ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN TOÁN RỜI RẠC NÂNG CAO

ỨNG DỤNG TÌM KIẾM TRA CỨU TÀI LIỆU TRONG LUẬT ĐẤT ĐAI 2013

Lớp: **CS521.L21**
Giảng viên: **Nguyễn Đình Hiển**
Nhóm thực hiện:

1. Trần Đỗ Quốc Khiêm 18520076
2. Nguyễn Dương Trúc Phương 18520133
3. Trần Hoàng Việt 18520192
4. Lê Đại Thành 18521404

TP. HỒ CHÍ MINH – THÁNG 6/2021

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU CHƯƠNG TRÌNH

1.1. Mô tả chương trình

Là một hệ thống tra cứu (là tài liệu về luật) từ một nguồn không có cấu trúc tự nhiên, chứa đựng một số thông tin nào đó từ một tập hợp lớn. Một trong những kỹ thuật phổ biến trong tra cứu đó là **Vector Space Model, Okapi BM25,…**

1.2. Ứng dụng chương trình

Dữ liệu trong các văn bản quá lớn để ta có thể tìm kiếm một cách nhanh chóng. Đó là lý do chương trình "tìm kiếm và tra cứu luật" được xây dựng. Chương trình là công cụ hỗ trợ cho chúng ta tìm kiếm, tra cứu luật nhanh chóng bằng các số hiệu, tiêu đề hoặc nội dung ngắn gọn của văn bản.

CHƯƠNG 2: THU THẬP VÀ XỬ LÝ DỮ LIỆU

2.1. Thu thập dữ liệu

Dữ liệu được thu thập từ trang: https://thuvienphapluat.vn/
Dữ liệu bao gồm 212 điều luật thuộc luật đất đai năm 2013
Dữ liệu được lưu vào 212 file được đặt tên là Dieu(x).txt

```
Chương 1.
QUY ĐỊNH CHUNG
Điều 5. Người sử dụng đất
Người sử dụng đất được Nhà nước giao đất, cho thuê đất, công nhận quyền sử dụng đất, nhận chuyển quyền sử dụng đất theo quy định của Luật này, bao gồm:
1. Tổ chức trong nước gồm cơ quan nhà nước, đơn vị vũ trang nhân dân, tổ chức chính trị, tổ chức chính trị - xã hội, tổ chức kinh tế, tổ chức chính trị xã hội - nghề nghiệp, tổ chức xã hội, tổ chức xã hội - nghề nghiệp, tổ chức sự nghiệp công lập và tổ chức khác theo quy định của pháp luật về dân sự (sau đây gọi chung là tổ chức);
2. Hộ gia đình, cá nhân trong nước (sau đây gọi chung là hộ gia đình, cá nhân);
3. Cộng đồng dân cư gồm cộng đồng người Việt Nam sinh sống trên cùng địa bàn thôn, làng, ấp, bản, buôn, phum, sóc, tổ dân phố và điểm dân cư tương tự có cùng phong tục, tập quán hoặc có chung dòng họ;
4. Cơ sở tôn giáo gồm chùa, nhà thờ, nhà nguyện, thánh thất, thánh đường, niệm phật đường, tu viện, trường đào tạo riêng của tôn giáo, trụ sở của tổ chức tôn giáo và cơ sở khác của tôn giáo;
5. Tổ chức nước ngoài có chức năng ngoại giao gồm cơ quan đại diện ngoại giao, cơ quan lãnh sự, cơ quan đại diện khác của nước ngoài có chức năng ngoại giao được Chính phủ Việt Nam thừa nhận; cơ quan đại diện của tổ chức thuộc Liên hợp quốc, cơ quan hoặc tổ chức liên chính phủ, cơ quan đại diện của tổ chức liên chính phủ;
6. Người Việt Nam định cư ở nước ngoài theo quy định của pháp luật về quốc tịch;
7. Doanh nghiệp có vốn đầu tư nước ngoài gồm doanh nghiệp 100% vốn đầu tư nước ngoài, doanh nghiệp liên doanh, doanh nghiệp Việt Nam mà nhà đầu tư nước ngoài mua cổ phần, sáp nhập, mua lại theo quy định của pháp luật về đầu tư.
```

2.2. Tiền xử lý dữ liệu

Tách từ: Sử dụng thư viện pyvi (https://pypi.org/project/pyvi/)

Xóa dấu tất cả các dấu câu.

Xóa tất cả stopwords có trong dữ liệu
Vietnamese-stopwords: https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt

```
chương 1 quy_định 5 đất đất nhà_nước giao đất thuê đất công_nhận quyền đất quyền đất quy_định luật bao_gồm 1 tổ_chức cơ_quan nhà_nước vũ_trang nhân_dân tổ_chức chính_trị tổ_chức chính_trị xã_hội tổ_chức kinh_tế tổ_chức chính_trị xã_hội nghề_nghiệp tổ_chức xã_hội tổ_chức xã_hội nghề_nghiệp tổ_chức sự_nghiệp công_lập tổ_chức quy_định pháp_luật dân_sự gọi tổ_chức 2 hộ gia_đình gọi hộ gia_đình 3 cộng_đồng dân_cư cộng_đồng việt_nam sinh_sống địa_bàn thôn làng ấp buôn phum sóc tổ dân_phố dân_cư tương_tự phong_tục tập_quán dòng_họ 4 cơ_sở tôn_giáo chùa nhà_thờ nhà_nguyện thánh_thất thánh_đường niệm phật đường tu_viện trường đào_tạo tôn_giáo trụ_sở tổ_chức tôn_giáo cơ_sở tôn_giáo 5 tổ_chức nước_ngoài chức_năng ngoại_giao cơ_quan đại_diện ngoại_giao cơ_quan lãnh_sự cơ_quan đại_diện nước_ngoài chức_năng ngoại_giao chính_phủ việt_nam thừa_nhận cơ_quan đại_diện tổ_chức liên_hợp quốc cơ_quan tổ_chức liên chính_phủ cơ_quan đại_diện tổ_chức liên chính_phủ 6 việt_nam định_cư nước_ngoài quy_định pháp_luật quốc_tịch 7 doanh_nghiệp vốn đầu_tư nước_ngoài doanh_nghiệp 100 vốn đầu_tư nước_ngoài doanh_nghiệp liên_doanh doanh_nghiệp việt_nam đầu_tư nước_ngoài mua cổ_phần sáp_nhập mua quy_định pháp_luật đầu_tư
```

CHƯƠNG 3: VECTOR SPACE MODEL

3.1. Ý tưởng

Với mỗi truy vấn, hệ thống tìm kiếm sẽ sử dụng một độ đo Rel(q,d) để tính độ tương đồng giữa truy vấn (query) đó với các tài liệu (docs), từ đó xếp dạng được kết quả trả về.

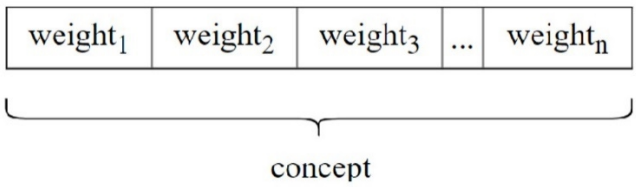
Ý tưởng của Vector Space Model là biểu diễn văn bản và các câu truy vấn dưới dạng Vector, Rep(d) của docs và Rep(q) của query sẽ cho kết quả là các vector. Sau đó tính độ tương đồng của query với từng documents theo công thức Sim(Rep(q), Rep(d)) để tìm ra docs vào phù hợp nhất với query.

- Rel(q, d): relevance giữa d và q
- Rep(d): hàm biểu diễn tài liệu d
- Rep(q): hàm biểu diễn truy vấn q
- Sim(Rep(q), Rep(d)): relevance giữa d và q

3.2. Concep vector

Biểu diễn documents và query bởi các concept vectors:

- Mỗi concept biểu diễn một chiều
- K concepts biểu diễn một không gian nhiều chiều.
- K concepts biểu diễn một không gian nhiều chiều.



3.3. Trọng số vector

Cách xác định và tính weights cho vector là hết sức quan trọng, ảnh hưởng đến độ chính xác của các thuật toán xếp hạng. Việc các từ có trọng số khác nhau là do không phải các từ đều có sự quan trọng giống nhau, sử dụng số lần xuất hiện của các từ làm vector không phải là một cách tối ưu. Ở phương diện các documents, một vài từ có thể mang nhiều thông tin hơn các từ còn lại.

Có nhiều kỹ thuật tính trọng số: TF, IDF, TF-IDF, …

3.3.1. TF: Term frequency

Từ nào xuất hiện nhiều trong câu thì quan trọng, công thức này sẽ đếm tuần suất xuất hiện các từ trong câu.

$$tf(t,d) = f(t,d) \sim Tần\ số\ xuất\ hiện\ của\ t\ trong\ d$$

TF normalization: Do tùy độ dài ngắn khác nhau của từng câu, mà việc đếm tần suất các từ có thể không công bằng. Nên ta chuẩn hóa bằng cách chia cho độ dài của d;

3.3.2. IDF: Inverse document frequency

Từ nào xuất hiện nhiều trong mọi câu thì không mang nhiều ý nghĩa (ví dụ như a, the, are, thì, là, …). Vì vậy trọng số IDF là nghịch đảo của tuần suất xuất hiện của các từ trong các documents.

$$idf(t,D) = log\frac{N}{1 + |\{d \in D: t \in d\}|}$$

- N: số tài liệu trong tập corpus N = |D|
- $|\{d \in D: t \in d\}|$ số docs mà từ tt xuất hiện. Cộng 1 cho mẫu số để tránh trường hợp chia cho 0 nếu từ đó không xuất hiện trong copus

3.3.3. TF-IDF: Term frequency - Inverse document frequency

Phép nhân giữa TD và IDF cho phép ta kết hợp cả 2 độ đo trên, từ vừa xuất hiện nhiều lần trong câu, vừa không phải là từ phổ biến xuất hiện trong mọi câu.

$$tfidf(t,d,D) = tf(t.d)idf(t,D)$$

3.4 Các độ đo:

Sau khi có được các vector cho query và docs, ta tính được similarity bằng cách tính khoảng cách giữa các vector.

3.4.1. Euclidean distance:

$$d(d,q) = \sqrt{\sum_{i=1}^{n}(d_i - q_i)^2}$$

3.4.2. Tích vô hướng:

$$\vec{d}\cdot\vec{q} = \sum_{i=1}^{n} d_i q_i$$

3.4.3. Cosine:

$$\cos\theta = \frac{A\cdot B}{\|A\|\|B\|}$$

Sau khi biểu diễn dưới dạng vector và tính khoảng cách, ta xếp hạng được các tài liệu tìm kiếm với từng query vector.

CHƯƠNG 4: MÔ HÌNH OKAPI

4.1. Okapi BM25

Trong tìm kiếm thông tin, Okapi BM25 là hàm tính thứ hạng được các công cụ tìm kiếm sử dụng để xếp hạng các văn bản theo độ phù hợp với truy vấn nhất định. Hàm xếp hạng này dựa trên mô hình xác suất, được phát minh ra vào những năm 1970 – 1980. Phương pháp có tên BM25 (BM – best match), nhưng người ta thường gọi "Okapi BM25", vì lần đầu tiên công thức được sử dụng trong hệ thống tìm kiếm Okapi, được sáng lập tại trường đại học London những năm 1980 và 1990.

BM25 là một phương pháp xếp hạng được sử dụng rộng rãi trong tìm kiếm. Trong Web search những hàm xếp hạng này thường được sử dụng như một phần của các phương pháp tích hợp để dùng trong machine learning, xếp hạng.

4.2. Bản chất của Okapi BM25

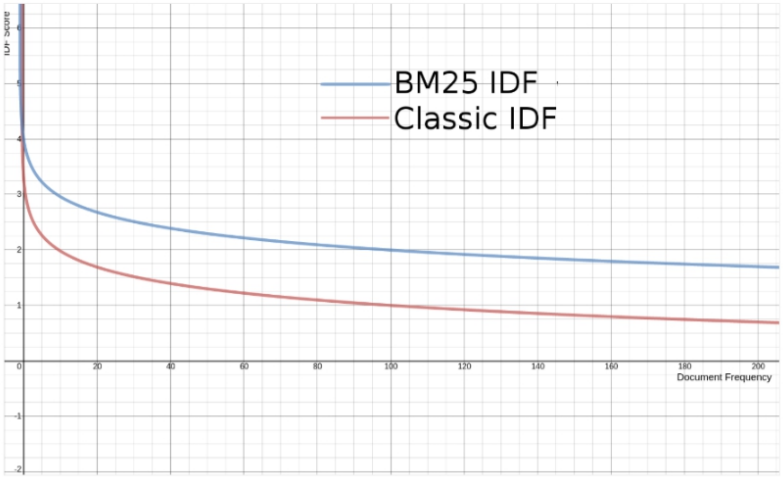
Một số thuật ngữ được sử dụng:

1. Relevance (độ liên quan)
2. Index (tức database)
3. Term (từ, từ khoa)
4. Field (trường)

Thực chất, BM25 dựa trên nền tảng của TF/IDF, và cải tiến dựa trên lý thuyết probabilitistic information retrieval. Từ đó điều chỉnh công thức để cho ra kết quả chinh xác hơn.

4.3. Công thức tính Okapi BM25 score

4.3.1. IDF trong Okapi BM25



Biểu đồ so sánh giá trị IDF score giữa TF/IDF cơ bản với BM25

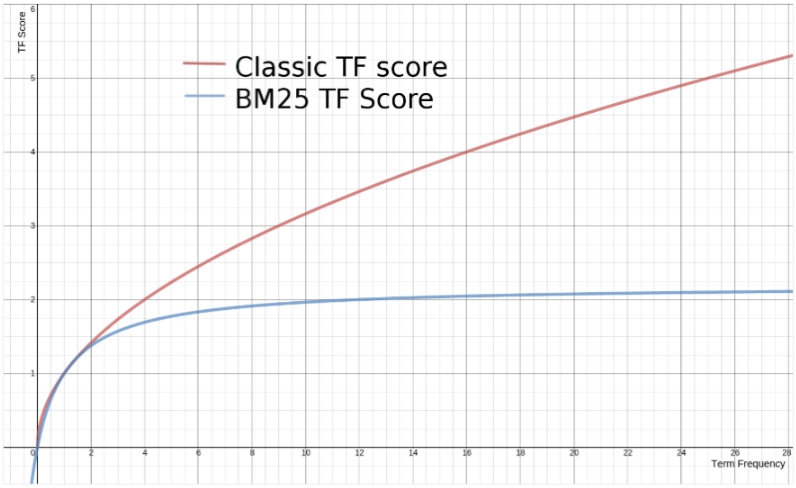
Như trên đồ thị, về cơ bản, cách tinh IDF score của Okapi BM25 tương đối giống với công thức tinh IDF score bình thường, tuy nhiên ở đây đã có một điều chỉnh nho nhỏ trong công thức IDF score của Okapi BM25.

$$IDF(t) = \log (1+ (docCount - docFreq +0.5)/(docFreq + 0.5))$$

Trong đó:

- docCount: số lượng document
- docFreq: số lượng document chứa term

4.3.2. TF trong BM25



Biểu đồ so sánh giá trị TF score giữa TF/IDF cơ bản với BM25

Đối với TF/IDF thì score từ TF sẻ tăng vô hạn khi TF tăng lên. Để giảm tác động của TF với relevance thì BM25 đã chỉnh sửa công thức của TF lại. Kết quả score của TF sẻ giới hạn tới 1 điểm cực đại, và chúng ta có thể tùy chỉnh giới hạn này.

$$TF = ((k+1) * freq) / (k + freq)$$

Trong đó:

- k: hằng số (thường là 1.2)
- freq: frequency của term trong document

4.3.3. Document Length trong BM25

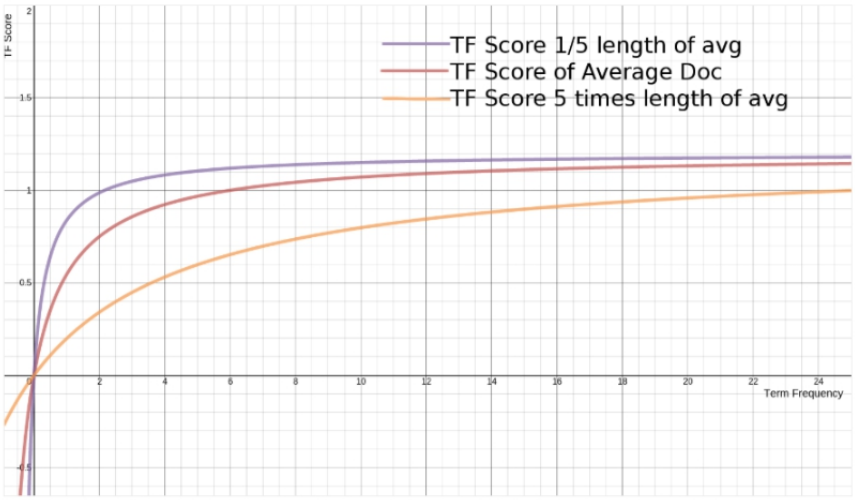
Thực ra công thức TF bên trên kia là chưa thực sự hoàn chỉnh, nó đúng với những document có độ dài trung bình trong toàn bộ index. Nếu độ dài document quá ngắn hoặc quá dài so với độ dài trung bình, thì công thức trên sẽ cho kết quả

thiếu chính xác. Vì thế người ta thêm vào trong công thức trên 2 tham số, một hằng số b và một giá trị độ dài L, công thức sẽ trở thành:

TF score = ((k + 1) * freq) / (k * (1.0 - b + b * L) + freq)

Trong đó:

- b=0.75 (mặc định).
- L là tỉ lệ giữa độ dài của document so với độ dài trung bình của tất cả documents: L = fieldLength / avgFieldLength



Biểu đồ so sánh TF score với 3 giá trị L khác nhau

Cũng như k, bạn có thể điều chỉnh b để phù hợp với mô hình bạn xây dựng. b càng gần 0 thì độ ảnh hưởng của document length càng nhỏ, và ngược lại, b càng lớn thì độ ảnh hưởng của document length càng lớn

4.3.4. Công thức cuối cùng:

Ta có công thức cuối cùng của BM25

IDF * (freq * (k1 + 1)) / (freq + k1 * (1 - b + b *(fieldLength/avgFieldLength)))

CHƯƠNG 5: ĐÁNH GIÁ VÀ KẾT LUẬN

5.1. Đánh giá

Sau khi chạy cả 2 model thì nhận thấy từ "điều", "chương", "mục" khi tìm kiếm sẽ bị sai do từ "điều" là stopword, từ "chương", "mục" là từ đơn nên đã tiền xử lý lại văn bản, sửa lại các từ trên thành từ ghép.
VD: điều 23 -> điều_23, chương 1-> chương_1, mục 1 -> mục_1, v.v

Sau khi tiền xử lí dữ liệu lại thì kết quả trả về chính xác hơn.

```
Search:  điều 23
After processing 23
-----------------------------------------------
 Kết quả tìm kiếm:
id document:  2
Chương 1.
QUY ĐỊNH CHUNG
Điều 2. Đối tượng áp dụng
1. Cơ quan nhà nước thực hiện quyền hạn và trách nhiệm đại diện chủ sở hữu toàn dân về đất đai, thực hiện nhiệm vụ thống nhất quản lý nhà nước về đất
```

```
Search:  điều 23
After processing điều_23
-------------------------------------------------
 Kết quả tìm kiếm:
id document:  23
Chương 2.
QUYỀN VÀ TRÁCH NHIỆM CỦA NHÀ NƯỚC ĐỐI VỚI ĐẤT ĐAI
MỤC 2. TRÁCH NHIỆM CỦA NHÀ NƯỚC ĐỐI VỚI ĐẤT ĐAI
Điều 23. Trách nhiệm quản lý nhà nước về đất đai
1. Chính phủ thống nhất quản lý nhà nước về
```

Để đánh giá hiệu quả của model nhóm đã tiến hành tính Precision, Recall và Average Precision dựa trên 5 docs trả về mỗi query

	Vector space model (cosine)			**BM25**		
Pro query	5^{th}-precision	5^{th}-recall	AP	5^{th}-precision	5^{th}-recall	AP
chương_10 mục_1	0.8	1.0	1.0	0.6	0.75	0.87
Đất rừng phòng_hộ	0.4	0.2	0.16	0.6	0.3	0.48
ổn_định lâu_dài	1.0	1.0	1.0	1.0	1.0	1.0
Điều_23	1.0	1.0	1.0	0.2	1.0	1.0
Định_giá đất	0.8	0,67	0.67	0.6	0.5	0.57
Đất trồng cây lâu năm	0.2	0.2	0.09	0.8	0.8	0.59
mAP			0.65			0.75

Vì dữ liệu của mô hình thuộc về chủ đề đất đai nên các từ đặc trưng (đất.., đất đai..) rất phổ biến và không có tác dụng trong việc truy vấn.
Các từ này sẽ được thêm vào stopword.

	Vector space model (cosine)			BM25		
	5th-precision	5th-recall	AP	5th-precision	5th-recall	AP
chương_10 mục_1	0.8	1.0	1.0	0.6	0.75	0.87
rừng phòng_hộ	0.8	0.33	0.3	0.6	0.3	0.48
ổn_định lâu_dài	1.0	1.0	1.0	1.0	1.0	1.0
Điều_23	1.0	1.0	1.0	1.0	1.0	1.0
Định_giá	1.0	0.67	0.67	0.8	0.67	0.93
trồng cây lâu năm	0.6	0.6	0.46	0.8	0.8	0.94
mAP			0.74			0.87

Nhận xét: Mô hình Okapi BM25 cho kết quả tốt hơn

5.2. Ưu điểm

- Dễ hiểu và dễ cài đặt.
- Phương pháp được sử dụng rộng rãi.
- Cho kết quả tốt và khả thi.

5.3. Nhược điểm

- Từ khóa tìm kiếm phải khớp chính xác với các từ trong văn bản.
- Về ngữ nghĩa, các tài liệu có ngữ cảnh tương tự nhưng từ ngữ khác nhau sẽ không được trả về.
- Các từ là độc lập với nhau.

CHƯƠNG 6: THAM KHẢO

- https://blog.duyet.net/2019/08/ir-vector-space-model.html
- https://viblo.asia/p/bm25-thuat-toan-xep-hang-cac-van-ban-theo-do-phu-hop-Az45bWGNKxY