TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỆ KHUYẾN NGHỊ

KHUYẾN NGHỊ TIN TỨC TRÊN VNEXPRESS

- **Sinh viên thực hiện:**
  - Trần Nhật Tiến
  - Nguyễn Tấn Phong
  - Kiên Tiến Đạt
- GVHD : Huỳnh Văn Tín

NỘI DUNG

01

GIỚI THIỆU

Bài toán Khuyến nghị tin tức ?

Internet phổ biến

Thời buổi dịch bệnh

Nắm bắt thông tin

→ Nhu cầu đọc tin tức tăng

- Input : nội dung tóm tắt (abstract) và tiêu đề (title) của một bài báo mà người dùng đọc

- Output : Danh sách top các tiêu đề của bài báo khuyến nghi dành cho người dùng

- ✓ **Nguồn thu thập**: vnexpress.net
- ✓ **Công cụ thu thập**: scrapy crawler
- ✓ **Số dòng** : 6315

02

DỮ LIỆU

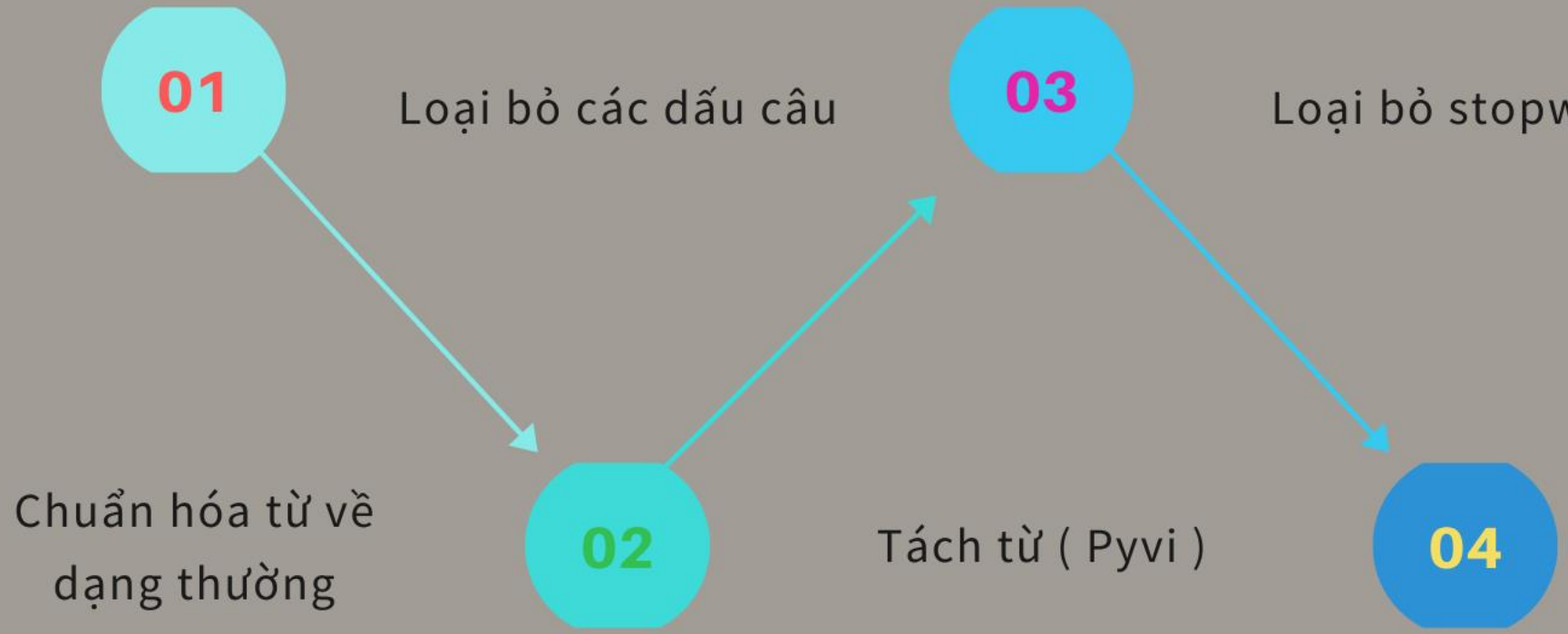
- ❑ **Các chủ đề thu thập :**
  - Giáo dục
  - Đời sống
  - Khoa học
  - Giải trí
  - Thể thao
  - Sức khỏe
  - Kinh doanh
  - Thời sự
  - Số hóa

Thuộc tính	Mô tả
NewsID	Số định danh của bài báo
NewsTitle	Tiêu đề của bài báo
NewsAbstract	Nội dung tóm tắt của bài báo
UserID	Danh sách ID của các user đã bình luận
Topic	Chủ đề của bài báo

NewsID	NewsTitle	NewsAbstract	UserID	Topic
4404506	Người đàn ông 'nghiện' chế tạo máy bay mô hình...	Gần 100 chiếc máy bay mô hình gắn động cơ, tro...	1043446290 1070564564 1058227852 1002629270 10...	Đời sống
4404938	Lời khuyên vàng khi chọn rèm cửa - VnExpress Đ...	Rèm cửa tưởng là thứ rất nhỏ trong tổng thể, n...	1041608906	Đời sống
4404916	Truyện tranh siêu nhân có giá 2,6 triệu USD - ...	Quyển truyện tranh đời đầu về Siêu nhân, xuất ...	1074373109 1005291929 1005634914 1046496997	Đời sống
4403372	Căn hộ chia không gian bằng cánh tủ - VnExpres...	Nhờ hệ tủ đa năng, nữ gia chủ hay đón khách tớ...	1057748718 1061840527 1014687651 1003052895 10...	Đời sống
4404061	10 phong cách thiết kế nhà nổi bật năm 2021 - ...	10 xu hướng về cách bố trí không gian, lựa chọ...	1056461424 1065962404 1071937430 1062696358 10...	Đời sống
...	...	...	...	...
4401427	Cứu cô gái 18 tuổi bị đột quỵ - VnExpress Sức ...	Cô gái 18 tuổi ở Hải Phòng, cảm thấy đau đầu, ...	1074268383 1028835601 1059803421 1049576733 10...	Sức khỏe
4400786	Covid-19 Hà Nội: Test nhanh dương tính, nhiều ...	Người phụ nữ 70 tuổi, ở đường Lê Thanh Nghị, s...	1002942675 1070849195 1005001771 1054575285 10...	Sức khỏe
4400930	TP HCM kêu gọi nhà thuốc tư nhân chống dịch Co...	Sở Y tế TP HCM ngày 9/12 kêu gọi các nhà thuốc...	1012746042 1072119819 1003023118 1002657154 10...	Sức khỏe
4401344	Bệnh xá 30 năm tuổi trên đảo Trường Sa - VnExp...	Từ tổ quân y 3 người trên vùng cát sỏi san hô ...	1002679291 1069731719	Sức khỏe
4400290	Cách tập thở, tăng thể lực cho F0 tại nhà - Vn...	F0 tại nhà có thể thực hiện các bài tập thở, t...	1033920745	Sức khỏe

03

TIỀN XỬ LÝ DỮ LIỆU

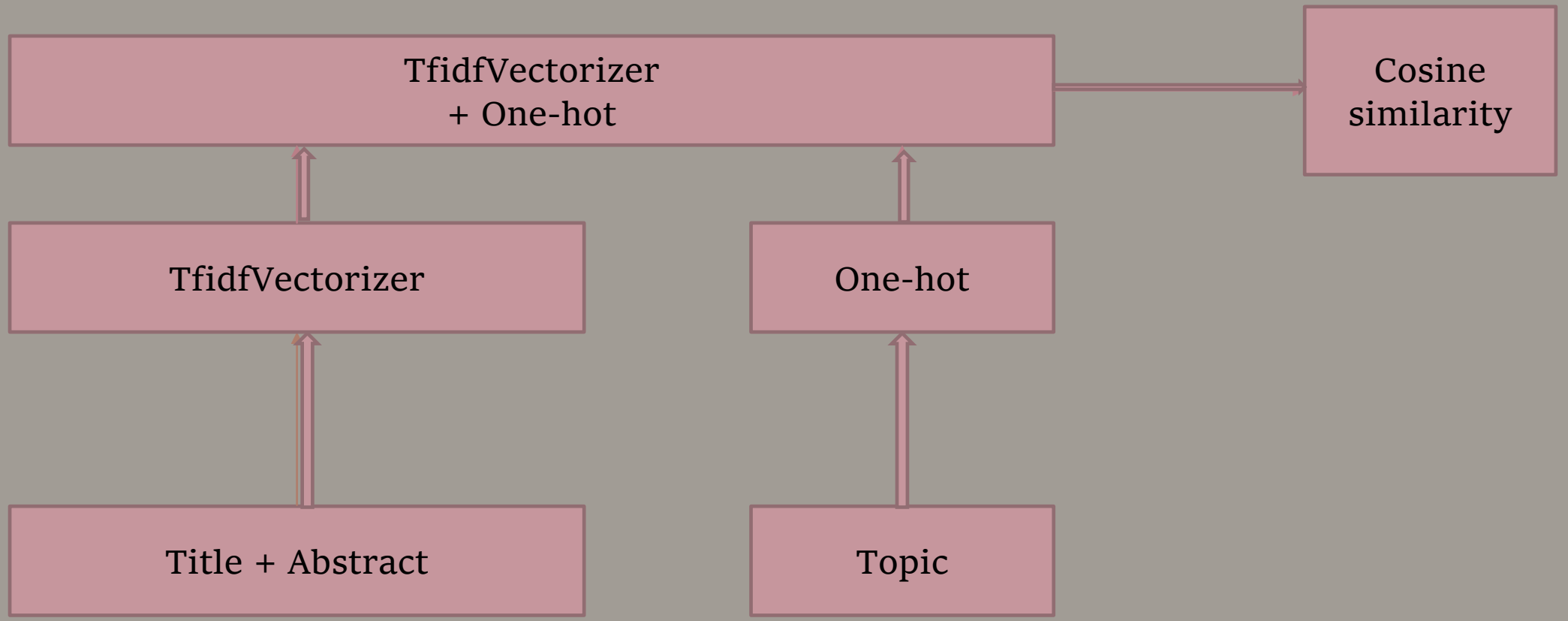


04

PHƯƠNG PHÁP

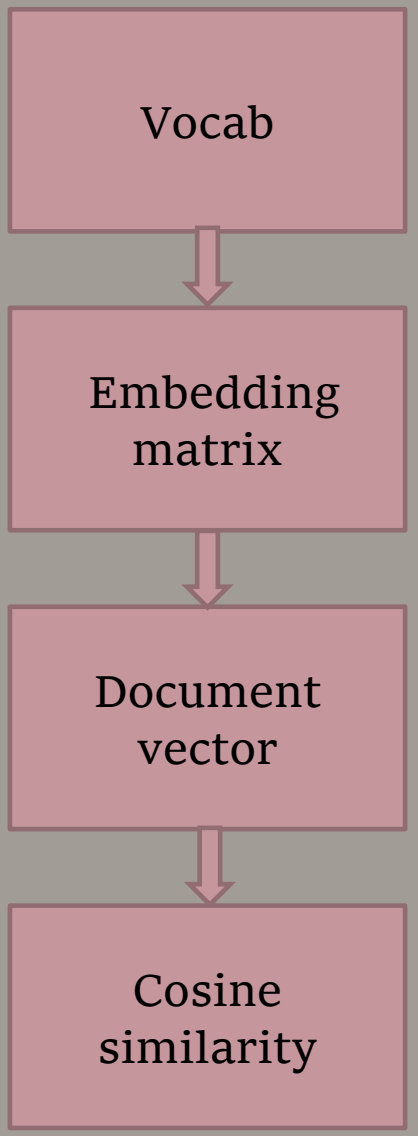
- **Lọc nội dung** (Content-based) : dựa trên nội dung tóm tắt (abstract) và tiêu đề của bài báo (title)
- **Các phương pháp Embedding** :
  - TF-IDF
  - Word2vec
  - FastText

TF-IDF


TfidfVectorizer
+ One-hot
Cosine
similarity
TfidfVectorizer
One-hot
Title + Abstract
Topic

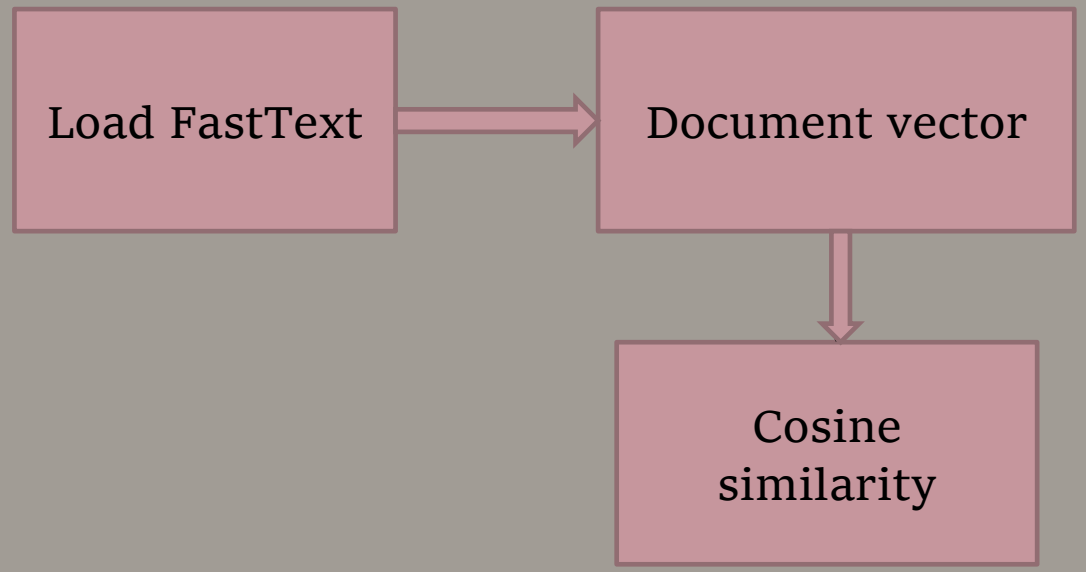
WORD2VEC

- **Word2Vec (Word to vector)** là phương pháp được Google công bố vào năm 2013. Mô hình Pre-trained Word2vec cho tiếng Việt dựa trên dữ liệu các bài báo từ baomoi.com
  - Số chiều : 300
  - Nguồn : GitHub - sonvx/word2vecVN: Pre-trained Word2Vec models for Vietnamese



FASTTEXT

- FastText là thư viện mã nguồn mở do Facebook tạo ra năm 2016, hỗ trợ việc huấn luyện phép nhúng từ và phân loại văn bản



	Precision@10	Recall@10
TF-IDF	46.26	46.87
Word2vec	55.68	46.21
FastText	45.54	46.31

05

ĐÁNH GIÁ

06

KẾT LUẬN

- Thu thập được một bộ dữ liệu cho bài toán khuyến nghị tin tức
- Áp dụng các phương pháp khuyến nghị dựa trên nội dung (Content-based) cho bài toán khuyến nghị tin tức.
- Xây dựng một hệ khuyến nghị đơn giản.

Cảm ơn thầy và các bạn đã lắng nghe !