ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Εισαγωγή

Ακαδημαϊκό Έτος 2023-2024

Ανάκτηση Πληροφορίας (Information Retrieval) - (IR)

είναι η εύρεση αντικειμένων (κυρίως εγγράφων) από μεγάλες συλλογές τα οποία ικανοποιούν μια ανάγκη για πληροφόρηση

Συνήθως αναζήτηση με λέξεις κλειδιά

Εφαρμογές;

Γιατί να μας ενδιαφέρει;

Παλιότερα,

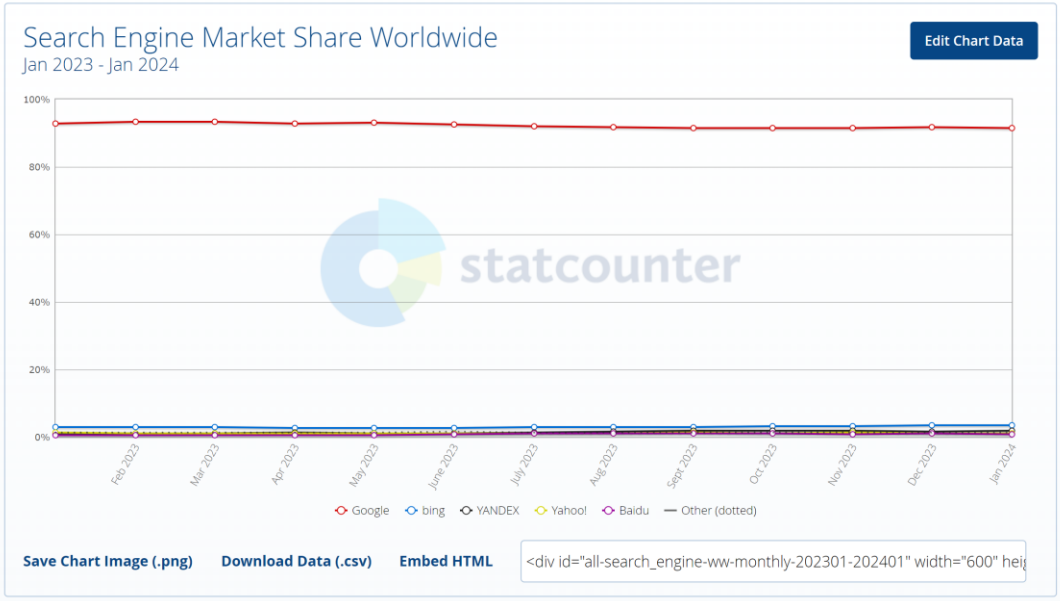
Βιβλιοθηκονόμους, βοηθούς νομικών επαγγελμάτων κλπ;

ISBN: 0-201-12227-8
Author: Salton, Gerard
Title: Automatic text processing: the transformation, analysis, and retrieval of information by computer
Editor: Addison-Wesley
Date: 1989
Content: <Text>

external attributes (metadata) and internal attribute (content)
Search by external attributes = Search in DB
IR: search by content

Εφαρμογές

Search Engines



Last 12 months, worldwide, all platforms

https://gs.statcounter.com/search-engine-market-share

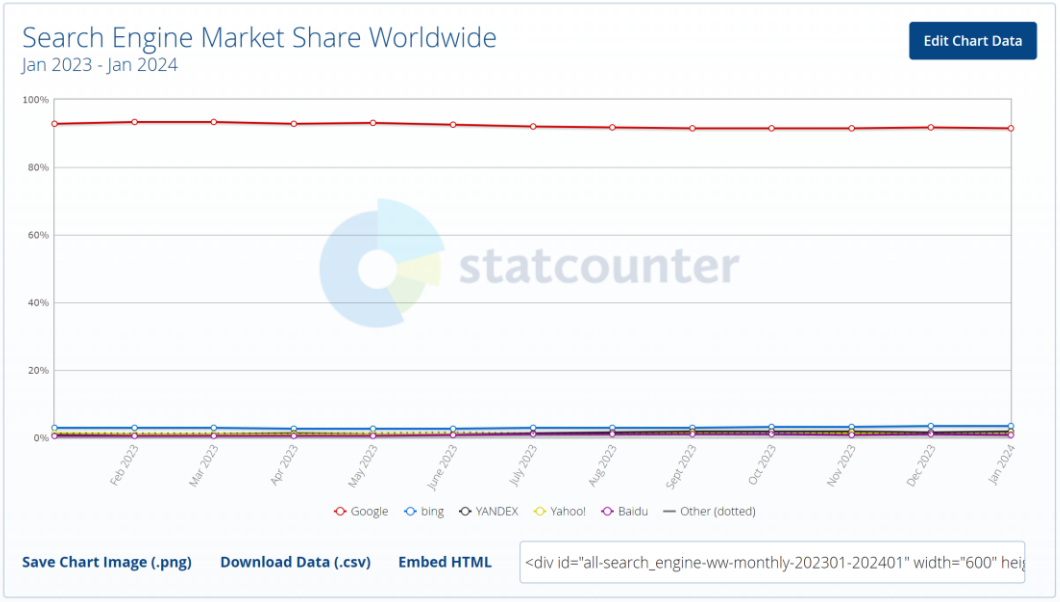
Google: 3.5 billion searches per day

- ***Bing:*** Microsoft
  Build on previous search engines (MSN Search, Windows Live Search, Live Search),
  In 2023, faster than Google to release an AI Chatbot service (Bing Chat) based on GPT4, but despite initial acceptance, market share remained at low levels.

- ***Yandex:*** ρωσική πολυεθνική τεχνολογική εταιρεία που ειδικεύεται σε υπηρεσίες και προϊόντα σχετικά με το Διαδίκτυο και θεωρείται η μεγαλύτερη εταιρεία τεχνολογίας της *Ρωσίας*. Λειτουργεί την μεγαλύτερη μηχανή αναζήτησης στη Ρωσία με μερίδιο αγοράς περίπου *65%*.

Πηγή: Wikipedia

Search Engines



Last 12 months, worldwide, all platforms

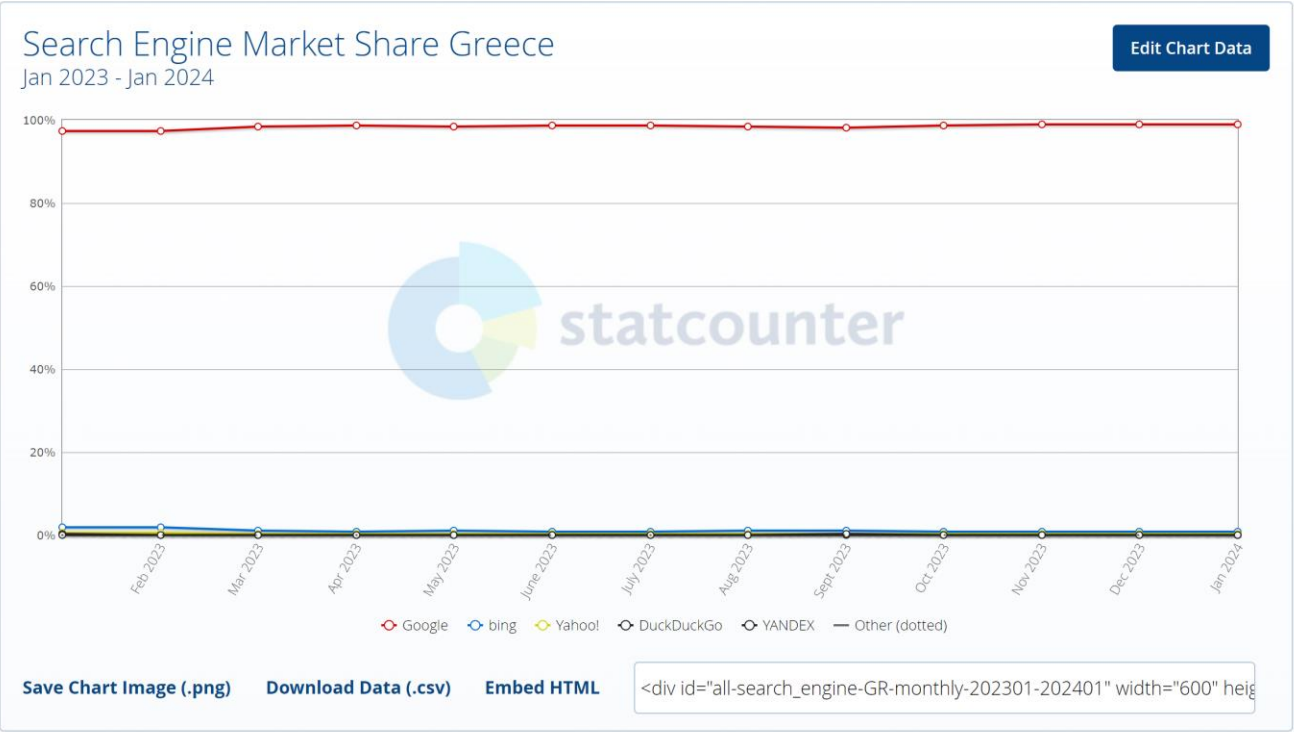
https://gs.statcounter.com/search-engine-market-share

- ***Baidu:*** 2η μεγαλύτερη μηχανή αναζήτησης στον κόσμο, και κατέχει το 76.05% του μεριδίου αγοράς στην αγορά μηχανών αναζήτησης της *Κίνας*.

- ***DuckDuckGo*** είναι διαδικτυακή μηχανή αναζήτησης που δίνει έμφαση στην *προστασία της ιδιωτικής ζωής* των χρηστών της και στην αποφυγή του "*φίλτρου φυσαλίδας*" των εξατομικευμένων αποτελεσμάτων αναζήτησης.
  - *Πηγή: Wikipedia*

Google	bing	Yahoo!	DuckDuckGo	YANDEX	Petal Search
98.83%	0.83%	0.21%	0.05%	0.04%	0.01%

Search Engine Market Share in Greece - January 2024



Petal search:

Developed by Huawei

https://gs.statcounter.com/search-engine-market-share

Desktop search

IR tools are designed to find information on a PC, including web browser history, e-mail archives, text documents, images, video, etc

Εφαρμογές

Email search

Social search

Enterprise search

helps people in an organization find the information they need to perform their jobs-- data extracted from inside the business, along with external data sources like document management systems, databases, etc

Domain specific search: Legal information retrieval, Digital libraries

Διαφορετικές απαιτήσεις ανάλογα με την εφαρμογή

Κατηγορίες εφαρμογών

- Στο web/διαδίκτυο (search engines)

  Δισεκατομμύρια έγγραφα σε εκατομμύρια υπολογιστές.

  Συλλογή εγγράφων, κλίμακα, διάταξη αποτελεσμάτων, ..

- Προσωπική ανάκτηση πληροφορίας

  (στον προσωπικό υπολογιστή, email, κλπ)

  Διαφορετικά είδη αρχείων, light-weight, maintenance-free, …

- Σε επίπεδο επιχείρησης, οργανισμού (enterprise, institutional)
- Αναζήτηση ειδικού σκοπού (domain-specific search) – πχ ερευνητικά άρθρα σε βιοχημεία

Ορισμός

Ανάκτηση Πληροφορίας (Information Retrieval) - (IR)

- είναι η εύρεση αντικειμένων κυρίως εγγράφων (**documents**) αδόμητης φύσης (*) (unstructured) που συνήθως έχουν τη μορφή κειμένου (**text**)
- από μεγάλες συλλογές (συνήθως αποθηκευμένες σε υπολογιστές)
- τα οποία ικανοποιούν μια ανάγκη πληροφόρησης (**information need**)

(*) όχι ακριβώς!

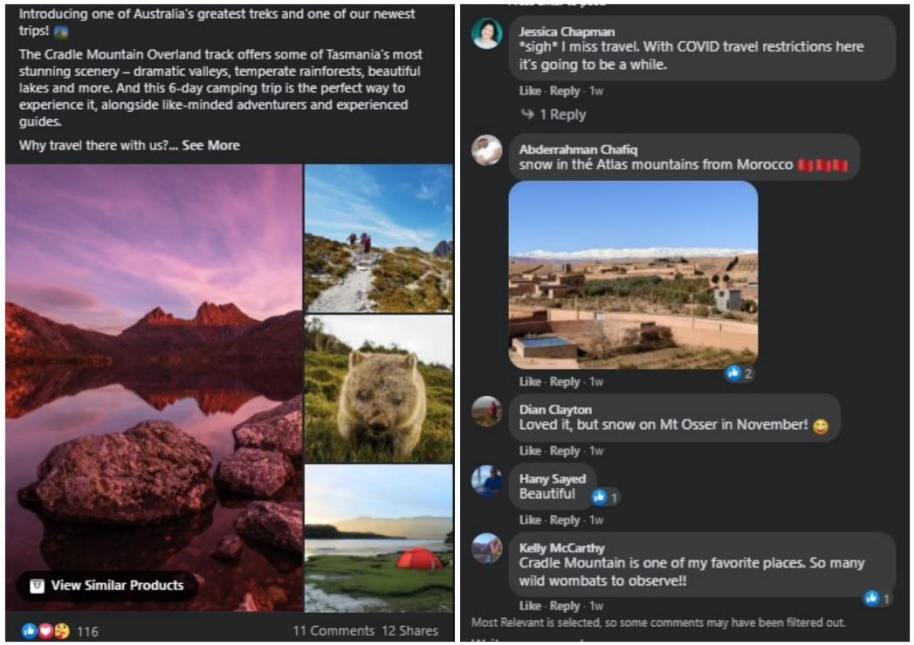
Αδόμητα δεδομένα

- Τυπικά αναφέρεται σε *ελεύθερο κείμενο*
- Επιτρέπει
  - Ερωτήματα με λέξεις κλειδιά (keyword) με πιθανούς τελεστές
  - Ποιο περίπλοκες ερωτήσεις για έννοιες: π.χ.,
    - Βρες όλες τις web σελίδες για την απελευθέρωση των Ιωαννίνων
- Κλασσικό μοντέλο για αναζήτηση σε έγγραφα κειμένου

structured

	A	B	C	D	E	F	G
1	Purchase ID	Last name	First name	Birthday	Country	Date of purchase	Amount of purchase
2	1	Davidson	Michael	04/03/1986	United States	10/12/2016	37
3	2	Vito	Jim	09/01/1994	United Kingdom	02/02/2016	85
4	3	Johnson	Tom	23/08/1972	France	02/11/2016	83
5	4	Lewis	Peter	18/10/1979	Germany	22/11/2016	27
6	5	Koenig	Edward	13/05/1983	Argentina	26/03/2015	43
7	6	Preston	Jack	16/06/1991	United States	06/11/2016	77
8	7	Smith	David	11/03/1965	Canada	15/11/2016	23
9	8	Brown	Luis	03/09/1997	Australia	03/07/2015	74
10	9	Miller	Thomas	07/01/1980	Germany	07/11/2016	13
11	10	Williams	Bill	26/07/1960	United States	20/11/2015	80
12	11	Gemini	Alexia	12/09/1995	Canada	11/03/2017	35
13	12	Bond	James	25/02/1975	United Kingdom	12/08/2017	40
14	13	Burgle	Patricia	01/12/1990	United States	18/01/2015	55
15	14	Reding	Michelle	07/04/1985	Canada	23/02/2017	28
16	15	Harvey	Billy	14/07/1971	United Kingdom	12/01/2016	41
17							

unstructured



https://www.altexsoft.com/blog/structured-unstructured-data/

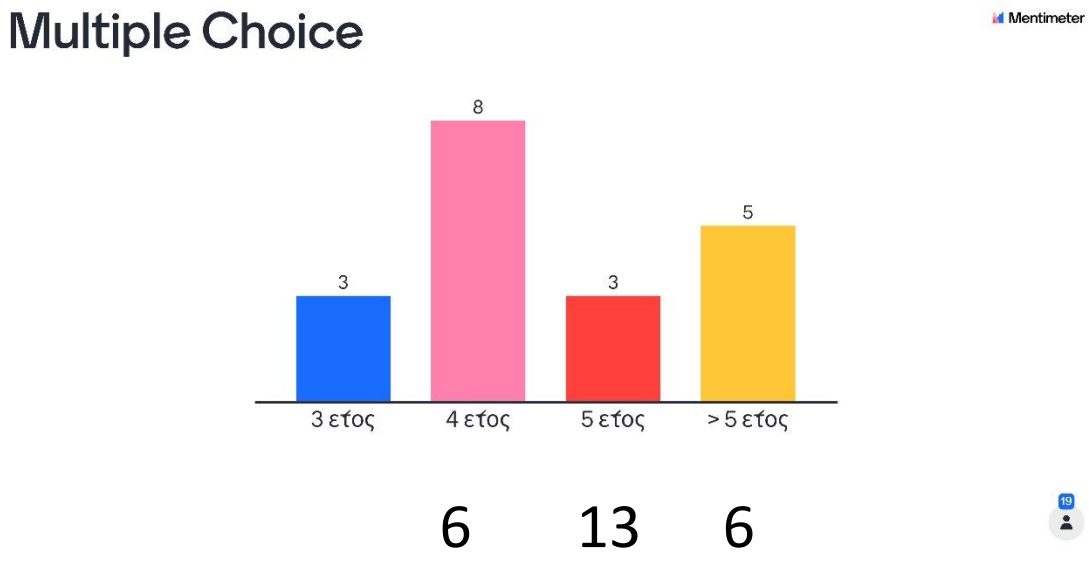
Ημιδομημένα δεδομένα

- Στην πραγματικότητα, δεν υπάρχουν αμιγώς μη δομημένα δεδομένα
  - π.χ., αυτή η διαφάνεια έχει διακριτές ζώνες όπως *Title* και *Bullets*
  - *Web pages?*
  - *Emails?*
- «Ημιδομημένη» αναζήτηση όπως:
  - *Title* contains ημιδομημένα AND *Bullets* contain αναζήτηση

… και βέβαια υπάρχει πάντα η γλωσσική δομή

Τι είναι η Ανάκτηση Πληροφορίας (Information Retrieval);

Poll 1: Ποιοι είστε;



LLMs (ChatGPT) και Ανάκτηση Πληροφορία

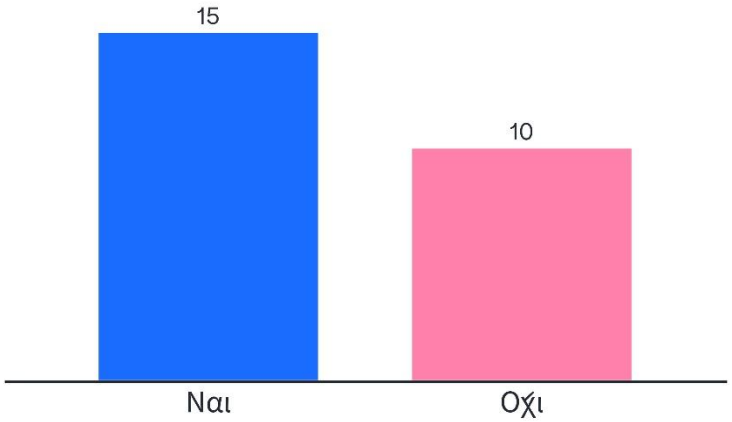
ChatGPT

Poll 2
Το έχετε χρησιμοποιήσει;

ChatGPT

Έχετε χρησιμοποιήσει το ChatGPT;

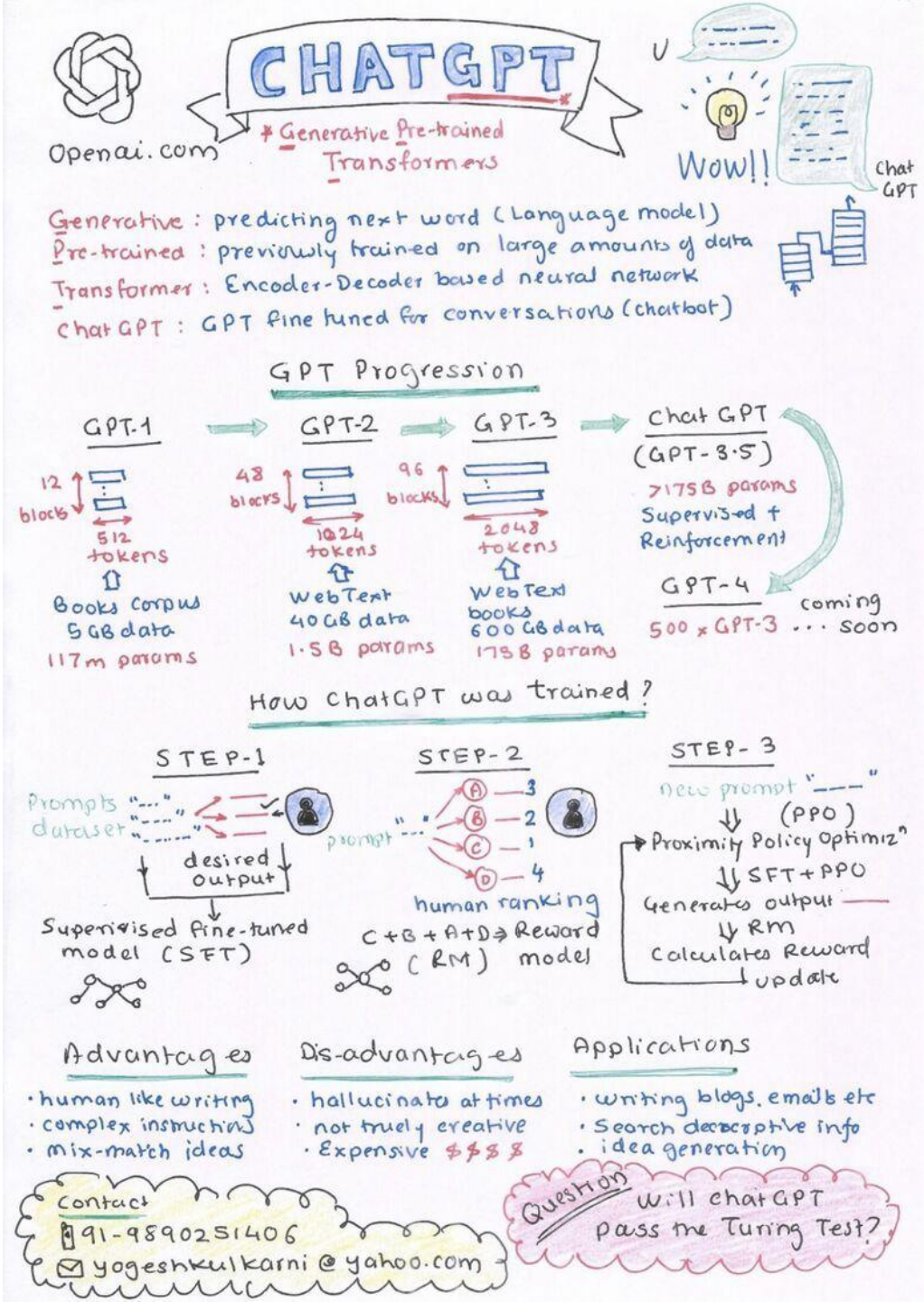
Mentimeter



20 1

ChatGPT

Τι είναι σε μια πρόταση


CHATGPT
Openai.com
* Generative Pre-trained Transformers
U
Wow!!
Chat GPT
Generative : predicting next word (Language model)
Pre-trained : previously trained on large amounts of data
Transformer : Encoder-Decoder based neural network
ChatGPT : GPT fine tuned for conversations (chatbot)
GPT Progression
GPT-1
12 blocks
512 tokens
Books corpus
5 GB data
117m params
GPT-2
48 blocks
1024 tokens
WebText
40 GB data
1.5B params
GPT-3
96 blocks
2048 tokens
WebText
books
600 GB data
175B params
Chat GPT
(GPT-3.5)
>175B params
Supervised + Reinforcement
GPT-4
500 x GPT-3
coming ... soon
How ChatGPT was trained ?
STEP-1
Prompts "..."
dataset "..."
desired output
Supervised fine-tuned model (SFT)
STEP-2
prompt "..."
A — 3
B — 2
C — 1
D — 4
human ranking
C+B+A+D ⇒ Reward (RM) model
STEP-3
new prompt "..."
(PPO)
Proximity Policy Optimiz"
SFT+PPO
Generates output
RM
Calculates Reward
update
Advantages
• human like writing
• complex instructions
• mix-match ideas
Dis-advantages
• hallucinates at times
• not truely creative
• Expensive $$$$
Applications
• writing blogs, emails etc
• Search descriptive info
• idea generation
contact
91-9890251406
yogeshkulkarni@yahoo.com
Question
Will chatGPT pass the Turing Test?

ChatGPT

Poll 2
Μέχρι 3 πιο θετικά
Μέχρι 3 πιο αρνητικά

ChatGPT

Μέχρι 3 πιο θετικά

Mentimeter

14

ChatGPT

Μέχρι 3 πιο αρνητικά

Συζήτηση

Λάθη
Λογοκλοπή (οικειοποιείται κείμενο άλλων)

ChatGPT και MYE003

Δύο οπτικές γωνίες/“disruptive” σε δύο άξονες

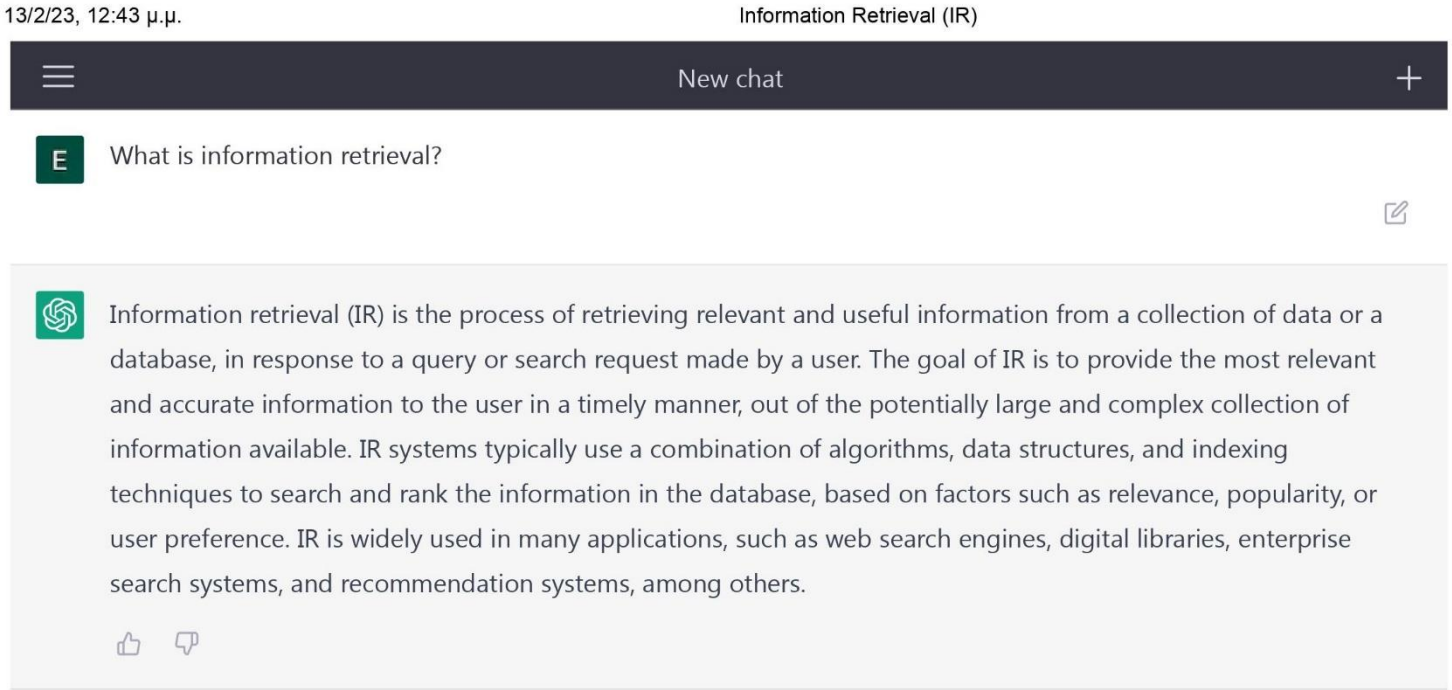
- Εκπαίδευση γενικά
- Στο αντικείμενο του μαθήματος: σχέση με αναζήτηση

ChatGPT και εκπαίδευση

Ως πηγή γνώσης (educational resource)

Ιδέες:

ChatGPT


13/2/23, 12:43 μ.μ.
Information Retrieval (IR)
New chat
E
What is information retrieval?
Information retrieval (IR) is the process of retrieving relevant and useful information from a collection of data or a database, in response to a query or search request made by a user. The goal of IR is to provide the most relevant and accurate information to the user in a timely manner, out of the potentially large and complex collection of information available. IR systems typically use a combination of algorithms, data structures, and indexing techniques to search and rank the information in the database, based on factors such as relevance, popularity, or user preference. IR is widely used in many applications, such as web search engines, digital libraries, enterprise search systems, and recommendation systems, among others.

ChatGPT και εκπαίδευση

Επίδραση στην αξιολόγηση των μαθητών/φοιτητών

Προσεγγίσεις:

1. Απαγόρευση

Πως:

Άλλου τύπου αξιολόγηση, αποκλεισμός πρόσβασης

2. Όχι απαγόρευση
   - Χρήση με αναφορά
   - Χρήση χωρίς αναφορά

The three models of satisfying information needs

	Structures queries	Search Engines	Large Language Models
Systems	**DBMS**	**Google search**	**ChatGPT**
Corpus	**Relational tables**	**Web+**	**Web texts, books, Wikipedia**
Input	**SQL**	**Keywords**	**Natural language**
Output	**Data in tables**	**Ranked sources**	**New content**

- Ο “κόσμος” παράγει περισσότερο από **2 exabytes** νέας πληροφορίας το χρόνο, 90% της οποίας είναι σε ψηφιακή μορφή και με 50% ετήσια αύξηση

- Τι θα γίνει με το κείμενο που παράγει το ChatGPT και γενικά από AI tools
  - Θα γίνει index?

Retrieval augmented generation (RAG)

- Rather than relying on a fixed LLM to deliver the answer to a query, if we first find relevant documents (online or elsewhere) and then use an LLM to process the query and the documents into an answer, this could provide an alternative to current web search.

- Executing this efficiently and at scale would be complex, but the effect would be akin to having an LLM do a web search and summarize the results.

Πίσω στο ΜΥΕ003: τι θα μάθουμε;

Αναζήτηση

- Μοντελοποίηση
- Επεξεργασία (φυσικές γλώσσες)
- Βασικές δομές
- Αξιολόγηση

Μηχανές Αναζήτησης

Δημιουργία της μηχανής αναζήτησης

- Δημιουργία συλλογής (αν δεν υπάρχει):
  - crawl, scrap, use specific APIs (social media)
- Ανάλυση του κειμένου
  - Ποιοι θα είναι οι όροι, περιλαμβάνει και γλωσσολογική επεξεργασία
- Κατασκευή ευρετηρίων
  - Η πιο απλή μορφή: για κάθε όρο σε ποια έγγραφα εμφανίζεται

(επεκτάσεις πχ με συχνότητα εμφάνισης, πληροφορία θέσης, συμπίεση, κλπ)

Μηχανές Αναζήτησης

Λειτουργία

Ο χρήστης υποβάλει ένα ερώτημα

Επεξεργασία του ερωτήματος

Χρήση του ευρετηρίου για να βρούμε (retrieve) τα συναφή έγγραφα και να τα διατάξουμε με βάση τη συνάφεια

Πως ορίζουμε τη συνάφεια:

- Boolean model: υπάρχει, δεν υπάρχει ο όρος
- tf-idf (βασισμένη σε κείμενο)
- Source importance: Pagerank (οι συνδέσεις), clickthrough (γενικά traffic), document age, authority (wikipedia)
- Personalization, contextualization
- User intent
- Learning to rank

Μηχανές Αναζήτησης

Αξιολόγηση

Δε μας ενδιαφέρει η διάταξη:
Recall/Precision/AUC

Μας ενδιαφέρει η διάταξη:
MAP, NDCG

Word embeddings + ML

Διανυσματικές αναπαραστάσεις όρων
word2vec (σε βάθος)
Transformers, LLMs (σε υψηλό επίπεδο)

Lucene

Open-source software για τη δημιουργία μηχανών αναζήτησης

Μηχανές Αναζήτησης

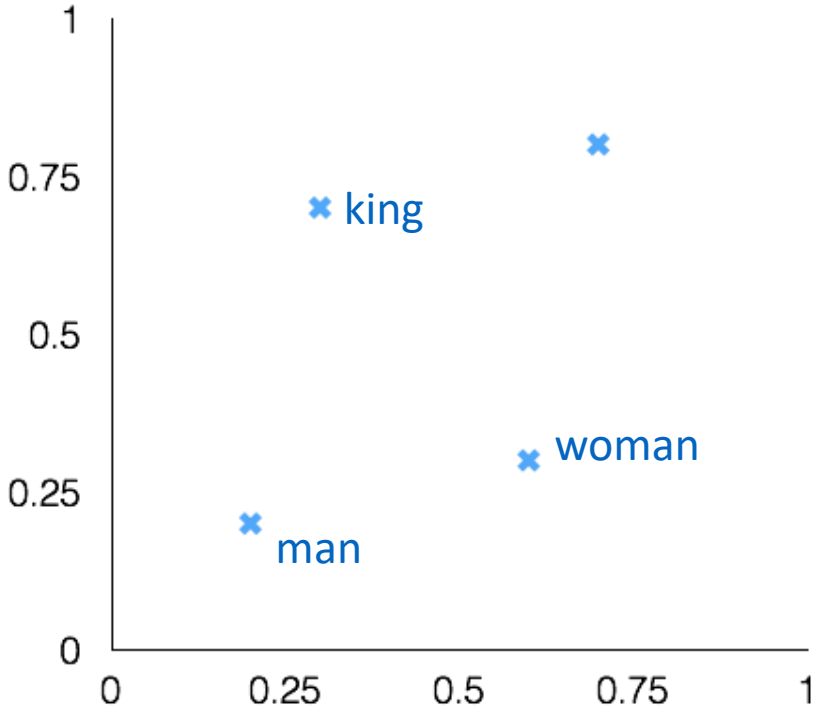
Lucene

Open-source software για τη δημιουργία μηχανών αναζήτησης

Μερικά Στοιχεία Μηχανικής Μάθησης

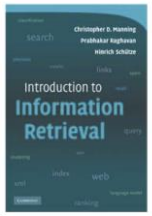
Word embeddings

+	king	[0.30 0.70]
−	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



Διαδικαστικά

- Ιστοσελίδα
- Βιβλίο
  - Cristopher D. Manning, Prabhakar Raghavan and Hinrich Schutze. *Εισαγωγή στην Ανάκτηση Πληροφοριών*, Εκδόσεις Κλειδάριθμος
  - Η αγγλική έκδοση διαθέσιμη δωρεάν



- Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Ανάκτηση Πληροφορίας*, 2η Έκδοση, Εκδόσεις Τζιόλα

Διαδικαστικά

- Βαθμολογία (μπορεί να αλλάξει):
  - Εργασία (έως 2 άτομα) – σε φάσεις: 50%
    - Προαιρετική εργασία – ανεξάρτητη μελέτη ML μοντέλων (20%) ή σύνδεση τους με τη μηχανή αναζήτησης
  - Τελικό Διαγώνισμα: 50% (αν όχι την προαιρετική εργασία)
    30% (αν την προαιρετική εργασία)

- Η εργασία δεν «κρατιέται»
- Για να περάσετε το μάθημα, βαθμός διαγωνίσματος ≥ 4

Λίγα λόγια για την εργασία

Μηχανή αναζήτησης

Lucene (solar)

2 καταληκτικές ημερομηνίες

Θεματικό (πέρυσι αναζήτηση τραγουδιών (στίχοι)

ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Εισαγωγή

Ακαδημαϊκό Έτος 2023-2024