



Đề tài:

PHÂN LOẠI VĂN BẢN DỰA TRÊN MÔ HÌNH ĐỒ THỊ

Học viên thực hiện: HOÀNG NGỌC DƯƠNG
Cán bộ hướng dẫn: PGS.TS VÕ ĐÌNH BẢY



- 1 **GIỚI THIỆU TỔNG QUAN**
- 2 **BÀI TOÁN PHÂN LOẠI VĂN BẢN**
- 3 **THỰC NGHIỆM, ĐÁNH GIÁ**
- 4 **KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**



- Phân loại văn bản là quá trình gán các văn bản vào một hay nhiều lớp văn bản đã được xác định từ trước.
- Có thể phân loại các văn bản một cách thủ công, tức là đọc nội dung từng văn bản và gán nó vào một loại nào đó.
- Hệ thống quản lý tập gồm nhiều văn bản cho nên cách này sẽ tốn nhiều thời gian, công sức và do đó là không khả thi.



GIỚI THIỆU TỔNG QUAN

UNIVERSITY OF TECHNOLOGY

- Để phân loại có nhiều phương pháp học máy trong trí tuệ nhân tạo như: máy vector hỗ trợ (SVM), Cây quyết định, Naïve Bayes, K láng giềng gần nhất ...
- Đối với những phương trên thường chuyển mô hình văn bản thành dạng véc tơ. **Nên độ lớn, số chiều của véc tơ lớn nên việc xử lý chậm.**
- Luận văn sử dụng mô hình đồ thị, khai thác dữ liệu đồ thị để chuyển văn bản sang dạng véc tơ.



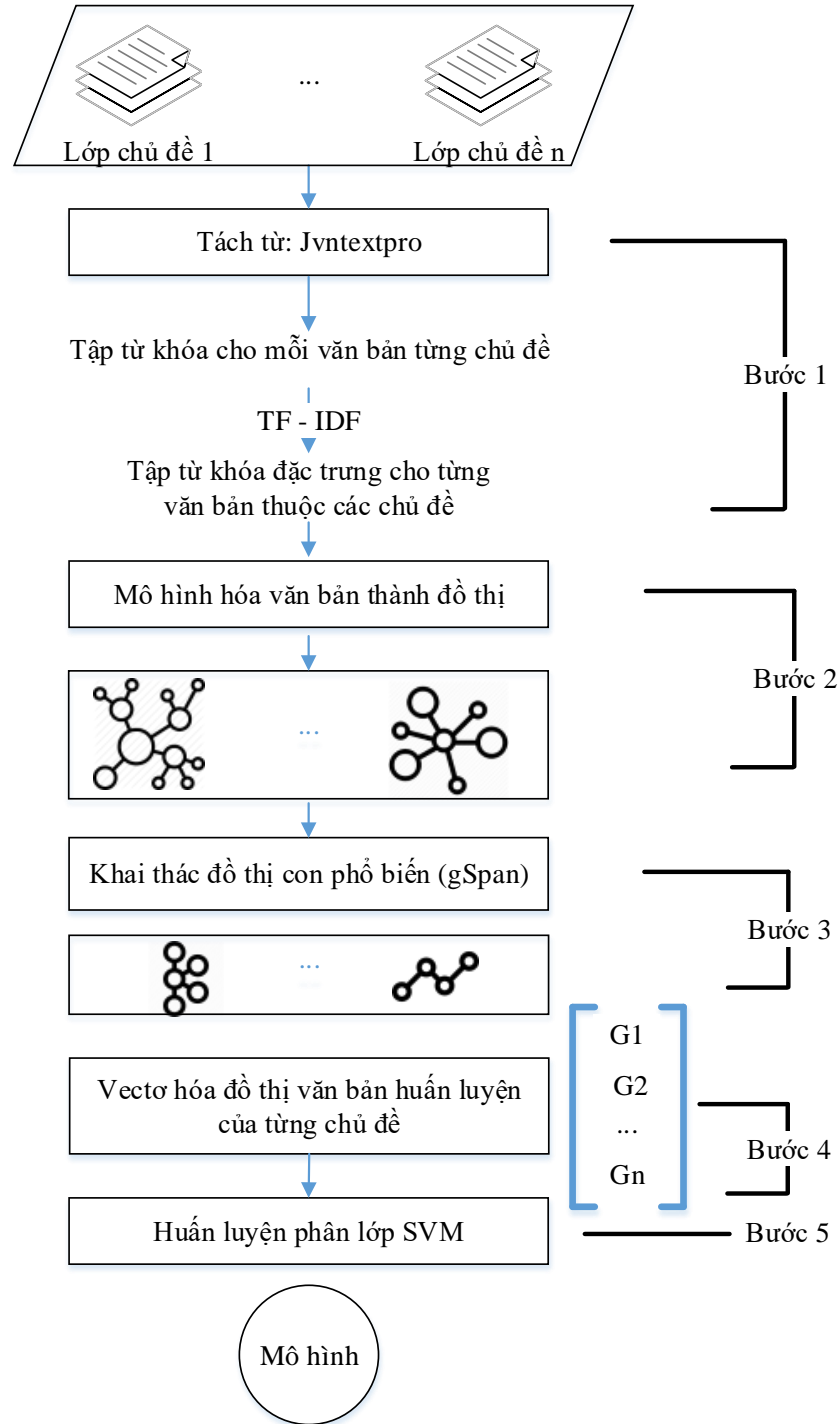
Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Các bước thực hiện:

- Bước 1: Thực hiện việc tách từ và tính TF – IDF
- Bước 2: Mô hình hóa văn bản thành đồ thị
- Bước 3: Khai thác đồ thị con phổ biến bằng thuật toán gSpan
- Bước 4: Vec tơ hóa đồ thị văn bản
- Bước 5: Huấn luyện phân loại bằng SVM



Mô hình huấn luyện phân loại văn bản



Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Bước 1: Thực hiện việc tách từ và tính TF – IDF

- Sử dụng bộ thư viện tách từ Jvntextpro được phát triển bởi tác giả Nguyễn Cẩm Tú.
- Tính tần suất xuất hiện của các từ trong tập văn bản bằng phương pháp TF - IDF.
- Loại bỏ những từ có tần suất xuất hiện thấp thu được bộ từ khóa cho từng văn bản.



```
graph TD;
    A((nghị quyết)) --- B((thông qua));
    B --- C((việt nam));
    B --- D((đảng cộng sản));
    C --- D;
    D --- E((XII));
    E --- F((thứ));
    F --- G((toàn quốc));
    G --- H((đại biểu));
    H --- I((hội));
    I --- J((đại));
```



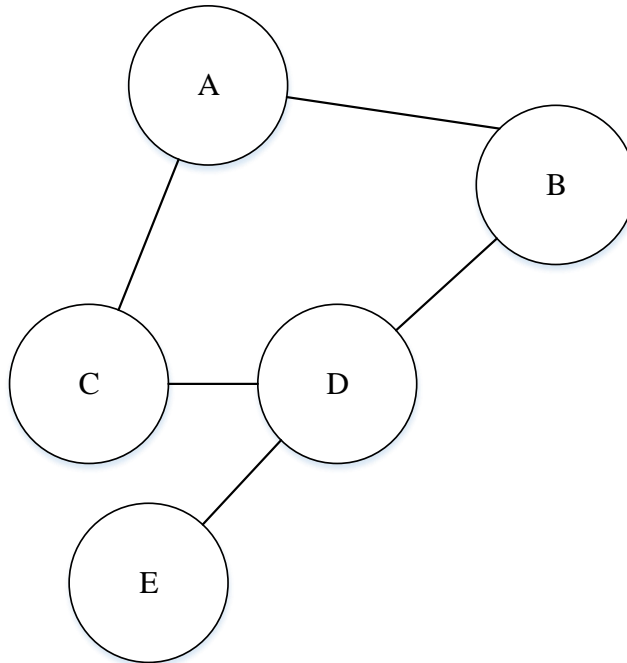
Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Bước 3: Khai thác đồ thị con phổ biến bằng thuật toán gSpan

Trong từng chủ đề, chúng ta tìm tập đồ thị con phổ biến có tần số xuất hiện lớn hơn ngưỡng phổ biến tối thiểu minsup.



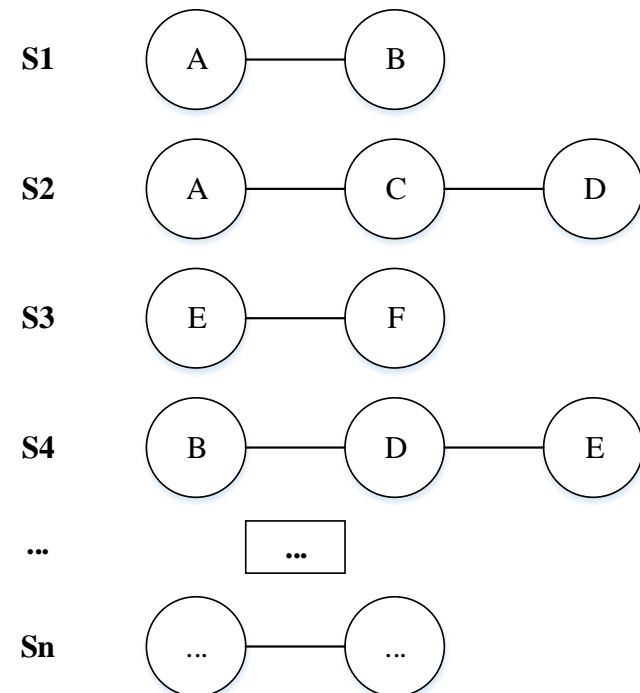
Bước 4: Vec tơ hóa đồ thị văn bản



Đồ thị văn bản G1

Vec tơ đồ thị văn bản G1 thu được
[1:1 2:1 3:0 4:1 ... n:?]

Tập đồ thị con phổ biến $S = \{S1, S2, \dots, Sn\}$

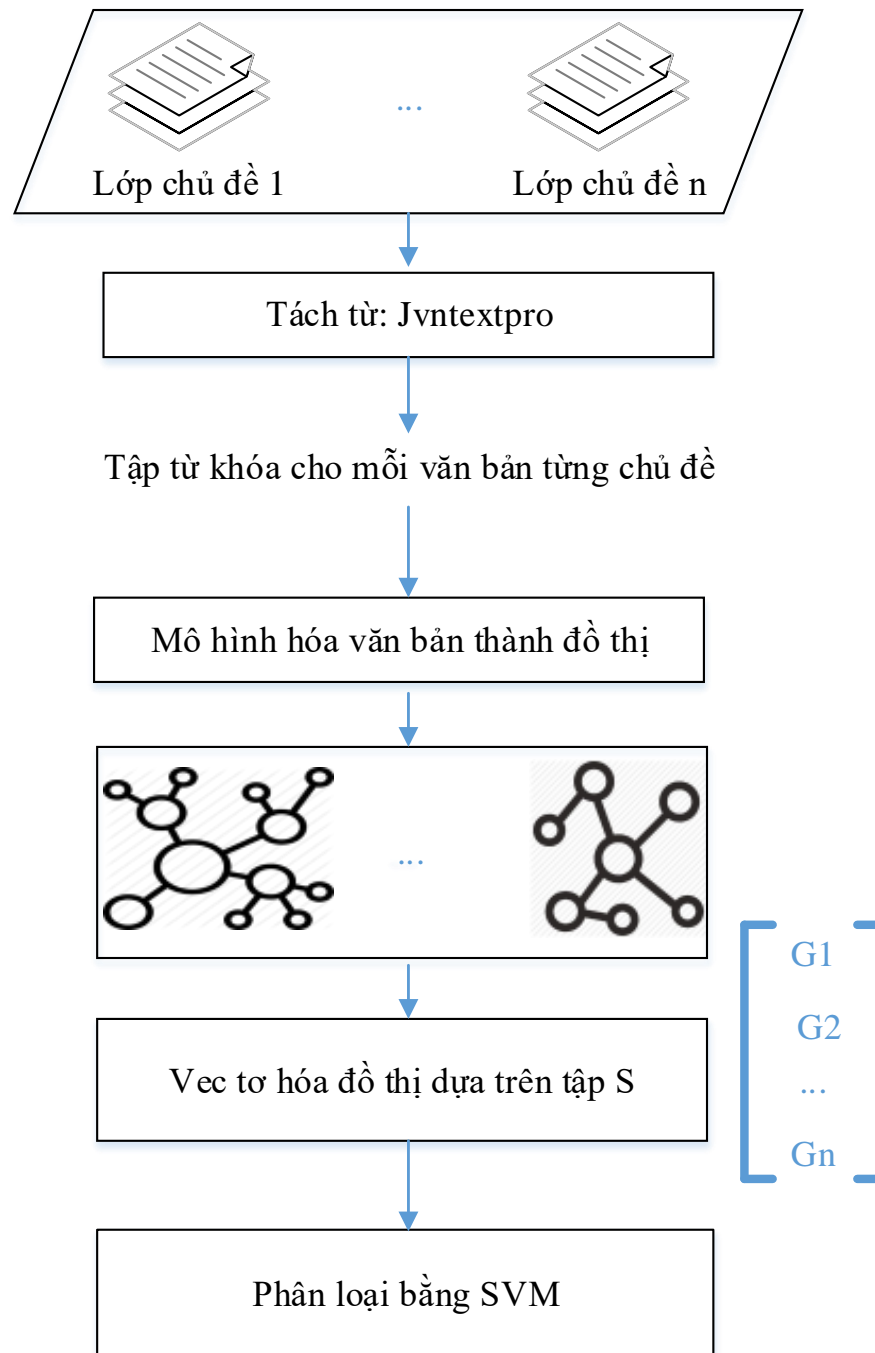


Bước 5: Huấn luyện phân loại bằng SVM

- SVM là một phương pháp học có giám sát để phân lớp dữ liệu.
- Công cụ khá mạnh trong nhiều lĩnh vực như: nhận dạng chữ viết, nhận dạng mặt người, phân loại văn bản, ...
- Ưu điểm của SVM là khả năng phân lớp với độ chính xác cao.



Mô hình phân loại văn bản



Phân loại văn bản dựa trên mô hình đồ thị

- Bước 1: Thực hiện công việc tách từ bằng bộ thư viện Jvntextpro.
- Bước 2: Mô hình hóa văn bản thành đồ thị, vector hóa văn bản bằng cách chiếu lên không gian đặc trưng S (như ở pha huấn luyện) để nhận được các vec tơ đặc trưng tương ứng.
- Bước 3: Đưa qua bộ phân loại SVM đã được huấn luyện để phân loại văn bản.



Hoàng Ngọc Dương (HUTECH, đề tài cao học) - Phần mềm phân lớp văn bản dựa trên Gspan & SVM (v.1.0)

Lựa chọn ngưỡng (minSup) & phương thức (*) Nội dung log chi tiết các tiến trình (*)

10 ☒ Tính TF-IDF

Nội dung log tổng quan (*)



Hoàng Ngọc Dương (HUTECH, đề tài cao học) - Phần mềm phân lớp văn bản dựa trên Gspan & SVM (v.1.0)

Lựa chọn folder chứa văn bản kiểm thử (*)

Tes_Chinh_Tri_Xa_Hoi

Nội dung log tiến trình (*)

Kết quả quá trình phân lớp theo chủ đề (*)

Tên văn bản	Chủ đề (dự đoán)	Tỷ lệ chính xác (%)
-------------	------------------	---------------------



Hoàng Ngọc Dương (HUTECH, đề tài cao học) - Phần mềm phân lớp văn bản dựa trên Gspan & SVM (v.1.0)

Phân lớp văn bản (Graph-based SVM) | Huấn luyện (Gspan, SVM) | Thực nghiệm Gspan (FreqGraph)

Lựa chọn folder chứa văn bản kiểm thử (*)

Tes_Chinh_Tri_Xa_Hoi

Tiến hành phân lớp chủ đề

Nội dung log tiến trình (*)

```

ten-lua-o-bien-dong-358105.txt]
Đọc văn bản -> [Xây dựng nghị định mới về chức năng, nhiệm vụ của Ban Quản lý Lăng.txt]
Đọc văn bản -> [Xúc động ngày đón các liệt sĩ về đất mẹ.txt]
Đọc văn bản -> [Xử lý nghiêm cả người đã hạ cánh toàn_đt.txt]
Đọc văn bản -> [Ông Nguyễn Thiện Nhân làm trưởng đoàn đại biểu Quốc hội TP.HCM_đt.txt]
Đọc văn bản -> [Ông Phí Thái Bình Tôi không vụ lợi_vn.txt]
Đọc văn bản -> [Ông Võ Kim Cự suy sụp khi bị kỷ luật_vn.txt]
Đọc văn bản -> [Điều hiu mua bán đất vùng ven TP. HCM, giá vẫn níu giữ_đt.txt]
Đọc văn bản -> [Đất thiêng vượt sông tời Trường Sa_đt.txt]
Đọc văn bản -> [Nhà thầu lắp không hồ sơ nhận tiền hỗ trợ của dân_đt.txt]
-----
Số lượng đồ thị phân tách -> [200] đồ thị
-----
Số lượng vectore phân tách -> [200] vectore
-----
Nạp SVM's model và tiến hành phân lớp văn bản...
-----
Thời gian hoàn thành: 74319 (ms)
[**]Kết thúc tiến trình phân lớp
  
```

Kết quả quá trình phân lớp theo chủ đề (*)

Tên văn bản	Chủ đề (dự đoán)	Tỷ lệ chính xác (%)
2-000-dai-bieu-du-hoi-nghi-dien-ho...	Chính trị & xã hội	97.38782514521262
491 đại biểu Quốc hội và bài toán c...	Chính trị & xã hội	96.98074962976276
50-cong-chuc-ngoi-boi-chu-hon-la-l...	Chính trị & xã hội	99.90842484026649
9-da-phuong-58-nguoi-nha-lam-qu...	Chính trị & xã hội	96.59190598950042
apec-2017-chuan-bi-can-chu-dao...	Chính trị & xã hội	98.67837358770964
apec-2017-hon-2-000-dai-bieu-tha...	Chính trị & xã hội	98.69533902481467
apec-2017-khai-mac-hoi-nghi-lan-t...	Chính trị & xã hội	90.43601288111918
apec-2017-phat-bieu-cua-chu-tich...	Chính trị & xã hội	98.1174835871019
asean-trung-quoc-ban-va-bien-don...	Chính trị & xã hội	81.27144911272391
& xã hội		92.74807088042016
& xã hội		98.67679112235234
& xã hội		99.94338198709947
& xã hội		98.96773757801078
& xã hội		98.64503312232257
& xã hội		97.4071152228979
& xã hội		98.98907479985084
& xã hội		99.91000095740019
& xã hội		94.99244378391569
& xã hội		99.99343883868916
& xã hội		93.51287305031732
& xã hội		99.43901144886324
& xã hội		98.62417306727733
& xã hội		99.03677228990702
& xã hội		99.15993495498286
& xã hội		97.5352184942659
& xã hội		96.9928736263487
& xã hội		96.9928736263487
& xã hội		75.27048373723237
& xã hội		97.55795215885928
& xã hội		65.21032085680906
& xã hội		60.89598851776089
& xã hội		75.82958255297703
& xã hội		92.2860846417191
& xã hội		81.60841656243589
& xã hội		96.99358318201196
Thể thao		42.119215125689045
& xã hội		75.37241527845435

Tiến trình phân lớp đồ thị đã được hoàn thành!

OK



Hoàng Ngọc Dương (HUTECH, đề tài cao học) - Phần mềm phân lớp văn bản dựa trên Gspan & SVM (v.1.0)

Phần lớp văn bản (Graph-based SVM) | Huấn luyện (Gspan, SVM) | Thử nghiệm Gspan (FreqGraph)

Lựa chọn folder chứa văn bản kiểm thử (*)
 Test_Suc_Khoe_Doi_Song

Tiến hành phân lớp chủ đề

Nội dung log tiến trình (*)

```

Đọc văn bản -> [_Tiêu_calo_voi_7_hoat_dong_hang_ngay.txt]
Đọc văn bản -> [_Viem_tai_giua_thanh_dich_benh_tham_lung_o_trẻ_nhỏ_Tuổi_Trẻ_Online.txt]
Đọc văn bản -> [_Đừng_chủ_quan_khi_người_lớn_bị_sốt_xuất_huyết_Tuổi_Trẻ_Online.txt]
Đọc văn bản -> [_Đi_khám_bệnh_123_lần_trong_4_tháng_đề_truoc_loi_bảo_hiểm_VnExpress_Sức_Khỏe.txt]
Đọc văn bản -> [_Điều_trị_thành_công_cơ_ung_thư_dương_vật_lệ_và_hiếm.txt]
Đọc văn bản -> [_Đầu_gối_phát_tiếng_kêu_Bảo_hiểm_sớm_của_viem_khớp_xương_mấn_tính.txt]
Đọc văn bản -> [_Để_“chuyện_ấy”_an_toàn.txt]
Đọc văn bản -> [_Đổi_phó_với_chúng_biến_ăn_ở_trẻ_nhỏ.txt]
Đọc văn bản -> [_Đột_nhiên_chóng_mặt_dữ_dội_kém_buôn_nôn_có_nguy_hiểm_Tuổi_Trẻ_Online.txt]
Đọc văn bản -> [_“Đẩy_nam_tính”_để_mắc_bệnh_thần_kinh_Tuổi_Trẻ_Online.txt]
-----
Số_lượng_đồ_thị_phân_tách -> [200]_đồ_thị
-----
Số_lượng_vectors_phân_tách -> [200]_vectors
-----
Nạp_SVM's_model_và_tiến_hành_phân_lớp_văn_bản...
-----
Thời_gian_hoàn_thành: 65209 (ms)
[**]Kết_thúc_tiến_trình_phân_lớp
  
```

Kết quả quá trình phân lớp theo chủ đề (*)

Tên văn bản	Chủ đề (dự đoán)	Tỷ lệ chính xác (%)
1.000 bác sĩ dự hội nghị sản phụ kh...	Sức khỏe	55.22040374080978
10 lợi ích tuyệt vời của _chuyện_ấy...	Sức khỏe	63.3621934873147
20-tuo-i-da-bi-ma-u-nhie-m-mo-28...	Sức khỏe	47.963382989079236
3 lưu ý triệu chứng, nguy cơ và ph...	Sức khỏe	47.06414519490959
3 sai lầm thường gặp của người về...	Sức khỏe	56.1279476058172
3-bi-quyet-len-dinh-de-dang-27653...	Sức khỏe	52.54835904431315
3-bi-quyet-tang-suc-de-khang-cho...	Sức khỏe	52.58706182184611
3-luu-y-vang-qup-bao-ve-tim-mac...	Sức khỏe	51.618980433141125
3-luu-chi-chon-can-pham-cham-so...	Sức khỏe	55.55253038385065
		51.02855342994209
		56.970790828646976
		53.628643274219314
		63.16651805927157
		64.20389147749245
		48.77197704266158
		52.35528534214021
		55.4688479030431
		64.65603106096907
		57.030229042664935
		49.13952446106172
		51.07415144576301
		59.05273537570701
		46.32659420634449
		63.393718819376296
		54.727316098355615
		53.31011878066418
		43.785410051767506
		62.85205863811674
		57.80922958140182
		63.272063548964894
		62.46328783545748
		60.72394795038834
		46.97786652573968
		51.07415144576367
		52.191057816133835
		61.996499213095845
		48.714410189276484

Tiến trình phân lớp dữ liệu đã được hoàn thành!



Hoàng Ngọc Dương (HUTECH, đề tài cao học) - Phần mềm phân lớp văn bản dựa trên Gspan & SVM (v.1.0)

Phân lớp văn bản (Graph-based SVM) | Huấn luyện (Gspan, SVM) | Thực nghiệm Gspan (FreqGraph)

Lựa chọn folder chứa văn bản kiểm thử (*)

Tes_Hoat_Dong_The_Thao

Tiến hành phân lớp chủ đề

Nội dung log tiến trình (*)

```

Đọc văn bản -> [Valverde tiến gần hơn đến ghế HLV
Barca - VnExpress Thể Thao.txt]
Đọc văn bản -> [vff-cam-hlv-hoang-anh-tuan-lam-bong-
da-376075.txt]
Đọc văn bản -> [vtv-binh-dien-long-an-bung-no-gianh-
ve-vao-chung-ket-370433.txt]
Đọc văn bản -> [vtv-binh-dien-long-an-ra-quan-da-
thang-368162.txt]
Đọc văn bản -> [vtv-binh-dien-long-an-vuot-kho-
chung-ket-370335.txt]
Đọc văn bản -> [vtv-binh-dien-long-an-xuat-sac-1-
ve-vao-ban-ket-369885.txt]
Đọc văn bản -> [xuan-truong-ve-hoi-quan-dtvn-tha-
nay-vi-xe-lam-doi-375471.txt]
Đọc văn bản -> [Đánh vượt chuẩn bốn gậy ở hố 16,
Spieth bị cắt loại tại AT&T Byron Nelson -
VnExpress Thể Thao.txt]
Đọc văn bản -> [Đại diện Thái Lan thua đau tại vòng
knock-out AFC Champions League.txt]
-----
Số lượng đồ thị phân tách -> [200] đồ thị
-----
Số lượng vectors phân tách -> [200] vectors
-----
Nạp SVM's model và tiến hành phân lớp văn bản...
-----
Thời gian hoàn thành: 62681 (ms)
[**]Kết thúc tiến trình phân lớp

```

Kết quả quá trình phân lớp theo chủ đề (*)

Tên văn bản	Chủ đề (dự đoán)	Tỷ lệ chính xác (%)
120-tay-golf-du-giai-vo-dich-trung-...	Thể thao	50.70015742361997
Allardye chia tay Crystal Palace - V...	Thể thao	52.236085856641566
Antoine Griezmann xác nhận khả n...	Thể thao	50.06868290202425
ban-gai-bung-to-bat-thuong-ronald...	Thể thao	53.19340681318414
bau-duc-da-xoay-hiv-hoang-anh-t...	Thể thao	49.39385752048561
bo-de-gea-va-smalling-cuc-nong-b...	Thể thao	50.39197571109339
bo-roi-bo-xinh-dele-all-to-chuc-sinh...	Sức khỏe	39.36973303879485
cay-bau-duc-hiv-hoang-anh-tuan-n...	Thể thao	49.99180674688303
cham-diem-noid-1-kienhuc-moi...	Thể thao	56.33478094115966
		55.77464884581633
		59.13744080396132
		53.39379673116408
		49.852733112699795
		63.365510875316424
		53.39379673116406
		52.630315490724676
		52.630315490724676
		49.15542776376372
		50.65701979880958
cua-ro-0-dong-xe-ao-vang-gai-xe...	Chính trị & xã hội	75.39527479112081
cuoc-chien-noi-bo-gianh-ve-vao-ba...	Thể thao	50.249222345027064
cuu-vo-dich-thai-lan-bangkok-glass...	Thể thao	50.3617568338793
dai-chien-chung-ket-cl-real-va-ron...	Thể thao	57.37291703876689
david-luz-choi-troi-tang-qua-cuc-d...	Thể thao	59.5553226121435
diem-10-chat-luong-cho-gai-bong...	Thể thao	41.40757001393548
Djokovic hủy diệt Thiem, vào chun...	Thể thao	50.11231292658539
Djokovic ở tuổi 30... không thể chậ...	Thể thao	52.2195063961146
dtvn-huu-thang-choi-nuoc-co-nhu...	Thể thao	48.95196125780545
dtvn-quan-u20-viet-nam-o-at-len-t...	Thể thao	53.54135755068293
eriksen-lap-sieu-phan-tottenham-b...	Thể thao	53.39364737132806
gai-bong-chuyen-nu-quoc-te-cup...	Chính trị & xã hội	75.06027725910218
gai-bong-chuyen-vtv9-binh-dien-t...	Chính trị & xã hội	43.8469451518723085
gai-hang-nhat-len-song-san-phu-h...	Thể thao	55.86883023497725
Giai cổ vua quốc tế HDBank 2017...	Thể thao	45.922594898455955
Giai cổ vua quốc tế HDBank thắng t...	Thể thao	51.994188913733964
Golfer Hàn Quốc lập nhiều kỷ lục khi...	Thể thao	41.40757001393535
he-lo-ban-gai-hoc-lua-cua-sao-leics...	Thể thao	58.81189930848686

Tiến trình phân lớp dữ liệu đã được hoàn thành!

OK



THU THẬP DỮ LIỆU

Chúng tôi tiến hành thu thập một số lượng lớn bài báo thuộc ba chủ đề: Chính trị - xã hội, Sức khỏe, Thể thao từ các nguồn tin tức điện tử:

<http://vnexpress.net/>,

<http://dantri.com.vn/>,

<http://tuoitre.vn/>



1. Thực nghiệm giảm số lượng đồ thị con phổ biến thông qua TF - IDF

Chủ đề	Số văn bản	Số lượng đồ thị phổ biến (FreqGraph)								
		minSup = 20%			minSup = 30%			minSup = 40%		
		Không TF-IDF	Có TF-IDF	%	Không TF-IDF	Có TF-IDF	%	Không TF-IDF	Có TF-IDF	%
Chính trị - xã hội	200	281	68	24.20	214	47	21.96	172	37	21.51
	300	402	94	23.38	294	75	25.51	235	59	25.11
	400	613	140	22.84	476	96	20.17	380	68	17.89
Sức khỏe	200	187	58	31.02	134	46	34.33	107	36	33.64
	300	354	97	27.40	281	77	27.40	225	58	25.78
	400	567	113	19.93	432	102	23.61	346	97	28.03
Thể thao	200	234	79	33.76	192	63	32.81	157	48	30.57
	300	456	85	18.64	378	68	17.99	297	52	17.51
	400	546	122	22.34	436	98	22.48	348	72	20.69



2. Thực nghiệm mức độ chính xác của phân lớp

Dữ liệu huấn luyện:

Tên chủ đề	Số lượng văn bản
Chính trị - xã hội	300
Sức khỏe	300
Thể thao	300



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

Tên chủ đề	Độ chính xác (Precision)	Độ phủ (Recall)	Độ đo F1 (F-measure)
Chính trị - xã hội	0.834	0.855	0.844
Sức khỏe	0.917	0.825	0.868
Thể thao	0.809	0.87	0.839
Trung bình	0.853	0.850	0.850



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

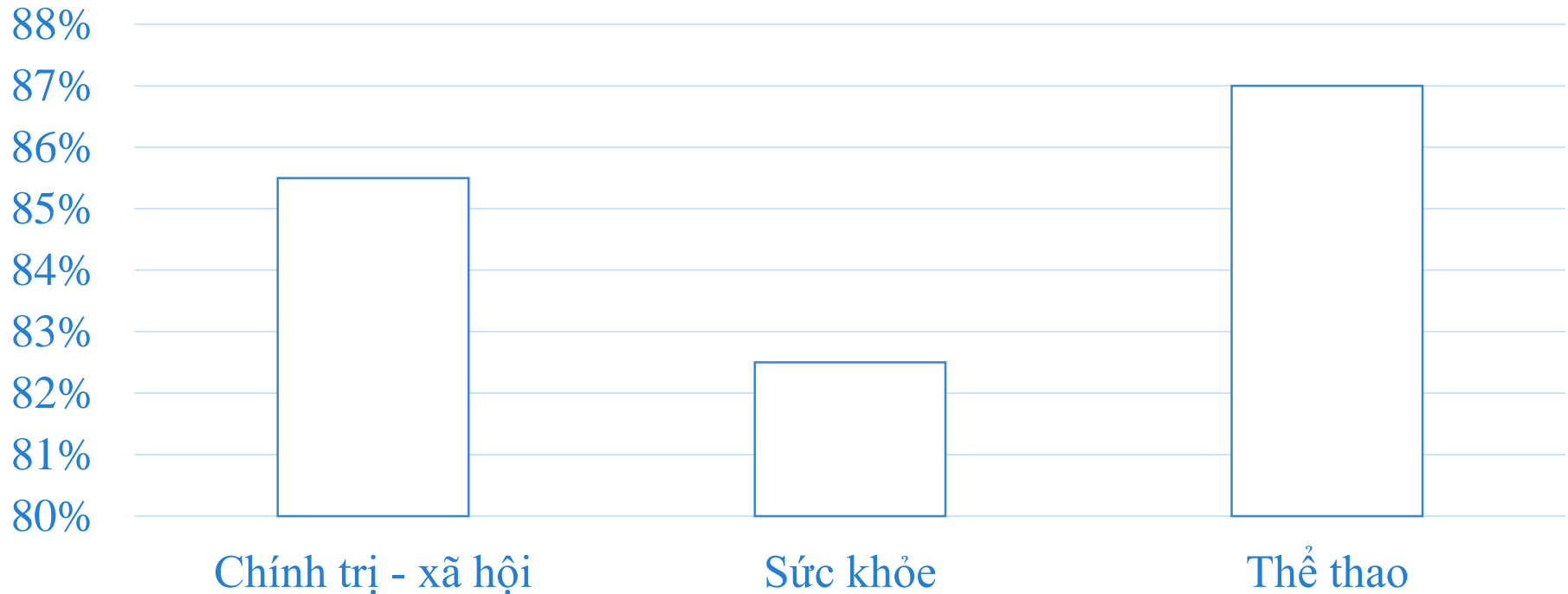
Tên chủ đề	Chính trị - xã hội	Sức khỏe	Thể thao
Chính trị - xã hội	171	8	21
Sức khỏe	15	165	20
Thể thao	19	7	174



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

Tỷ lệ (%) độ chính xác của mô hình phân loại trên số lượng văn bản kiểm thử



□ Tập dữ liệu huấn luyện: 300 văn bản, số tài liệu kiểm tra: 200 văn bản



Thực nghiệm mức độ chính xác của phân lớp khi tăng dữ liệu huấn luyện:

Tên chủ đề	Số lượng văn bản
Chính trị - xã hội	500
Sức khỏe	500
Thể thao	500



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

Tên chủ đề	Độ chính xác (Precision)	Độ phủ (Recall)	Độ đo F1 (F-measure)
Chính trị - xã hội	0.87	0.9	0.885
Sức khỏe	0.955	0.845	0.897
Thể thao	0.843	0.91	0.875
Trung bình	0.889	0.885	0.886



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

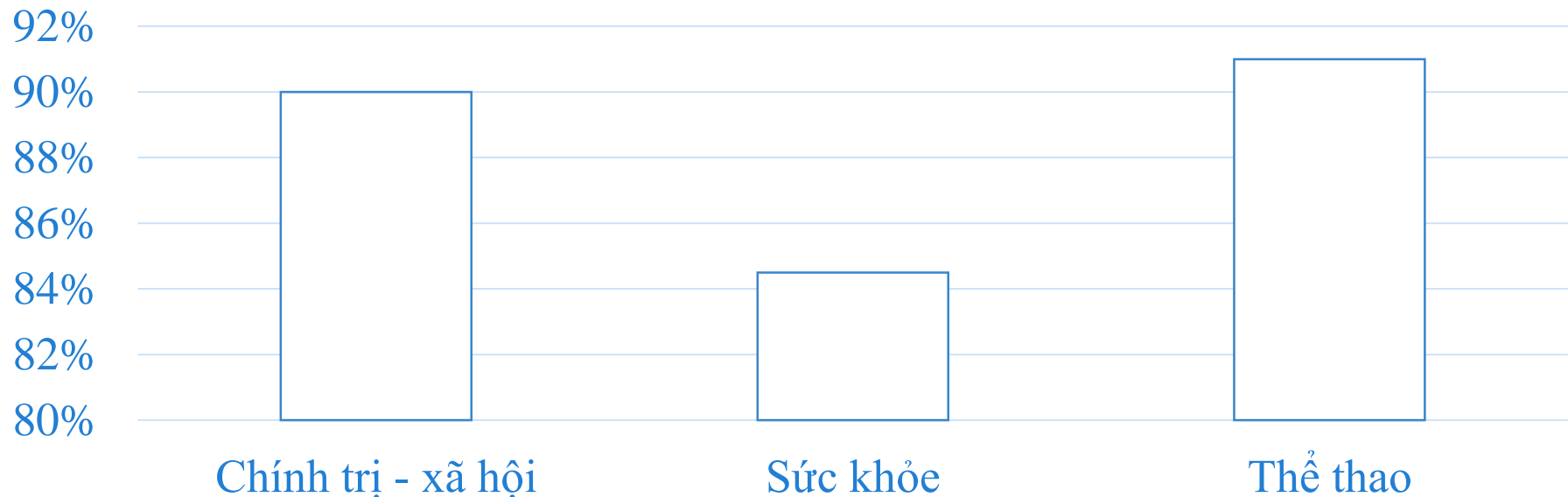
Tên chủ đề	Chính trị - xã hội	Sức khỏe	Thể thao
Chính trị - xã hội	180	5	15
Sức khỏe	12	169	19
Thể thao	15	3	182



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

Tỷ lệ (%) độ chính xác của mô hình phân loại trên số lượng văn bản kiểm thử



□ Tập dữ liệu huấn luyện: 500 văn bản, số tài liệu kiểm tra: 200 văn bản



Thực nghiệm phân lớp khi gộp chung các văn bản thuộc ba chủ đề:

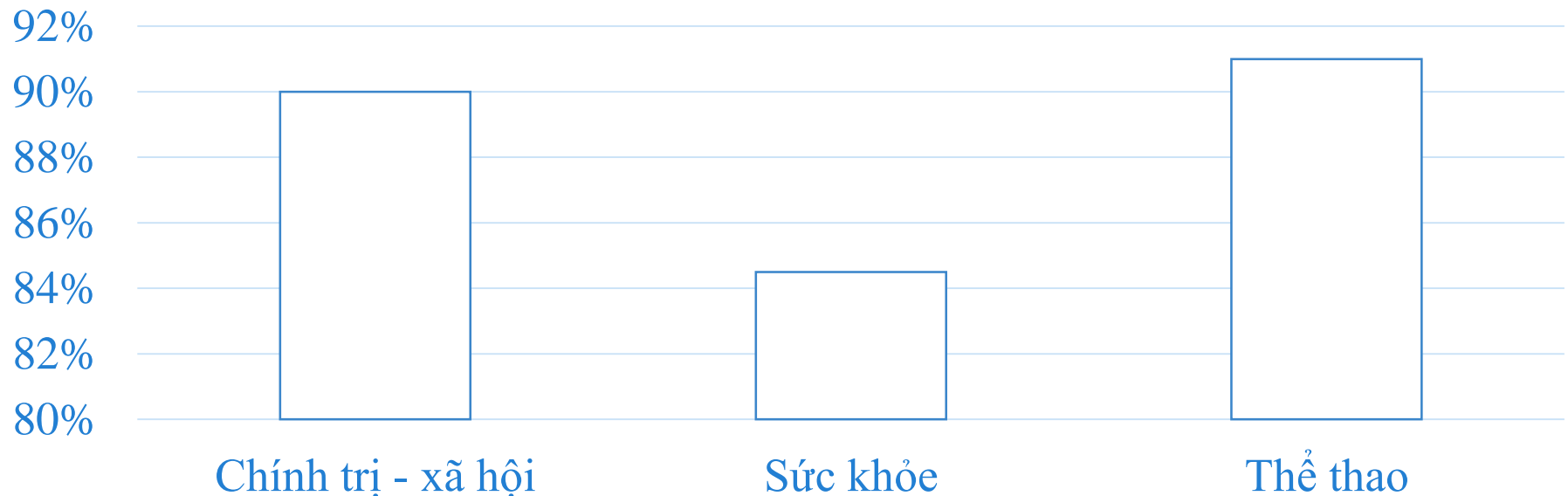
Tên chủ đề	Chính trị - xã hội	Sức khỏe	Thể thao	Số lượng
Chính trị - xã hội	180	5	15	600
Sức khỏe	12	169	19	
Thể thao	15	3	182	



KẾT QUẢ THỰC NGHIỆM

UNIVERSITY OF TECHNOLOGY

Tỷ lệ (%) độ chính xác của mô hình phân loại trên số lượng văn bản kiểm thử



□ Tập dữ liệu huấn luyện: 500 văn bản, số tài liệu kiểm tra: 200 văn bản



Đóng góp của luận văn

- Tính TF - IDF nhằm loại bỏ các hư từ trong tiếng Việt, bằng cách loại bỏ những từ có trọng số thấp hơn ngưỡng trung bình. Làm giảm số lượng đồ thị con phổ biến nên số chiều của vec tơ văn bản cũng giảm theo. Từ đó việc huấn luyện và phân lớp sẽ nhanh và chính xác hơn.
- Đóng góp một hướng tiếp cận mới cho bài toán phân loại văn bản tiếng Việt.
- Xây dựng mô hình phân loại văn bản dựa trên mô hình đồ thị có thể ứng dụng trong thực tế.



HẠN CHẾ & KHÓ KHĂN

- Chưa xây dựng được nhiều bộ dữ liệu dùng để huấn luyện và đánh giá.
- So sánh mô hình xây dựng với các mô hình khác như: phân loại văn bản dựa vào mô hình không gian vector truyền thống, sử dụng cây quyết định, ...



- Làm giàu thêm bộ dữ liệu huấn luyện để nâng cao độ chính xác trong phân lớp văn bản.
- Xây dựng và đánh giá mô hình với nhiều bộ dữ liệu thực tế hơn nữa.
- Đưa mô hình đồ thị có hướng vào trong bài toán xây dựng đồ thị văn bản.
- Áp dụng mô hình đề xuất cho các bài toán thực tế khác như: khai phá các mạng xã hội, phân loại email, ...



Cảm ơn Quý Thầy Cô và các bạn đã
lắng nghe

