

Proiect Data Mining (Watson)

1. Membrii echipei:

- Balo Alexandra-Elena
- Chirică Paula-Maria
- Duca Bianca-Ecaterina
- Titi Carmen-Andreea
- Slevoacă Viorica

2. Specializare: Baze de Date

3. Tema aleasa: Proiectul Default

1. Indexare și căutare

1.a) Crearea indexului

Clasa IndexCreator se ocupă de crearea indexului folosind Lucene. Pentru fiecare pagină Wikipedia se creează un document separat în index.

Am decis ca indexul să conțină două câmpuri de tipul TextField: "title" pentru titlul paginii și "content" pentru conținutul paginii.

Documentele din fisierul wikidata au fost preprocesate astfel:

- am eliminat stop-words folosind `org.apache.lucene.analysis.core.StopAnalyzer`

lista stop-words urilor cuprinde urmatoarele cuvinte:

```
"a", "an", " and ", " are ", "as", "at", "be"," but ", "by", "
for ", "if", "in", " into ","is", "it", "no", " not ", "of",
"on", "or"," such "," that ","the", " their ", " then ",
"there ", " these ", " they ", " this ", "to", " was ","
will"," with "
```

- am lematizat conținutul paginilor de Wikipedia folosind librăria StanfordCoreNLP
- am tokenizat textul folosind `Standard Tokenizer`

Crearea indexului a durat aproximativ 3 ore.

Probleme legate de paginile Wikipedia:

- în conținutul unor pagini se regăseau secvențele "[[" și "]]" care nu delimitau titlurile unor pagini

Exemplu: `</ref> and [[Vertebrate trachea|trachea]]s from human stem cells towards this end. Several [[artificial urinary bladder]]s have been grown in laboratories`

- existau rânduri care încep cu "[[" și se termină cu "]]", dar semnalau prezența unor imagini sau fișiere

Exemplu **1:**

`[[File:1964RepublicanPresidentialPrimaries.svg|thumb|300px|Republican primaries results by state In South Dakota and Florida, Goldwater finished second to "unpledged delegates", but he finished before all other candidates.]]`

Exemplu **2:**

`[[Image:Vegetation-no-legend.PNG|thumb|center|800px|Terrestrial biomes classified by vegetation|]]`

Am rezolvat asta considerând că rândul care conține titlul începe cu "[[" și se termină cu "]]" și nu conține secvențele "File" și "Image".

1. b) Căutarea

Am construit query-ul concatenând categoria și întrebarea, iar din întrebare am eliminat stop words. Am folosit un QueryParser pentru a căuta doar în conținutul paginii.

2. Măsurarea performanței

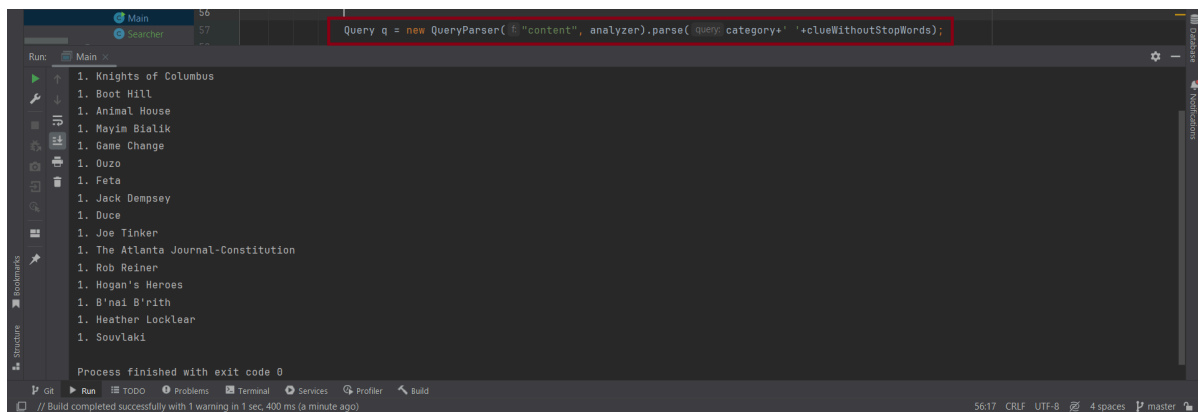
Am ales să folosim Precision at 1 pentru evaluarea performanței sistemului deoarece este relevant în situațiile în care ne interesează să oferim utilizatorului un singur răspuns

corect și acesta să fie pe prima poziție. Valoarea obținută este 0.16 (am obținut 16 răspunsuri corecte din 100).

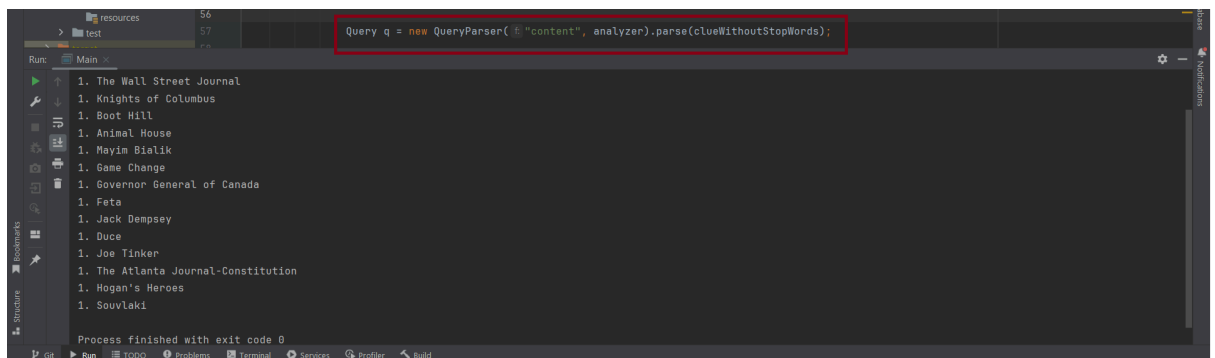
3. Analiza erorilor

a) În analiza rezultatelor vom discuta următoarele cazuri:

- ❖ am căutat doar în conținut și am folosit atât categoria cât și întrebarea pentru a construi query-ul, astfel am obținut 16 răspunsuri corecte din totalul de 100 de întrebări;



- ❖ am căutat doar în conținut și am folosit numai întrebarea pentru a construi query-ul, obținând 14 răspunsuri corecte din 100;



- ❖ am făcut căutarea pe titlu si conținut folosind un *MultiFieldQueryParser* cu query-ul format din categorie și întrebare, în acest caz am obținut 10 răspunsuri corecte

```
String[] fields = {"title", "content"};
MultiFieldQueryParser queryParser = new MultiFieldQueryParser(fields, analyzer);
Query q = queryParser.parse(query.category+" "+clueWithoutStopWords);
```

1. Boot Hill
1. Game Change
1. Feta
1. Jack Dempsey
1. Duce
1. The Atlanta Journal-Constitution
1. Rob Reiner
1. Hogan's Heroes
1. Ottoman Empire
1. Souvlaki

Process finished with exit code 0

❖ am făcut căutarea pe titlu si conținut cu query-ul format doar din întrebare și sistemul a răspuns corect la 7 întrebări

```
String[] fields = {"title", "content"};
MultiFieldQueryParser queryParser = new MultiFieldQueryParser(fields, analyzer);
Query q = queryParser.parse(clueWithoutStopWords);
```

1. Knights of Columbus
1. Game Change
1. Governor General of Canada
1. Jack Dempsey
1. Hogan's Heroes
1. Ottoman Empire
1. Souvlaki

Process finished with exit code 0


b) Considerăm că un sistem atât de simplu răspunde corect la unele întrebări deoarece întrebările conțin secvențe foarte asemănătoare cu conținutul paginii Wikipedia. Un exemplu ar fi:

Knights of Columbus

From Wikipedia, the free encyclopedia

Not to be confused with [Knights of St Columba](#).

The **Knights of Columbus (K of C)** is a global [Catholic fraternal service order](#) founded by [Fr. Michael J. McGivney](#) on March 29, 1882.^{[1][2]} Membership is limited to practicing Catholic men.^[3] It is led by Patrick E. Kelly, the order's 14th Supreme Knight.^{[3][4]} The organization is named after the explorer [Christopher Columbus](#).



SERVICE ORGANIZATIONS

Father Michael McGivney founded this fraternal society for Catholic laymen in 1882

Knights of Columbus

c) Legat de întrebările a căror răspuns este greșit, am observat câteva cauze importante, cum ar fi: unele întrebări au indicii care ajută la găsirea răspunsului corect, puse în ghilimele, cum sunt citate, fragmente importante, etc., dar acestea sunt eliminate la partea de lematizare; o altă problemă ar fi că mai multe pagini de Wikipedia conțin informații mult prea asemănătoare cu întrebarea, iar acest lucru duce la poziționarea pe primul loc a unui răspuns relevant care nu este corect, iar pe cel corect îl poziționează pe poziții mai slabe(Exemplu: "UCLA CELEBRITY ALUMNI: This woman who won consecutive heptathlons at the Olympics went to UCLA on a basketball scholarship" Pentru această întrebare poziționează răspunsurile ca "University of California, Los Angeles", "Cy Young (athlete)", "George Stanich", pe primele poziții, înaintea "Jackie Joyner-Kersey", care este răspunsul corect)

4. Îmbunătățirea rezultatelor

În urma primelor analize făcute, am putut observa că la 16 dintre întrebări, răspunsul corect nu se află pe prima poziție, dar se află în primele 10 răspunsuri, în majoritatea cazurilor, chiar în primele 5. Astfel, am folosit MRR (Mean Reciprocal Rank) care este o metrică utilizată pentru a evalua performanța unui sistem de căutare și de recomandare în găsirea răspunsurilor corecte într-o listă de rezultate, rolul său fiind să ofere o măsură a eficienței sistemului în prioritizarea rezultatelor relevante pentru utilizatori. MRR este calculat prin luarea inversului mediei pozițiilor primei răspunsuri corecte pentru fiecare întrebare sau căutare. Cu cât valoarea MRR este mai mare, cu atât sistemul este considerat mai bun, deoarece furnizează răspunsuri corecte mai aproape de partea de sus a

listei de rezultate. Am calculat MRR pentru cele 16 întrebări în care răspunsul corect se afla în top 10, astfel:

MRR=

$$\frac{1}{16} * \left(\frac{1}{3} + \frac{1}{4} + \frac{1}{2} + \frac{1}{6} + \frac{1}{2} + \frac{1}{6} + \frac{1}{5} + \frac{1}{5} + \frac{1}{8} + \frac{1}{4} + \frac{1}{3} + \frac{1}{3} + \frac{1}{2} + \frac{1}{7} + \frac{1}{6} + \frac{1}{3} \right) = 0.321$$

Deci, în medie, răspunsul corect este găsit la aproximativ a treia poziție în lista de rezultate pentru fiecare întrebare sau căutare. Am folosit ChatGPT pentru a reierarhiza primele 10 răspunsuri pe care le-a generat sistemul nostru, în funcție de probabilitatea de a fi răspunsul corect pe prima poziție ținându-se cont de informațiile pe care le deține chatul.

Pentru că am observat că la 16 dintre întrebări, sistemul pune în primele 10 răspunsuri pe cel corect, acestea 16 interogări i le-am dat la ChatGPT împreună cu răspunsurile generate, fără a specifica pe cel corect, astfel acesta l-a pus pe cel bun pe prima poziție în noua ierarhie în toate cazurile.

Un exemplu de întrebare dată de noi, cu ierarhizarea inițială:

I'M BURNIN' FOR YOU

France's Philip IV--known as "The Fair"--had Jacques De Molay, the last Grand Master of this order, burned in 1314

1. Philip IV of France
2. Knights Templar
3. Ferdinand IV of Castile
4. Victor Amadeus I, Duke of Savoy
5. Humbert de Pairaud
6. Louis X of France
7. Charles IV of France
8. James de Molay
9. Philip V of France
10. Tomar

Răspunsul de la ChatGPT cu noua ierhizare:

I'M BURNIN' FOR YOU:

1. Knights Templar
2. Philip IV of France
3. Jacques De Molay
4. Louis X of France
5. Charles IV of France
6. Philip V of France
7. James de Molay
8. Ferdinand IV of Castile
9. Humbert de Pairaud
10. Tomar

În urma acestei actualizări, performanța sistemului nostru a crescut cu 50%, valoarea metricii în momentul actual fiind 0.32, iar înainte era 0.16.

Discuția cu ChatGPT se poate consulta la linkul <https://chat.openai.com/c/4b03e3d5-b1a4-4c2a-9e6b-8fdee4c1c9d2>.