

**BÁO CÁO ĐỒ ÁN – PBL7**  
**HỆ THỐNG DỰ BÁO XU HƯỚNG**  
**CHỦ ĐỀ BÀI BÁO KHOA HỌC**

**Giảng viên hướng dẫn**

**TS. HUỲNH HỮU HƯNG**

**Sinh viên thực hiện**

**Nguyễn Văn Hoàng Phúc**

**Nguyễn Văn Mạnh**

**Nguyễn Công Cường**

**Nhóm**

**20Nh10**

**Lớp**

**20T1**

# Table of contents

**01**

**Giới thiệu**

**02**

**Công việc liên quan**

**03**

**Dữ liệu**

**04**

**Giải pháp về  
công nghệ**

**05**

**Kết quả**

**06**

**Kết luận và  
hướng phát triển**

# 01. Giới thiệu


Trong thời đại công nghệ số hiện nay, việc dự báo xu hướng của chủ đề bài báo khoa học đã trở thành một lĩnh vực quan trọng, giúp các nhà nghiên cứu, học giả và các tổ chức khoa học định hướng và phát triển nghiên cứu một cách hiệu quả hơn.

- **Mục đích:** Xây dựng một hệ thống dự đoán xu hướng và tìm kiếm bài báo khoa học dựa trên nguồn dữ liệu đáng tin cậy. Giúp cho các nghiên cứu sinh dễ dàng tiếp cận các công nghệ hiện đại nhất và các xu hướng công nghệ.
- **Mục tiêu:**
  - Trích xuất keyword từ một bài báo khoa học nói chung và một đoạn văn bản nói riêng.
  - Cung cấp các sơ đồ, biểu đồ thể hiện các keywords (chủ đề) đang thịnh hành một cách trực quan.
  - Cho phép nghiên cứu sinh tìm kiếm thông qua mô tả về công việc muốn thực hiện, hiển thị danh sách các bài báo khoa học liên quan nhất, có kèm link bài báo.
  - Gợi ý các chủ đề đang thịnh hành và bài báo khoa học về chủ đề đó.

# 02. Công việc liên quan


Các công nghệ đã tồn tại cho keyword extraction  
có 3 nhóm: textual, graph-based and hybrid models

## 01. Textual



Textual sẽ trích xuất keyword trực tiếp từ văn bản gốc bằng các phương pháp xử lý ngôn ngữ tự nhiên

## 02. Graph-based



Graph-based chuyển tài liệu đến a co-occurrence graph, khi đó các nodes đại diện cho các từ và các cạnh biểu diễn mối liên hệ giữa 2 từ trong một không gian bối cảnh.

## 03. Hybrid



Hybrid thì sẽ là kết hợp của cả 2 phương pháp trên

# 02. Công việc liên quan

## Textual model

- Trong textual model, mục tiêu là tạo ra các từ khóa trực tiếp từ văn bản gốc.
- Một mô hình đơn giản trong loại này sử dụng kỹ thuật TF-IDF để trích xuất các từ khóa.
- Sau đó, các nghiên cứu đã tập trung vào các phương pháp học máy để huấn luyện một bộ phân loại để nắm bắt các từ khóa.
- Với sự ra đời của các phương pháp học sâu như CNNs, LSTM, và các giải pháp hiện đại như Transformers.
- Các phương pháp chính bao gồm KEA, KP-Miner, RAKE, YAKE, TNT-KID, BERT.

# 02. Công việc liên quan

## Textual model

- **KEA (Keyphrase Extraction Algorithm):** sử dụng một số đặc điểm của văn bản như tần suất xuất hiện của các từ và vị trí của chúng để đề xuất các từ khóa.
- **KP-Miner:** tập trung vào việc tìm ra các từ khóa chính trong các văn bản. Nó sử dụng các kỹ thuật phân loại để xác định các từ khóa có ý nghĩa nhất trong văn bản.
- **RAKE:** tập trung vào việc phân tách các cụm từ trong văn bản dựa trên các ký tự phân tách như dấu cách và dấu câu.
- **YAKE:** tập trung vào việc đánh giá sự quan trọng của các từ khóa dựa trên ngữ cảnh của chúng trong văn bản.
- **TNT-KID:** có thể học từ dữ liệu và tự động trích xuất các từ khóa quan trọng từ văn bản.

## 02. Công việc liên quan

### Graph-based model

- Graph-based models sẽ xây dựng a co-occurrence graph từ tài liệu. Nó thể hiện sự tương tác giữa các từ trong một tập hợp tài liệu.
- Trong đồ thị này, các từ được đại diện bởi các nút, và có một cạnh giữa hai từ nếu những từ này xuất hiện cùng nhau trong một cửa sổ ngữ cảnh.
- Sau khi xây dựng đồ thị đồng xuất hiện, một số đo lường trung tâm như degree, closeness, betweenness và vector riêng được áp dụng để tìm từ khóa.
- Các phương pháp sử dụng lý thuyết đồ thị để lựa chọn từ khóa bao gồm TextRank, CollabRank, DegExt, NE-Rank, TopicRank, Positionrank, M-GCKE.

## 02. Công việc liên quan

### Graph-based model

- **TextRank:** biến thể của thuật toán PageRank để xác định sự quan trọng của các từ trong một tập hợp tài liệu dựa trên mối liên kết
- **CollabRank:** CollabRank tập trung vào việc tính toán sự quan trọng của các từ dựa trên sự cộng tác (collaboration) giữa chúng trong tài liệu.
- **DegEx:** xác định các từ khóa bằng cách sử dụng mức độ kết nối của các từ trong đồ thị đồng xuất hiện.
- **NE-Rank:** NE-Rank tập trung vào việc xác định sự quan trọng của các thực thể có tên trong tài liệu.
- **TopicRank:** Xác định các chủ đề quan trọng trong tài liệu và chọn từ khóa tương ứng với các chủ đề đó.
- **PositionRank:** PositionRank xem xét vị trí của các từ trong tài liệu để xác định sự quan trọng của chúng, với giả định rằng các từ khóa xuất hiện ở các vị trí quan trọng trong tài liệu.
- **M-GCKE:** kết hợp đồ thị đồng xuất hiện với ngữ cảnh đa cấp độ để xác định các từ khóa có ý nghĩa trong tài liệu.



## 02. Công việc liên quan

### Hybrid model

- The hybrid models nỗ lực kết hợp 2 loại đã đề cập trước đó.
- Trong bài báo tham khảo, tác giả chia sẻ: “chúng tôi cố gắng phát triển một phương pháp trích xuất từ khóa hiệu quả bằng cách kết hợp các mô hình dựa trên đồ thị và văn bản, và sử dụng các kỹ thuật gán nhãn chuỗi và phân loại.”

# 03. Dữ liệu

## NGUỒN DỮ LIỆU BÀI BÁO

Website: <https://proceedings.neurips.cc/>

### ➤ Mô tả dữ liệu crawl

- Thời gian công bố, tác giả
- Abstract
- Link bài báo
- Kết quả bài báo

NeurIPS Proceedings ➡ ↻

#### Federated Submodel Optimization for Hot and Cold Data Features

Part of [Advances in Neural Information Processing Systems 35 \(NeurIPS 2022\)](#) Main Conference Track

Bibtex

Paper

Supplemental

#### Authors

Yucheng Ding, Chaoyue Niu, Fan Wu, Shaojie Tang, Chengfei Lyu, yanghe feng, Guihai Chen

#### Abstract

We focus on federated learning in practical recommender systems and natural language processing scenarios, while each client's local data tend to interact with part of features, updating only a small submodel. Features normally involve different numbers of clients, generating the differentiation of hot and cold features. Randomly selecting clients to participate and uniformly averaging their submodel updates, will be severely slow. Specifically, the model parameters related to hot (resp., cold) features will be updated quickly (resp., slowly). We use feature-related clients as the metric of feature heat to correct the aggregation of submodel updates. We also work as a suitable diagonal preconditioner. We also rigorously analyze FedSubAvg's convergence rate to evaluation results demonstrate that FedSubAvg significantly outperforms FedAvg and its variants.

NeurIPS Proceedings ➡ ↻

### Advances in Neural Information Processing Systems

Edited by: *S. Koyejo and S. Mohamed and A. Agarwal and D. Belgrave*  
ISBN: 9781713871088

● Main Conference Track ● Datasets and Benchmarks Track

- Federated Submodel Optimization for Hot and Cold Data Features
- On Kernelized Multi-Armed Bandits with Constraints *Xingyao*
- Geometric Order Learning for Rank Estimation *Seon-Ho Lee*
- Structured Recognition for Generative Models with Explainable
- NAS-Bench-Graph: Benchmarking Graph Neural Architectures
- Fast Bayesian Coresets via Subsampling and Quasi-Newton
- What You See is What You Classify: Black Box Attributions
- Adaptive Interest for Emphatic Reinforcement Learning *Mu*
- Scaling & Shifting Your Features: A New Baseline for Efficient

# 03. Dữ liệu

STT	Feature	Description
1	Year	Năm công bố bài báo [1987,2024]
2	Volume	Bắt đầu từ 0 ứng với 1987
3	Pages	Số trang { 1--12 }
4	Status	Main Conference Track   Datasets and Benchmarks Track (Bộ dữ liệu theo dõi hội nghị chính và theo dõi điểm chuẩn)
5	Book Title	Tiêu đề lớn
6	Title	Tiêu đề của bài báo
7	Authors	Danh sách các tác giả
8	Editors	Danh sách các người chỉnh sửa
9	Publishers	Danh sách nhà xuất bản
10	Main Url	Địa chỉ chính của toàn bộ thông tin bài báo
11	Metadata Url	Địa chỉ json chứa toàn bộ thông tin bài báo
12	Paper Url	Địa chỉ pdf bài báo gốc
13	Supplemental Url	Địa chỉ pdf bài bổ sung
14	Reviews Url	Địa chỉ trang web review bài báo
15	MetaReview Url	Địa chỉ meta review
16	AuthorFeedback Url	Địa chỉ pdf bài nhận xét của tác giả (2019)
17	Reviews And Public Comment	Tương tự Địa chỉ trang web review bài báo
18	Abstract	Tóm tắt bài báo (Quan trọng)

# 03. Dữ liệu

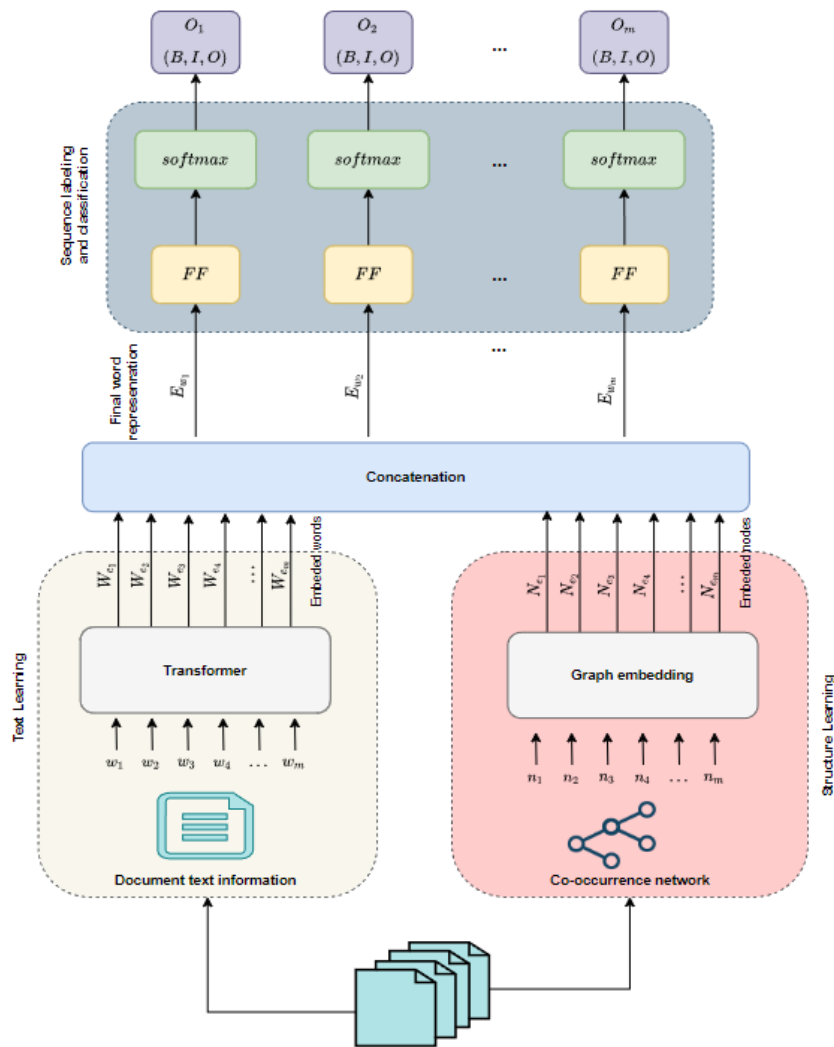
## NGUỒN DỮ LIỆU HUẤN LUYỆN BAN ĐẦU

Dataset	Loại doc	Doc	Keywords (mỗi doc)	Tokens
Inspec	Abstract	2000	29230 (14.62)	128.20
SemEval2010	Abstract	243	4002 (16.47)	149.34
SemEval2017	Paragraph	493	8969 (18.19)	178.22

**Nhận xét tập dữ liệu:** Số lượng dữ liệu của tập Inspec chiếm đa số (2000 rows) có một chút khác biệt về số lượng tokens mỗi doc, tuy nhiên nếu nhìn tổng quan có thể thấy, số lượng keywords mỗi doc cũng có xu hướng tăng, nếu số lượng tokens tăng lên.

# 04. Giải pháp

## BERTGRAPH



## EXEM MODEL

# Abstracts

one aim of the study was to assess the occurrence of a correlation between the body twist and the shooting performance of paratroopers in subsequent shooting series in the Laser Run event. The study involved 25 paratroopers (15 males and 10 females) who were selected on the basis of their performance in the event both the national and international levels. The shooting took place at a laser shooting range. During each shooting series (data from 2 to 20 shots) the shooting performance was recorded in terms of the number of hits and the score obtained by the highest level of designation of Laser Run organized by the United International de Penetration Moderne. The correlation analysis did not show any significant relationship between postural balance and performance in the Laser Run event. The correlation between the scores obtained in the first and second shooting series was positive, but not significant. The lowest score was significantly higher for the second, third and fourth shooting series compared to the first series. In the Laser Run event, the postural balance disturbances caused by the body twist were not related to the shooting performance.

With the initiation of globalization, developing countries, particularly those in Asia, have been witnessing an immense surge of FDI inflows during the past two decades. Even though India has been a latecomer to the FDI scene compared to other East Asian countries, its considerable market potential and a liberalized policy regime has sustained its attraction as a favorable destination for foreign investors. This research paper aims to examine the impact of FDI on the Indian economy, particularly after two decades of economic reforms, and analyzes the challenges to position itself favorably in the global competition for FDI.

*Foreign Direct Investment (FDI) refers to an investment made by a company based in one country in another company based in another country. FDI is often preferred over Foreign Institutional Investments (FII) as it is considered being the most beneficial form of foreign investment for an economy. FDI plays a multidimensional role in the overall development of any economy. It provides a new source for capital, can lead to technological up gradation, skill enhancement and allocates efficiency efforts. While FDI is expected to create positive impact on economic growth, it also brings in certain negative impact on Indian economy during the past few years. The present study is conducted to study the relationship and analyze the impact of FDI on Indian exports.*

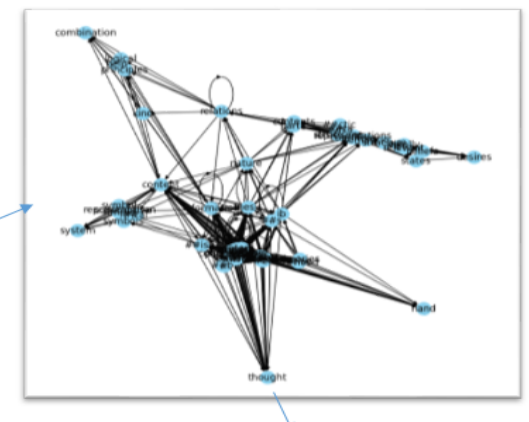
**B1: Xây dựng cấp từ và mối liên hệ**

Word pairs + relations

['conflict', 'thought'], SE  
['language', 'atom'], SE  
['language', 'information'], PN  
['representation', 'system'], PN  
['theories', 'hand'], SE  
['false', 'conflict'], SE  
['false', 'ib'], SE

['conflict', 'thought'], SE  
 ['language', 'atom'], SE  
 ['language', 'information'], PN  
 ['representation', 'system'], PN  
 ['theories', 'hand'], SE  
 ['false', 'conflict'], SE  
 ['false', 'ib'], SE

B2: Xây dựng Co-occurrence network



B3: Tìm tập chủ đạo (dominating set)

```
graph TD; Word2Vec[Word2Vec] --> Combine[Combine Word2Vec + FastText]; FastText[FastText] --> Combine; Combine --> Model[Model]
```

The diagram illustrates a model architecture. It features three blue rounded rectangular boxes stacked vertically. The top box is labeled "Word2Vec", the middle box is labeled "FastText", and the bottom box is labeled "Combine (Word2Vec + FastText)". Arrows from the "Word2Vec" and "FastText" boxes point to the "Combine" box. An arrow from the "Combine" box points to a larger box labeled "Model" at the top of the slide.

Word2Vec

FastText

FastText

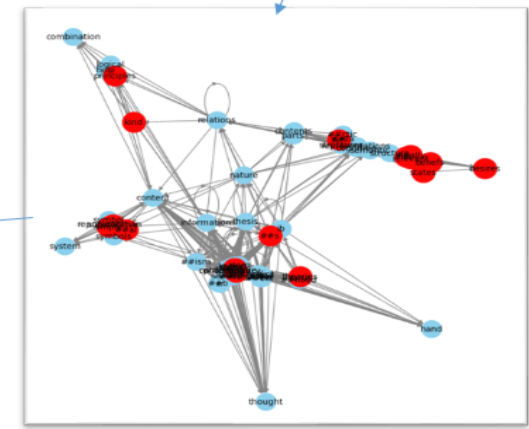
Combine (Word2Vec + FastText)

## B5: Huấn luyện model Skipgram + CBOW

# Random walks

['relevant', 'representations', 'structures', 'structure', 'beliefs']  
 ['syn', 'contents', 'semantic', 'structure', 'desires']  
 ['lot', 'semantic', 'cognitive', 'semantic', '##ally']  
 ['simple', 'content', 'atom', 'content', '##al']  
 ['simple', 'symbols', 'representation', '##al', 'system']  
 ['fred', 'i', '##s', 'mind', 'i']  
 ['lot', 'structure', 'cognitive', 'representations', 'beliefs']  
 ['fred', 'i', '##hood', '##s', 'relations']  
 ['syn', 'structures', 'structure', 'beliefs', 'desires']  
 ['fred', 'thesis', 'symbols', '##al', 'system']

B4: Xây dựng random walks



B3: Tìm tập chủ đạo (dominating set)

# EXEM MODEL

## **Bước 1: Xây dựng cặp từ và mối liên hệ**

**Tiền xử lý dữ liệu và xây dựng quan hệ cụm danh từ:**

- Loại bỏ stopword, viết thường
- Phân tách token

**Giữ lại các cụm danh từ:**

- Chỉ giữ lại các token nằm trong cụm danh từ, thường làm chủ ngữ hoặc tân ngữ trong câu.

**Xây dựng hai quan hệ chính:**

- PN (Phrase Noun)
- SE (Sentence)

**Ví dụ minh họa:**

Câu: "Việt Nam là một quốc gia xinh đẹp."

Quan hệ PN: ['Việt', 'Nam', 'PN'], ['quốc', 'gia', 'PN']

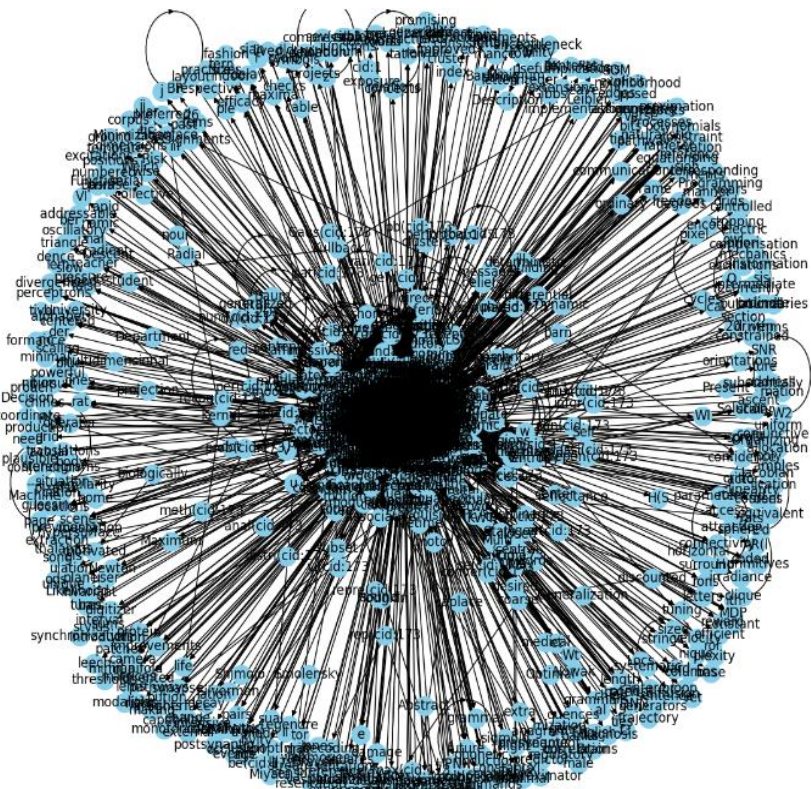
Quan hệ SE: ['Việt', 'quốc', 'SE']

# EXEM MODEL

## Bước 2: Xây dựng co-occurrence network

- Để có thể lưu trữ và thực hiện các bước tính toán một cách dễ dàng chúng tôi tổ chức thông qua một thư viện là NetworkX, đối tượng được khởi tạo với source, target, edge.

Ví dụ: ['Việt', 'Nam', 'PN'] tương ứng chính là 3 giá trị cần truyền cho đối tượng NetworkX.





# EXEM MODEL

## Bước 3: Tìm tập chủ đạo (dominating set)

### Algorithm 1 Finding a dominating set

**Require:** Đồ thị có kết nối  $G = (V, E)$

$D = \emptyset$

**loop**

**if**  $\text{IsEmpty}(V - [D \cup \text{Neighbors}(D)])$  **then**

        STOP

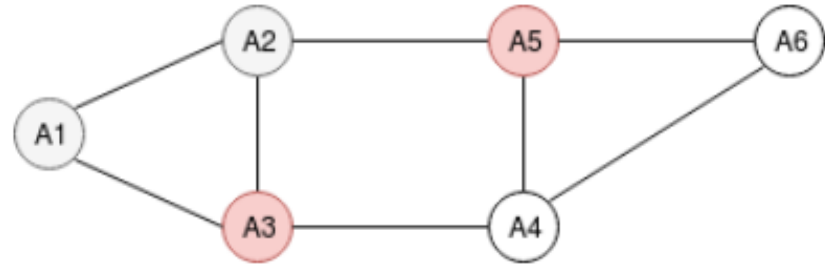
**end if**

    Chọn ngẫu nhiên một vector  $w \in V - [D \cup \text{Neighbors}(D)]$

$D \leftarrow D \cup \{w\}$

**end loop**

**return**  $D$



# EXEM MODEL

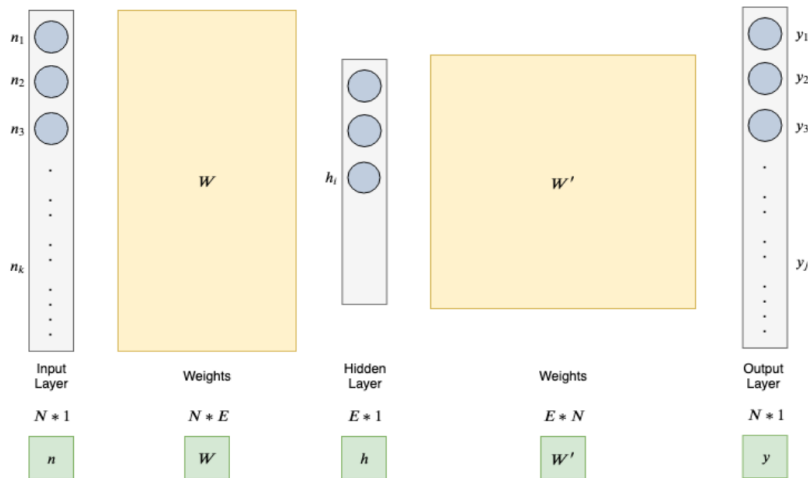
## Bước 4: Xây dựng random walks

- Quy trình:
  - Bắt đầu từ một nút thống trị.
  - bước đi bắt đầu từ nút thống trị và chọn ngẫu nhiên một nút lân cận và thêm vào bước đi.
  - Tiếp tục đến khi có ít nhất một nút thống trị khác trong đường dẫn hoặc đạt độ dài cố định LR.
- Ví dụ: Trong hình dưới, nút đỏ là nút thống trị, độ dài bước đi là 5. Mỗi hể có nhiều hơn hai nút thống trị trong đường dẫn.

```
['relevant', 'representations', 'structures', 'structure', 'beliefs']  
['syn', 'contents', 'semantic', 'structure', 'desires']  
['lot', 'semantic', 'cognitive', 'semantic', '##ally']  
['simple', 'content', 'atom', 'content', '##al']  
['simple', 'symbols', 'representation', '##al', 'system']  
['fred', 'l', '##s', 'mind', 'l']  
['lot', 'structure', 'cognitive', 'representations', 'beliefs']  
['fred', 'f', '##hood', '##s', 'relations']  
['syn', 'structures', 'structure', 'beliefs', 'desires']  
['fred', 'thesis', 'symbols', '##al', 'system']
```

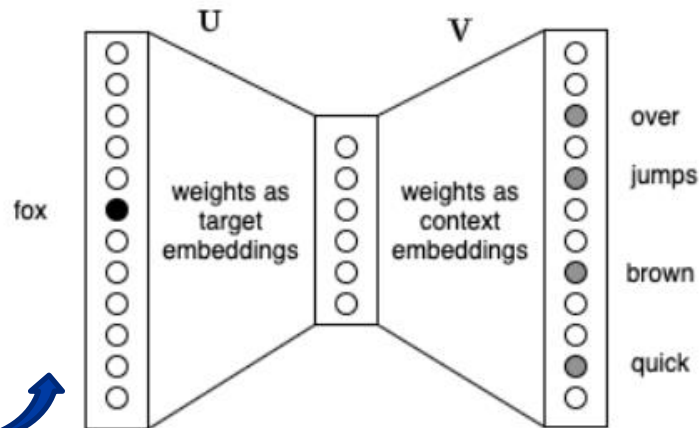
# EXEM MODEL

## Bước 5: Huấn luyện model Skipgram + CBOW

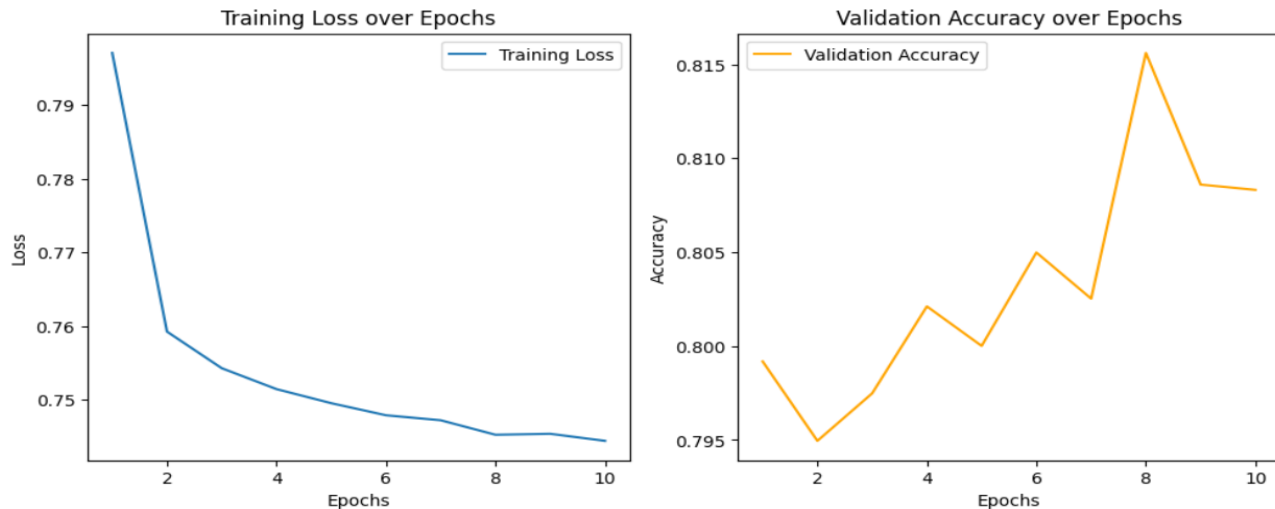


Trong sơ đồ bên, với chuỗi "n1 n2 n3 n4 n5", mô hình skip-gram nhận "n3" làm đầu vào và dự đoán các đầu ra là "n1", "n2", "n4" và "n5"

The quick brown fox jumps over the lazy dog.



# 05. Kết quả Huấn luyện



**Nhận xét:** Quá trình huấn luyện cho thấy tốc độ hội tụ nhanh trong vòng chưa đến 10 epochs, sau đó độ chính xác trên tập validation biến động nhưng kết quả không cải thiện. Còn đồ thị loss của tập train vẫn có xu hướng giảm đều và hiệu quả.

# 05. Kết quả

## F1-score on test

Kiến trúc tổng	Weight của loss	kq (d=200+th=3+ walks=1e4)	kq (d=400+th=3+ walks = 1e4)	kq (d=600+th=3+ walks=1e4)	kq (d=600+th=3+ walks = 1e5)	kq (d=600+th=2+ walks = 1e5)
BERT	[1, 9, 9]	0.3021	x	x	x	x
	[1, 5, 5]	<b>0.3051</b>	x	x	x	x
	[1, 19, 19]	0.2549	x	x	x	x
BERT + WV	[1, 9, 9]	0.2947	0.3224	0.3274	0.3203	0.3290
	[1, 5, 5]	0.2959	0.3185	<b>0.3308</b>	0.3237	0.3251
	[1, 19, 19]	0.2801	0.3163	0.3175	0.3151	0.3223
BERT + FT	[1, 9, 9]	0.2855	0.3239	0.3296	0.3232	0.3291
	[1, 5, 5]	0.2905	0.3168	<b>0.3325</b>	0.3267	0.3256
	[1, 19, 19]	0.2760	0.3021	0.3164	0.3167	0.3220
BERT + WVFT	[1, 9, 9]	0.2922	0.3289	0.3308	0.3273	0.3324
	[1, 5, 5]	0.2975	0.3205	<b>0.3341</b>	0.3297	0.3324
	[1, 19, 19]	0.2778	0.3082	0.3192	0.3221	0.3275

# 06. Kết luận và hướng phát triển

## Kết quả đạt được

- ❖ Đề xuất một phương pháp đa phương thức gọi là BERTGraph kết hợp học tập ngữ cảnh từ văn bản bằng BERT Transformer và học tập đại diện ngữ cảnh từ đồ thị tương tác bằng kỹ thuật nhúng đồ thị.
- ❖ Giải quyết bài toán trích xuất từ khóa như một nhiệm vụ phân đoạn phụ đề, sử dụng mã hóa BIO.
- ❖ Đánh giá hiệu quả của BERTGraph trên kết hợp 3 tập dữ liệu và so sánh với phương pháp dựa trên ngôn ngữ duy nhất (chỉ sử dụng BERT).
- ❖ Kết quả cho thấy BERTGraph vượt trội hơn đáng kể so với các phương pháp dựa trên một phương thức.

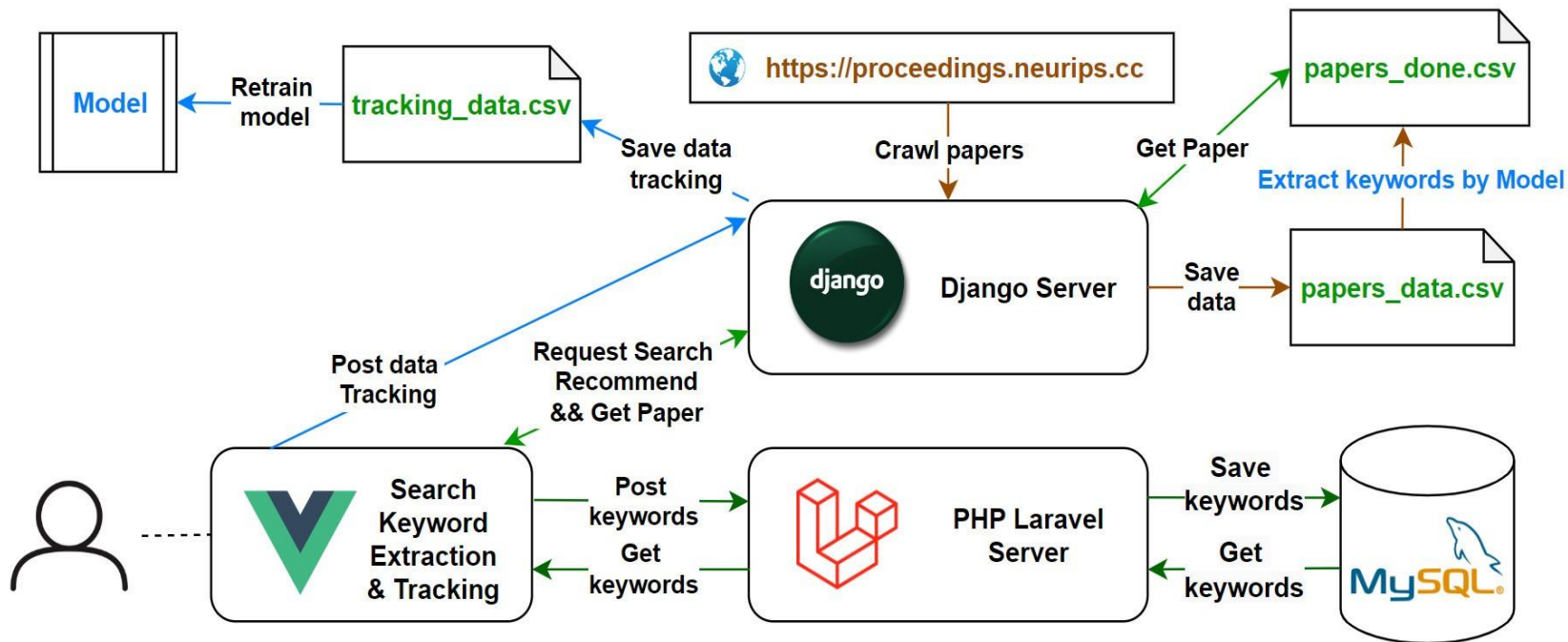
# 06. Kết luận và hướng phát triển

## Hướng phát triển

- ❖ Đổi mới các hướng tiếp cận của Graph embedding thay vì chỉ lọc ra các cụm danh từ, chuyển sang các phương pháp dùng cả động từ, trạng từ,...
- ❖ Phát triển model học ngữ cảnh tốt hơn thay BERT thành Roberta, DistilBert,...
- ❖ Không ngừng mở rộng tập dữ liệu và từ đó phát triển model trên dữ liệu chuẩn.
- ❖ Hướng mới: giải quyết bằng các model multi-task trên bộ dữ liệu lớn (xu hướng hiện nay) và fine-tune trên nhiệm vụ trích xuất keyword

# 07. Ứng dụng

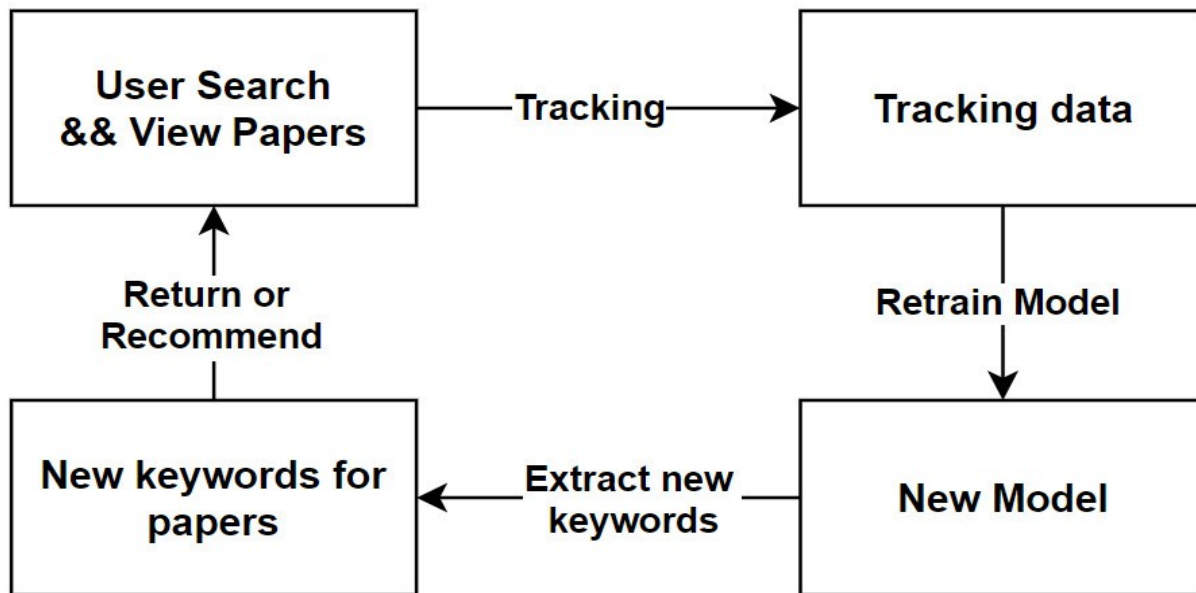
## Tổng quan hệ thống





# 07. Ứng dụng

## Tracking & Crawl



The slide features a light blue background with decorative hexagonal shapes in the corners. The top-left corner has a dark blue hexagon and a light blue one. The top-right corner has a light blue hexagon and a medium blue one. The bottom-left corner has a light grey hexagon and a light blue one. The bottom-right corner has a light blue hexagon and a dark blue one.

**Thanks !**

**Demo and Q&A**