NOTE ĐI HỌC CS336 – TRUY VẤN THÔNG TIN ĐA PHƯƠNG TIỆN

21/9/2022:

- 1**. Khác nhau giữa information retrieval và search?**
  + Search là 1 loại IR. IR System là một giải pháp khá rộng được sử dụng để truy xuất nội dung từ kho dữ liệu phi cấu trúc như thư viện kỹ thuật số, trang web, nguồn cấp tin tức, bài đăng trên mạng xã hội. Search Engine chỉ hoạt động tốt nhất cho nội dung trang web.

S.No	Differentiator	Web Search	IR
1	Languages	Documents in many different languages. Usually search engines use full text indexing, no additional subject analysis	Databases usually cover only one language or indexing of documents written in different languages with the same vocabulary. Ex: Adversary-Opposing
2	File Types	Several file types, some hard to index because of a lack of textual information.	Usually all indexed documents have the same format (e.g. PDF) or only bibliographic information is provided.
3	Document length	Wide range from very short to very long. Longer documents are often divided into parts.	Document length varies, but not to such a high degree as with the Web documents
4	Document structure	HTML documents are semi structures.	Structured documents allow complex field searching
5	Spam	Search engines have to decide which documents are suitable for indexing.	Suitable document types are defined in the process of database design.

- **2. Define IR Problem**
  IR có thể được định nghĩa là một chương trình phần mềm xử lý việc tổ chức, lưu trữ, truy xuất và đánh giá thông tin từ các kho tài liệu, đặc biệt là thông tin văn bản. IR là hoạt động thu thập tài liệu thường có thể được ghi lại ở dạng unstructure, tức là thường là văn bản đáp ứng nhu cầu thông tin từ bên trong các bộ sưu tập lớn được lưu trữ trên máy tính. Ví dụ: IR có thể là khi người dùng nhập truy vấn vào hệ thống.
- **3. Trợ lí ảo , máy tìm kiếm: vinfast, siri, kiki của vng**
- **4. Input, output:**
  có 2 input là query và info resource, output: info relevant to an info need
- **5. Mục đích IR**
  là khoảng cách intension gap giữa query và info need
- **6. why: info overloaf**
- Buổi sau: Boolean và vector space model

28/9/2022:

- đề CS336: cho 3,4 queries. Hệ thống nào tốt nhất? tính metrics. Chứng minh độ đo đó với nhu cầu (nhiều nhu cầu). Rank, precision, recall, thứ tự đúng đầu tiên. PHẢI XÁC ĐỊNH NHU CẦU ĐO GÌ R DÙNG ĐỘ ĐO GÌ
- concepts đc xác định dựa trên tần suất xuất hiện từ khóa. Bao nhiêu từ khóa để represent documents. 1 từ khóa là 1 concepts. giá trị của vecto là trọng số biểu diễn tầm quan trọng từ khóa . Mức độ quan trọng đc xác định dựa trên tần suất frequency, số lượng document chứa từ khóa. Từ khóa nào càng ít xuất hiện trong các documents thì weight càng cao. TF-IDF

5/10/2022:

- token là semantic unit, stopword là selected type, type là list of dictionary
- query expansion

19/10/2022:

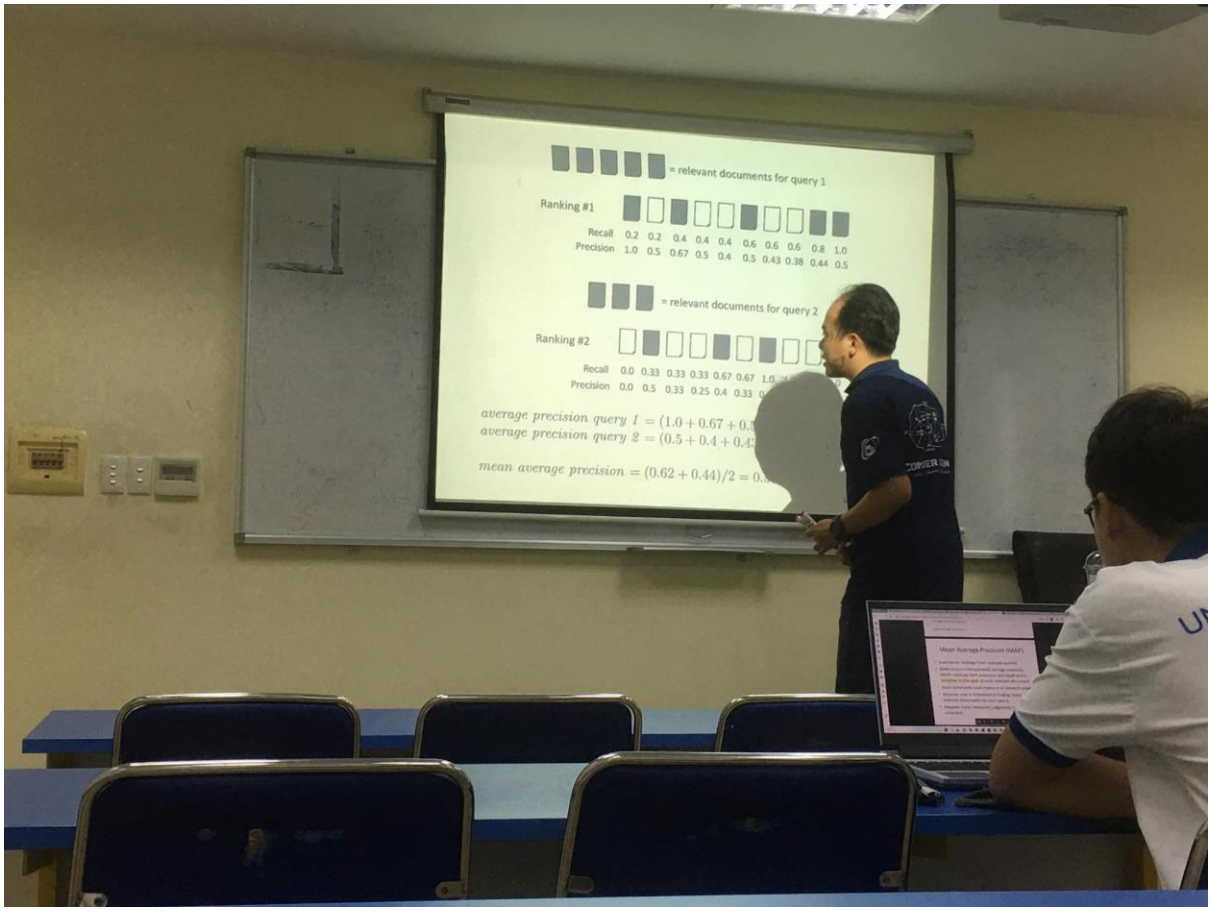
- input và output của index construction
- Bản chất của Heap Law, Zip Law, và nó vận dụng trong IR như thế nào
- Index construction làm gì: mối liên hệ giữa term và document, term trong dictionary và document trong corpus/collection. Mỗi term có 1 posting list, mỗi document parse ra, retrieve query có ko thì đánh 1,0. Chương này bàn cái làm gì với indexx khi tài nguyên bị giới hạn. Vấn đề: information overload
- non-positional posting là 1 dòng của bảng gồm term và Doc #
- BSBI : chia block . Tại sao chi 10 chia 100 block? 1entry 12 bytes gồm termID (1 số nguyên-4bytes) - . Lưu luôn từ gốc term thay vì term ID. Vấn đề term dài khoảng 8 bytes tốn hơn termID là 1 số nguyên.
- SPIMI: idea1 có dictionary cho từng block, idea2: Ko sort mà accumulate các posting -> merge kết quả lại
    - -> tự chạy thử
- Ưu điểm SPIMI: có thể dùng compression. Phần cuối xác định parallel tasks, phải độc lập nhau, ko phải đợi này đợi kia.
- index construction có các task song song: parsing, inverting,

26/10/2022:

- 1. Tại sao evaluation: Có nhiều giải pháp/ mô hình khác nhau, cần xem cái nào cần nhu cầu của mình, ở mức độ nào, [SO SÁNH]. giúp optimize giải pháp hiện tại. Evaluation System cần những gì : Annotation (bộ dữ liệu có đáp án), metrics (mục đích đánh giá, giá trị đó thể hiện đặc tính abilities hệ thống yêu cầu)
- 2.Human labeled Corpora: collection, set of queries, labels/annotation.. expected/correct answers. dấu tích v (relevance). Có 2 loại labels: Binary(relevant, irrelevant) và non-binary(thông thường là score).
- Annotation do nhiều người gán (expert), mang tính khách quan, số đông mang tính tổng quát cao. Giải quyết conflict trong data labelers.

2/11/2022:

- [Evaluation] rất quan trọng sẽ thi, ý nghĩa từng metrics, có khả năng vận dụng.
- Đi thi: Câu hỏi vận dụng cho 1 hệ thống vs 1 kết quả cụ thể, đánh giá so sánh. NDCG dành đo hệ thống, k phải nhị phân. trước đó là có metrics dành cho binary (relevant, irrelevant).
- Precison-at-k : do mình xét j và và bản chất query, precison thấp mà nói k tốt là sai (2 trong 100 cái đúng).
- R-precison: đo độ chính xác của hệ thống với top kết quả trở về tuy nhiên R thay đổi theo query k, khắc phục so với precision-at-k là k cố định. Hệ thống tốt là tìm hết và đẩy lên top (idea của system). Đo khả năng ko bỏ sót và đúng.
- MAP là sensitive rank,



Hình ảnh 1 đề sẽ ra thi CK

chú ý : chỉ tính vị trí có precision, ko tính hết là đi thi ko đủ time :))
7/12/2022
Input -> BoF -> f_input (vector này là Histogram of Visual Words)
BoF: là local region

Feature Extractor

- Trong 1 hệ thống tìm kiếm, 2 module quan trọng nhất là Feature Extraction và Comparision
- BoF là 1 trong những kĩ thuật/ giải pháp feature extraction
  - TF là histogram of words
  - ----------
  - Input -> Interest Point (IP) Detector -> Location of IP -> IP Descriptor(Thường dùng SIFT) -> f_IP (feature)
  - Việc build dictionary chỉ build 1 lần độc lập với query, dictionary phải đủ lớn để bao hết query
  - Vấn đề chọn K-clusters. K nhỏ thì bao quát thiếu , K lớn thì phân mảnh không gian
    LARGE SCALE IMAGE SEARCH
- Bắt nguồn từ information OverLoad
- Large : gồm 2V (volume, variation)
- Phải xác định như thế nào là similar và relevant
- Chi phí lớn ở bước compare 2 vector (bị ảnh hưởng bởi scale collections)

3 vấn đề liên quan đến scale ability : Tree based structure, Hashing, Binary Small…

Kd-tree (d là dimension,) bản chất là binary tree nhưng với dữ liệu nhiều chiều