

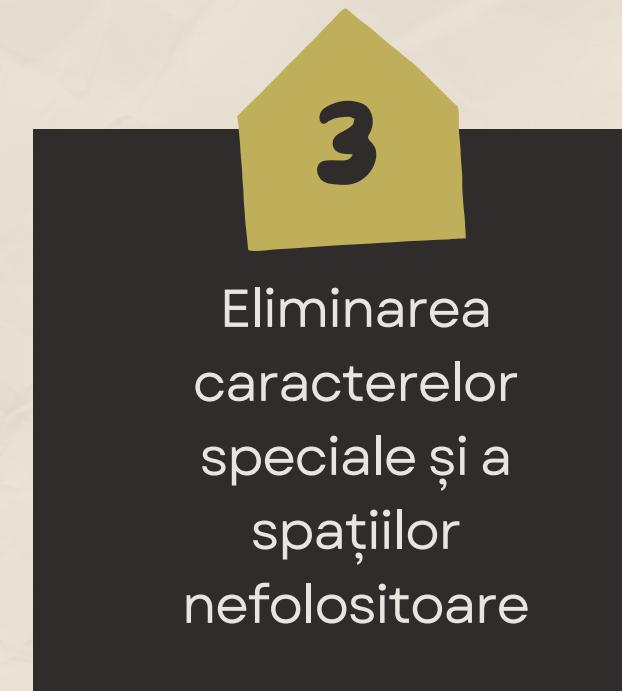
DATA MINING: BUILDING (A PART OF) WATSON

Acest proiect constă în dezvoltarea unui sistem de tip Information Retrieval ce are ca scop returnarea răspunsurilor pentru o serie de întrebări din cadrul unui concurs Jeopardy.

INDEXAREA

Se pornește de la un set de date format din 80 de fișiere, fiecare dintre acestea conținând o suita de pagini Wikipedia. Ulterior, fiecare fișier este prelucrat, astfel încât sunt extrase articole, care sunt transformate în documente ale librăriei Apache Lucene și adăugate în index. La final procesului, indexul este salvat pentru utilizare ulterioară.

Procesarea termenilor



INDEXAREA

Probleme întâlnite

- Existența unor tag-uri specifice Wikipedia, și anume: [[File:...]], [[Image:...]].
- Tagurile specifice Wikipedia pot apărea fără să fie închise corespunzător

Pentru rezolvarea acestor probleme, am utilizat o serie de expresii regulate pentru identificarea și eliminarea tag-urilor specifice.

```
data = data.replaceAll( regex: "\\\\[File:.*]]]", replacement: "" );
data = data.replaceAll( regex: "\\\\[File:.*]", replacement: "" );

data = data.replaceAll( regex: "\\\\[Image:.*]]]", replacement: "" );
data = data.replaceAll( regex: "\\\\[Image:.*]", replacement: "" );
```

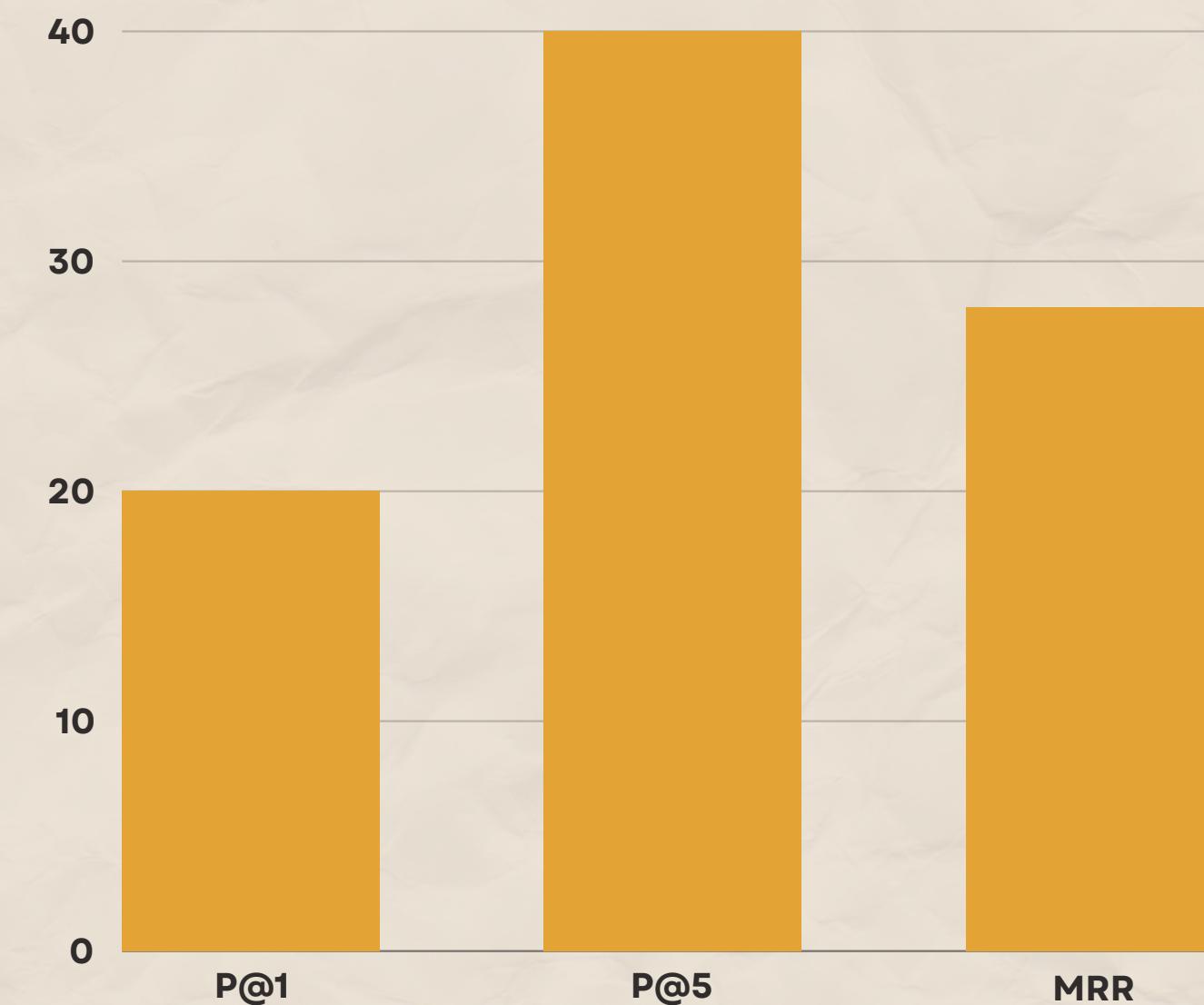
Construirea interogării

Pentru realizarea textului interogării, am alipit categoria și indicul, ulterior aplicând normalizarea asupra stringului obținut.

```
query(parser, searcher, clue: category.trim() + " " + clue, directoryReader.maxDoc());
```

MASURAREA PERFORMANTEI

- P@1 (precision at 1) - acest tip de metrică este relevant în contextul unui joc de tip Jeopardy, deoarece dorim să obținem răspunsul corect pentru fiecare întrebare de cultură generală.
- P@5 (precision at 5) - această metrică este relevantă în contextul în care vrem să folosim un element de Machine Learning pentru îmbunătățirea rezultatelor.
- MRR (Mean Reciprocal Rank) - în contextul unui concurs de tip Jeopardy, în care scopul este obținerea răspunsului corect pentru fiecare întrebare, capacitatea unui sistem de a returna cel mai bun răspuns este crucială. Metrica MRR măsoară exact această capacitate, făcând-o esențială pentru determinarea eficienței sistemului nostru.



Valorile din grafic reprezintă procente

ANALIZA ERORILOR

Rezultate

Din 100 de întrebări:

- 20 corecte
- 80 greșite (dintre care 20 cu răspuns în top 5)

Răspunsuri corecte cu un sistem atât de simplu

Însăși structura pe care o au întrebările concursului Jeopardy reprezintă unul dintre motivele pentru care un sistem atât de simplu poate da răspunsuri corecte. Aceste întrebări reprezintă, în multe cazuri, chiar titlul sau primele fraze din articolele Wikipedia.

Probleme întâlnite

Eliminarea caracterelor speciale poate face ca înțelesul unei fraze să se piardă.

Sistemul poate să interpreteze greșit întrebarea datorită unor construcții complexe sau a unor nuanțe specifice

În fraze complexe, sistemul identifică doar segmente, ignorând contextul general.

Ambiguitatea mare complică identificarea răspunsului corect, uneori agravată de adăugiri la indiciu.

IMBUNATATIREA EXTRAGERII

Pentru jocul Jeopardy, idealul este să obținem răspunsul corect pe prima poziția, scopul final fiind de a atinge un scor cât mai mare. Așadar, ne-am gândit să încercăm o metodă de optimizare pentru cele 20 de întrebări cu răspunsurile în top 5, în speranța de a îmbunătăți scorul anterior.

Tehnologia aleasă pentru această optimizare este ChatGPT (GPT 4.0). Pentru fiecare dintre cele 20 de întrebări am generat un string pentru interogările ulterioare. În acest string am solicitat reordonarea fișierelor date ca input, astfel încât răspunsul corect să poată fi identificat, oferind și indicul pentru acesta.

Rezultate după îmbunătățire:
P@1 - 0.4

ChatGPT a reușit să plaseze pe primul loc răspunsul corect la toate întrebările date.