



Alternate names for Wiki pages

Peter Bokor



Motivation

- Articles on Wikipedia may only have 1 title
- In the real world - multiple names for the same thing
- People - pseudonyms, name changes
- Countries, Cities - Name changes throughout history

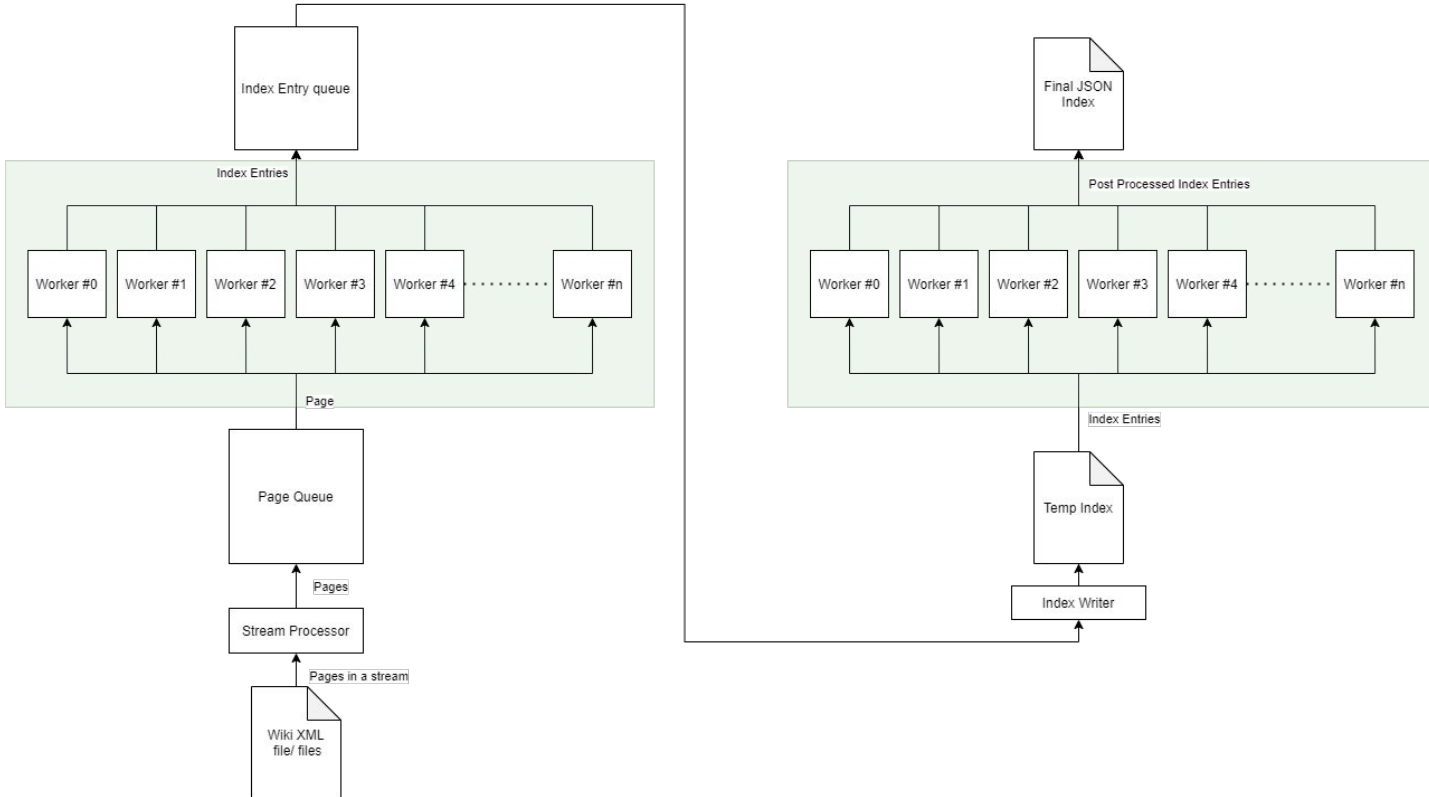
If one index contains titles and alt. Names:

- Less requests to wiki -> Faster loading times for everyone
- Any alt. name search query leads to the main page -> ease of access for people



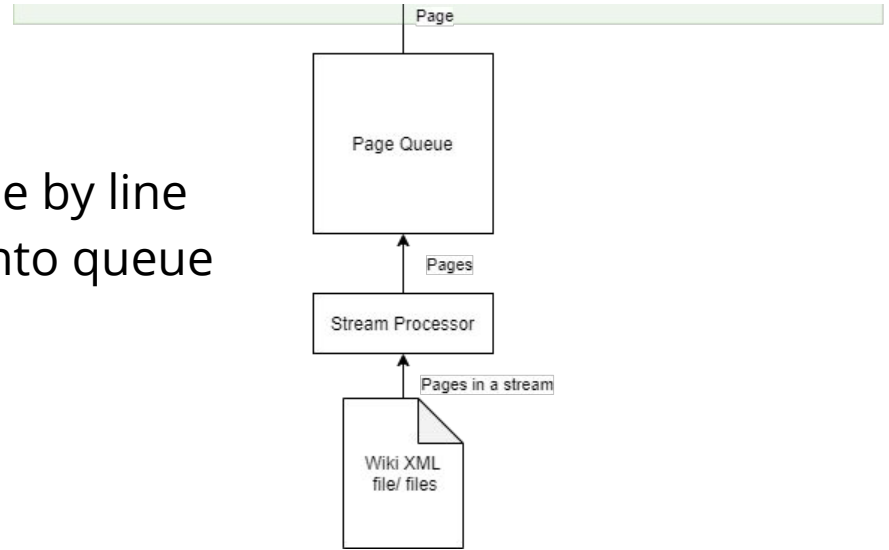
Solution

Overview of the system



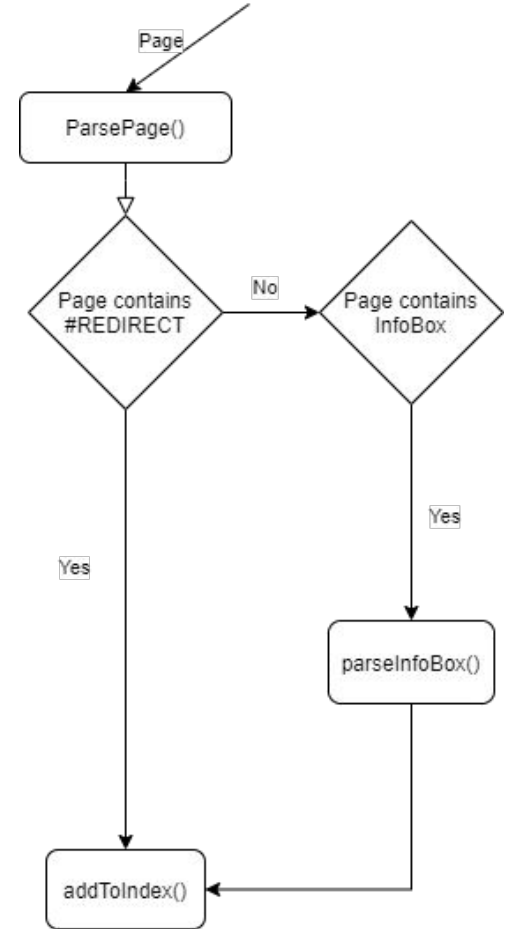
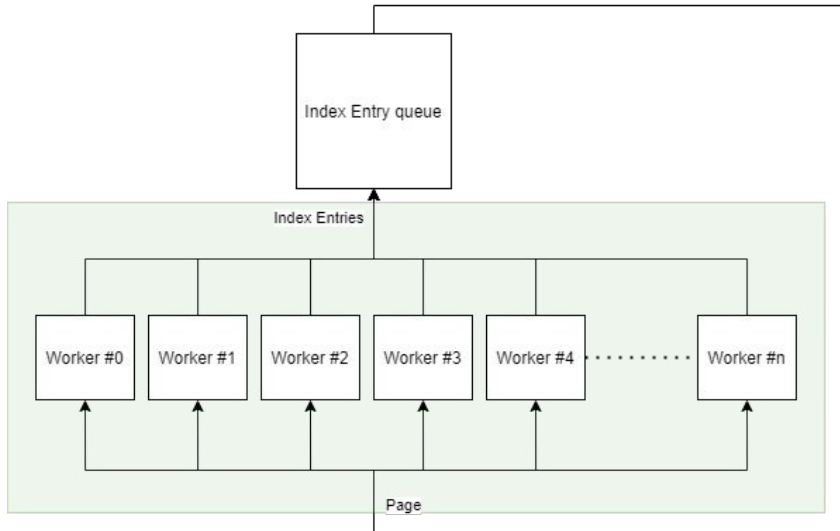
Stream Processing Wiki Dumps

- Dumps in .bz2 format
- BZ2File library in python
- Stream output to lxml.etree lib.
- Lxml.etree.iterparse to parse file line by line
- Extract Title, Text and insert them into queue



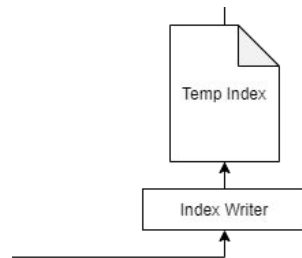
Text Processing

- Entries taken from queue
- Multiprocessing used
- Multiple workers parse text
- Result (alt. title) is put into queue



Temporary Index

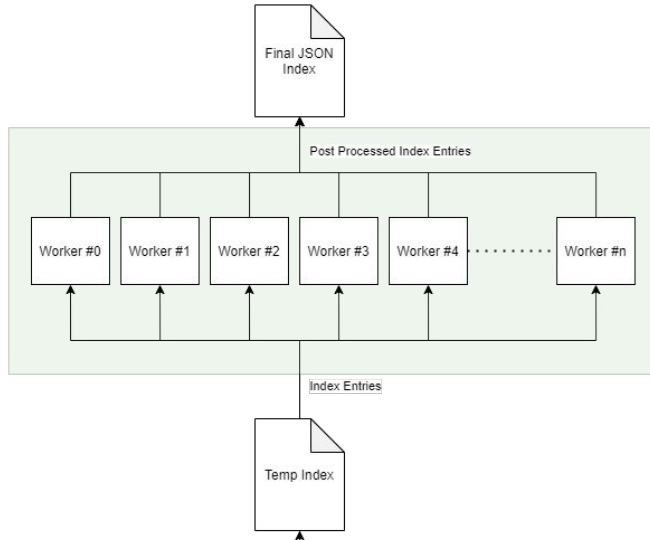
- Built from worker processed entries
- Mapping = alt. Title: title
- Needs to be cleaned - e.g. 1st row
- Needs to be “reversed” to map title: alt. titles




```
Wikipédia:Pomoc:Pomoc:Obsah
Platon:Platón
Neoplatonizmus:novoplatonizmus
Buddhizmus:Budhizmus
Empatia:vcitenie
Neotomizmus:novotomizmus
Buddha:Budha
Kynická škola:kynizmus
Atiša:Atiša
Belzebub:Baal-zebúb
Neofreudizmus:neopsychoanalýza
Neomarxizmus:Neomarxizmus
Neoštrukturalizmus:Postštrukturalizmus
Neohegelianstvo:novohegelovstvo
Novopozitivismus:neopozitivismus
Komutatívna operácia:Komutatívnosť
DVD Video:DVD
Postgresql:PostgreSQL
Monako (štát):Monako
Reykjavik:Reykjavík
Andorra (štát):Andorra
Mogadiišo:Mogadišo
Barma:Mjanmarsko
Myanmar:Mjanmarsko
Andora:Andorra
Maurícius:Maurícius
Venezuelská bolívarovská republika:Venezuela
Čilská republika:Čile
Špicbergy:Svalbard
```

Index Post Processing

- Changes mapping of the temporary index
- Creates the final .JSON file
- Multiprocessing to speed things up



```
...  
"Briansk": [   
  "Brjansk"  
],  
"Arsakovci": [   
  "Aršakovci"  
],  
"Dramaturg": [   
  "Dramaturgička"  
],  
"eweština": [   
  "Ewejščina"  
],  
"Macao": [   
  "Osobitná správná oblasť Čínskej ľudovej republiky Macao",  
  "Osobitná administratívna oblasť Čínskej ľudovej republiky Macao"  
],  
"sklovina": [   
  "Zubná sklovina",  
  "Enamelum",  
  "Adamantin"  
],  
"e-mail": [   
  "Majl",  
  "Meil",  
  "E-pošta",  
  "Elektronický list",  
  "E-mailová správa",  
  "Emailová správa",  
  "E-poštová správa",  
  "Mailová správa",  
  "Mejlová správa"  
],  
"polostrov": [   
  "Peninsula",  
  "Poloostrov"  
],  
...
```


Evaluation of the solution

Manual evaluation

- Obviously nice looking examples - we as an expert can evaluate these
 - `"Tirolsko": ["Tyrolsko", "Tyroly", "Tiroly", "Tirolská zem"]`
 - `"elektromobil": ["Elektrický automobil", "Automobil na elektrický pohon", "E-mobil", "Emobil", "Elektrické auto"]`
 - `"Ekvádor": ["Ekuador", "Ecuádor", "Ekvador", "Ekvádorská republika"]`
- Abbreviations
 - `"KDU-ČSL": ["Křesťanská a demokratická unie - Československá strana lidová"]`
 - `"Monsignor": ["Msgr."]`
- Special terminology e.g. Latin names / Eng. names - we can use google search and see if we are lead to the main article
 - `"Bifenyl": ["Difenyl", "Fenylbenzén", "E 230"]`
 - `"Hrtan": ["Larynx"]`
 - `"trilobitovce": ["Trilobitoidea"]`
- Slang
 - `"Špekáčik": ["Safaládka"]`

Automatic evaluation

- If we have a search API we can search on the internet for the alternate titles and look for the main titles in the results

Evaluation examples

E 230

Všetko Obrázky Mapy Nákupy Videá Víac Nastavenia Nástroje

Približne 2 260 000 000 výsledkov (0,44 sekúnd)

Obrázky pre dopyt E 230

w124 mercedes benz kompressor interior model e class wagon



Nahľadanie obrázkov



Zobrazíť všetky

www.dtest.sk > Ěčka

E 230 Bifenyly - Testy a recenzie výrobkov - dTest

Bifenyly je umělý konzervant. Může vyvolat alergické reakce u citlivých osob. Používá se také jako pesticid u citrusů a při ošetřování citrusových plodů. Použití v ...

Trilobitoidea

Všetko Obrázky Mapy Videá Správy Víac Nastavenia Nástroje

Približne 12 000 výsledkov (0,58 sekúnd)

Tip: [Vyhľadávajte len výsledky v slovenčine](#). Jazyk vyhľadávania môžete určiť tu: [Nastavenia](#)

sk.wikipedia.org > wiki > Trilobitovce

Trilobitovce – Wikipédia

triedy **Trilobitoidea** definovanej ako trieda v tých starých systémoch, v ktorých sa podkmeň

Trilobitomorpha delil na 2 triedy: Trilobita a **Trilobitoidea**, išlo teda o ...

```
["Nicopolis"], "Genderqueer": ["Non-binary"], "Celestini": ["Celestináni"], "Epsilon": ["E"], "Lambda": ["Λ"], "Mí": ["M"],  
"Ksí": ["Ξ"], "Omikron": ["O"], "Ró": ["P"], "Sigma": ["Σ"], "Tau": ["T"], "Fí": ["Φ"], "Chí": ["X"], "Héta": ["H"],  
"Sampi": ["Տ", "T"], "Irak": ["Iracká republika"], "Kremenčuk": ["Kremenčug"], "Ryv": ["Jerk"], "Višap": ["Višapakar"],  
"Komitas": ["Komitas Vardapet", "Soghomon Soghomonian"], "Bederņovci": ["Bederņovci z Bederņu"], "Septuaginta": ["LXX"]
```

Was the multiprocessing needed?

| | |
|---|---------------|
| Slovak wiki sequential (1 worker) - | 3.75 minutes |
| Slovak wiki sequential (4 worker) - | 2.67 minutes |
| Slovak wiki sequential (8 worker) - | 2.67 minutes |
| Slovak wiki with 1 worker thread - | 2.4 minutes |
| Slovak wiki with 4 workers (final solution) - | 2.31 minutes |
| Slovak wiki with 8 workers - | 2.32 minutes |
| English wiki with multiprocessing (4 workers) - | 115 minutes |
| English wiki no multiprocessing estimate - | 179.7 minutes |

```
WORKER 3: AbbesS : AbbesS]]
INDEX_WRITER: Title: AbbesS
WORKER 1: AbbevilleFrance : Abbeville]]
INDEX_WRITER: Title: AbbevilleFrance
WORKER 3: AbbeY : Abbey]]
INDEX_WRITER: Title: AbbeY
WORKER 1: AbboT : Abbot]]
INDEX_WRITER: Title: AbboT
WORKER 2: Abbreviations : Abbreviation]]
INDEX_WRITER: Title: Abbreviations
WORKER 3: AbeceDarians : AbeceDarian]]
INDEX_WRITER: Title: AbeceDarians
WORKER 2: AbensbergGermany : Abensberg]]
INDEX_WRITER: Title: AbensbergGermany
WORKER 1: AnarchY : Anarchy]]
INDEX_WRITER: Title: AnarchY
WORKER 0: AndorrA : Andorra]]
INDEX_WRITER: Title: AndorrA
WORKER 2: AnarchoCapitalism : Anarcho-capitalism]]
INDEX_WRITER: Title: AnarchoCapitalism
WORKER 0: AnarchoCapitalists : anarcho-capitalism]]
INDEX_WRITER: Title: AnarchoCapitalists
WORKER 1: AutoMorphism : Automorphism]]
INDEX_WRITER: Title: AutoMorphism
WORKER 2: Africa : Africa]]
INDEX_WRITER: Title: Africa
WORKER 1: AppliedStatistics : Statistics]]
INDEX_WRITER: Title: AppliedStatistics
WORKER 2: Ameboid stage : Amoeba
INDEX_WRITER: Title: Ameboid stage
WORKER 0: Auteur Theory Film : Auteur
INDEX_WRITER: Title: Auteur Theory Film
WORKER 3: Actress : Actor]]
INDEX_WRITER: Title: Actress
WORKER 2: Allotrope : Allotropy]]
INDEX_WRITER: Title: Allotrope
WORKER 2: Archaeology/Broch : Broch]]
INDEX_WRITER: Title: Archaeology/Broch
WORKER 3: Aa River : AA#Rivers]]
INDEX_WRITER: Title: Aa River
WORKER 2: Astronomers and Astrophysicists : Astronomer
INDEX_WRITER: Title: Astronomers and Astrophysicists
WORKER 1: Afghanistan (1911 Encyclopedia) : Afghanistan
INDEX_WRITER: Title: Afghanistan (1911 Encyclopedia)
WORKER 0: Anglican Church : Anglicanism
INDEX_WRITER: Title: Anglican Church
WORKER 1: AU : Au]]
INDEX_WRITER: Title: AU
WORKER 1: Ancient civilization : Civilization]]
INDEX_WRITER: Title: Ancient civilization
WORKER 1: Accountancy : Accounting]]
INDEX_WRITER: Title: Accountancy
WORKER 0: Awk : AWK]]
INDEX_WRITER: Title: Awk
WORKER 3: AgoraNomic : Nomic
INDEX_WRITER: Title: AgoraNomic
WORKER 2: Anti-semitism : Antisemitism]]
INDEX_WRITER: Title: Anti-semitism
WORKER 0: Anti-semitic : Antisemitism
INDEX_WRITER: Title: Anti-semitic
WORKER 2: Alumna : Alumnus]]
INDEX_WRITER: Title: Alumna
WORKER 0: A priori and a posterior knowledge : Knowledge
INDEX_WRITER: Title: A priori and a posterior knowledge
WORKER 2: Anarcho-capitalists : anarcho-capitalism
INDEX_WRITER: Title: Anarcho-capitalists
WORKER 2: Applied statistics : ru:Прикладная статистика
INDEX_WRITER: Title: Applied statistics
```

Possible improvements

3 stage system:

1. Multiprocessed XML reader -> xml queue
 - a. Problem - if the file is too large, not enough RAM for the whole queue
2. Multiprocessed Workers -> Index queue
3. Multiprocessed Index construction

Stages 1, 2 and 3 wait for each other to finish

Possible solution to a. - Use smaller batches, not the whole file