# Pristopi k iskanju informacij - od klasičnih do jezikovnih modelov

Zaključna predstavitev pri predmetu raziskovalni seminar

Andrej Erjavec

# Kazalo vsebine

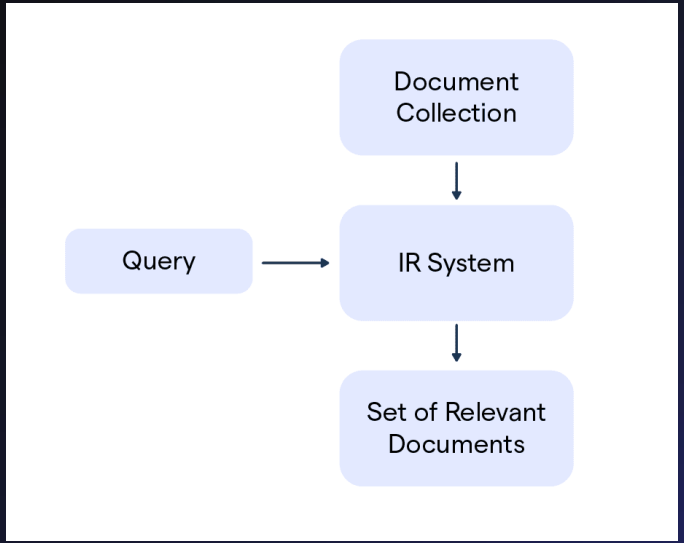
# Uvod

- Naraščajoča količina digitalno shranjenih podatkov
- Iskanje informacij (IR) – veja računalniške znanosti
- >80% vseh podatkov je nestrukturiranih
- Ni vnaprej definirane sheme
- Več modelov za procesiranje besedil

# Iskanje informacij - definicija

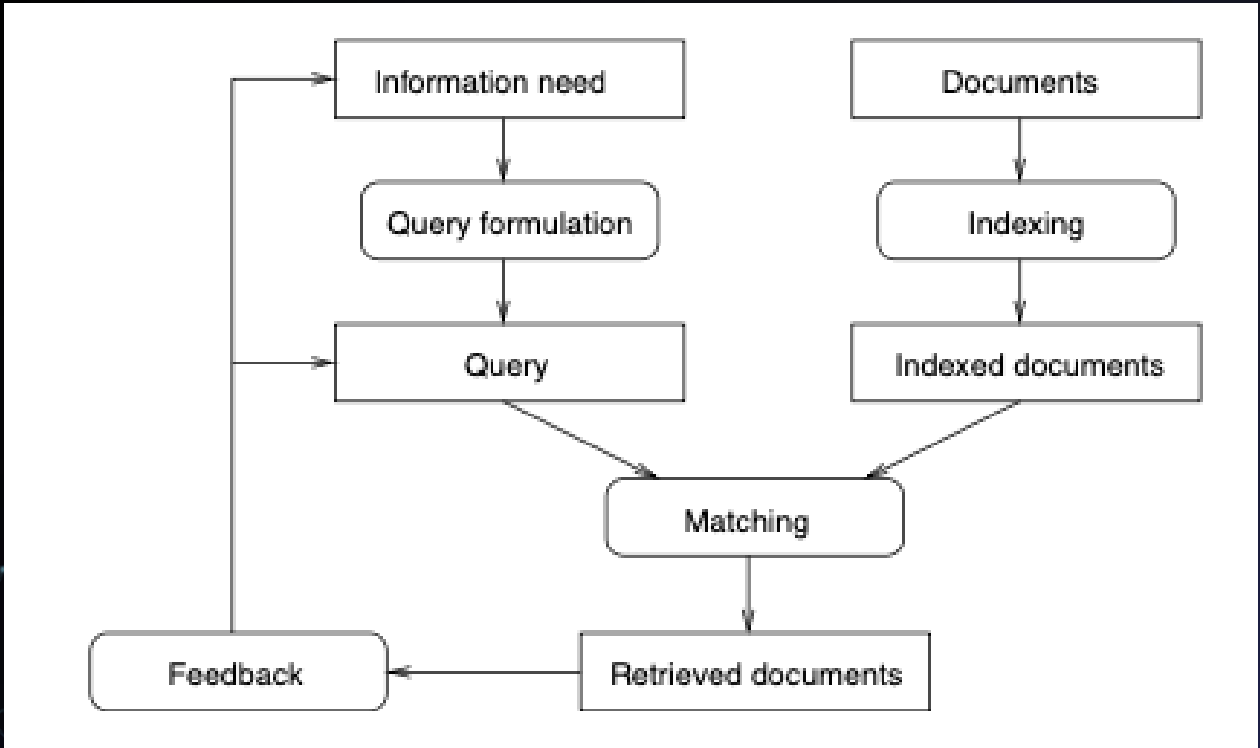
- Veja računalniške zananosti
- Hranjenje, predstavitev, iskanje informacij
- Tesno povezano z NLP in rudarjenjem besedil

- **Cilj: pridobiti množico relavantnih dokumentov glede na uporabnikovo poizvedbo.**




Poenostavljena shema procesa iskanja informacij
(vir: https://botpenguin.com/glossary/information-retrieval)

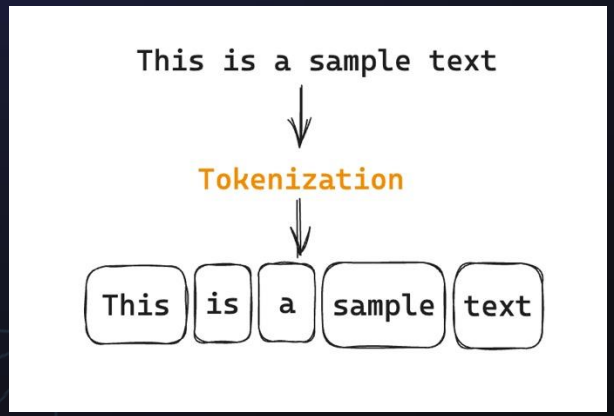
# Proces iskanja informacij



# Proces iskanja informacij (1)

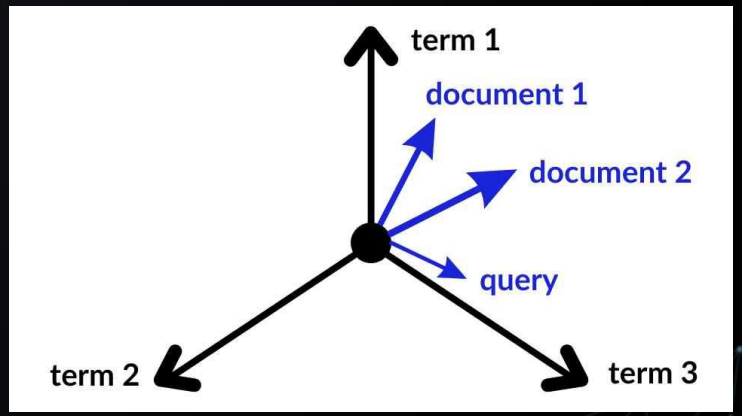
## 1. Predobdelava

- A ➔ a
- Tokenizacija
- Odstranjevanje pomensko manj pomembnih besed
- Stemming (Studying → Study)
- Lematizacija (Caring → Care)




## 2. Indeksiranje

- Pretvorba tokenov
- Inverzni indeks
- Vektorske predstavitve
- Vgnezditve



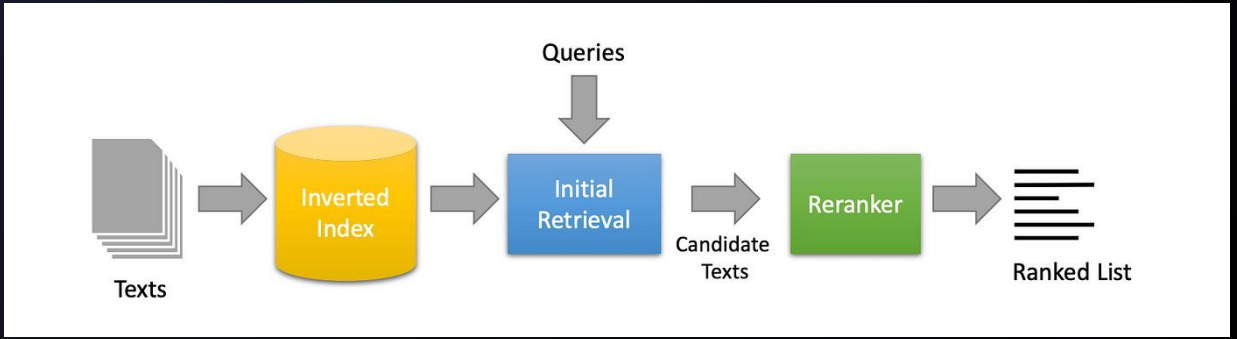
# Proces iskanja informacij (2)

## 3. Iskanje

- Poizvedba
- Rezultat je nabor ustreznih dokumentov
- Ad-hoc in regex neučinkovita
- Enostavni algoritni (BM25)
- Rezultat je naurejena množica dokumentov
- (N=1000)

## 4. Rangiranje

- Urejanje glede na relavantnost
- Kompleksnejši modeli
- Rezultat je urejena množica dokumentov
- (N=20)




Fazi iskanja in rangiranja
(vir: https://neerajku.medium.com/document-ranking-using-bert-a4b00eb258c4)

# Izzivi pri iskanju informacij

## Semantična neskladja

Uporabnik ne pozna structure ter besedišča dokumentov, zato prihaja do neskladij med poizvedbo ter vsebino dokumentov.

## Polisemantičnost

Beseda ima lahkospremenjen pomen glede na kontekst. Model mora prepoznati kontekst okoli besede za določanje pomena.

## Semantika besedila

Podatko razpršeni po besedilu, relacije med koncepti definirane s semantiko + nepoznavanje strukture.

## Problem neskladja v besedišču

*angl. Vocabulary mismatch problem.*
Različni uporabniki za isti koncept uporabljajo različna poimenovanja.

# Modeli za iskanje informacij

## Boolovi modeli

- Boolova logika

## Probabilistični model

- Binarni vektorji
  BM25

## Vektorski modeli

- Večdimanzionalni vektorksi proctor
- Gosti vektorji
- Bag of Words
- TF-IDF

## Vgnezditve

- Redki vektorji
- Word2Vec
- GloVe

## Jezikovni modeli

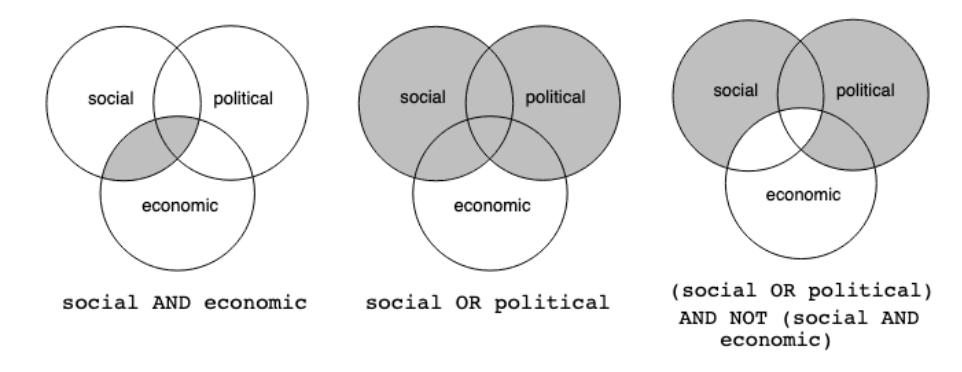
- Zajem konteksta
- BERT
- modeli GPT

# Boolov in razširjeni Boolov model

- Nizanje ključnih besed z logičnimi operatorji
- Lahko predstavimo z Venn diagramom
- Jasna struktura poizvedb
- Visoka stopnja nadzora nad pridobljenimi dokumenti
- Dokument ustreza ali pa ne ustreza (binaren rezultat)

**Razširjeni Boolov model = klasični + algoritem za rangiranje**

- Uteževanje poizvedb in dokumentov
- P-norm metrika
- Dokumenti rangirani po verjetnost, da so ustrezni



# Probabilistični model

- Predstavitev z binarnimi vektorji.
- Komponenta vektorja = izraz v dokumentu
- Predp: pojavljanje izrazov v besedilu je statistično neodvisno.

Za vsak dokument izračunamo $P(R \mid d, q)$
Verjetnost, da je dokument $d$ relavanten za poizvedbo $q$

BM25 (Best Match 25) uporabljen za iskanje in rangiranje.

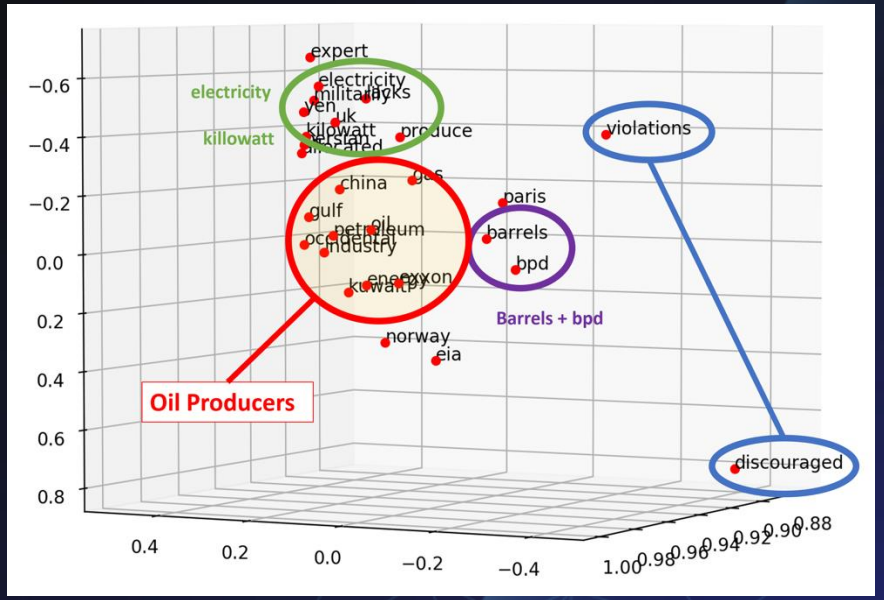
- tf, idf + normalizacija dolžine besedila (preprečevanje prevlade daljših besedil)

# Vektorski modeli

- Dokumenti = vektorji
- Večdimenzionalen Evklidski vektorski prostor
- Podbni izrazi so blizu v vektorskem prostoru

- Redka predstavitev
  (angl. sparse vecotrs; one-hot kodiranje)
  - dim = |korpus|
- Gosta predstavitev (angl. Dense vectors)
  - dim ≈ 300

- Kosinusna in evklidska razdalja



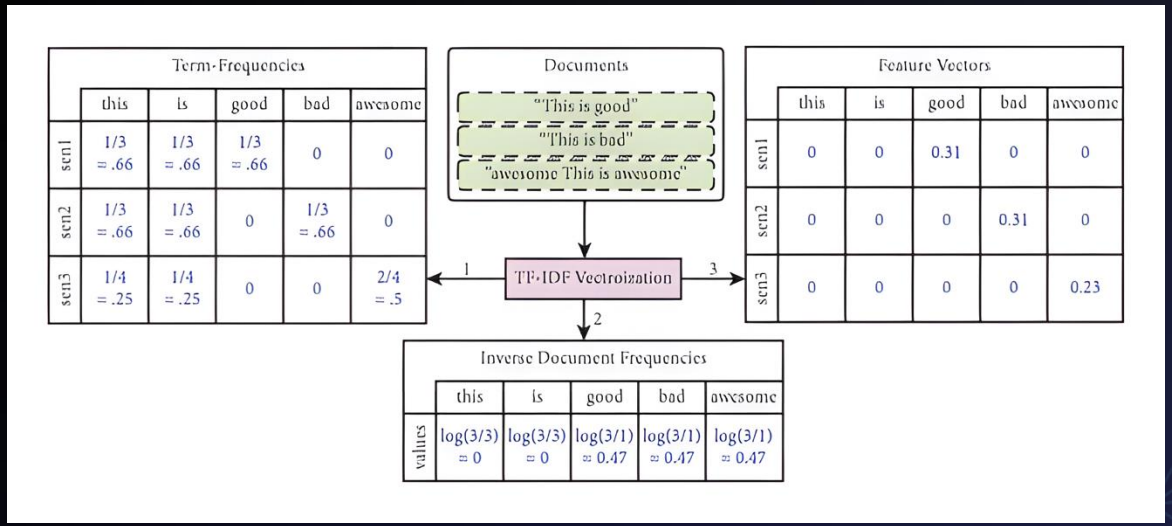

Vektorska predstavitev besed
(vir: https://aegis4048.github.io/understanding_multi-dimensionality_in_vector_space_modeling)

# TF-IDF

- Tf – število pojavitev besede v dokumentu
- Idf – normalizacija pogostih izrazov

$$TFIDF = tf(t,d) * IDF(t)$$

\+ Učinkovit, enostaven za implementacijo
\− Linearna rast utži s pogostostjo izraza (prioritizacija pogostih izrazov)
\− Ni normalizacije dolžine dokumenta (BM25)



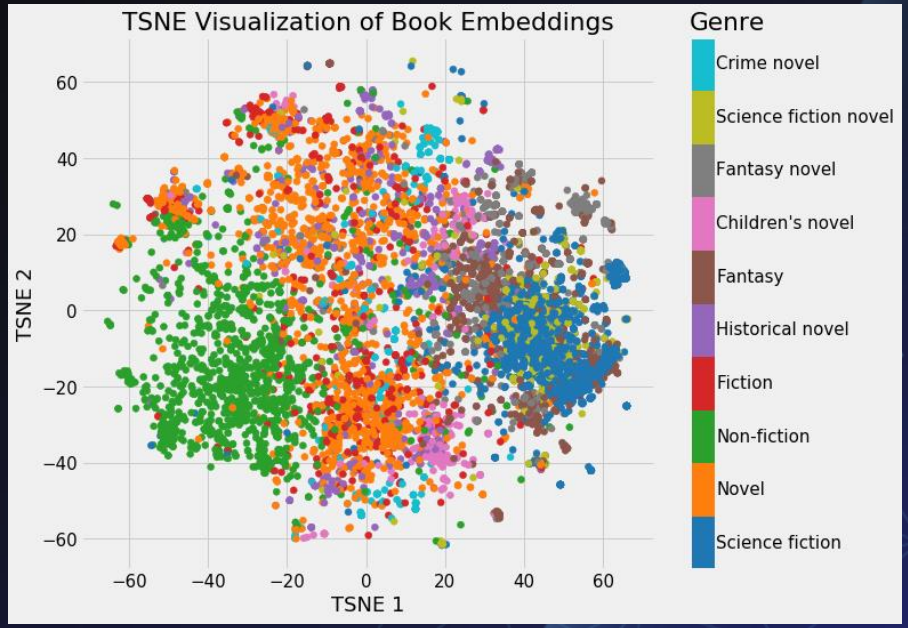

TF-IDF vektorizacija (vir: https://www.researchgate.net/figure/Illustration-of-TFIDF-vectorization_fig3_354354484)

# Vgnezditve

Do sedaj:

- redki vektorji ($\dim = |korpus|$)
- Ni zajema semantike besed

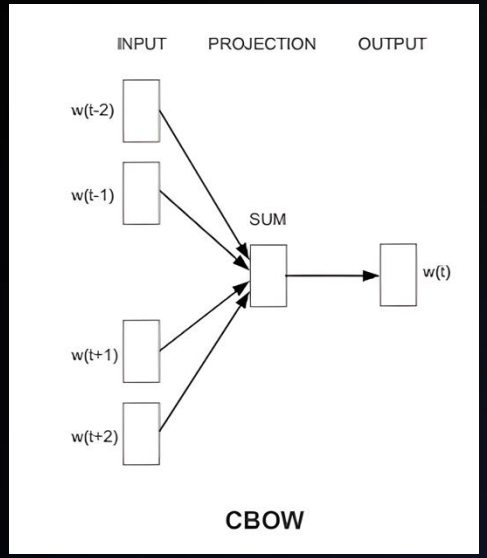
- Vgnezditve = predstavitev bedes z gostimi vektorji
- Semantično podobne besede (dokumenti) so si blizu v vektorskem prostoru

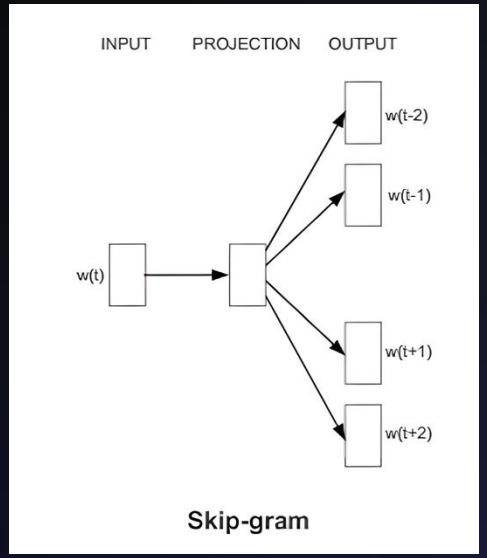


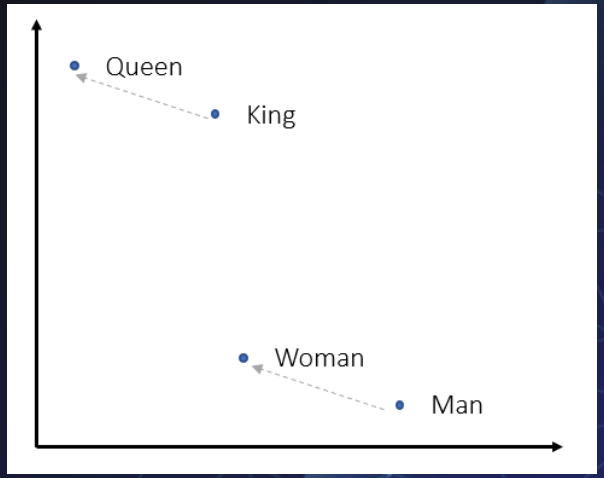

Grafična predstavitev vgnezditev (vir: https://devopedia.org/word-embedding)

# Word2Vec

- Google, 2013




Kontekst ➔ ciljna beseda
(Boljši za redke besede)

Beseda ➔ kontekst
(Hitrejši)

$$King - man + woman = queen$$

# GloVe

- Model nenedzorovanega učenja
- Izboljšava Word2Vec
- Analiza so-pojavitve besed (kolikokrat se beseda pojavi v kontekstu druge besede v korpusu)

- Boljši v hitrosti napovedovanja kot CBOW in Skip-gram na enakem korpusu

- Ne upošteva polisemije besed (identični vektorji)
- Ne prepoznava odvisnosti med besedami, pripon, predpon

# Jezikovni modeli - BERT

- Razvoj globokega učenja in transformerskih modelov
- Modeli kontekstualnih vgnezditev

- Google, 2018
- Jedro Google spletnega iskalnika
- Dvosmerno branje besedila – zajem konteksta v obeh smereh od ciljne besede
- Predhodno treniran (pre-trained) na besedilih Wikipedia

- Pomen besede se dimamično prilagaja glede na kontekst (polisemija)
- Številne izpeljake (Docbert – klasifikacija dokumentov)

## Maskirano jezikovno modeliranje

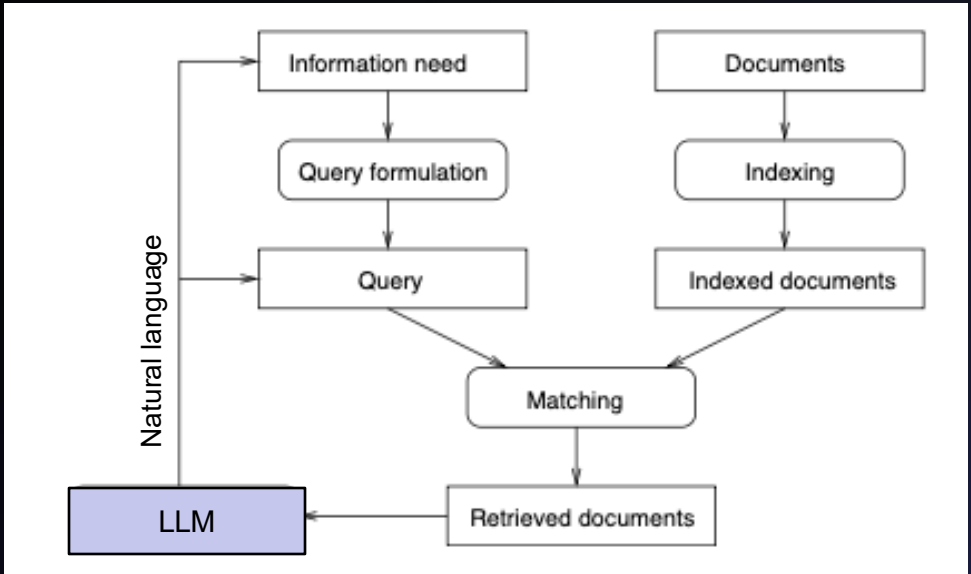
Naključno maskiranje besed in učenje rekonstrukcije iz širšega konteksta.

## Napovedovanje naslednjega stavka

Ali si dva stavka logično sledita

# Jezikovni modeli - GPT

- Generative pre-trained transformers
- Generacija tekočega, semantično smiselnega besedila
- Običajno niso uporabljeni za vektorsko predstavitev




Integracija LLM v proces iskanja informacij

# Zaključek

- Velika količina digitalno shranjenih dokumentov ➔ potreba po učinkovitem iskanju informacij
- Skozi leta razvoj v kompleksno znanstveno področje.
- Od klasičnih do jezikovnih modelov, nevronske mreže, globoko učenje

# Hvala za pozornost!

Vprašanja?

# Reference


[1] B. Mitra and N. Craswell, “Neural Models for Information Retrieval,” May 03, 2017, *arXiv*: arXiv:1705.01509. doi: 10.48550/arXiv.1705.01509.

[2] Y. Zhu *et al.*, “Large Language Models for Information Retrieval: A Survey,” Sep. 04, 2024, *arXiv*: arXiv:2308.07107. doi: 10.48550/arXiv.2308.07107.

[3] K. A. Hambarde and H. Proença, “Information Retrieval: Recent Advances and Beyond,” *IEEE Access*, vol. 11, pp. 76581–76604, 2023, doi: 10.1109/ACCESS.2023.3295776.

[4] G. Salton, E. A. Fox, and H. Wu, “Extended Boolean information retrieval,” *Commun. ACM*, vol. 26, no. 11, pp. 1022–1036, Nov. 1983, doi: 10.1145/182.358466.

[5] P. Bafna, D. Pramod, and A. Vaidya, “Document Clustering: TF-IDF approach”.

[6] D. Hiemstra and A. P. de Vries, “Relating the new language models of information retrieval to the traditional retrieval models”.

[7]

A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, “Text Clustering with Large Language Model Embeddings,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100–108, Dec. 2025, doi: 10.1016/j.ijcce.2024.11.004.

[8]

M. Sanderson and W. B. Croft, “The History of Information Retrieval Research,” *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444–1451, May 2012, doi: 10.1109/JPROC.2012.2189916.

[9]

K. W. Church, “Word2Vec,” *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, Jan. 2017, doi: 10.1017/S1351324916000334.

[10]

N. T. James and R. Kannan, “A Survey on Information Retrieval Models, Techniques and Applications,” *IJARCSSE*, vol. 7, no. 7, p. 16, Sep. 2017, doi: 10.23956/ijarcsse.v7i7.90.

[11]

Q. Jiao and S. Zhang, “A Brief Survey of Word Embedding and Its Recent Development,” in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China: IEEE, Mar. 2021, pp. 1697–1701. doi: 10.1109/IAEAC50856.2021.9390956.