Hệ Khuyến Nghị

Research Paper Recommendation

GVHD:
Thầy Nguyễn Văn Kiệt
Thầy Huỳnh Văn Tín

Nhóm 3:

Tiêu Kim Hảo	*19521480*
Văn Kim Ngân	*19520177*
Nguyễn Thị Bảo Hân	*19520071*

Overview

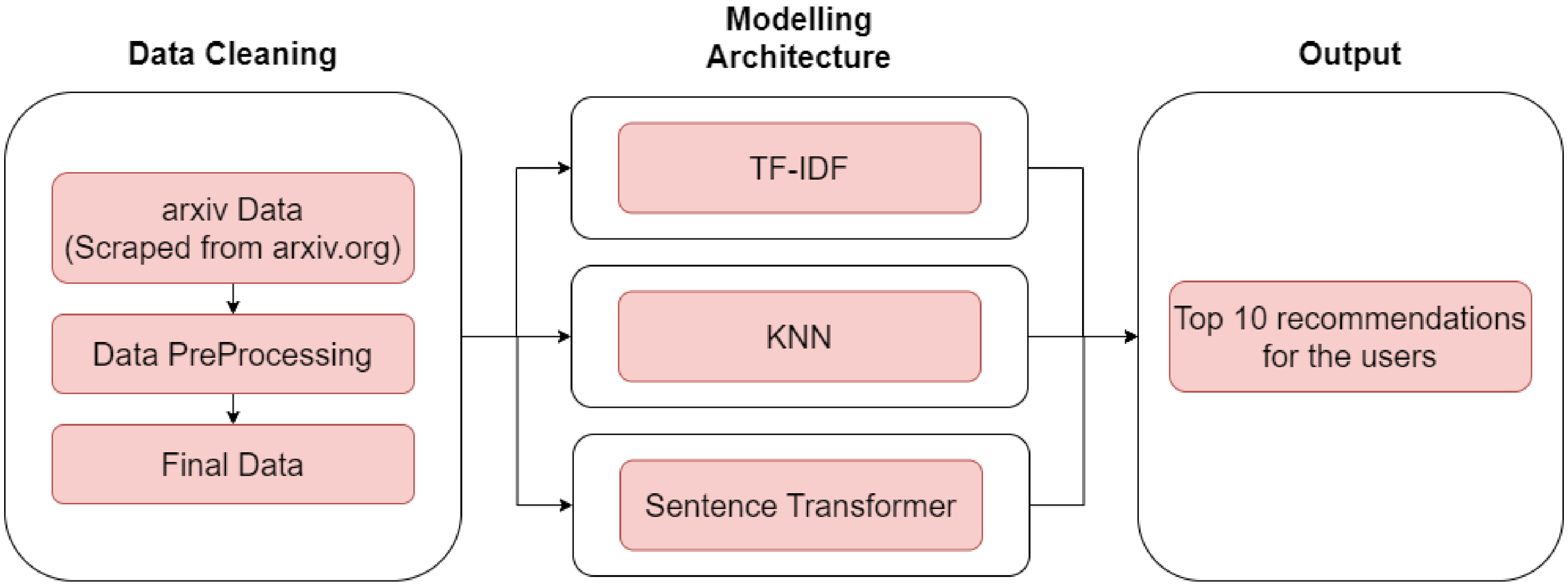

1
Introduction
2
Dataset &
Preprocessing
3
Model
4
Experiment
5
Conclusion

Introduction

Mô tả bài toán:

- Xây dựng hệ thống khuyến nghị những bài báo nghiên cứu khoa học trong lĩnh vực công nghệ.
- Giúp việc tìm kiếm những bài báo liên quan trở nên dễ dàng hơn.

Quy trình


Data Cleaning
arxiv Data
(Scraped from arxiv.org)
Data PreProcessing
Final Data
Modelling
Architecture
TF-IDF
KNN
Sentence Transformer
Output
Top 10 recommendations
for the users

Dataset & Preprocessing

Dataset

- Nhóm thu thập được **10.000** điểm dữ liệu từ **arxiv**.

Bảng mô tả thuộc tính dataset

No.	Attribute name	Data Types	Description
1	Title	string	Tiêu đề của bài báo
2	Date	datetime	Thời điểm bài báo được đăng
3	Id	string	ID của bài báo
4	Summary	string	Tóm tắt nội dung của bài báo
5	URL	string	Đường dẫn URL của bài báo

Dataset


```
{"Title":{"0":"Mapping Tropical Forest Cover and Deforestation
with Planet NICFI Satellite Images and Deep Learning in Mato Gr
osso State (Brazil) from 2015 to 2021"},
"Date":{"0":"2022-11-17 18:59:44+00:00"},
"Id":{"0":"2211.09806v1"},
"Summary":{"0":"Monitoring changes in tree cover for rapid asse
ssment of deforestation is considered the critical component of
 any climate mitigation policy for reducing carbon."},
"URL":{"0":"http://arxiv.org/pdf/2211.09806v1"}}
```

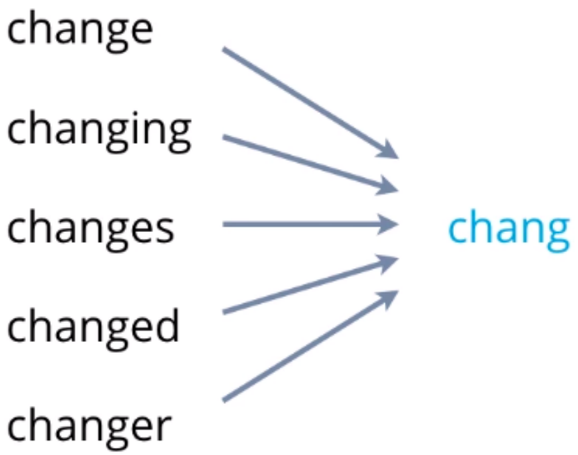
Preprocessing

- Kiểm tra các giá trị **Null**.
- Loại bỏ các **ký tự đặc biệt**.
- Áp dụng **Stemming Algorithm**.

Monitoring changes in tree cover for rapid assessment of deforestation is considered the critical component of any climate mitigation policy for reducing carbon.

monitor chang in tree cover for rapid assess of deforest is consid the critic compon of ani climat mitig polici for reduc carbon.

Stemming



Preprocessing

31,713 unigrams

Stemming algorithm được sử dụng để reduce unigram size còn **23,178 unigrams**

Nhận xét:

- Giảm **~8500 unigrams**.
- Giúp việc tính toán nhanh và cải thiện hiệu suất.

Model

TF-IDF &
Cosine Similarity →

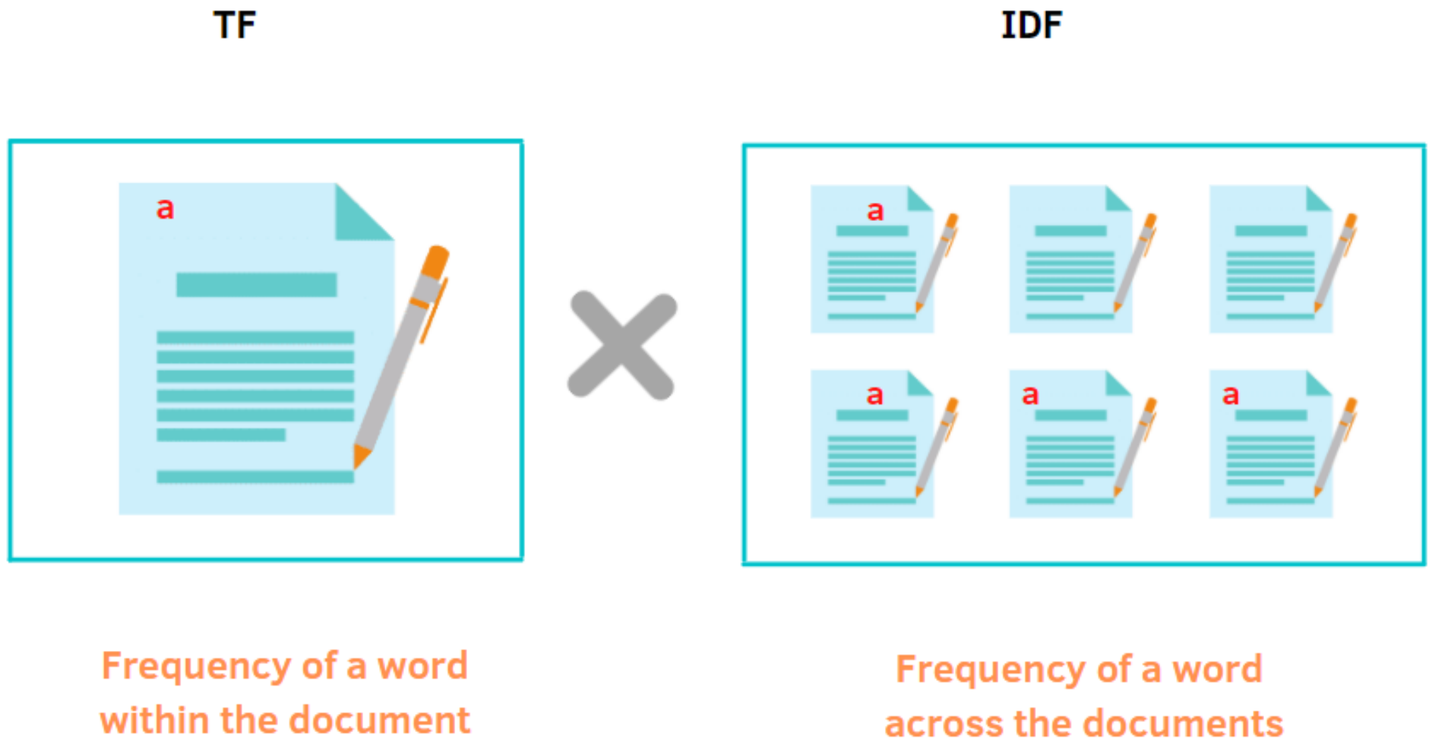
KNN →

Sentence
Transformer →

1.TF-IDF & Cosine Similarity

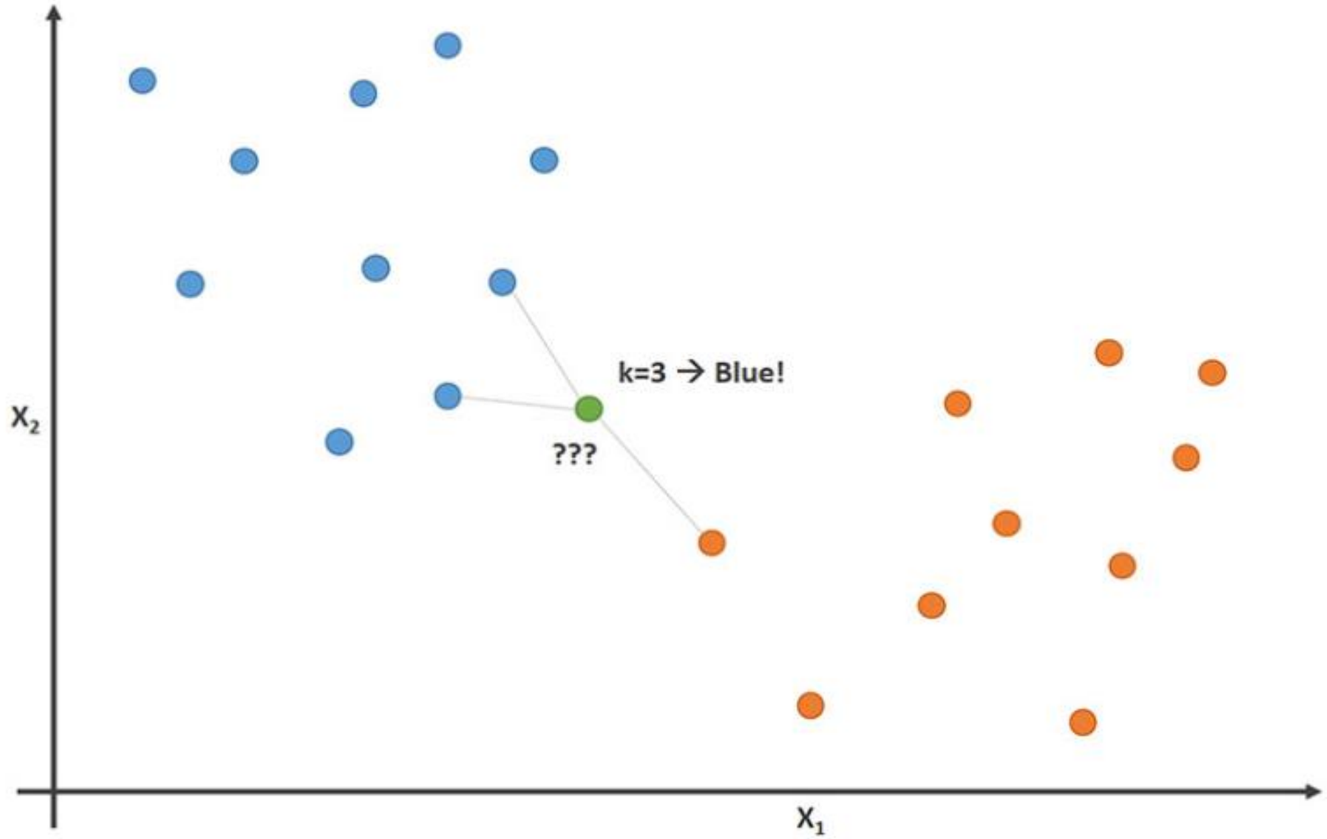
Using Unigram

TF-IDF matrix: (9990, 23178)

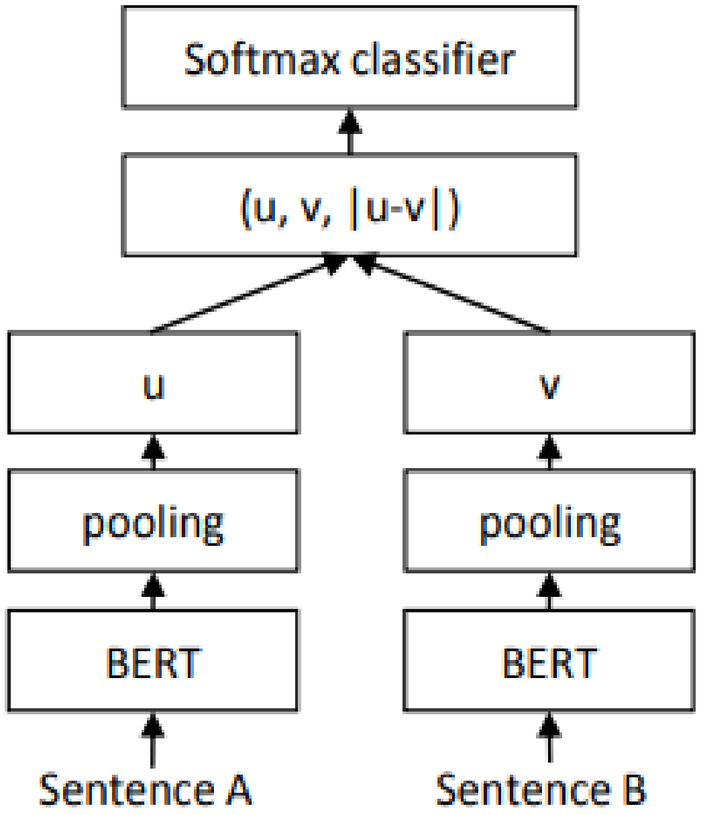


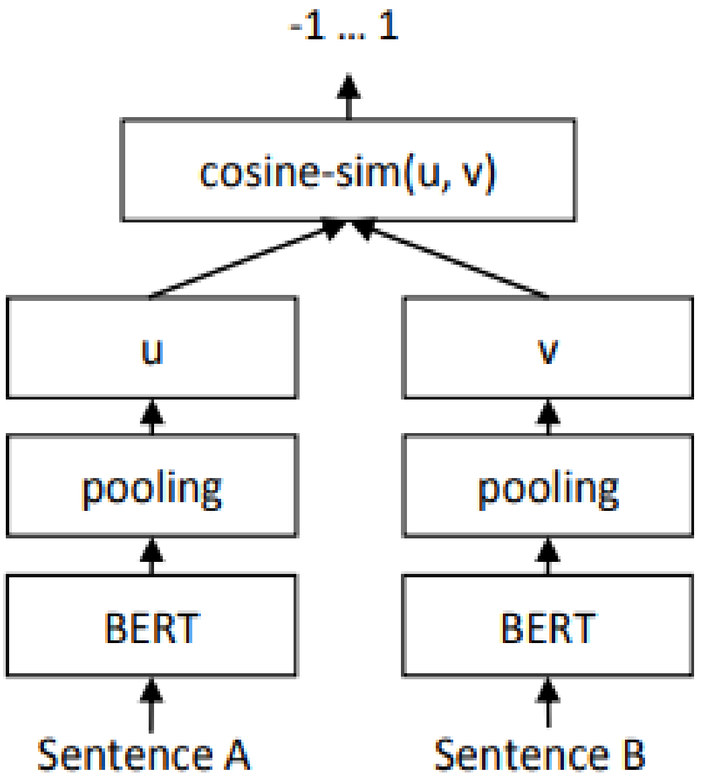
2. KNN

Input matrix: (9990, 23178)



3. Sentence Transformer





Experiment

- Thực nghiệm trên 3 mô hình TF-IDF, KNN, và Sentence Transformer áp dụng phương pháp stemming với n-gram là 1 và độ đo tương đồng cosine similarity.
- Chọn ra 10 paper cho tập test.

Đánh giá

- Cho 10 người đánh giá mức độ liên quan giữa bài báo gốc với những bài báo được khuyến nghị.

0: không liên quan 1: liên quan

Top 10 bài báo được khuyến nghị

A New Kind of Adversarial Example	MORA: Improving Ensemble Robustness Evaluation with Model-Reweighing Attack	0.689346194
	Adaptive Smoothness-weighted Adversarial Training for Multiple Perturbations with Its Stability Analysis	0.673604965
	AccelAT: A Framework for Accelerating the Adversarial Training of Deep Neural Networks through Accuracy Gradient	0.67120713
	Adversarial Coreset Selection for Efficient Robust Training	0.667131126
	Enhancing Targeted Attack Transferability via Diversified Weight Pruning	0.656535685
	ADDMU: Detection of Far-Boundary Adversarial Examples with Data and Model Uncertainty Estimation	0.655549645
	Stateful Detection of Adversarial Reprogramming	0.649802744
	Scaling Adversarial Training to Large Perturbation Bounds	0.644129276
	Approximate better, Attack stronger: Adversarial Example Generation via Asymptotically Gaussian Mixture Distributic	0.643854737
	White-Box Adversarial Policies in Deep Reinforcement Learning	0.635106623

Bảng kết quả độ tương đồng của 3 thuật toán

TF-IDF		
Paper RCM	STEMMING	KHÔNG STEMMING
1	0.336843	0.309826
2	0.378926	0.296853
3	0.472766	0.35188
4	0.296526	0.270421
5	0.268042	0.23245
6	0.349001	0.287358
7	0.34887	0.321103
8	0.257381	0.233424
9	0.323475	0.278197
10	0.29321	0.236436
	0.332504	**0.2817948**

KNN		
Paper RCM	STEMMING	KHÔNG STEMMING
1	0.286149888	0.285572101
2	0.360650346	0.349597842
3	0.28261254	0.28029342
4	0.229613741	0.242760213
5	0.275099055	0.268552846
6	0.44261648	0.423564217
7	0.384431429	0.380502157
8	0.487117583	0.471257252
9	0.269686374	0.256509796
10	0.396515267	0.363013336
	0.34144927	**0.332162318**

Sentence Transformer		
Paper RCM	STEMMING	KHÔNG STEMMING
1	0.689243442	0.691350436
2	0.700947571	0.682801461
3	0.711578	0.706672275
4	0.697959232	0.682360959
5	0.606046963	0.543643534
6	0.713009602	0.658626813
7	0.773307478	0.694353259
8	0.706452495	0.673009169
9	0.720546949	0.696698451
10	0.653846371	0.635615587
	0.69729381	**0.666513194**

TF-IDF		
Paper RCM	STEMMING	KHÔNG STEMMING
1	0.336843	0.309826
2	0.378926	0.296853
3	0.472766	0.35188
4	0.296526	0.270421
5	0.268042	0.23245
6	0.349001	0.287358
7	0.34887	0.321103
8	0.257381	0.233424
9	0.323475	0.278197
10	0.29321	0.236436
	0.332504	**0.2817948**

KNN		
Paper RCM	STEMMING	KHÔNG STEMMING
1	0.286149888	0.285572101
2	0.360650346	0.349597842
3	0.28261254	0.28029342
4	0.229613741	0.242760213
5	0.275099055	0.268552846
6	0.44261648	0.423564217
7	0.384431429	0.380502157
8	0.487117583	0.471257252
9	0.269686374	0.256509796
10	0.396515267	0.363013336
	0.34144927	**0.332162318**

Sentence Transformer		
Paper RCM	STEMMING	KHÔNG STEMMING
1	0.689243442	0.691350436
2	0.700947571	0.682801461
3	0.711578	0.706672275
4	0.697959232	0.682360959
5	0.606046963	0.543643534
6	0.713009602	0.658626813
7	0.773307478	0.694353259
8	0.706452495	0.673009169
9	0.720546949	0.696698451
10	0.653846371	0.635615587
	0.69729381	**0.666513194**

Tỉ lệ bài báo được RCM phù hợp chính xác

KNN	0.34
TF-IDF	0.36
Sentence transformer	0.46

Conclusion

- Xây dựng được 1 mô hình khuyến nghị các bài báo khoa học dựa trên summary.
- Mô hình Sentences Transformer ghi nhận được kết quả tốt nhất (~46% số bài báo được recommend phù hợp) trên cả 3 mô hình.
- Phương pháp Stemming làm giảm số lượng unigram -> Tính toán nhanh và cải thiện hiệu suất.

Hướng Phát triển

- Tăng mức độ chính xác của các paper được khuyến nghị.
- Train được trên số n-gram cao hơn.
- Phát triển được thành một ứng dụng thực tế.

DEMO

Cảm ơn thầy và các bạn đã lắng nghe.