BUILDING (A PART OF) WATSON - DEFAULT PROJECT

Seturile de date utilizate

Datele de intrare sunt extrase dintr-o arhivă ce conține mai multe pagini de Wikipedia. Întrebările concursului Jeopardy au fost extrase de pe site-ul **j-archive.com,** cu date de la show-uri difuzate între 01.01.2013 - 07.01.2013.

Parsarea datelor

Fiecare fișier text din arhiva cu paginile de Wikipedia conține mai multe pagini ce vor fi interpretate ulterior ca și documente de către indexul construit.

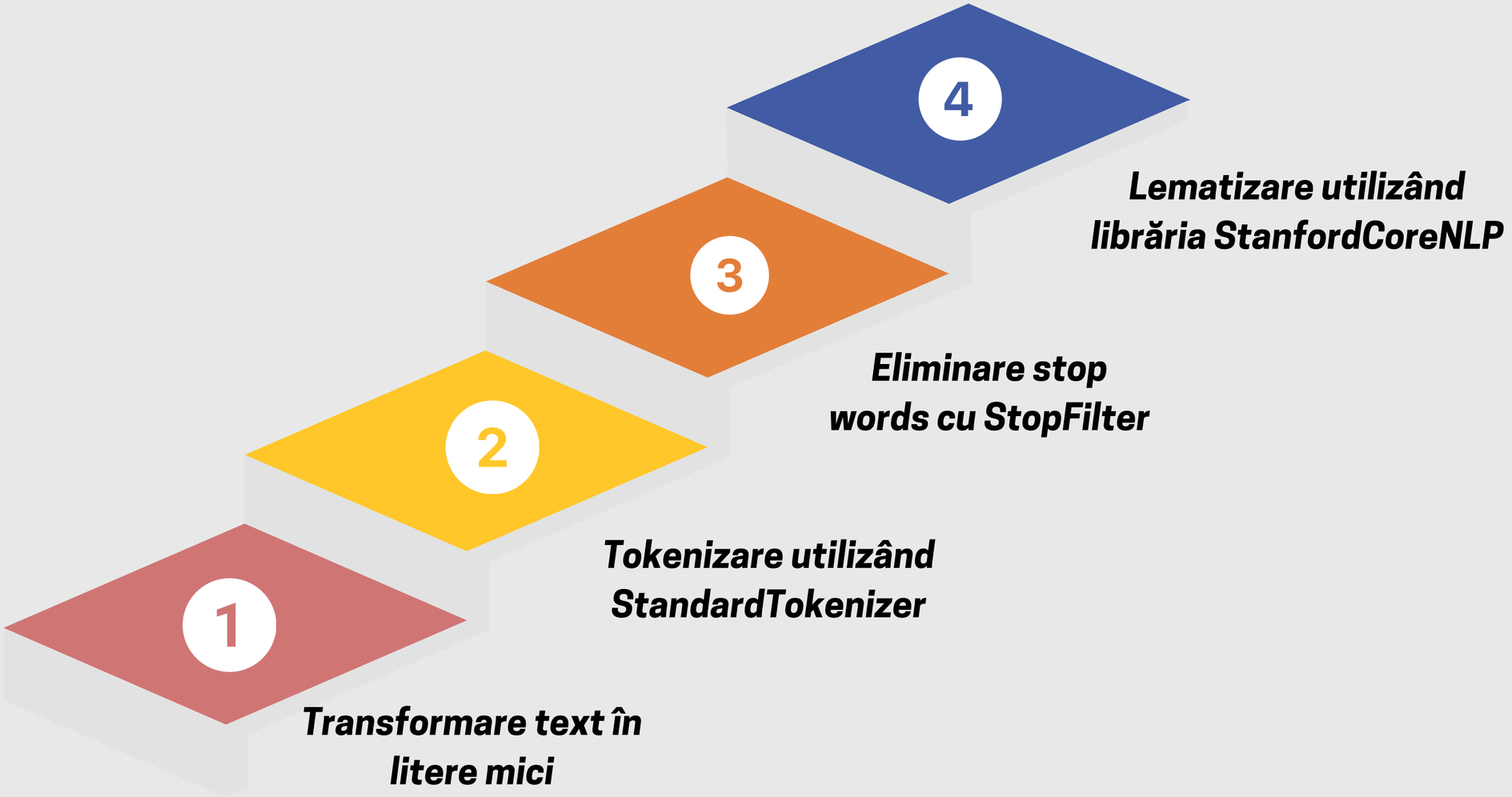
Probleme observate la parsare

- Prezența structurilor / tagurilor de forma: [ref]...[/ref]
- Prezența structurilor / tagurilor de forma: [tpl]...[/tpl]
- Prezența structurilor / tagurilor de forma: [[File: ...]]
- Prezența structurilor / tagurilor de forma: [Image:]

Rezolvarea problemelor la parsarea documentelor

Pentru fiecare dintre problemele menționate anterior am creat un regex care să elimine din fișier structurile problematice (ex. imagini)

PREPROCESAREA DATELOR



CONSTRUIREA INDEXULUI

În procesul de construire a indexului am urmat următorii pași:

- Am citit conținutul fiecărui fișier din cele 80
- Am eliminat elementele problematice (menționate anterior)
- Am folosit un regex pentru a separa conținutul fiecărui fișier în funcție de titlu
- Am adăugat fiecare pagină identificată ca un document separat în index

Fiecare document are 2 atribute:

- Content - conținutul paginii de Wikipedia
- Filename - titlul paginii

REZULTATELE OBȚINUTE INIȚIAL

După construirea indexului de căutare, acesta a fost testat folosind **100 de indicii Jeopardy.**

Modelul nostru a obținut următoarele rezultate:

23
răspunsuri corecte

77
răspunsuri greșite

Din cele 77 de răspunsuri greșite am observat că în 20 de cazuri indexul nostru plasa răspunsul corect pe pozițiile 2-5.

Pentru măsurarea performanței modelului am utilizat **Precision at one (P@1).**

ÎMBUNĂTĂȚIREA REZULTATELOR OBȚINUTE

În urma observării faptului că în 20 de cazuri indexul plasează răspunsul corect pe pozițiile 2-5, ne-am propus să îmbunătățim performanța indexului dezvoltat folosind **ChatGPT.**

Astfel, pentru întrebările ale căror răspunsuri se regăseau în primele 5 documente furnizate de către indexul nostru am decis să îl întrebăm pe ChatGPT care ar fi titlul documentului cel mai potrivit pentru indiciul trimis.

În urma îmbunătățirii cu ChatGPT am reușit să creștem performanța modelului de la o precizie de P@1 = 0.23 la P@1 = 0.41 ceea ce înseamnă că performanța sistemului s-a îmbunătățit cu aproximativ 78%.