Information Retrieval

Course presentation

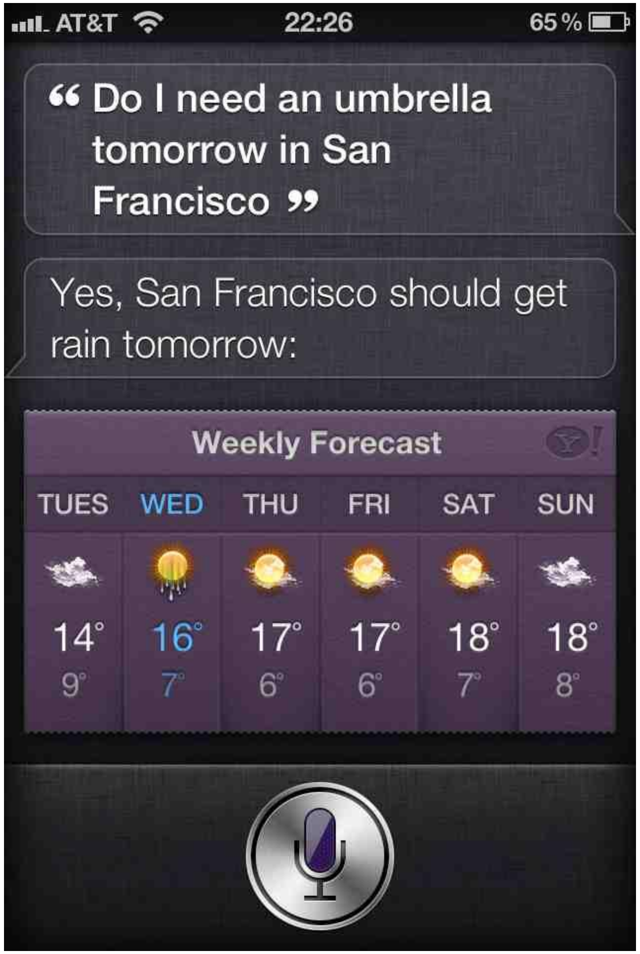
João Magalhães

Information retrieval

Question Answering

Move to mobile favors a move to **speech** which favors **natural language information search**

- Will we move to a time when over half of searches are spoken?



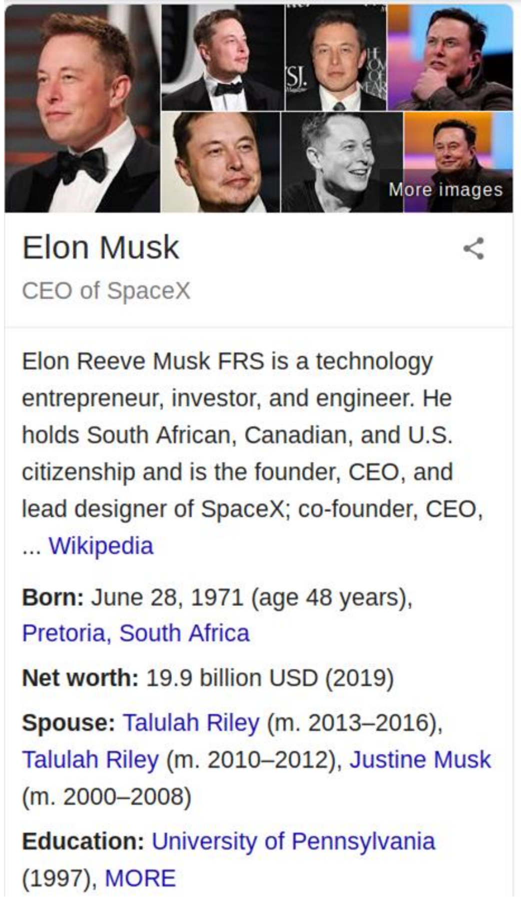
“ Do I need an umbrella tomorrow in San Francisco ”

Yes, San Francisco should get rain tomorrow:

Weekly Forecast

TUES	WED	THU	FRI	SAT	SUN
14°	16°	17°	17°	18°	18°
9°	7°	6°	6°	7°	8°

Named entities



Elon Musk

CEO of SpaceX

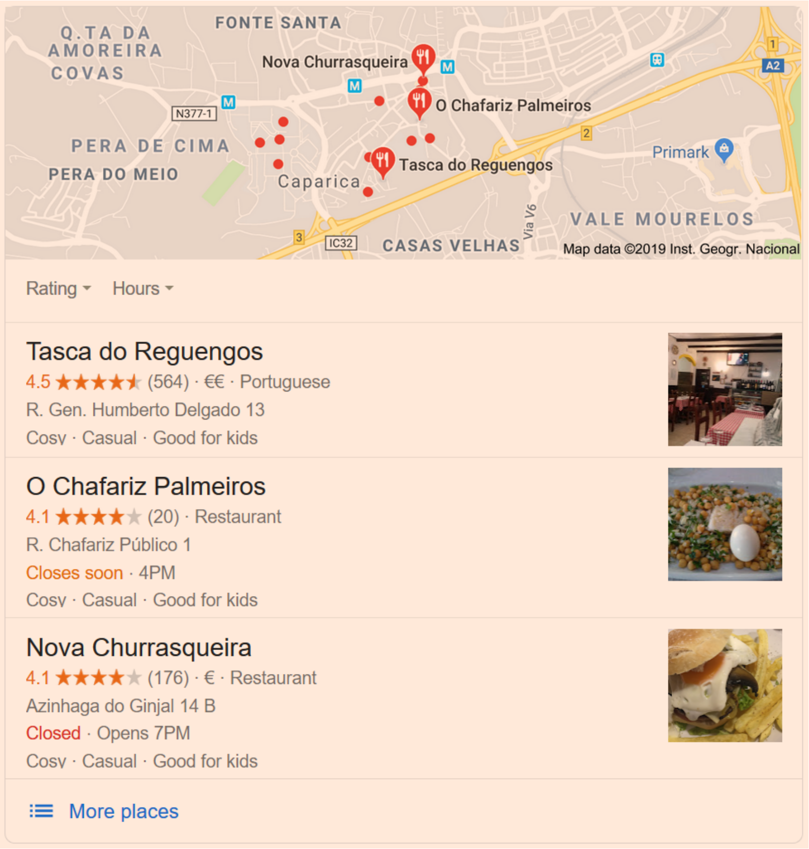
Elon Reeve Musk FRS is a technology entrepreneur, investor, and engineer. He holds South African, Canadian, and U.S. citizenship and is the founder, CEO, and lead designer of SpaceX; co-founder, CEO, ... Wikipedia

Born: June 28, 1971 (age 48 years), Pretoria, South Africa

Net worth: 19.9 billion USD (2019)

Spouse: Talulah Riley (m. 2013–2016), Talulah Riley (m. 2010–2012), Justine Musk (m. 2000–2008)

Education: University of Pennsylvania (1997), MORE



Tasca do Reguengos
4.5 ★★★★★ (564) · €€ · Portuguese
R. Gen. Humberto Delgado 13
Cosy · Casual · Good for kids

O Chafariz Palmeiros
4.1 ★★★★★ (20) · Restaurant
R. Chafariz Público 1
Closes soon · 4PM
Cosy · Casual · Good for kids

Nova Churrasqueira
4.1 ★★★★★ (176) · € · Restaurant
Azinhaga do Ginjal 14 B
Closed · Opens 7PM
Cosy · Casual · Good for kids

Conversational Search

- Alexa, Siri, Google Assistant…
- CS methods need to track the evolution of the information need in the conversation;
- It needs to identify salient information needed for the current turn in the conversation;
- Retrieval methods are required to retrieve the relevant information from a knowledge base (e.g. Wikipedia).

U: Tell me about the **Neverending Story film**.
A: ...

U: What is **it** about?
A: ...

U: Who was the author and when **it** was published?
A: ...

U: Who are the **main characters**?
A: ...

U: Did the horse **horse Artax** really die?
A: ...

Recommendation methods

- Recommender systems aim at suggesting new products to users based on their preferences

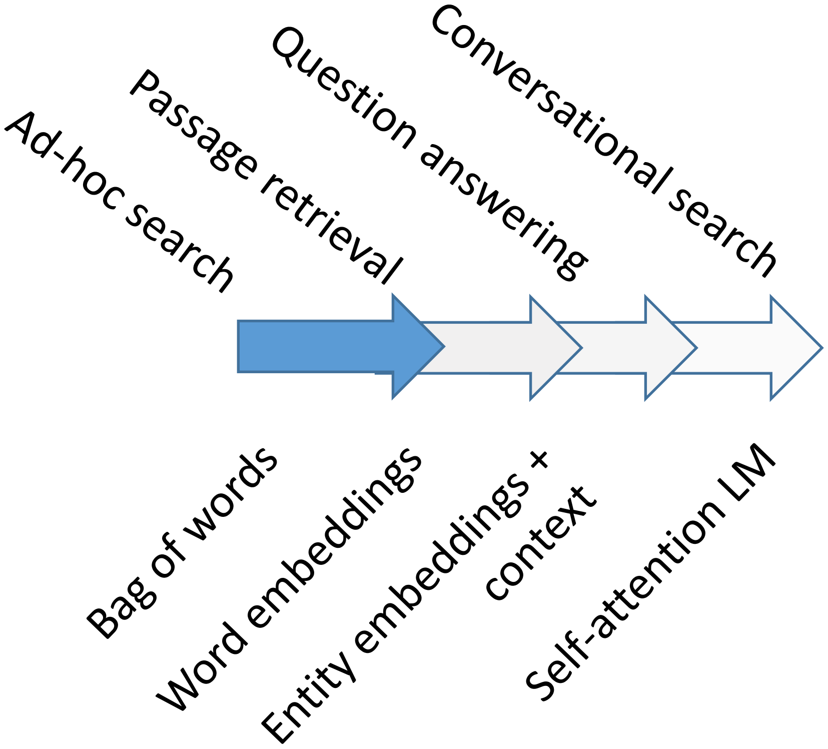
- Recommendations can be computed from two different type of inputs:
  - Product characteristics
  - Collective user ratings

NETFLIX tripadvisor Spotify amazon.com

Search in 2025?

What will people do in 2025?

- Type key words into a search box?
- Ask questions to their computer in natural language?
- Use social or “human powered” search?



Ad-hoc search

Passage retrieval

Question answering

Conversational search

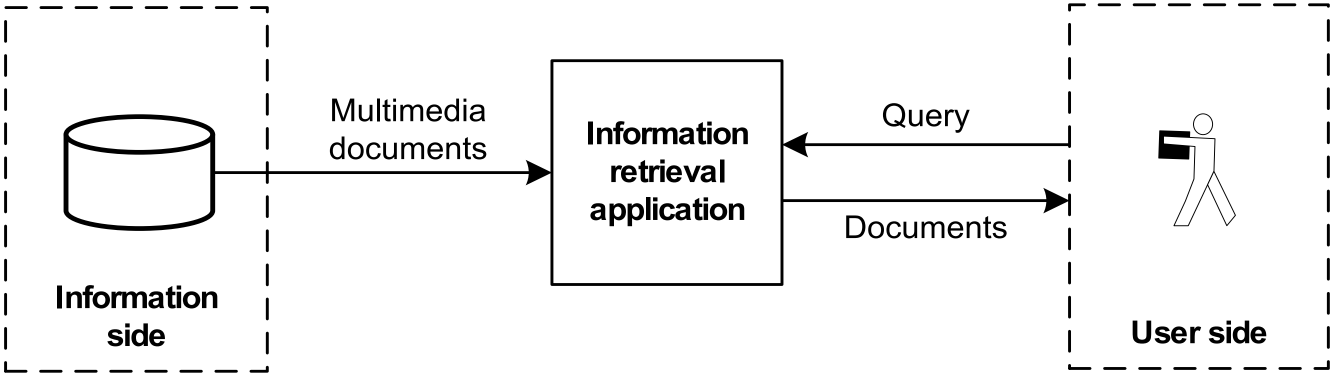
Bag of words

Word embeddings

Entity embeddings + context

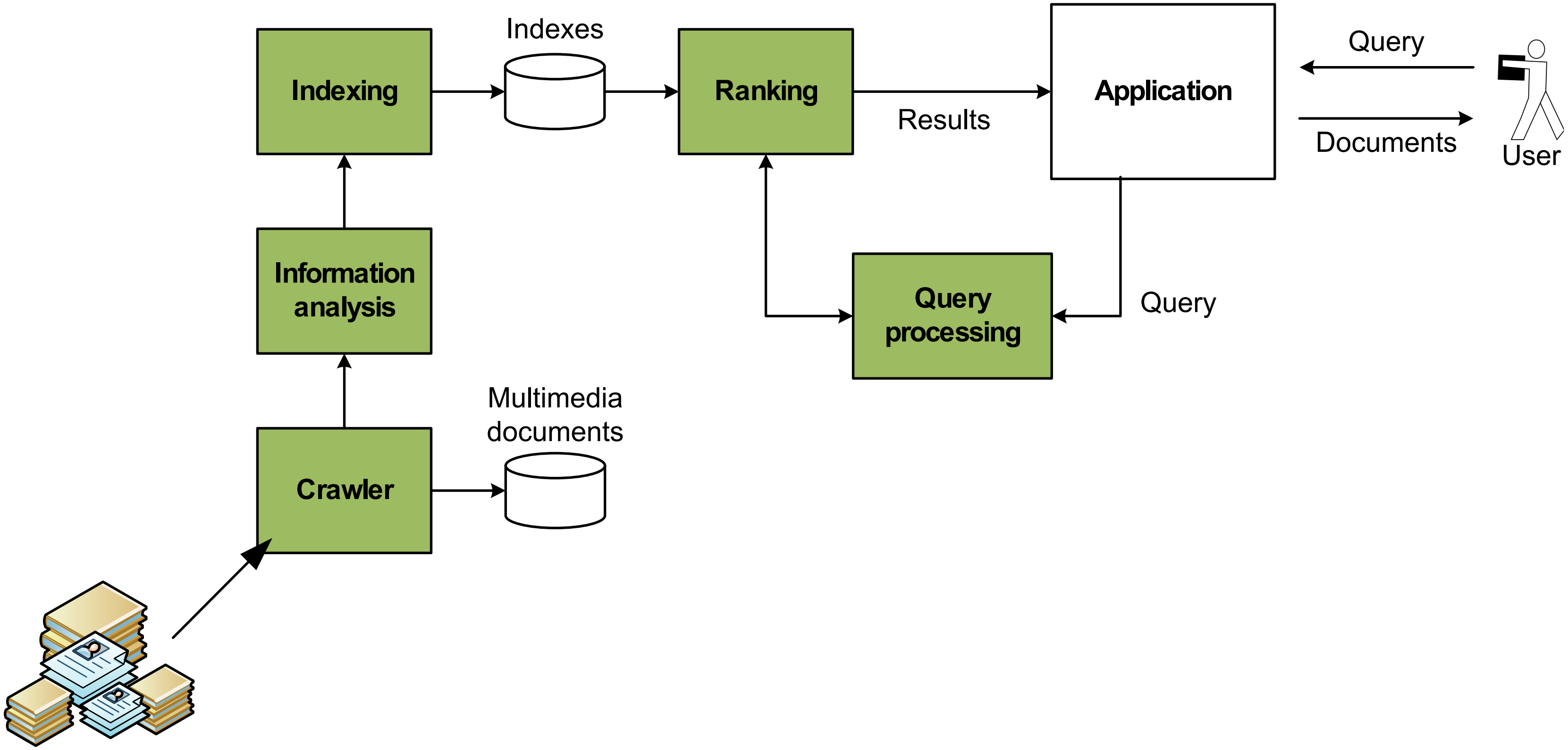
Self-attention LM

Relevance vs similarity



What is the best algorithm to compute the relevance of documents for a given user information need?

Putting all together...



The tasks of a search application

- **Collect** data for storage
  - Crawler
- Analyse collected data and compute the **relevant information**
  - Information analysis
- Store data in an **efficient** manner
  - Indexing
- Process **user** information needs
  - Querying
- Find the documents that best **match** the user information need
  - Ranking

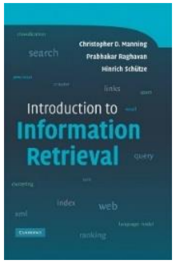
Schedule

Information Retrieval and Natural Language Processing			
Week	#	Lecture	In-class labs
16/set/20	1	Introduction	
23/set/20	2	Text processing, NGRAMS, cosine distance	Selecting answers
30/set/20	3	Language models	
07/out/20	4	Evaluation	
14/out/20	5	Classification tasks: sentiment, category, spa	
21/out/20	6	Pseudo relevance models	Re-ranking answers
28/out/20	7	Learning to rank	
04/nov/20	8	Word embeddings	
11/nov/20	9	Information extraction	
18/nov/20	10	Question answering	
25/nov/20	11	Conversational search	Conversational context
02/dez/20	12	Recommendation and personalization	
09/dez/20		Project support	
16/dez/20			

References

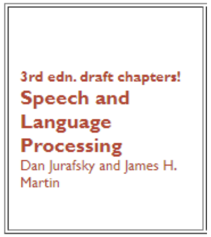
- Slides and articles provided during classes.

- Books:



C. D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008.

https://nlp.stanford.edu/IR-book/information-retrieval-book.html



3rd edn. draft chapters!
Speech and
Language
Processing
Dan Jurafsky and James H.
Martin

Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed. draft)

https://web.stanford.edu/~jurafsky/slp3/

Course grading

- The course has two mandatory components:
  - Project (groups of 3 students): 60% **(minimum grade > 9.0)**
    - (three submissions, on the 20th of each month)
  - Theoretical part (1 test or 1 exam): 40% **(minimum grade > 9.0)**

- Theory test/exam:
  - Test: January 4 to 16
  - Exam: To be defined

- Additional rules:
  - You may use one sided A4 sheet handwritten by you with your notes.
  - It must be handed in at the end of the test.

Project: Conversational search

- Track the evolution of the information need in the conversation;
- Identify salient information needed for the current turn in the conversation;
- Retrieval methods are required to retrieve the relevant information from a knowledge base (e.g. Wikipedia).

- A search end-point will be provided with the Wikipedia corpus index.

Project phases

- **Phase 1: Selecting answers (20%)** (20 October)
  - Searching with Language Model
  - Data inspection of conversational search sessions
  - Evaluation
- **Phase 2: Re-Ranking answers (20%)** (20 November)
  - Learning to rank
  - Neural Language Models
- **Phase 3: Conversational context (20%)** (20 December)
  - Modeling conversational context

Summary

- Context
- Objectives and plan
- Grading
- Labs