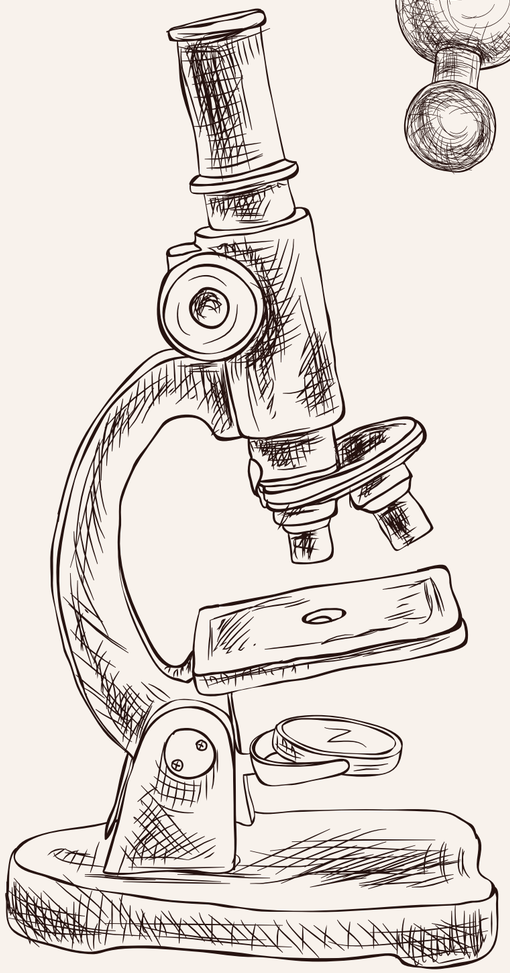
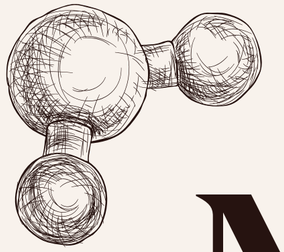
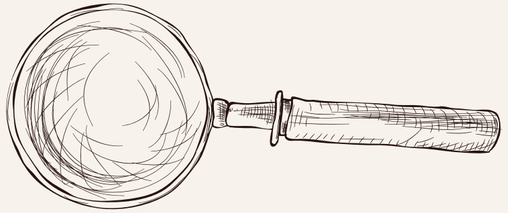
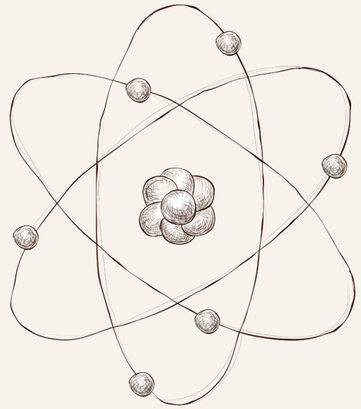
DATA MINING



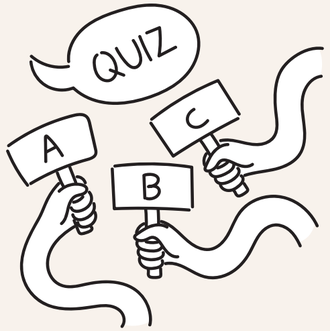


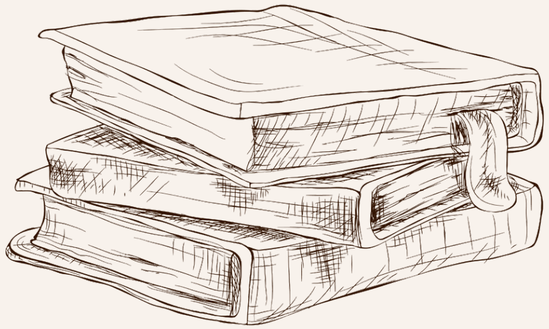




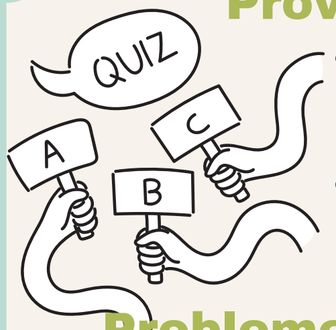
Bulding (a part of) Watson

Balo Alexandra
Chirica Paula
Duca Bianca
Titi Carmen
Slevoaca Viorica





Provocarea proiectului



- **Seturile de date utilizate** provin dintr-o arhiva ce contine mii de pagini Wikipedia, fiecare pagina avand titlu, categorie și conținut.
- Fiecare fișier conține mai multe pagini Wikipedia, care urmează a fi idexate în documente separate.

Probleme legate de parsarea paginile Wikipedia

- în conținutul unor pagini se regăseau secvențele "[[" și "]]" care nu delimitau titlurile unor pagini
- existau rânduri care încep cu "[[" și se termină cu "]]", dar semnalau prezența unor imagini sau fișiere

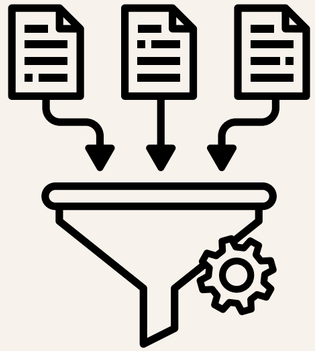
Am rezolvat asta considerând că rândul care conține titlul începe cu "[[" și se termină cu "]]" și nu conține secvențele "File" și "Image".

Structura

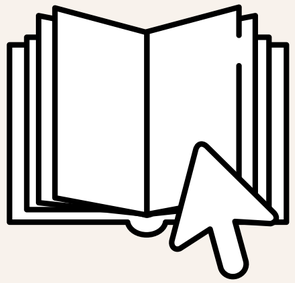
- Clasa IndexCreator se ocupă de crearea indexului folosind Lucene. Pentru fiecare pagină Wikipedia se creează un document separat în index.
- Am decis ca indexul să conțină două câmpuri de tipul TextField: "**title**" pentru titlul paginii și "**content**" pentru conținutul paginii.

Preprocesare

- am **tokenizat** textul, utilizand StandardTokenizer
- am eliminat **stop-words** folosind org.apache.lucene.analysis.core. StopAnalyzer
- am **lemmatizat** conținutul paginilor de Wikipedia folosind librăria StanfordCoreNLP



Construirea indexului



Pentru a realiza construirea indexului am procedat astfel:

1. Am parcurs toate cele 80 de fișiere și le-am citit conținutul
2. Am analizat problemele legate de conținut și le-am rezolvat
3. Am adăugat fiecare pagină într-un document separat

Analiza erorilor

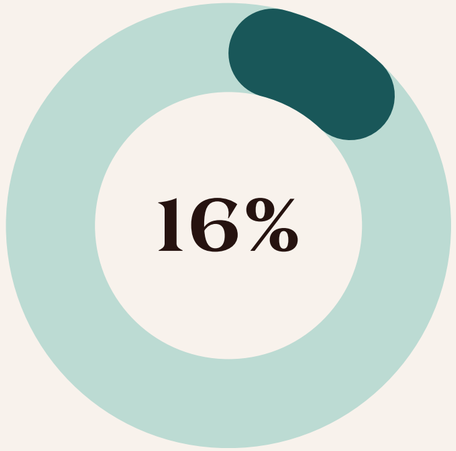
Am încercat mai multe situații de căutare cu ajutorul sistemului, pentru a analiza posibilele rezultate:

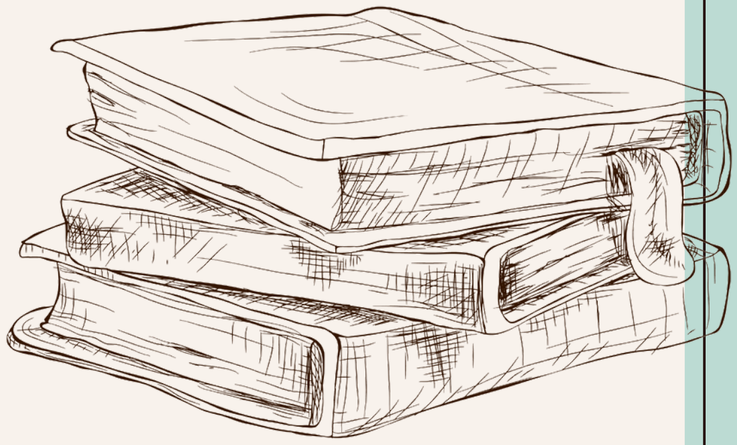
1. Am căutat doar în conținut, folosind doar categoria și/sau întrebarea
2. Am căutat pe titlu și continut, folosid categoria și/sau întrebarea

Întrebarile conțin secvente asemanatoare cu conținutul paginilor Wikipedia și de aceea un sistem atat de simplu răspunde corect la unele întrebări.

Rezultate

Dupa construirea indexului de cautare, modeulul nostru a obtinut urmatoarele rezultate:

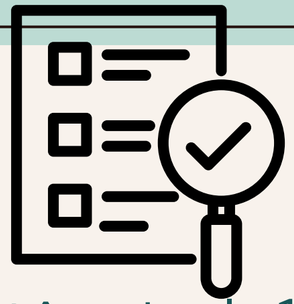




Am obtinut 16 raspunsuri corecte si 84 gresite.

Din cele 84 de raspunsuri gresite, am observat ca 16 au raspunsul corect in primele 10 pozitii, majoritatea in primele 5. Pentru masurarea performantei, am folosit Precision At One

Îmbunătățirea Rezultatelor:



La 16 întrebari sistemul nostru a plasat raspunsul corect în primele 10 pozitii și astfel ne-am gandit să îmbunătățim rezultatele obținute folosindu-ne de ChatGpt.

Pentru aceasta, i-am dat la ChatGpt cele 16 întrebări și ierarhia inițială, fară a-i specifica răspunsul corect, iar acesta ne-a dat o noua ierarhizare, punand răspunsul corect pe prima pozție, ținand cont de probabilitatea pe care a calculat-o și de indiciile din întrebare. În urma aceste modificări, performanța sistemului nostru a crescut cu 50%. Astfel, P@1 = 0.32, în urma rezultatelor inițiale fiind 0,16.