

WikiTranslator

Viktor Modroczký

Vyhľadávanie informácií
November 2022

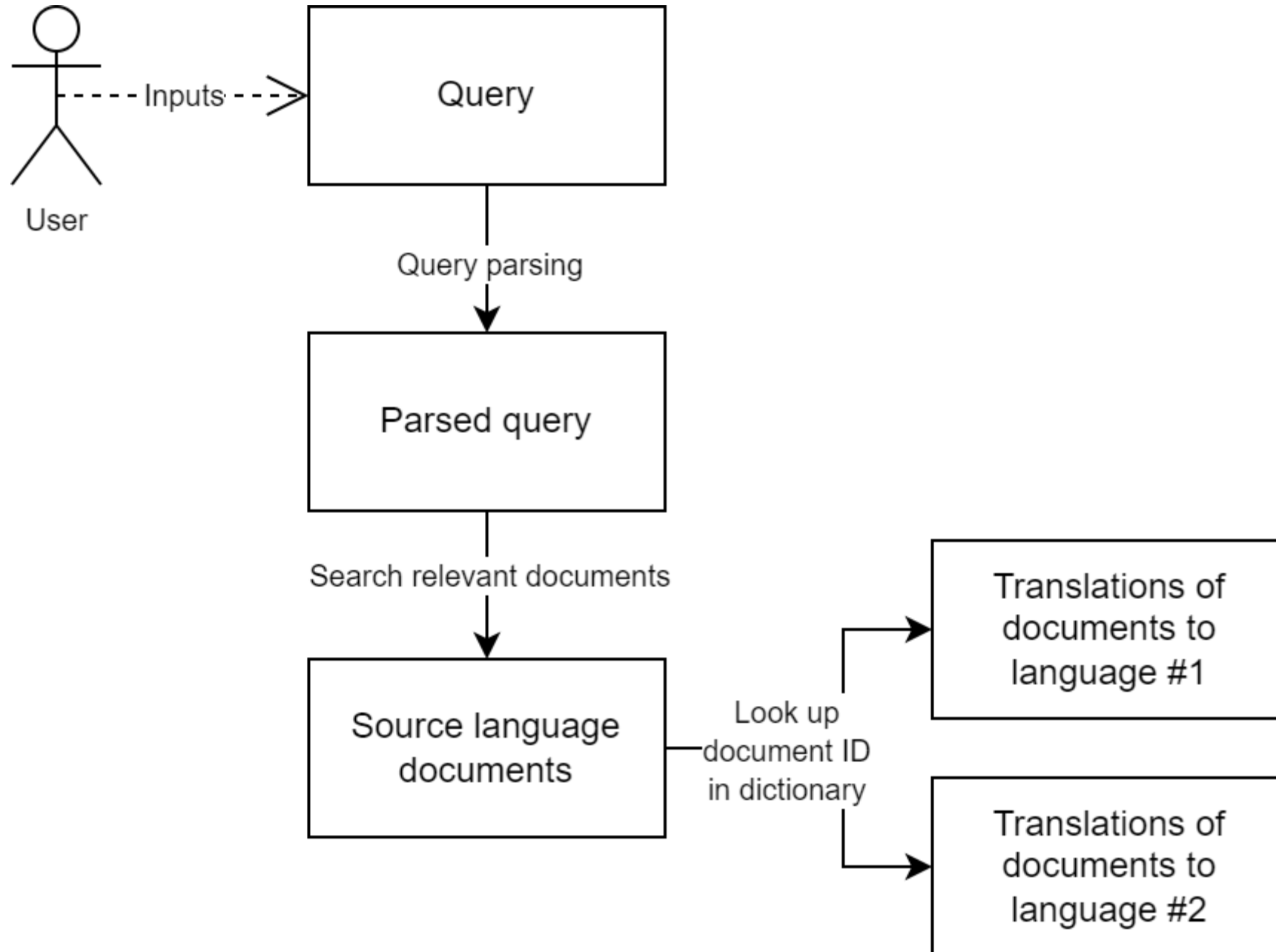
WikiTranslator

- Slovník spájajúci jazyky s možnosťou vyhľadávania
 - Paralelné texty SK-CS-HU
 - Cross-language Information Retrieval
-
- Java 17 + Maven
 - Apache Spark – distribuované spracovanie pomocou Datasetov
 - Apache Lucene – indexovanie a vyhľadávanie

Postup

1. Stiahnutie Wikipedia XML dumpov
2. Využitie nástroja WikiExtractor v Dockeri => JSON články
3. Vyhľadanie SK-CS a SK-HU párov ID článkov z SQL DB
4. Vytvorenie prieniku SK-CS-HU ID pomocou Apache Spark
5. Vyhľadanie článkov v JSON súboroch podľa ID v SK-CS-HU prieniku pomocou Apache Spark
6. Uloženie z nich vytvorené dokumenty (id, title, text) pomocou Apache Spark
7. Indexovanie pomocou Apache Lucene
8. Vytvorenie mapovania pre všetky jazyky:

```
{ „lang_id“: [ „translation_1_id“, „translation_2_id“ ], ... }
```
9. Vyhľadávanie



Používateľské rozhranie

1. `exit`
2. `find article ID pairs`
3. `create sk-cs-hu ID conjunction with Spark`
4. `create docs with Spark`
5. `create Lucene index`
6. `create ID mapping`
7. `use translation search (type 'exit' for quitting search mode)`

Apache Spark

```
spark-class org.apache.spark.deploy.master.Master --host localhost
```

```
spark-class org.apache.spark.deploy.worker.Worker  
spark://localhost:7077 --cores 4 --memory 1G
```

```
spark-class org.apache.spark.deploy.worker.Worker  
spark://localhost:7077 --cores 4 --memory 1G
```

- GUI na localhost:8080



Spark Master at spark://localhost:7077

URL: spark://localhost:7077

Alive Workers: 2

Cores in use: 8 Total, 8 Used

Memory in use: 2.0 GiB Total, 2.0 GiB Used

Resources in use:

Applications: 1 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20221117150325-192.168.1.109-56953	192.168.1.109:56953	ALIVE	4 (4 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20221117150338-192.168.1.109-56969	192.168.1.109:56969	ALIVE	4 (4 Used)	1024.0 MiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20221117151557-0000	(kill) WikiTranslator	8	1024.0 MiB		2022/11/17 15:15:57	Viktor	RUNNING	2 s

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

ID: app-20221117151557-0000

Name: WikiTranslator

User: Viktor

Cores: Unlimited (8 granted)

Executor Limit: Unlimited (2 granted)

Executor Memory: 1024.0 MiB

Executor Resources:

Submit Date: 2022/11/17 15:15:57

State: RUNNING

[Application Detail UI](#)

▼ **Executor Summary (2)**

ExecutorID	Worker	Cores	Memory	Resources	State	Logs
1	worker-20221117150325-192.168.1.109-56953	4	1024		RUNNING	stdout stderr
0	worker-20221117150338-192.168.1.109-56969	4	1024		RUNNING	stdout stderr

sk:T&t:covid test

ID Language Title

667655 SK Rýchly antigénový test COVID-19
1688511 CS Rychlý antigenní test na covid-19
1773397 HU Covid19-antigén gyorsteszt

667137 SK Rýchly antigénový test
1688600 CS Rychlý antigenní test
1774498 HU Antigén gyorsteszt

645499 SK COVID-19
1564791 CS Covid-19
1683177 HU Covid19

102017 SK Test
242872 CS Testování
11604 HU Test (egyértelműsítő lap)

646838 SK Pandémia ochorenia COVID-19
1559185 CS Pandemie covidu-19
1679312 HU Covid19-pandémia

905 SK Turingov test
4218 CS Turingův test
182823 HU Turing-teszt

444545 SK Kolmogorovov-Smirnovov test
324710 CS Kolmogorovův?Smirnovův test
979361 HU Kolmogorov?Szmirnov-próba

339758 SK Klasická teória testov
913998 CS Klasická testová teorie
493883 HU Klasszikus tesztelmélet

652542 SK Vakcína proti chorobe COVID-19
1651518 CS Vakcína proti covidu-19
1689580 HU Covid19-vakcina

526502 SK Crash Test Dummies
694195 CS Crash Test Dummies
1830033 HU Crash Test Dummies
