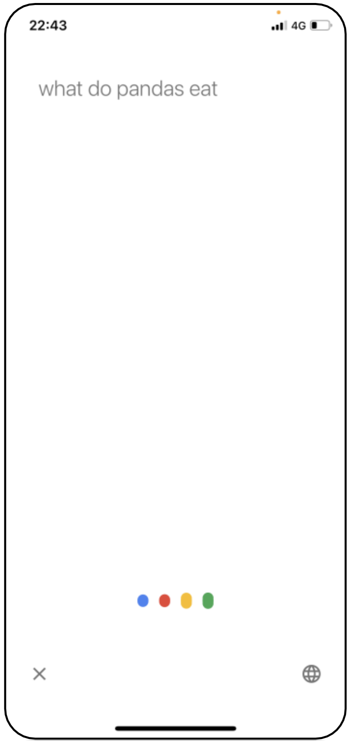
# Information Retrieval

Course presentation

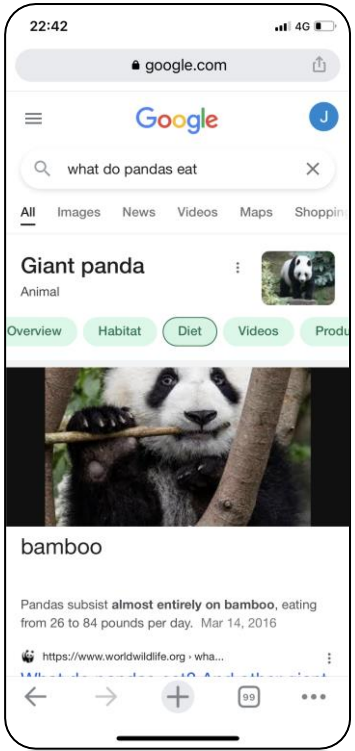
João Magalhães

# Information retrieval

# Voice based Search



what do pandas eat



**Giant panda**

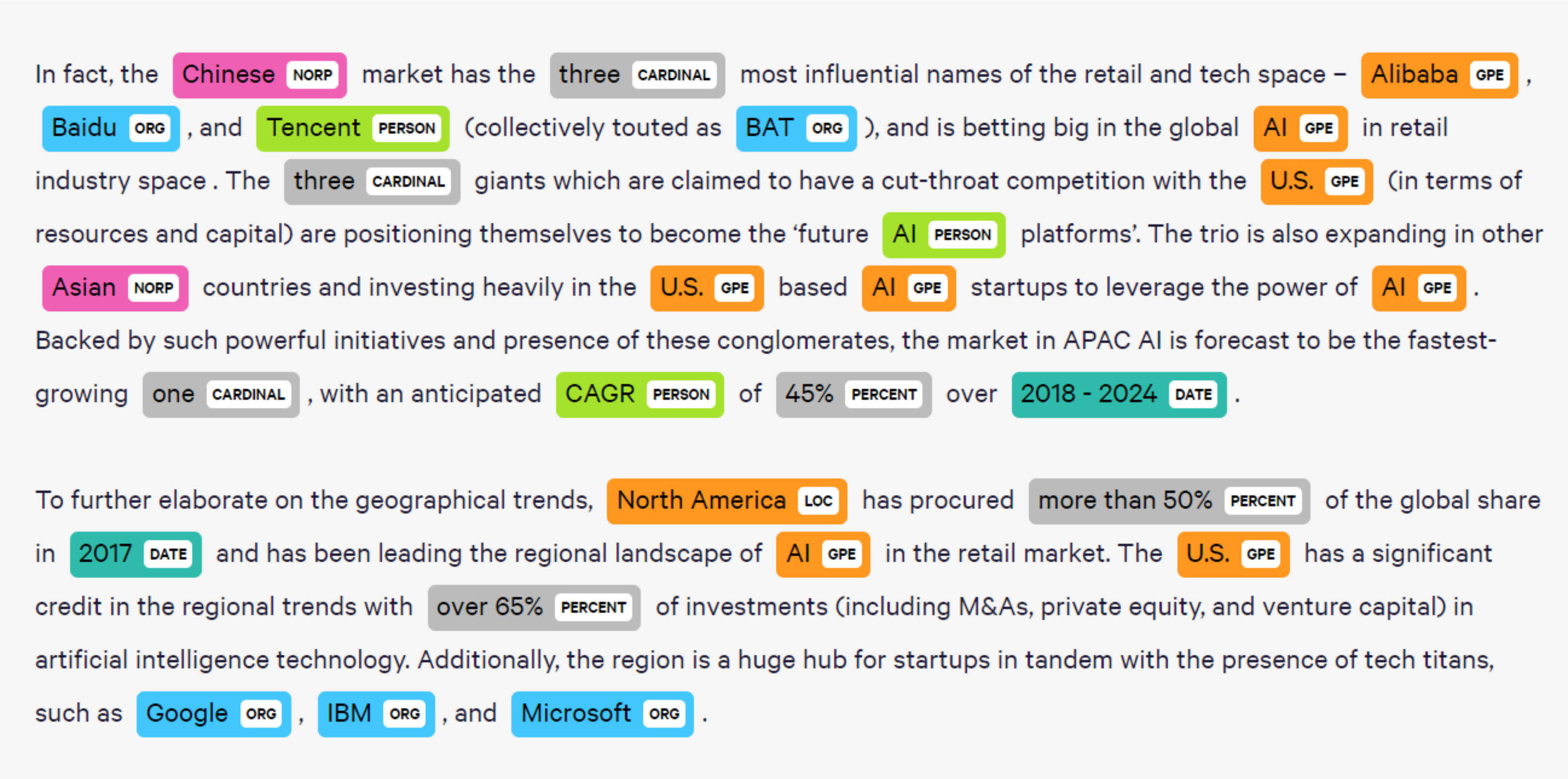
Animal

bamboo

Pandas subsist **almost entirely on bamboo**, eating from 26 to 84 pounds per day. Mar 14, 2016

https://www.worldwildlife.org › wha...

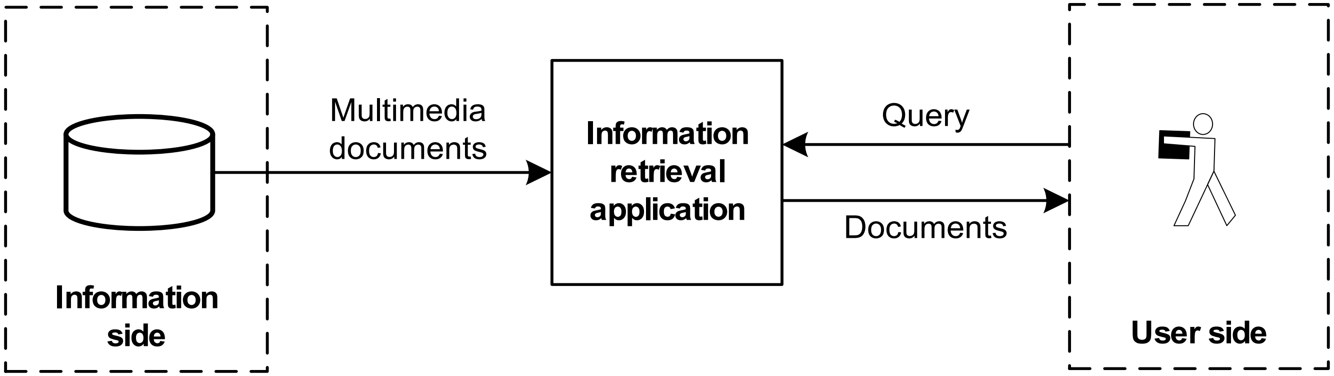
# Natural language understanding



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG ), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

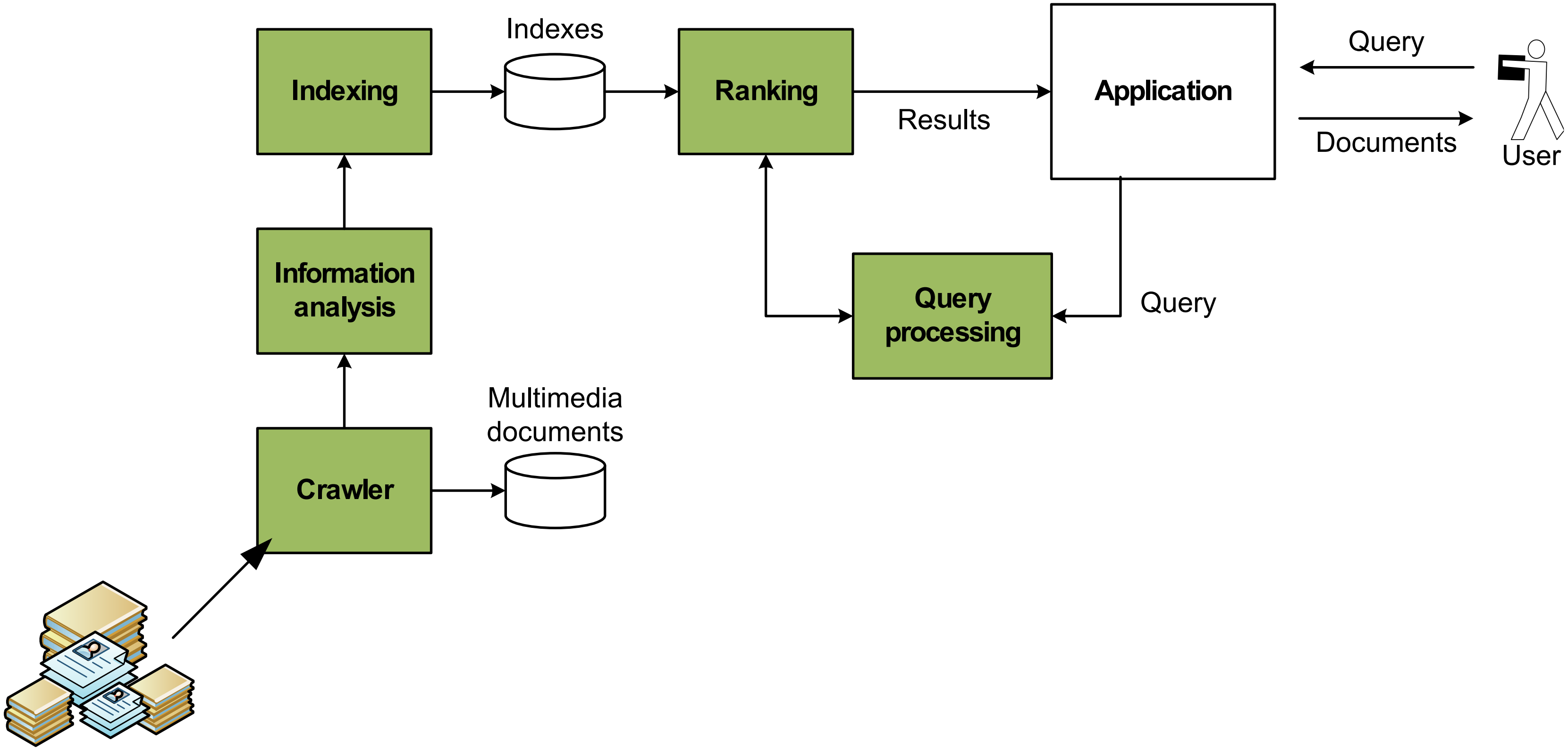
To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

# Relevance vs similarity



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

# Search architecture

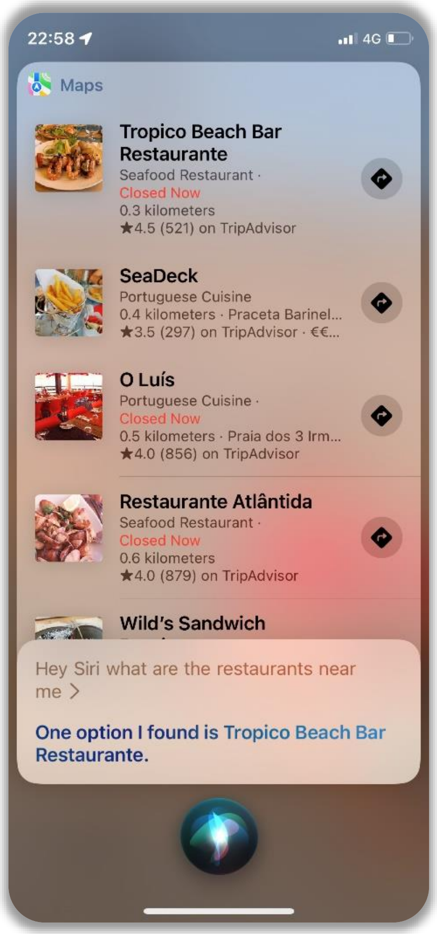


# Search architecture components

- **Crawl** data for storage
- Analyse documents and compute **meaningful representations of natural language**
- Store data in an **efficient** manner
- Process **user information needs**
- **Find answer** to user request

# Mobile QA

- Hands-free devices favor a move to speech-based interaction.
- Voice-interaction favors dialogs.
- Voice interaction is nowadays a commodity.



**Tropico Beach Bar Restaurante**
Seafood Restaurant ·
Closed Now
0.3 kilometers
★4.5 (521) on TripAdvisor

**SeaDeck**
Portuguese Cuisine
0.4 kilometers · Praceta Barinel...
★3.5 (297) on TripAdvisor · €€...

**O Luís**
Portuguese Cuisine ·
Closed Now
0.5 kilometers · Praia dos 3 Irm...
★4.0 (856) on TripAdvisor

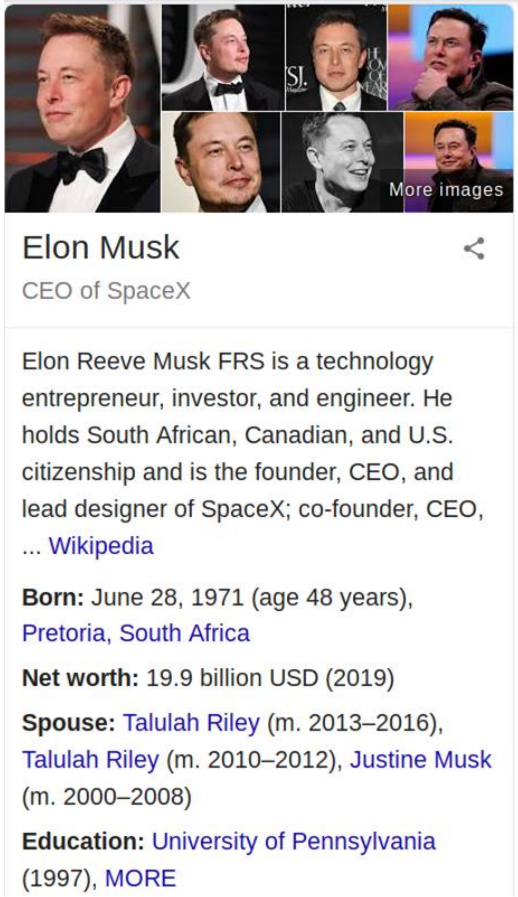
**Restaurante Atlântida**
Seafood Restaurant ·
Closed Now
0.6 kilometers
★4.0 (879) on TripAdvisor

**Wild's Sandwich**

Hey Siri what are the restaurants near me >

One option I found is Tropico Beach Bar Restaurante.

# Named entities



Elon Musk

CEO of SpaceX

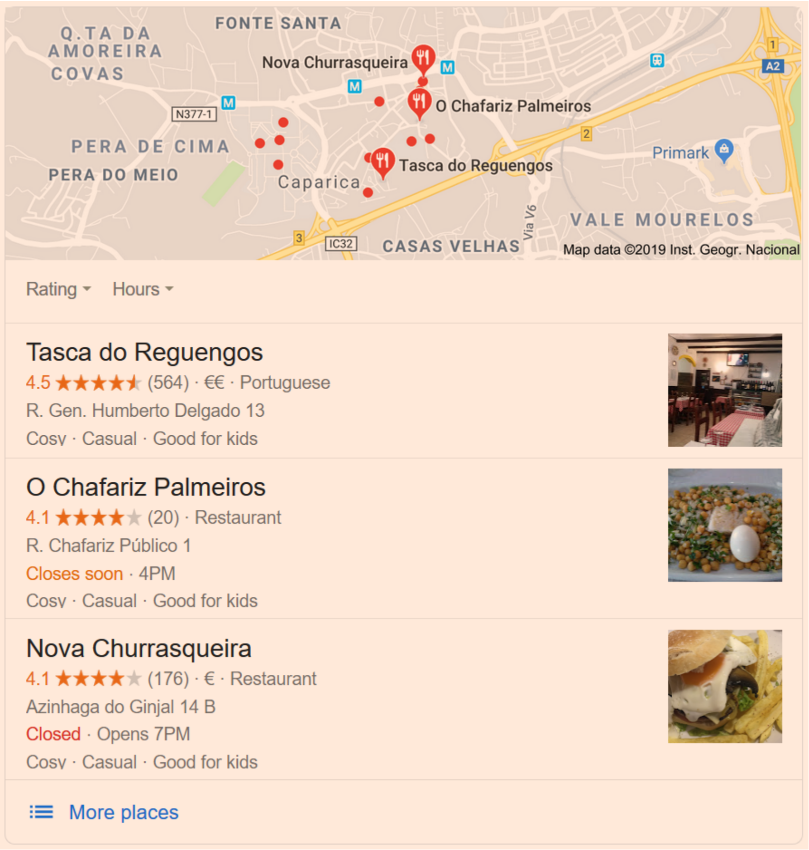
Elon Reeve Musk FRS is a technology entrepreneur, investor, and engineer. He holds South African, Canadian, and U.S. citizenship and is the founder, CEO, and lead designer of SpaceX; co-founder, CEO, ... Wikipedia

**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

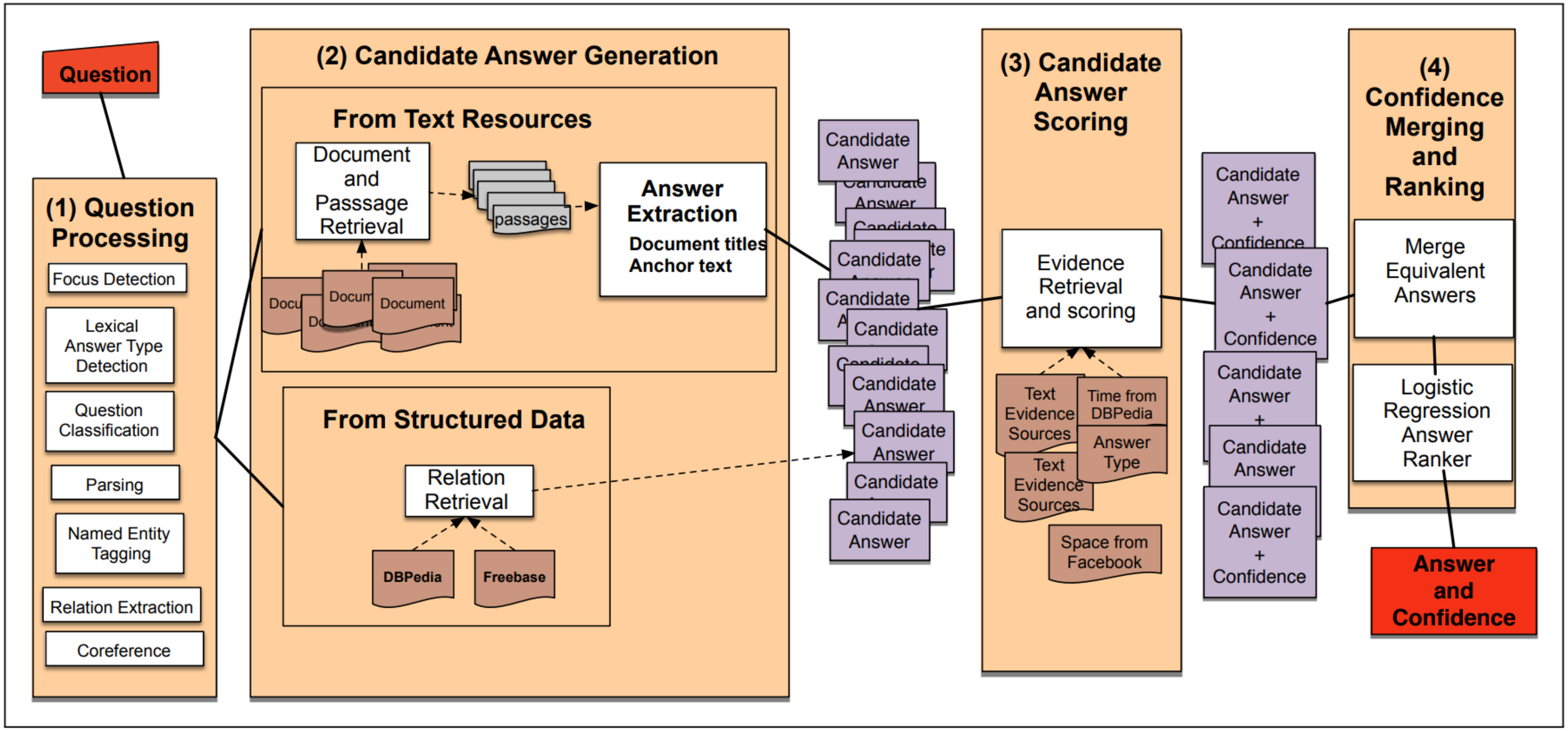
Azinhaga do Ginjal 14 B

Closed · Opens 7PM

Cosy · Casual · Good for kids

More places

# Question answering architecture



**Question**

**(1) Question Processing**

Focus Detection

Lexical Answer Type Detection

Question Classification

Parsing

Named Entity Tagging

Relation Extraction

Coreference

**(2) Candidate Answer Generation**

**From Text Resources**

Document and Passsage Retrieval

passages

Document

**Answer Extraction**

**Document titles**
**Anchor text**

**From Structured Data**

Relation Retrieval

**DBPedia**

**Freebase**

Candidate Answer

**(3) Candidate Answer Scoring**

Evidence Retrieval and scoring

Text Evidence Sources

Time from DBPedia

Answer Type

Space from Facebook

Candidate Answer + Confidence

**(4) Confidence Merging and Ranking**

Merge Equivalent Answers

Logistic Regression Answer Ranker

**Answer and Confidence**

# Question answering components

- **Collect** data for storage
- Analyse documents and compute **meaningful representations of natural language**
- Store data in an **efficient** manner
- Process **user question**
- **Find candidate answers**
- Extract correct answer

# 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: ...

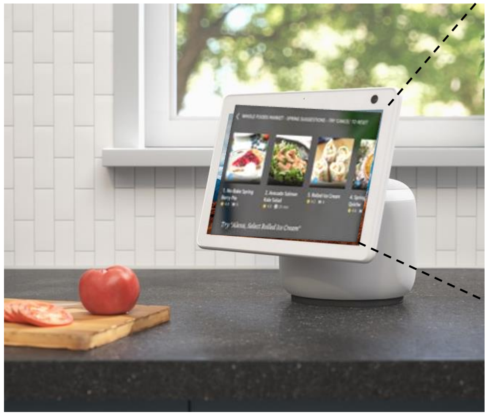
U: What is **it** about?
A: ...

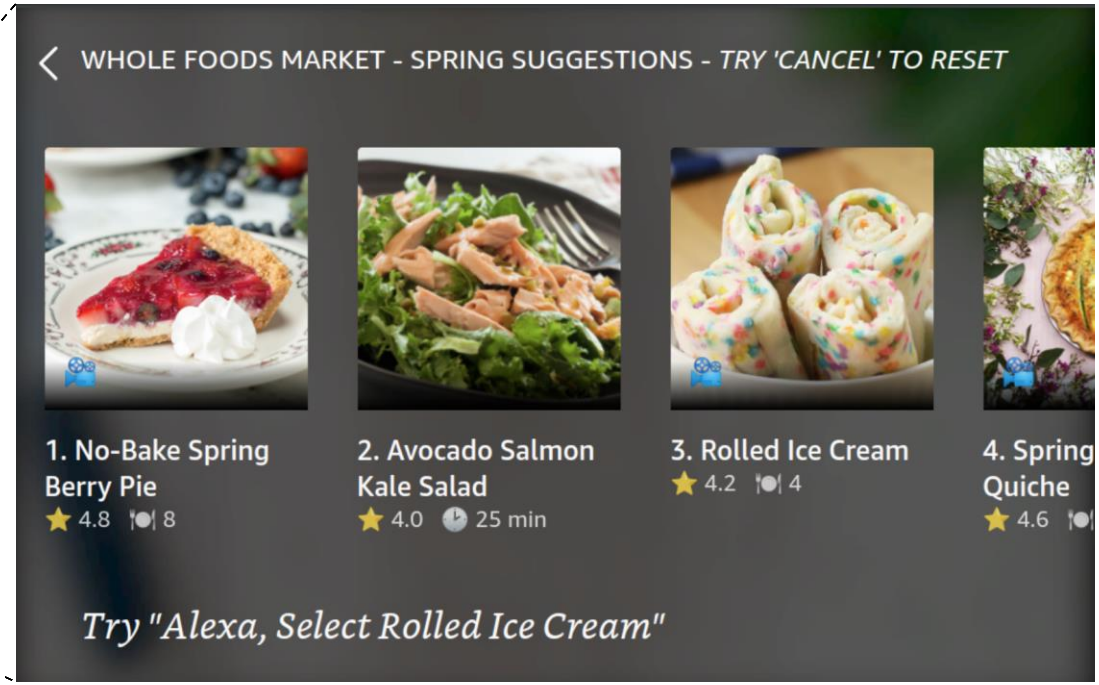
U: Who was the author and when **it** was published?
A: ...

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

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

# Alexa TaskBot for Cooking





WHOLE FOODS MARKET - SPRING SUGGESTIONS - *TRY 'CANCEL' TO RESET*

1. No-Bake Spring Berry Pie
★ 4.8 🍽 8

2. Avocado Salmon Kale Salad
★ 4.0 🕑 25 min

3. Rolled Ice Cream
★ 4.2 🍽 4

4. Spring Quiche
★ 4.6

*Try "Alexa, Select Rolled Ice Cream"*

# Alexa TaskBot for DIY

https://www.amazon.science/academic-engagements/ten-university-teams-selected-to-participate-in-alexa-prize-taskbot-challenge

Alexa, help me fix a scratch on my car

Sure, I am happy to help. I learned how to do this on WikiHow.com Scratches are easy to remove. Ready to get started?

Sure!

Ok, Hose down the area to rinse off dirt from the surface. Take your hose and aim a concentrated jet directly onto the scratches.

Won't it make the scratches worse?

Don't worry, spraying the fine scratches won't make them worse or damage your paint

Ok, done. What next?

Scrub the scratched area using soapy water and a sponge.

I did that, and rinsed off the soap. Easy!

That's great! Now Choose a rubbing compound and polish designed for cars.

Oh no! I don't have a rubbing compound. Is there something else I could use?

That's OK. I found this on a Website eurotechcarcare.com "Old-fashioned shoe polish is a great homemade recipe for fixing scratches in your car's exterior.". Would like to use it or try something else?

https://www.youtube.com/watch?v=0nuCILb6VlI

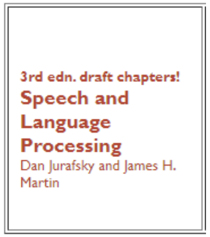
# Schedule

| Natural Language Processing and Search | | | |
|---|---|---|---|
| **Week** | **#** | **Lecture** | **In-class labs** |
| 16-Sep-22 | 1 | Introduction | **Project phase 1** |
| 23-Sep-22 | 2 | Text processing, PoS, NGRAMS, cosine distar | |
| 30-Sep-22 | 3 | Evaluation | |
| 7-Oct-22 | 4 | Language models | |
| 14-Oct-22 | 5 | Document categorization and ranking | **Project phase 2** |
| 21-Oct-22 | 6 | PoS and NE | |
| 28-Oct-22 | 7 | Word embeddings | |
| 4-Nov-22 | 8 | Contextual embeddings | |
| 11-Nov-22 | 9 | Question answering | **Project phase 3** |
| 18-Nov-22 | 10 | Expalinable NLP | |
| 25-Nov-22 | 11 | Project tips and feedback | |
| 2-Dec-22 | 12 | Computational Ethics for NLP | |
| 9-Dec-22 | | | |
| 16-Dec-22 | | | |

# References

- Slides and articles provided during classes.

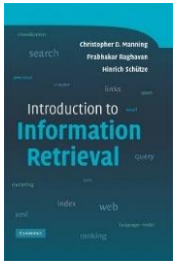
- Books:



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/](https://web.stanford.edu/~jurafsky/slp3/)



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

# Lecturers

- Lectures: João Magalhães (jmag@fct.unl.pt)
- Labs: David Semedo (df.semedo@fct.unl.pt)
- When sending messaging lecturers always identify the course name and your group ID.

# Course grading

- The course has two mandatory components:
  - Laboratory (groups of 3 students): 60% **(minimum grade > 10)**
    - Three phases, 20% per phase
  - Theoretical part (1 test or 1 exam): 40% **(minimum grade > 8)**

- Theory test/exam:
  - 1 Test or 1 Exam (dates do 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.

# Lab submission dates

- **Phase 1: Basic language models (20%)** (13 October)
  - Basic notions of text processing and similarity
  - Language models
  - Evaluation

- **Phase 2: Info. extraction and search (20%)** (10 November)
  - Entities extraction
  - Learning to rank documents

- **Phase 3: Neural language models (20%)** (8 December)
  - Word embeddings and context embeddings
  - Self-attention

# Summary

- Context
- Objectives and plan
- Grading
- Labs