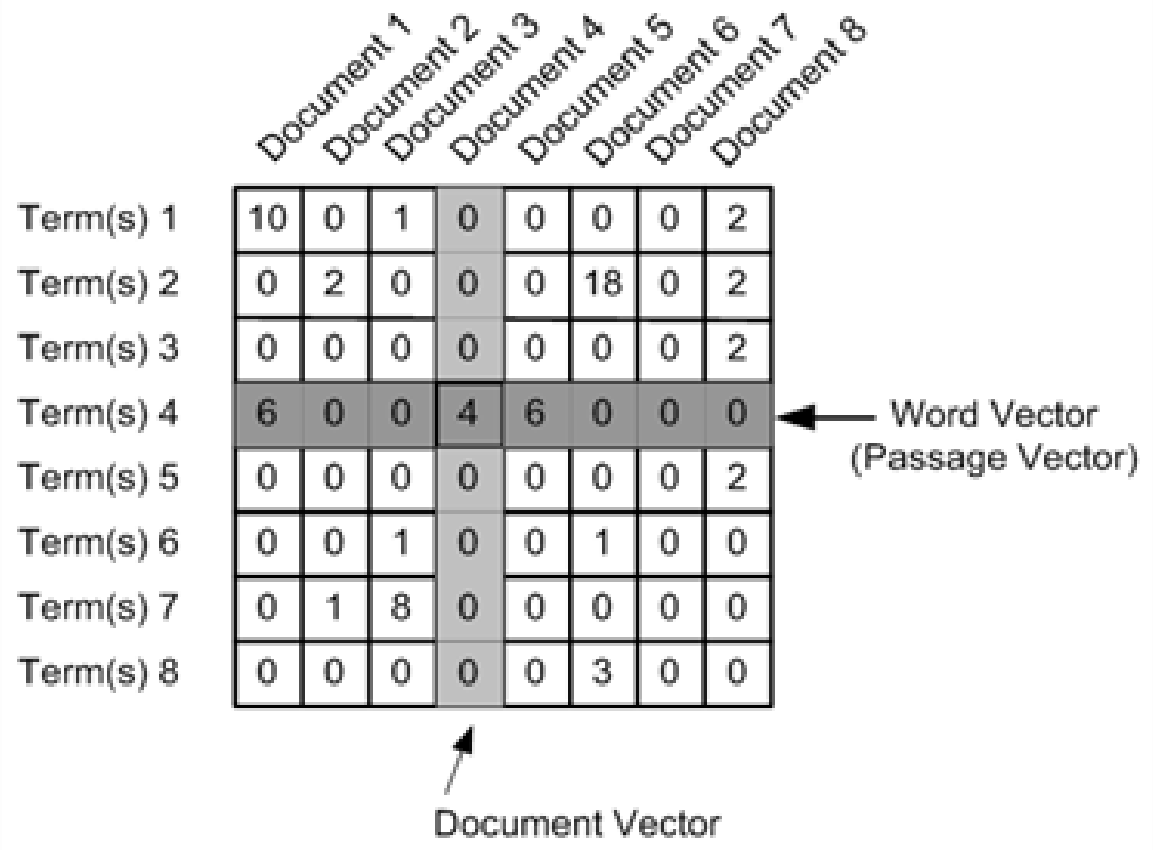
- **Absence of space** which hard in identifying the boundary of the words that lead to improper word segmentation.
  - I go to school.
  - **ខ្ញុំទៅសាលារៀន**

Original text (untoken Khmer): "ខ្ញុំទៅសាលារៀន"
Correct token output: ["ខ្ញុំ", "ទៅ", "សាលារៀន"]

IV. Literature Review

2. Vector and Documents

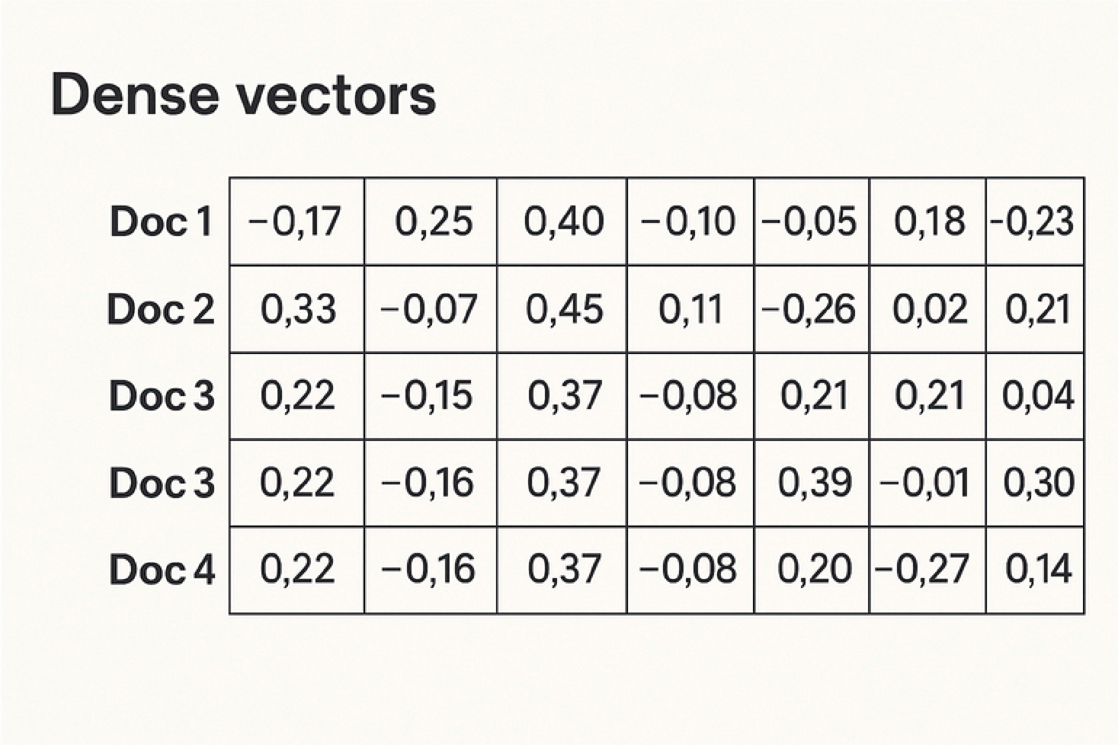
- Words, documents represented by vectors in natural language processing,
- Vectors and documents can be represented as a **term-document matrix, Row represents word or vocabulary and Column represents a document.**
- **E**ach cell tells how often and many times the words appear in each document.



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Word Vector
(Passage Vector)
Document Vector

Sparse Vector: representation method that represents text-like **one-hot encoding**, has the **element zero** with only **few non-zero values**.



Dense vectors

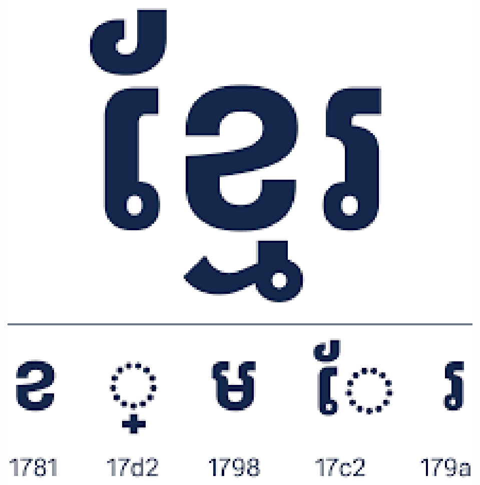
Doc 1	−0,17	0,25	0,40	−0,10	−0,05	0,18	-0,23
Doc 2	0,33	−0,07	0,45	0,11	−0,26	0,02	0,21
Doc 3	0,22	−0,15	0,37	−0,08	0,21	0,21	0,04
Doc 3	0,22	−0,16	0,37	−0,08	0,39	−0,01	0,30
Doc 4	0,22	−0,16	0,37	−0,08	0,20	-0,27	0,14

Dense Vector: is a **non zero element** that stores meaningful numerical value across all dimensions. It **generated by embedding model**.

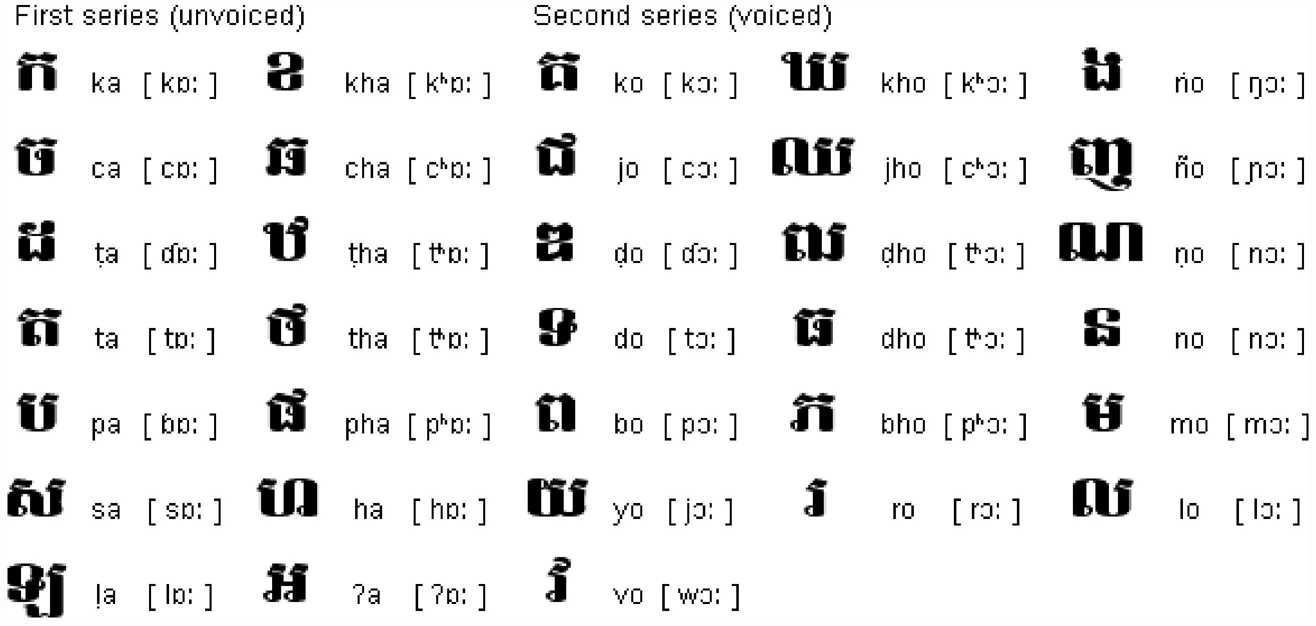
IV. Literature Review

3. Related Work: Semantic Word Search

Previous Reserch on semantic search that conducted by **Buoy et al. (2021)** also proposed solution of the preprocessing step before applying semantic model like **FastText** to extract feature or turn to vector representation. The proposed Solutions are:



Character Sequence Normalization:
Break & reorder syllables into base consonant order, subscripts, diacritics, vowels.



First series (unvoiced)						Second series (voiced)								
ក	ka	[kɒː]	ខ	kha	[kʰɒː]	គ	ko	[kɔː]	ឃ	kho	[kʰɔː]	ង	ṅo	[ŋɔː]
ច	ca	[cɒː]	ឆ	cha	[cʰɒː]	ជ	jo	[cɔː]	ឈ	jho	[cʰɔː]	ញ	ño	[ɲɔː]
ដ	ṭa	[ɗɒː]	ឋ	ṭha	[tʰɒː]	ឌ	ḍo	[ɗɔː]	ឍ	ḍho	[tʰɔː]	ណ	ṇo	[nɔː]
ត	ta	[tɒː]	ថ	tha	[tʰɒː]	ទ	do	[tɔː]	ធ	dho	[tʰɔː]	ន	no	[nɔː]
ប	pa	[ɓɒː]	ផ	pha	[pʰɒː]	ព	bo	[pɔː]	ភ	bho	[pʰɔː]	ម	mo	[mɔː]
ស	sa	[sɒː]	ហ	ha	[hɒː]	យ	yo	[jɔː]	រ	ro	[rɔː]	ល	lo	[lɔː]
ឡ	ḷa	[lɒː]	អ	ʔa	[ʔɒː]	វ	vo	[wɔː]						

Khmer G2P Spell Checker:
Converts graphemes to phonemes → finds correct phoneme sequences → reconverts to correct spelling

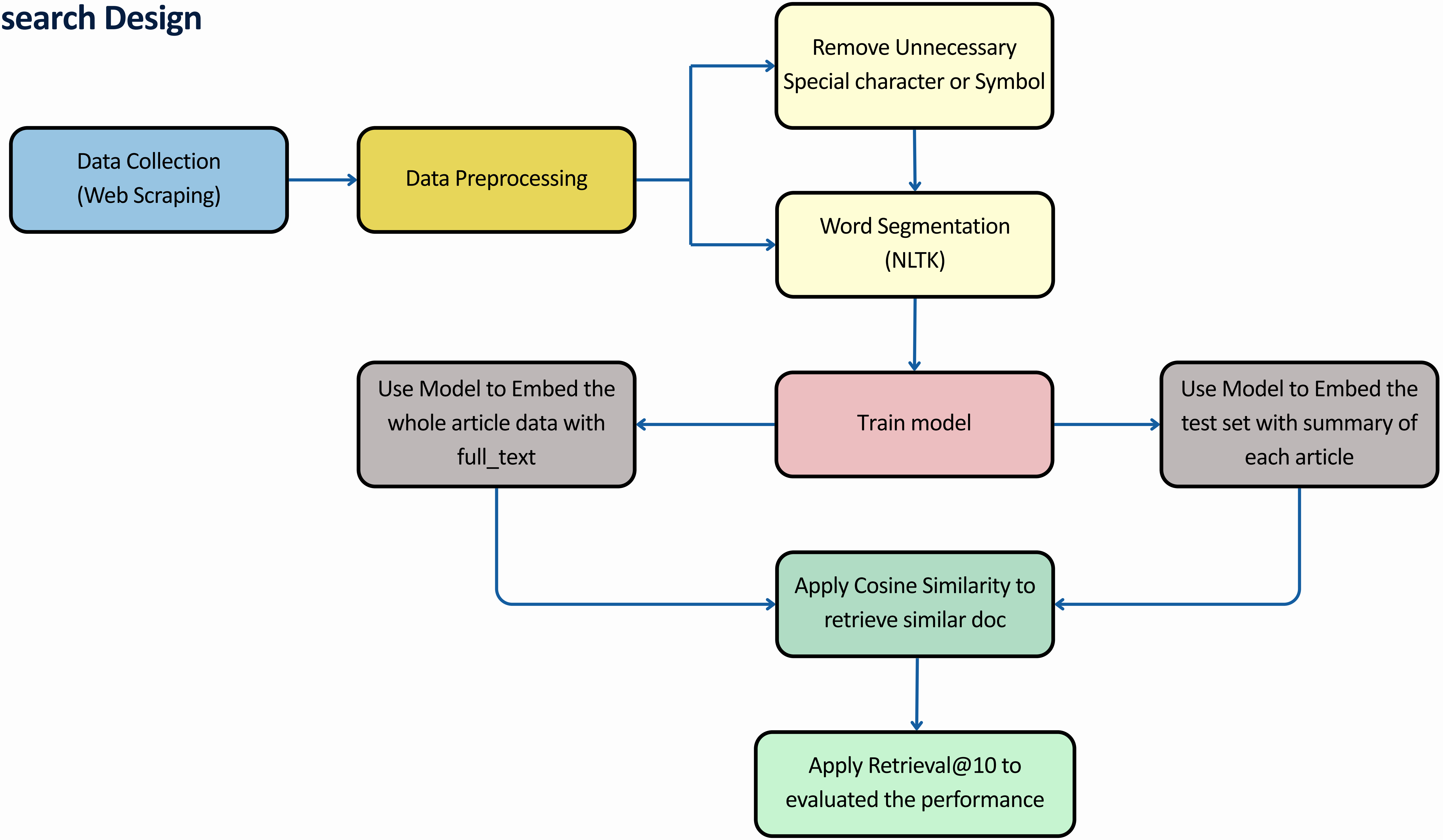
FastText for
Semantic Model

Gap in Existing Research and Data:

- Data: Not much text datasets in Khmer
- Comparison of Embedding model or text representation model with other model to see which model perform best.

V. Methodology

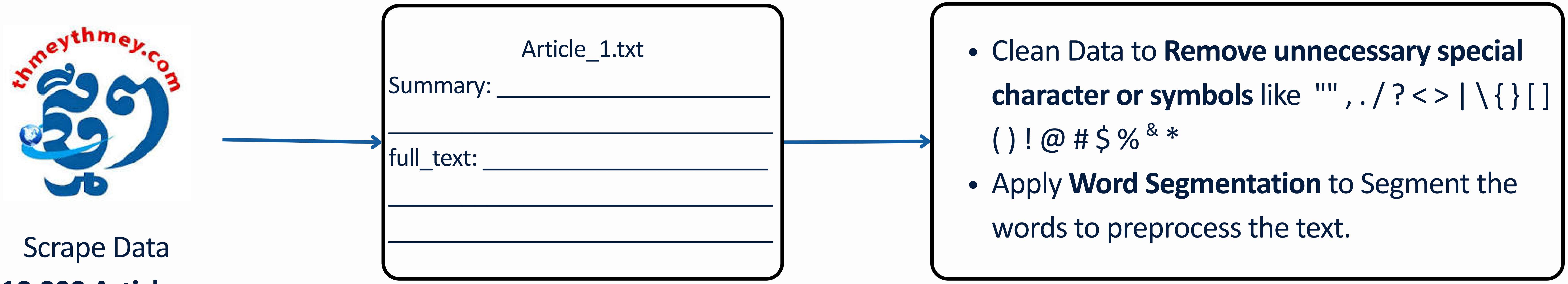
1. Research Design


Data Collection (Web Scraping)
Data Preprocessing
Remove Unnecessary Special character or Symbol
Word Segmentation (NLTK)
Train model
Use Model to Embed the whole article data with full_text
Use Model to Embed the test set with summary of each article
Apply Cosine Similarity to retrieve similar doc
Apply Retrieval@10 to evaluated the performance

V. Methodology

2. Data Collection & Preprocessing

- Scrape Data from "**Thmey Thmey News"** with **10K** articles that include summary and full_text for each article.



Scrape Data
10,000 Articles

Article_1.txt
Summary: ____________________
____
full_text: ____________________
____
____

Format Structure after Scraped (txt file)

- Clean Data to **Remove unnecessary special character or symbols** like "" , . / ? < > | \ { } [ ] ( ) ! @ # $ % & *
- Apply **Word Segmentation** to Segment the words to preprocess the text.

Data Cleaning & Preprocessing

- **Word Segmentation** is a preprocess to segment the text into meaningful words to identify the word boundaries for Khmer.

Original text (unsegmented Khmer):
"ចំណេះដឹងគឺជាកម្លាំង"

After Segmented

Segmented output:
["ចំណេះដឹង", "គឺជា", "កម្លាំង"]

V. Methodology

3. Models (TF-IDF : Term Frequency - Inversed Document Frequency)

Term Frequency Inverse Document or TF-IDF is a **statistical method** that **measures how important the word** is within the document in the collection of a corpus.

$$w_{t,d} = \boxed{\mathrm{tf}_{t,d}} \times \boxed{\mathrm{idf}_t}$$

$$tf_{t,d} = \mathrm{count}(t, d)$$

$$idf_t = log_{10}(\frac{N}{df_t})$$

tf term frequency which is the frequency of the word or ***term t*** in the ***document d***

idf give a **higher weight** to **words** that **appear** only in a **few documents**.

- These two weights above **reflect** both the **term's importance** in the document and **its rarity** across the corpus.
- The final **TF-IDF weight** for a term t in document d is the **product** of its **term frequency (TF)** and **inverse document frequency (IDF)**.

V. Methodology

3 Models (BoWs: Bag of Words)

Bag of words is a high-dimensional, sparse vector that represents the text based on the frequency of the words or term in a document without adding weight to the rare word like tf-idf.

Document D1	កម្ពុជាជាផ្ទះរបស់ប្រជាជនកម្ពុជាជាច្រើន	កម្ពុជា ជា ផ្ទះ ប្រជាជន កម្ពុជា កម្ពុជា: 2, ជា: 1, ផ្ទះ: 1, ប្រជាជន: 1
Document D2	សេដ្ឋកិច្ចកម្ពុជាបន្តកើនឡើង	សេដ្ឋកិច្ច កម្ពុជា បន្ត កើន ឡើង សេដ្ឋកិច្ច: 1, កម្ពុជា: 1, បន្ត:1, កើន: 1, ឡើង: 1

	កម្ពុជា	**ជា**	**ផ្ទះ**	**ប្រជាជន**	**សេដ្ឋកិច្ច**	**បន្ត**	**កើន**	**ឡើង**	**BoW Vector Representation**
D1	2	1	1	1	0	0	0	0	[2,1,1,1,0,0,0,0]
D2	1	0	0	0	1	1	1	1	[1,0,0,0,1,1,1,1]

Sparse Vector Representation

V. Methodology

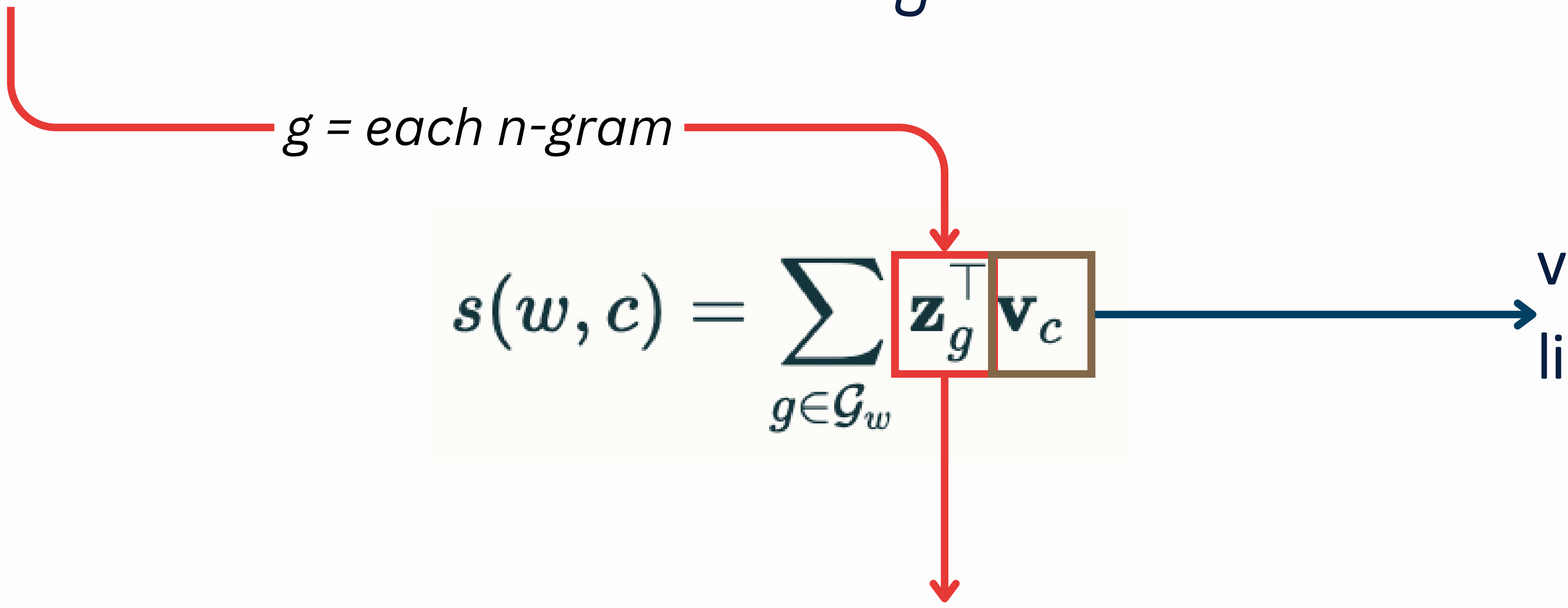
3 Models (FastText)

- FastText is a **word embedding** technique developed by Facebook AI Research **(Bojanowski et al., 2017)**. Like Word2Vec, FastText derived from the Skip gram base model that its learning objective is to predict the surrounding words.
- The **scoring function** between **word** and its **context(surrounded words)** is calculated as below:

Example: "ខ្ញុំទៅសាលារៀន"

The word ***denoted as w =*** "សាលារៀន"
with n-gram = 3
Gw = <សា, សាល, ◌ាលា, លារ, ◌ារៀ, រៀន, ◌ៀន>

Sample of Scoring function between words
"សាលារៀន" *and context word* **"ទៅ"**
S(សាលារៀន, ទៅ)



g = each n-gram

$$s(w, c) = \sum_{g \in \mathcal{G}_w} \boxed{\mathbf{z}_g^\top} \boxed{\mathbf{v}_c}$$

vector of context word (surrounded words).
like context word denoted as C = "ខ្ញុំ" or "ទៅ"

Zg : Vector transpose of the n-gram vector representation of the center words.

V. Methodology

3 Models (gte-multilingual-base)

GTE-Multilingual-Base is an open-source model for **text retrieval in different languages**. The model was developed by **Zhang et al. (2024)**, which supports multilingual long context and could **query up to 8192 tokens.** It consists of two components are **Text Representation Model (TRM)** and **Reranker.**

Text Encoder:

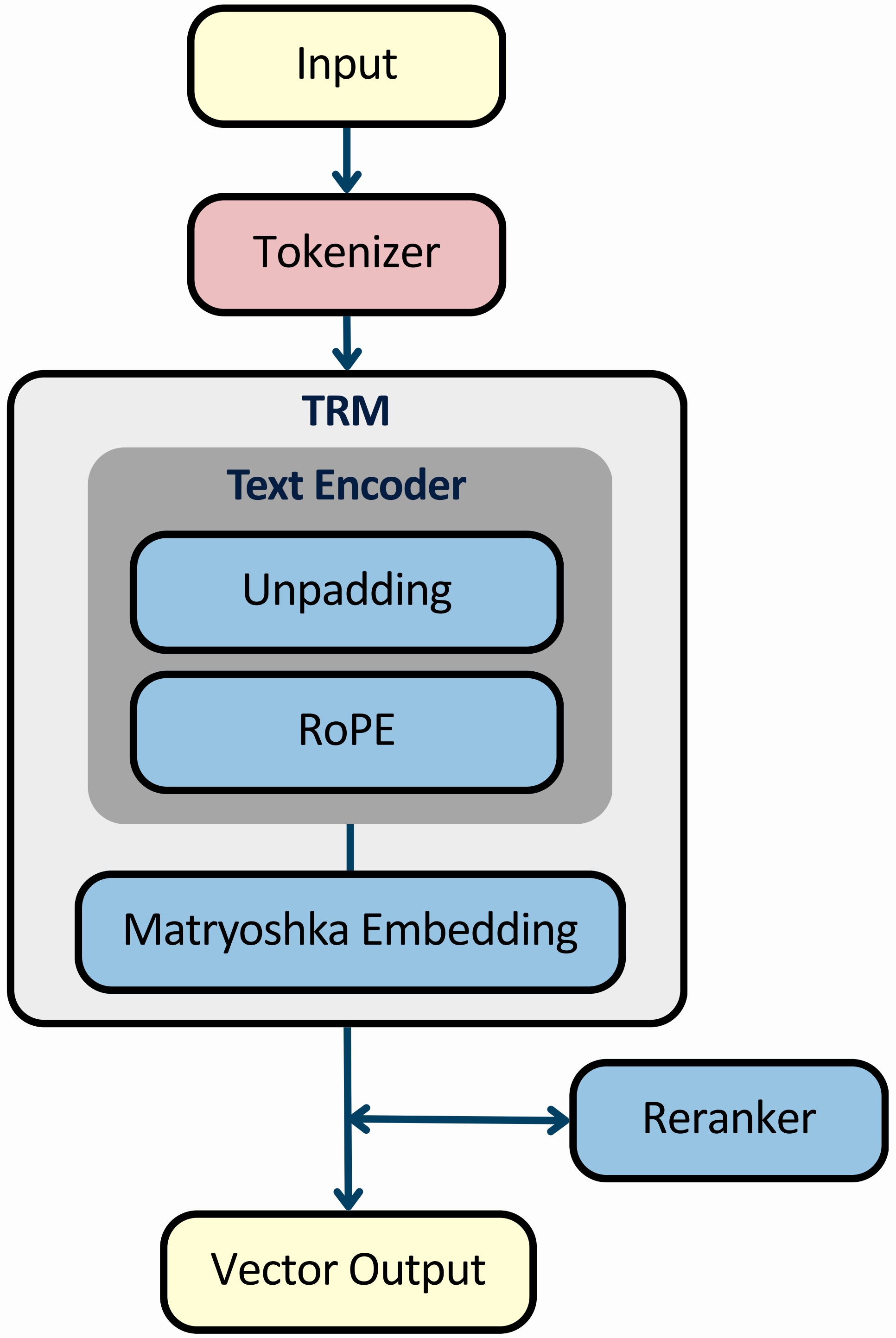
- Built on **Bert** Architecture (Encoder Only)
- Enhanced with **RoPE** and **Unpadding** for **efficiency**
- **Train** Encoder with **two** stages:
  - Stage 1: **2048** Tokens
  - Stage 2: **8192** Tokens

TRM (Text Representation Model)

- Use **Matryoshka** embedding for **elastic vector size**
- Trained to **shorten** the embedded vector, but still **meaningful**.

Reranker

- Reorders retrieved candidates to refine the results



GTE multilingual architecture

V. Methodology

4. Retrieval Method (Cosine Similarity)

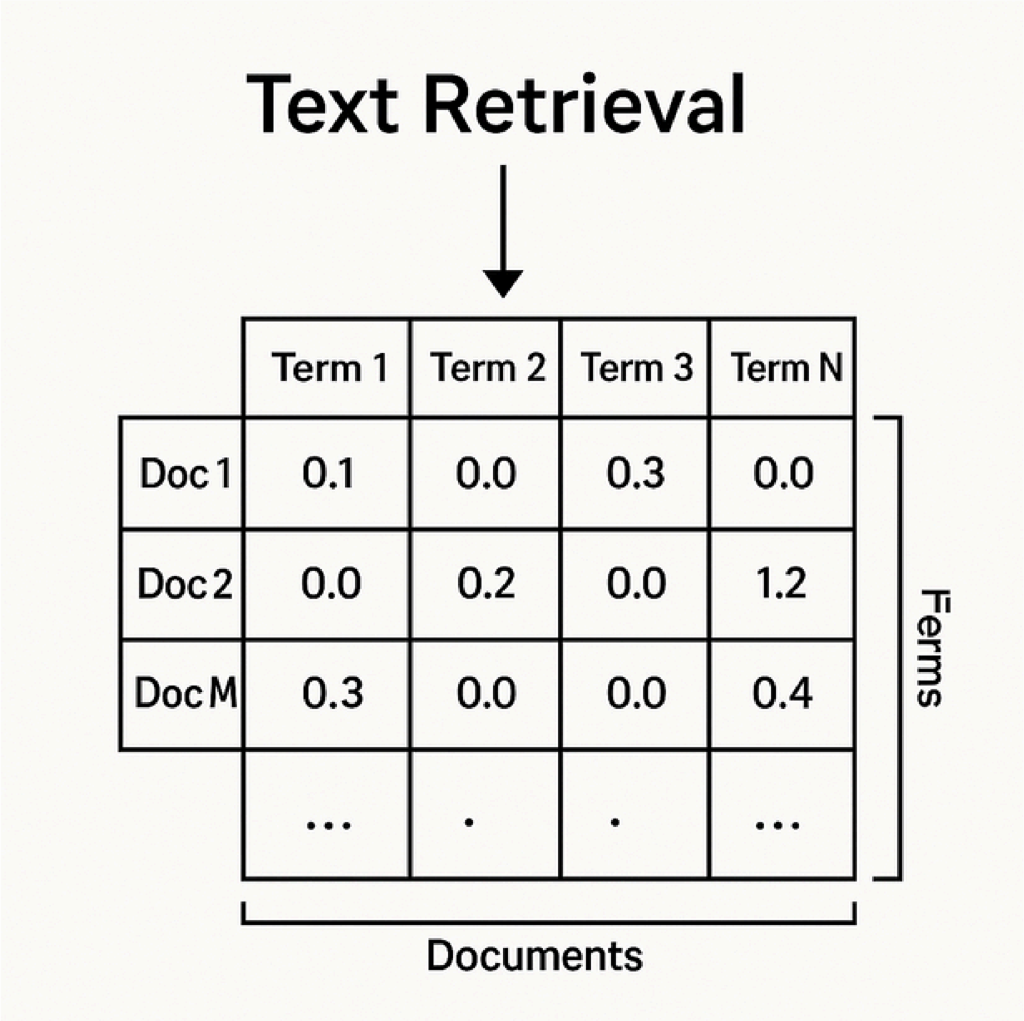
After the documents were turned into **vector representation** as the figure below, to **measure** how similar **two vectors** are, this study uses **cosine similarity** to **retrieve document vectors** which are **similar to the query.**

Cosine Similarity between two vectors is calculated as below:

$$\text{Cosine Similarity} = \frac{\boxed{\vec{A}} \cdot \boxed{\vec{B}}}{||\vec{A}|| \times ||\vec{B}||}$$

Vector of the query = Vector of the document

Query Vector and **Document Vector** must have the same dimensions


Text Retrieval
Terms
Documents

	Term 1	Term 2	Term 3	Term N
Doc 1	0.1	0.0	0.3	0.0
Doc 2	0.0	0.2	0.0	1.2
Doc M	0.3	0.0	0.0	0.4
	...	.	.	...

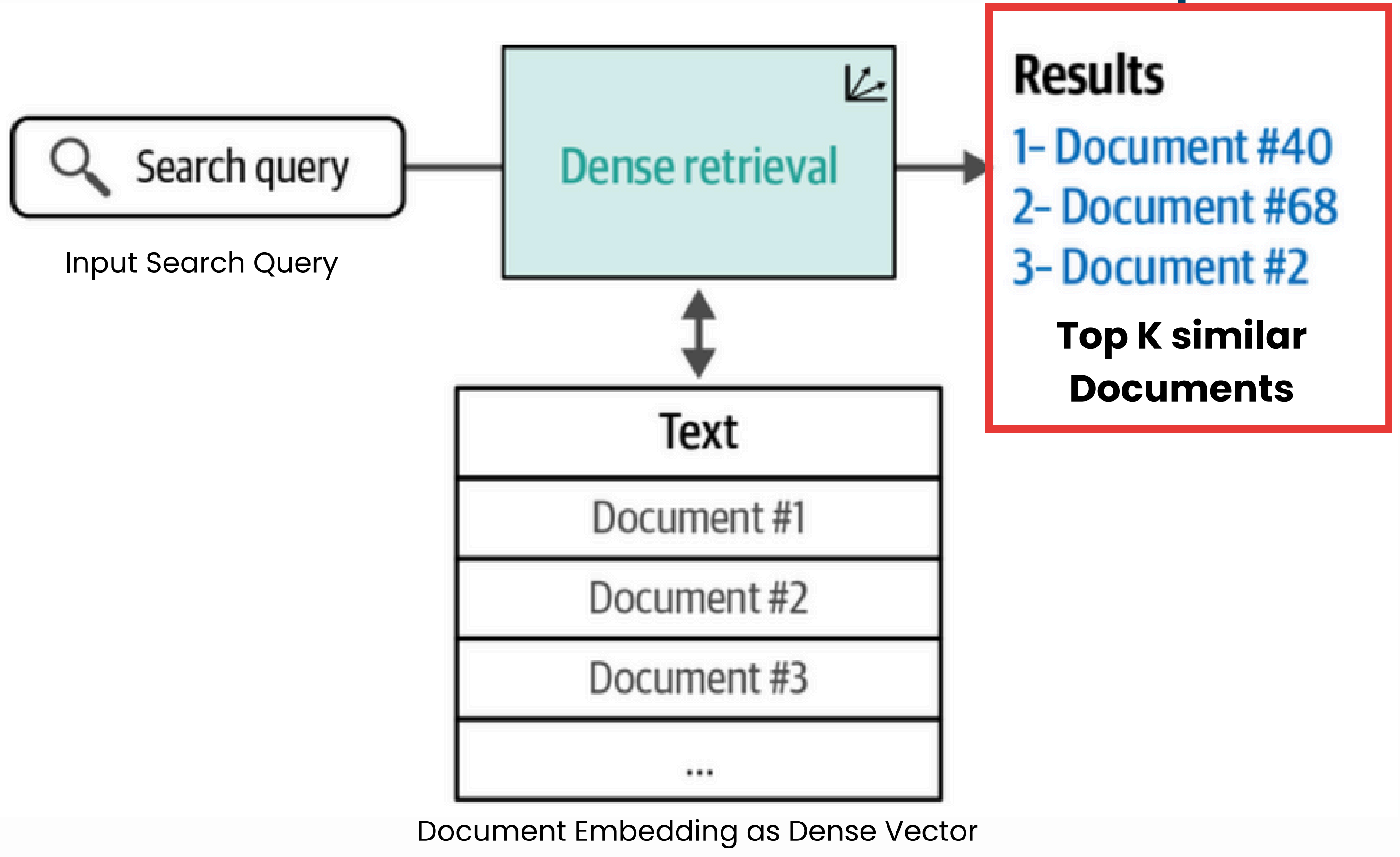
V. Methodology

5. Evaluation Metrics

- To evaluate the effectiveness of the article retrieval system, **Retrieve@10** was used as the main metric. This metric **checks** whether the **correct article appears** within the **top 10 results** returned by the system when a query is made.
- **Measures** the percentage of queries where the **correct article** is **found in** the **top 10** most similar results.

$$\text{Retrieve@10} = \frac{\text{Number of correct results in top 10}}{\text{Total number of queries}}$$

total number test query where each query is use to retrieve the each document

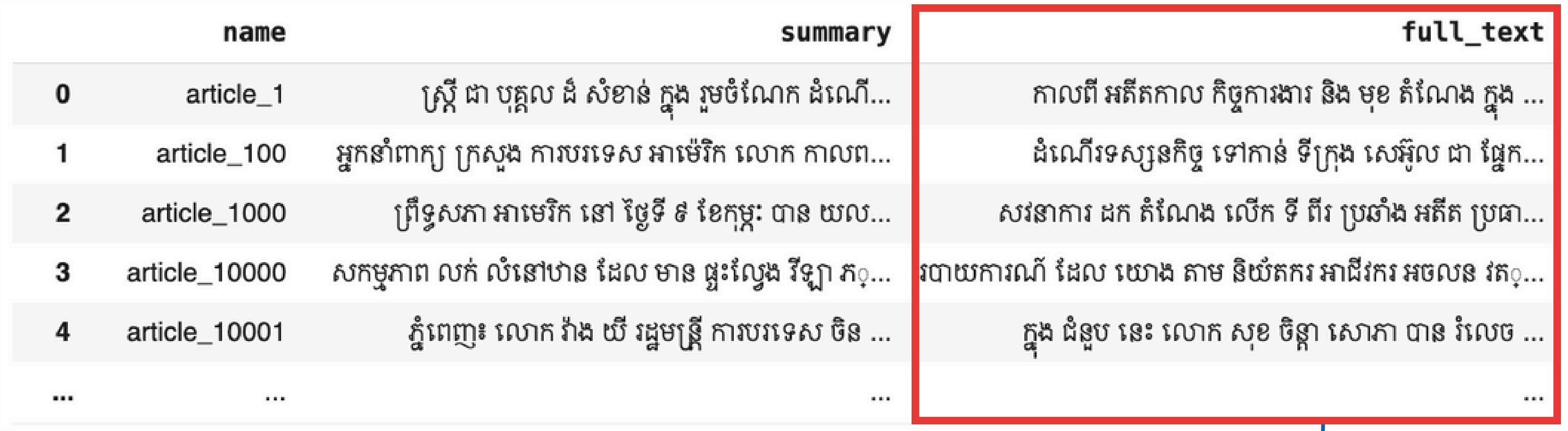


VI. Experiment

During the Experiment the Kaggle will be used for this research experiment and evaluating each model performance.

- After preprocessing and cleaning, **10, 000 Khmer articles** will be splitted into training sets **80%** and testing sets **20%** and put in dataframe as picture below for easy access.

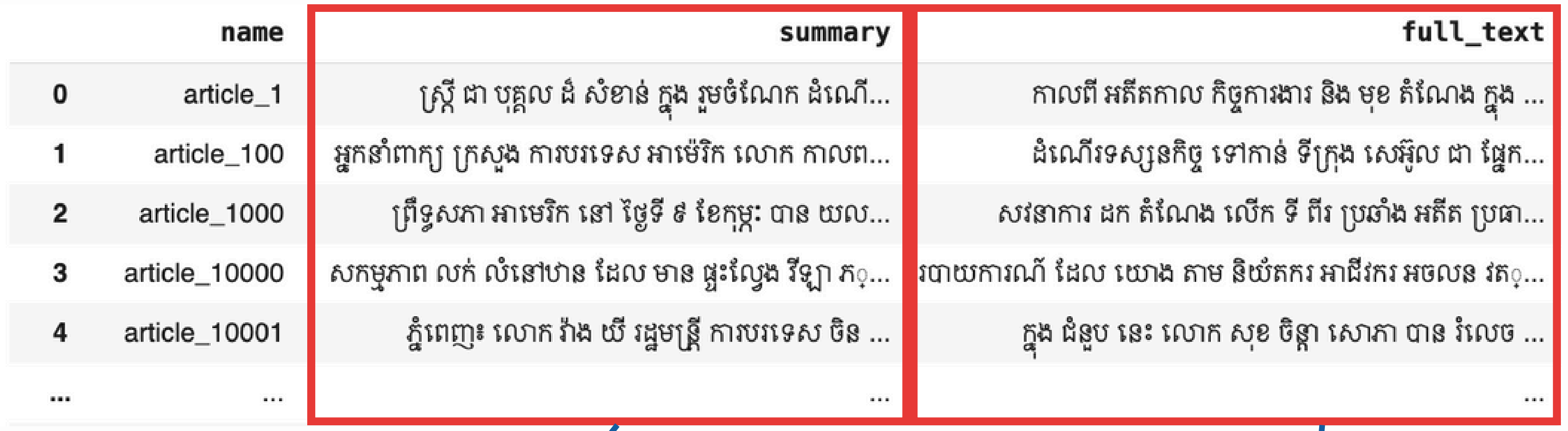
Training Phase



	name	summary	full_text
0	article_1	ស្ត្រី ជា បុគ្គល ដ៏ សំខាន់ ក្នុង រួមចំណែក ដំណើ...	កាលពី អតីតកាល កិច្ចការងារ និង មុខ តំណែង ក្នុង ...
1	article_100	អ្នកនាំពាក្យ ក្រសួង ការបរទេស អាម៉េរិក លោក កាលព...	ដំណើរទស្សនកិច្ច ទៅកាន់ ទីក្រុង សេអ៊ូល ជា ផ្នែក...
2	article_1000	ព្រឹទ្ធសភា អាមេរិក នៅ ថ្ងៃទី ៩ ខែកុម្ភៈ បាន យល...	សវនាការ ដក តំណែង លើក ទី ពីរ ប្រឆាំង អតីត ប្រធា...
3	article_10000	សកម្មភាព លក់ លំនៅឋាន ដែល មាន ផ្ទះល្វែង វីឡា ភ្...	របាយការណ៍ ដែល យោង តាម និយ័តករ អាជីវករ អចលន វត្...
4	article_10001	ភ្នំពេញ៖ លោក វ៉ាង យី រដ្ឋមន្ត្រី ការបរទេស ចិន ...	ក្នុង ជំនួប នេះ លោក សុខ ចិន្តា សោភា បាន រំលេច ...
...	...	...	...

Train the model with **full_text** of **training set** to let it learn the words

Testing Phase

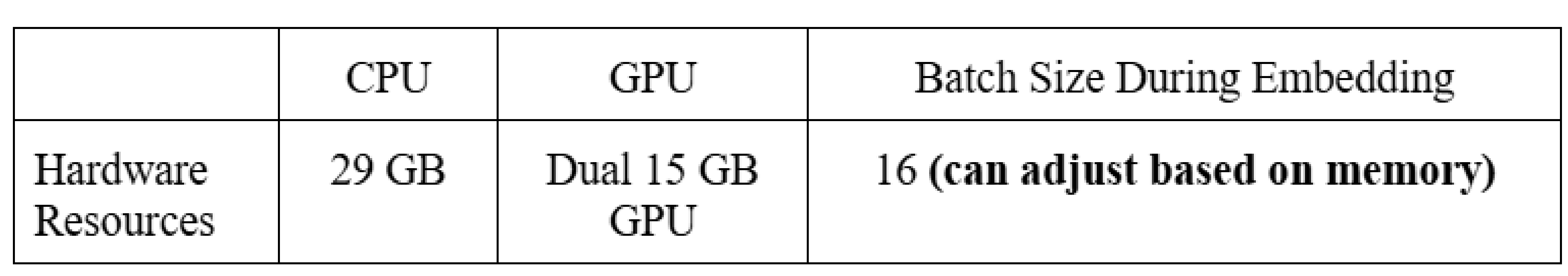


	name	summary	full_text
0	article_1	ស្ត្រី ជា បុគ្គល ដ៏ សំខាន់ ក្នុង រួមចំណែក ដំណើ...	កាលពី អតីតកាល កិច្ចការងារ និង មុខ តំណែង ក្នុង ...
1	article_100	អ្នកនាំពាក្យ ក្រសួង ការបរទេស អាម៉េរិក លោក កាលព...	ដំណើរទស្សនកិច្ច ទៅកាន់ ទីក្រុង សេអ៊ូល ជា ផ្នែក...
2	article_1000	ព្រឹទ្ធសភា អាមេរិក នៅ ថ្ងៃទី ៩ ខែកុម្ភៈ បាន យល...	សវនាការ ដក តំណែង លើក ទី ពីរ ប្រឆាំង អតីត ប្រធា...
3	article_10000	សកម្មភាព លក់ លំនៅឋាន ដែល មាន ផ្ទះល្វែង វីឡា ភ្...	របាយការណ៍ ដែល យោង តាម និយ័តករ អាជីវករ អចលន វត្...
4	article_10001	ភ្នំពេញ៖ លោក វ៉ាង យី រដ្ឋមន្ត្រី ការបរទេស ចិន ...	ក្នុង ជំនួប នេះ លោក សុខ ចិន្តា សោភា បាន រំលេច ...
...	...	...	...

Use **Summary as the Query** to **retrieve** the **documents** .

Use the model to **Embed** the **full_text of entire data**

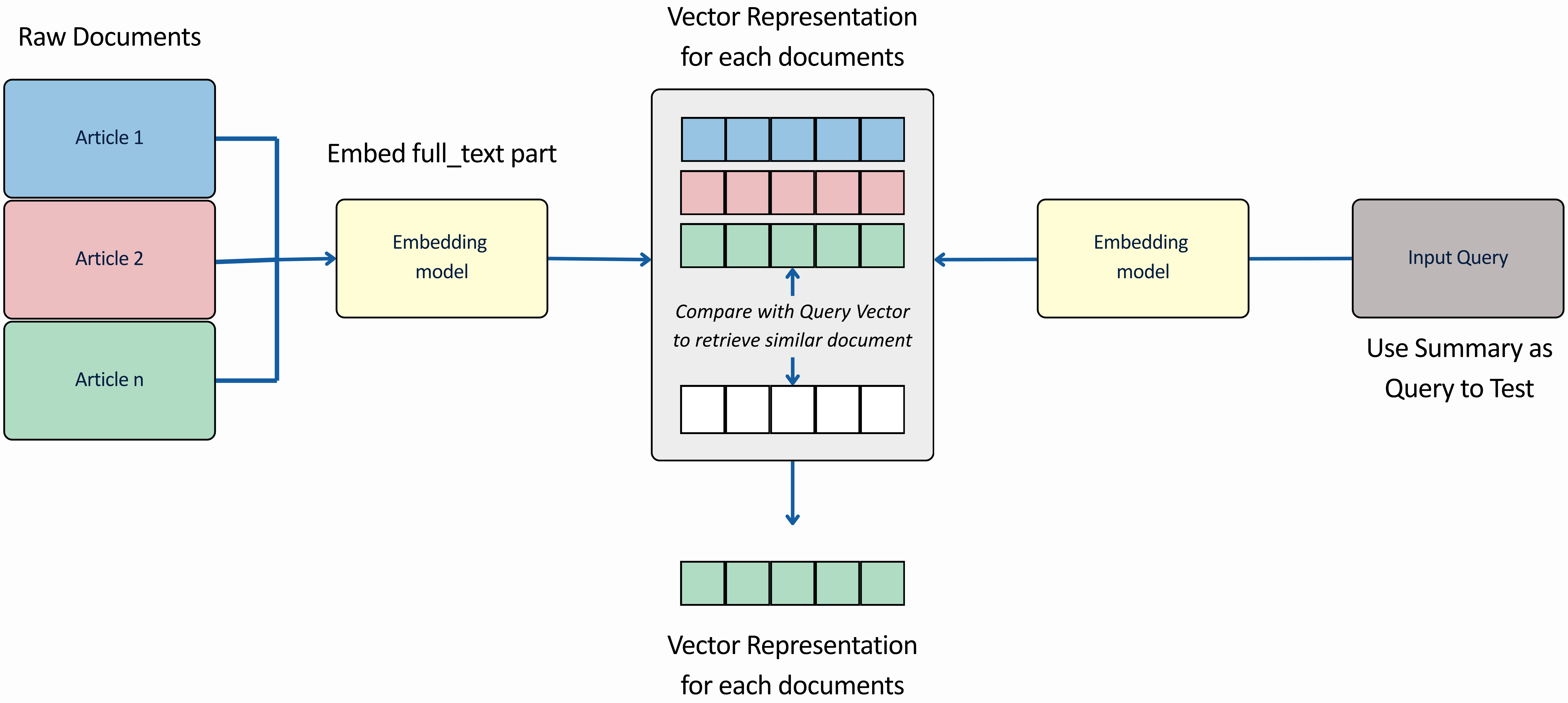
The experiment the **requirement resources** that might needed to embed the documents is **Kaggle GPU TP4 x 2.**



	CPU	GPU	Batch Size During Embedding
Hardware Resources	29 GB	Dual 15 GB GPU	16 **(can adjust based on memory)**

VI. Experiment

Detail Process of Retrieving the relevance document with summary as query during testing.


Raw Documents
Article 1
Article 2
Article n
Embed full_text part
Embedding model
Vector Representation for each documents
Compare with Query Vector to retrieve similar document
Embedding model
Input Query
Use Summary as Query to Test
Vector Representation for each documents

VII. Result

Model	Total Query	Correctly Retrieve	Not in top 10	Retrieval@10
TF-IDF	2005	1725	280	86%
BoWs	2005	1164	841	58,05%
FastText	2005	1188	817	59,25%
gte-multilingual-base	2005	1904	101	94,96%

- The results show that **GTE-Multilingual-Base performed the best** among the four methods, especially in understanding the semantic meaning of the queries.
- On the other hand, **TF-IDF** also gave a strong result with an **86% score**, showing that even without deep semantic understanding, traditional methods can still be very effective by focusing on important keywords.
- **Bag of Words** and **FastText embeddings** had lower performance, likely **due to limitations** in **capturing context** and **meaning** or **the way** they **were trained** on the data.

VIII. Limitations & Future Work

Limitations

- Limited Khmer corpus may **not reflect diverse domain** usage.
- Few **NLP tools** or **pre-trained models** support Khmer.
- **Broader comparisons** with **more methods** and languages **can reveal** Khmer-specific **retrieval challenges**.

Future Works

- **Expand** the Khmer text **corpus** to cover more **diverse domains**.
- Develop **more NLP** tools and **pre-trained models** for Khmer.
- **Compare** with more **retrieval methods** and **other languages** to better understand Khmer-specific challenges.

IX. Conclusion

- This study **compared vectorization methods** for Khmer text retrieval using both **traditional** and **embedding-based** approaches.
- GTE-Multilingual-Base **outperformed** others in **capturing semantic meaning**, while **TF-IDF** also performed well (86%), highlighting the strength of **keyword-based methods**.
- Bag of Words and FastText showed lower accuracy, likely due to their **limited context understanding** and the way they **were trained** (FastText is more suited to work with word semantics for searching with words or classification tasks, rather than sentences, for semantic understanding in long queries).
- Overall, the results suggest that combining various embedding methods can improve search systems for low-resource languages, such as Khmer, **as seen in the GTE architecture**.

References

[1] Thuon, N. (2024). Khmer Semantic Search Engine (KSE): Digital Information Access and Document Retrieval. arXiv, June 16, 2024. https://doi.org/10.48550/arXiv.2406.09320.

[2] Buoy, R., Taing, N., & Chenda, S. (2021). Khmer word search: Challenges, solutions, and semantic-aware search. arXiv preprint arXiv:2112.08918. https://arxiv.org/abs/2112.08918

[3] Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2024. All rights reserved. Draft of January 12, 2025. https://web.stanford.edu/~jurafsky/slp3/6.pdf

[4] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5(1), 135–146. https://doi.org/10.1162/TACL_A_00051

[5] Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., & Zhang, M. (2024). mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. https://doi.org/10.48550/arxiv.2407.19669

[6] Su, Jianlin, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. “RoFormer: Enhanced Transformer with Rotary Position Embedding.” arXiv, November 8, 2023. https://doi.org/10.48550/arXiv.2104.09864.

Appendices

Appendix 1: Khmer Text Data on Huggingface: For future research or study purpose
https://huggingface.co/datasets/Vanda10/khmer_text_articles_data

Appendix 2: Comparative Study Source Code
https://github.com/Vanda10/comparative_study_on_khmer_text_retrieval/tree/main

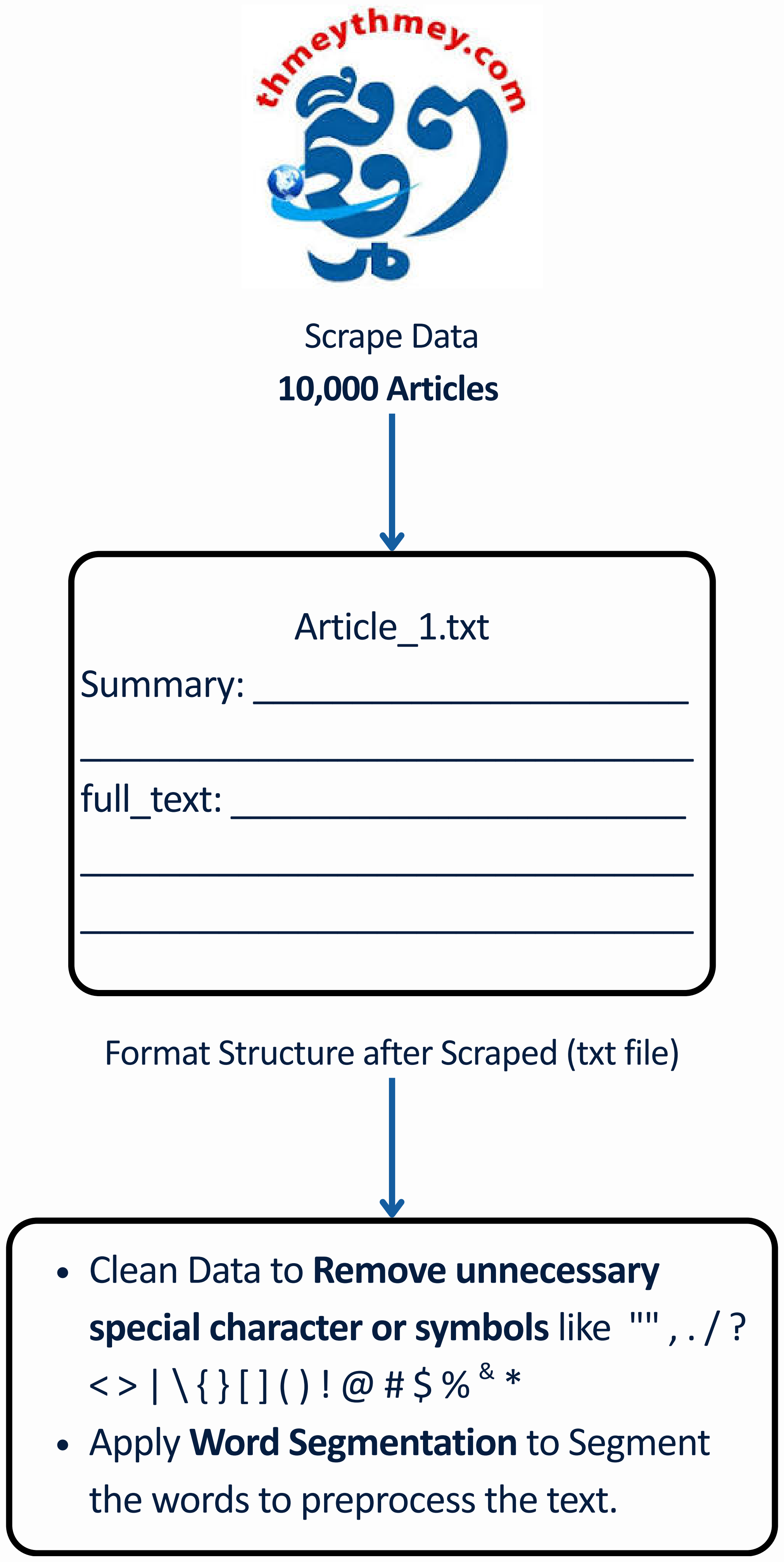
Royal University of Phnom Penh
Faculty of Engineering

Congratulations Class of 2025!

THANK YOU

Date: 21 Jun, 2025

Presented by: Vanda Loeurm



Scrape Data

10,000 Articles

Article_1.txt

Summary: ______________________

____

full_text: ______________________

____

____

Format Structure after Scraped (txt file)

- Clean Data to **Remove unnecessary special character or symbols** like "" , . / ? < > | \ { } [ ] ( ) ! @ # $ % & *
- Apply **Word Segmentation** to Segment the words to preprocess the text.

Data Cleaning and Preprocessing