TC3 Information Retrieval

Master 1 AI, Upsay
T4, March to April

Kim Gerdes
Lisn

Last time

- introduction
- terms and domains
- tutorial:
    - gentle introduction on new textual dataset
    - indexes and counting

Planning

1. 10/3 gentle introduction
2. 17/3 big dataset, binary evaluation
3. 24/3 improvements: embeddings
4. 28/3 NER
5. 31/3 taxonomy, Hearst
   NO 7/4!
6. 14/4 work on projects, discussions
7. 21/4 project presentation

Today

- Review of last week’s notebook
- evaluation without ranking
- trying to find answers in big dataset

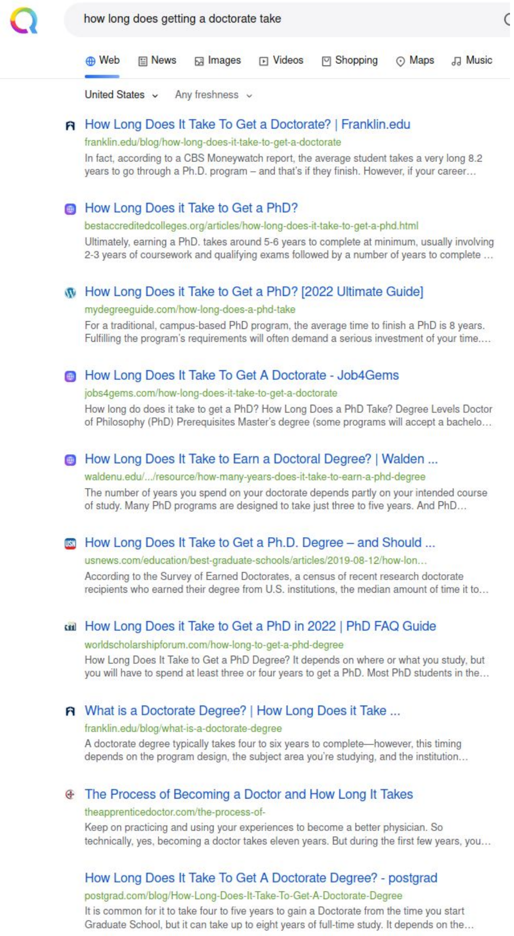
first thing now:

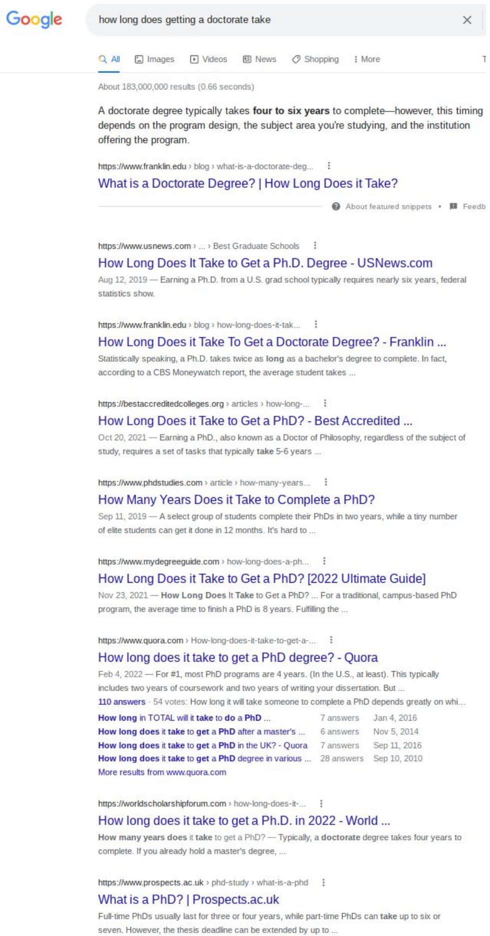
1. get the notebook on ecampus
2. go to the “reading in our smaller files” section and start the first cell or download manually
   so that we can get started right away later!

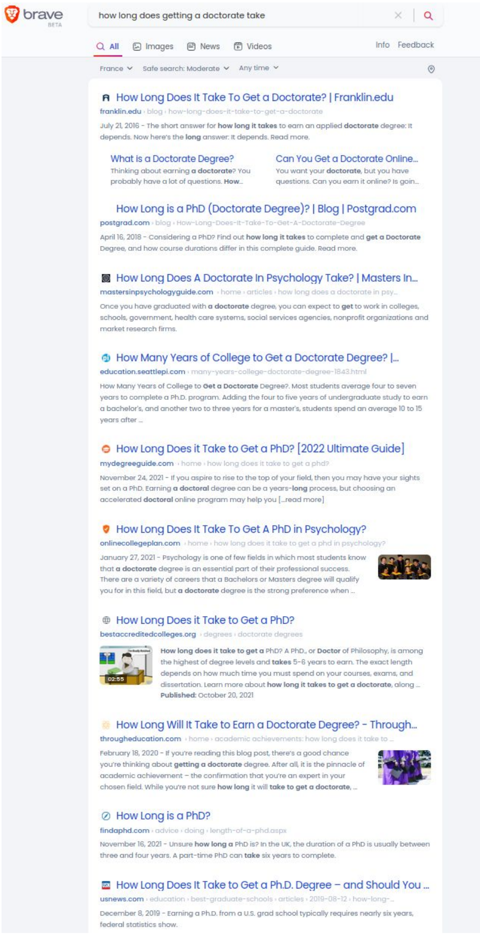
science?

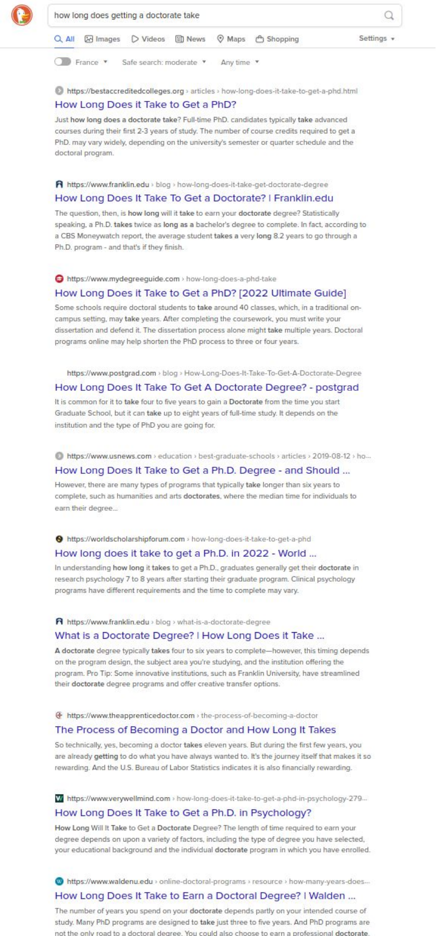
- last time:
    - very subjective
        - annotation of entities
        - relevance to a query
            - depending on many personal factors, points of view, …
- we can tweak many screws
- goal:
    - having heard of some of the screws
- there is no golden bullet

How to evaluate the quality of a retrieval system?









How to evaluate the quality of a retrieval system?

- many factors
  - speed
  - interface
  - price
  - user adaptation
  - relevance
    - who decides?
  - precision & recall
    - how to compute?
    - how does the user know that there aren’t any better documents
      - that are not shown?
      - that have not been crawled?

How to evaluate the quality of a retrieval system?

- We need a gold standard:
  - a biiiiig set of documents
  - many ‘typical’ queries
  - manual (semi-automatic?) evaluation of the relevance of each document to each query
    - ouch. that’s hard.

Information Retrieval Test Collections

Each IR test collection is comprised of:

1. Document collection
2. Set of information needs (descriptions + queries)
   a. A common requirement is to have at least 50 information needs
3. Set of relevance judgements for each query-document pair
   a. Binary relevance judgements (document relevant or non-relevant)
   b. Graded relevance judgements (less common, more difficult for human annotators)
   c. Q: Is it possible to annotate all query-document pairs for relevance?

Test collections are used for

- Evaluating retrieval effectiveness w.r.t. different settings
- Quantifying effects of e.g., different preprocessing methods, different ranking functions
- Comparing performance against other systems (usually in evaluation campaigns)
- Fine-tuning of system parameters, done on a development test collection

(Goran Glavaš, class on IR and WS)

Information Retrieval Test Collections

Some standard test collections:

- Cranfield – first IR test collection (from 1957)
  - 1,398 abstracts of aerodynamics journal articles
  - 225 queries, complete relevance judgements (1,398 x 225 annotations!)
- TREC collections – NIST Text Retrieval Conferences (1992 – today)
  - Ad-hoc retrieval task: 1.89M docs, 450 inf. needs, incomplete rel. judg.
  - Many other tasks: blog track, cross-lingual track, QA track, ...
- CLEF collections – Conference and Labs of the Evaluation Forum
  - Focus on European languages
  - Mono-lingual and cross-lingual ad-hoc retrieval tasks, QA tasks, ..

(Goran Glavaš, class on IR and WS)

Confusion matrix, unranked (binary) evaluation

- gold:
  - all relevant documents for each query

	relevant	not relevant
retrieved	tp	fp
not retrieved	fn	tn

- Accuracy doesn't work:
  - (tp+tn)/(tp+tn+fp+fn)
  - because most documents are irrelevant
  - a search engine that returns nothing gets a high accuracy

Confusion matrix, unranked evaluation

	relevant	not relevant
retrieved	tp	fp
not retrieved	fn	tn

- Precision tp/(tp+fp)
- Recall tp/(tp+fn)
- F-measure?

Confusion matrix, unranked evaluation

	relevant	not relevant
retrieved	tp	fp
not retrieved	fn	tn

- Precision tp/(tp+fp)
- Recall tp/(tp+fn)
- F-measure?
  - harmonic mean
  - general case

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

  - how to change β if recall is important such as in prior art search for patents?

Example

For some query q, there are in total 4 relevant documents (R) documents in the collection, whereas all other documents are not relevant (N).

- Some IR system returns 6 documents for the query q:
  - N,
  - R,
  - N,
  - R,
  - N,
  - N
- Compute precision, recall, and F1-measure
- Note that we don't need the false negative / total number of documents

(Goran Glavaš, class on IR and WS)

Example

For some query q, there are in total 4 relevant documents (R) documents in the collection, whereas all other documents are not relevant (N).

- Some IR system returns 6 documents for the query q:
    - N,
    - R,
    - N,
    - R,
    - N,
    - N

	relevant	not relevant
retrieved	tp: 2	fp: 4
not retrieved	fn: 2	tn: x

- Compute precision, recall, and F1-measure
    - p: 2/6=1/3, r:2/4=1/2, F1:$2 * \frac{1}{3} * \frac{1}{2} / (\frac{1}{3} + \frac{1}{2}) = \frac{1}{3} * 6/5 = \frac{2}{5} = 0.4$
- Note that we don't need the false negative / total number of documents
    - say tn = x = 100
    - accuracy = (tp+tn)/(tp+tn+fp+fn)=102/108=0.94

Confusion matrix, unranked evaluation

	relevant	not relevant
retrieved	tp	fp
not retrieved	fn	tn

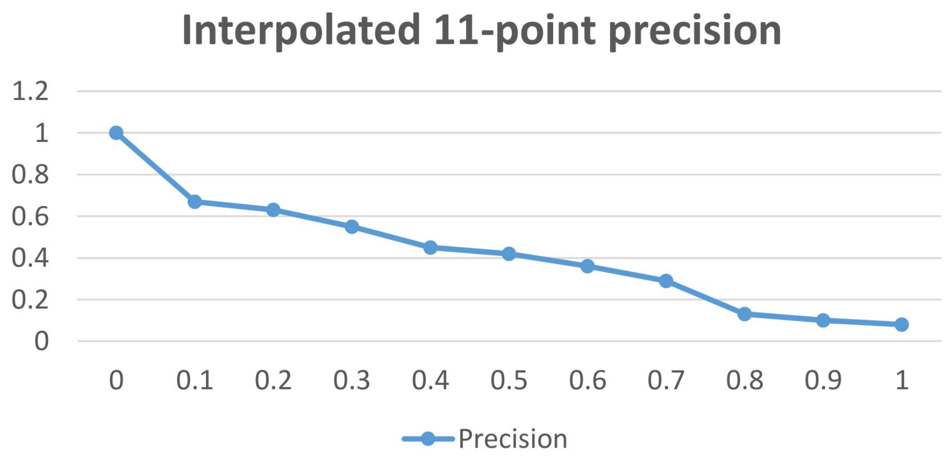
- good summary:
  - https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)

Ranked evaluation

- for now: [N, R, N, R] is equally good as ranking [R, R, N, N]
- Rank-based metrics:
    - Precision-recall curve
    - 11-point precision
    - MAP
    - P@k
    - R-precision
    - nDCG

11-point precision

- Interpolated 11-point precision describes performance of an IR system through precision measured at 11 different levels of recall:
  - Measuring precision at ranks where recall is:
    0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
  - For each recall level, average precisions measured over different queries



Mean average precision

- We would like to have a single-figure measure of retrieval effectiveness across all recall levels
- Average precision (AP) for a query q with relevant documents $\{d_1, ..., d_m\}$ is computed by averaging the precision scores measure at ranks of relevant docs:

$$\mathrm{AP}(q) = \frac{1}{m}\sum_{k=1}^{m} P(R_k)$$

R_k is the rank at which we find the k-th relevant document

- Mean average precision is AP averaged over the set of queries Q

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} P(R_{jk})$$

P@k

- MAP takes into account all recall levels, even at very low ranks
  - This is inappropriate for web search:
    - Less than 6% users look at the second page of results
- Precision at rank k (**P@k**) is precision at the fixed rank k in the ranking (e.g., P@5, P@10, P@20)
  - we will look into P@10

- **R-precision** is the P@k where k equals to the number of relevant documents for the query
  - E.g., if there are 5 relevant documents for the query in total, then R-precision = P@5

Other scores

- All methods so far assumed that we have binary relevance annotations
    - Sometimes we have graded relevance annotations
      E.g., from 1 (marginally relevant) to 5 (highly relevant)

  → Normalized Discounted Cumulative Gain (nDCG), …

Next steps

- retrieval techniques and evaluation
- grouping similar terms
  - making use of embeddings
- NER
- Hearst patterns

let’s break and move over to the practical part

- grab the notebook on ecampus
  - start the download right away…
- today: harder notebook
  - hopefully not too big data for your computer
  - if too hard: try colab and share your experience

Merci de votre

attention
considération
intérêt
écoute
présence
curiosité
question

!