Jezikovna primerjava dveh romanov – z jezikovnim modelom *Google Gemini*

Raziskovati želimo osnovne lastnosti povedi (npr. dolžino povedi), nekatere oblikoskladenjske vidike besed in besednih zvez.

Besedili:

Mark Twain: Tom Sawyer

Franz Kafka: Prozess

Kje dobim besedilni datoteki ?

https://raw.githubusercontent.com/tpetric7/raj2022-book/refs/heads/master/data/books/tom.txt

https://raw.githubusercontent.com/tpetric7/raj2022-book/refs/heads/master/data/books/prozess.txt

Naloge

1. Namesti potrebne pakete / module!
2. Naloži pakete v pomnilnik (import v Pythonu)!
3. Preberi datoteke (pot do mape? Oblika datoteke?) in shrani besedila v podatkovnem nizu (dataframe) z imenom po lastni izbiri (npr. df)! Stolpec »text« vsebuje besedilo, stolpec »doc_id« pa ime prebrane besedilne datoteke.
4. Razdeli besedila na povedi (sentence tokenize)!
5. Izračunaj povprečno dolžino povedi in standardni oklon za obe besedili (doc_id)!
6. Nariši diagram (npr. barplot), ki prikazuje povprečno dolžino povedi obeh besedil!
7. Razdeli povedi na besede (tokenize, tokens)!
8. Izloči nemške funkcijske in druge nezaželene besede (remove stopwords)!
9. Preštej besede v obeh besedilih!
10. Nariši diagram (npr. barplot), ki prikazuje po deset najpogostejših besed iz obeh besedil!
11. Preštej dvočlenske besedne zveze (bigrame) v obeh besedilih!
12. Nariši diagram (npr. barplot), ki prikazuje po deset najpogostejših besednih zvez (ngramov) iz obeh besedil!
13. Lematiziraj besede v povedih obeh besedil (npr. z nltk, spacy)!
14. Nariši diagram (npr. barplot), ki prikazuje najpogostejše slovarske enote!
15. Besedne vrste: ...