DIGITAL CONTENT RETRIEVAL - MOD.B

Part 1 - Search Facility

Andrea Frigatti

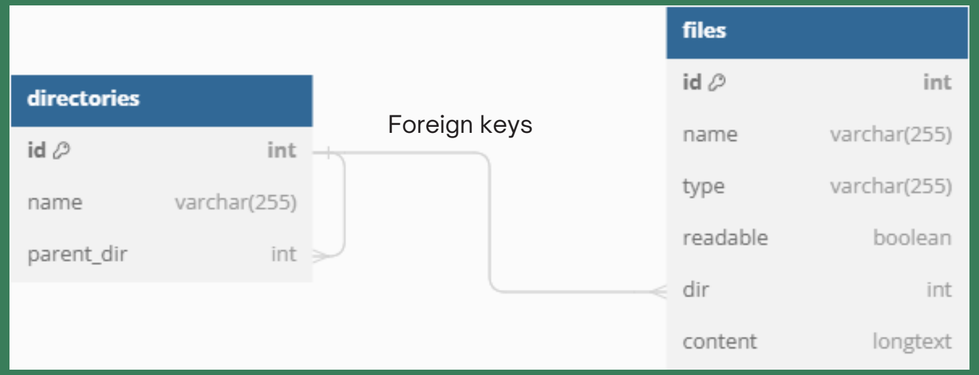
19 June 2024

Overview

Project goals:

- Index pages downloaded from Wikipedia and others files from file system on a persistence level
- Retrieve all documents matching a given string
- Count number of occurencies of the given string inside retrieved files

MySQL Database



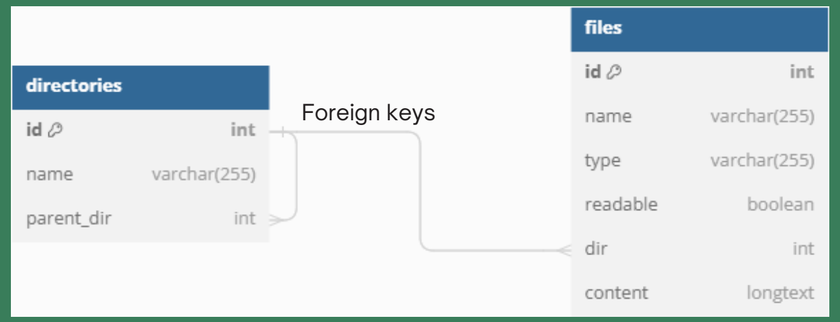
Advantages:

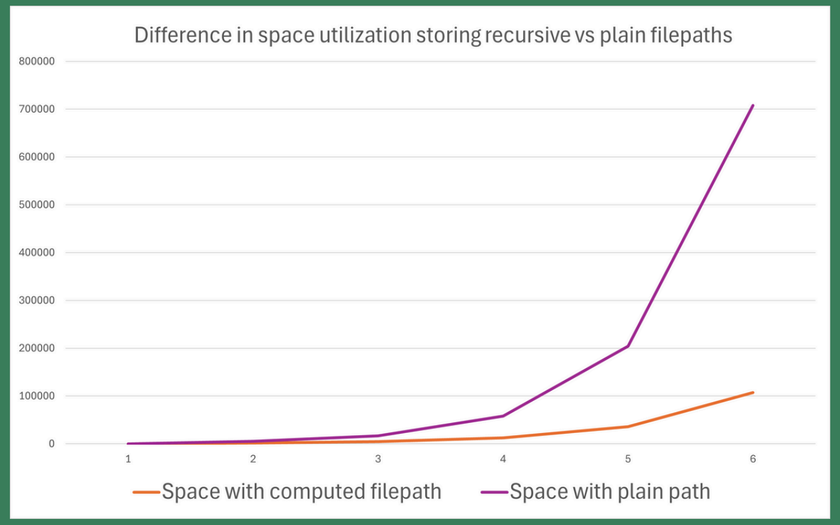
- *Foreign keys ensure referential integrity between directories themselves and files.*
- *On Delete Cascade automates the process of elimination of all files and subdirectories when a directory is deleted.*
- *When a directory name changes, only the directory name needs to be updated, not the entire path of each file.*
- *Storing only the parent directory ID instead of the full path for each file ensures saving space.*

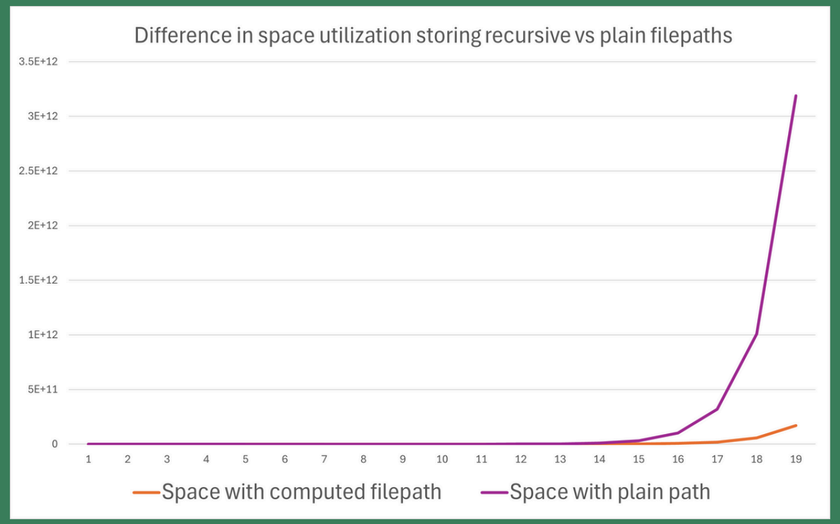
Indexes:

- *Full text index on “content” for table “files”*
- *Index on “name” for table “files”*

MySQL Database







Plots computed on the assumptions of 3 subfolders per folder, 10 files per folder and an average file/directory name length of 10 characters

Wikipedia Pages Ingest

- Use of Wikipedia API to download pages based on a Python library called “wikipediaapi”
- Starting from a Root Category (e.g. Music) recursively download articles with a maximum depth of 5 levels
- Subcategories become folders and pages are downloaded in HTML or txt format and saved into local file system

WIKIPEDIA
The Free Encyclopedia

Workflow

Inserting of files into DB

- *Directories and files are read from the local file system keeping track of parent-child hierarchy*
- *Insertion into DB is made with chunks of files and directories to reduce I/O computational cost*

Retrieval of files using local cache

- *Simple local cache based on FIFO Logic*
- *Query searches files inside local cache before querying the DB*
- *System shows cache hit rate*

Retrieval of files

- *Perform a query to table “files” to find all files whose name or content match a given string (case sensitive search)*
- *Perform a recursive query to table “directories” to retrieve files absolute path*
- *Count number of occurrences of the given string in readable files*

Search Results

```
PS C:\Users\andre\Desktop\DCRB-Part1> python .\db_search.py
Enter string to search: music
Occurences: 16                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Ancient Greek music theory/Arsis and thesis.txt
Occurences: 14                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Ancient Greek music theory/Genus (music).txt
Occurences: 51                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Ancient Greek music theory/Hagiopolitan Octoechos.txt
Occurences: 3                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Ancient Greek musicians/Aglais (musician).txt
Occurences: 1                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Ancient Greek musicians/Aristocleides.txt
Occurences: 2                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Metrical feet/Amphibrach.txt
Occurences: 3                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Metrical feet/Anaclasis (poetry).txt
Occurences: 1                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Ancient Greek music/Metrical feet/Metrical foot.txt
Occurences: 2                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Carnatic classical music festivals/Birmingham Thyagaraja Festival.txt
Occurences: 7                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Carnatic classical music festivals/Chembai Sangeetholsavam.txt
Occurences: 11                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Carnatic classical music festivals/Chennaiyil Thiruvaiyaru.txt
Occurences: 3                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Celtic music festivals/Austin Celtic Association.txt
Occurences: 1                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Celtic music festivals/Brandivy.txt
Occurences: 4                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Christian music festivals/Agape Music Festival.txt
Occurences: 1                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Christian music festivals/Christian Festival Association.txt
Occurences: 44                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by culture/Music festivals by culture/Christian music festivals/Christian music festival.txt
Occurences: 1                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by date/Music by year/Classical music by year/2009 in classical music.txt
Occurences: 2                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by date/Music by year/Classical music by year/2010 in classical music.txt
Occurences: 1                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by date/Music by year/Classical music by year/2011 in classical music.txt
Occurences: 25                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American music in Africa/Afro rock.txt
Occurences: 15                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American music in Africa/Boomba music.txt
Occurences: 17                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American music in Africa/G.V. Series.txt
Occurences: 4                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American musical groups/2 Live Crew.txt
Occurences: 6                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American musicians/Adia (musician).txt
Occurences: 136                 File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American musicians/African-American music.txt
Occurences: 54                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/African-American music/African-American musicians/Black conductors.txt
Occurences: 20                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/African-American music/Acid house.txt
Occurences: 6                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/African-American music/African-American Music Appreciation Month.txt
Occurences: 136                 File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/African-American music/African-American music.txt
Occurences: 102                 File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/Afro-Caribbean music/Afro-Caribbean music.txt
Occurences: 16                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/Afro-Caribbean music/Martha Ellen Davis.txt
Occurences: 28                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/Afro-Caribbean music/Taina Asili.txt
Occurences: 12                  File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/Black British music/Afroswing.txt
Occurences: 7                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/Black British music/BBC Radio 1Xtra.txt
Occurences: 6                   File: /Users/andre/Desktop/DCRB-Part1/DCRB/Music by ethnicity/Music of the African diaspora/Black British music/British Black music.txt
```

```
Occurences: 6                   File: /Users/andre/AppData/Local/Programs/Python/Python310/Lib/site-packages/gensim/test/test_data/head500.noblanks.cor_wordids.txt
File from: cache

Occurences: 1                   File: /Users/andre/AppData/Local/Programs/Python/Python310/Lib/site-packages/gensim/test/test_data/toy-data.txt
File from: cache

File type reading not supported.    File: /Users/andre/AppData/Local/Programs/Python/Python310/Lib/site-packages/nltk/corpus/reader/indian.py
File from: cache

Occurences: 4                   File: /Users/andre/Documents/Università/Digital Content Retrieval Mod.B/Project/DCRB/Music by date/Music by century/Music genres by century/16th-century music genres/Odissi music/Bhikari Bal.txt
File from: cache

Occurences: 2                   File: /Users/andre/Documents/Università/Digital Content Retrieval Mod.B/Project/DCRB/Music by ethnicity/Musicians by ethnicity/Classical musicians by ethnicity/Opera singers by ethnicity/Jewish opera singers/Mario Ancona.txt
File from: cache

Occurences: 2                   File: /Users/andre/Documents/Università/Digital Content Retrieval Mod.B/Project/DCRB/Music by ethnicity/Musicians by ethnicity/Singers by ethnicity/Opera singers by ethnicity/Jewish opera singers/Mario Ancona.txt
File from: cache

File type reading not supported.    File: /Users/andre/AppData/Local/Programs/Python/Python310/Lib/site-packages/nltk/corpus/reader/__pycache__/indian.cpython-310.pyc
File from: cache

File type reading not supported.    File: /Program Files/Huawei/HMS Core/res/drawable/HWFrameworkUI/btn_gray_indialog_disable.png
File from: database
```

Conclusions

01 The search function allows users to find matches based on file names or content

02 Using of integrity constraints to guarantee consistency

03 Efficiently ingestion of files using chunk insertion under single transaction

04 Uses indexes to speed up query execution time

DIGITAL CONTENT RETRIEVAL - MOD.B

Part 2 - Information Retrieval System

Andrea Frigatti

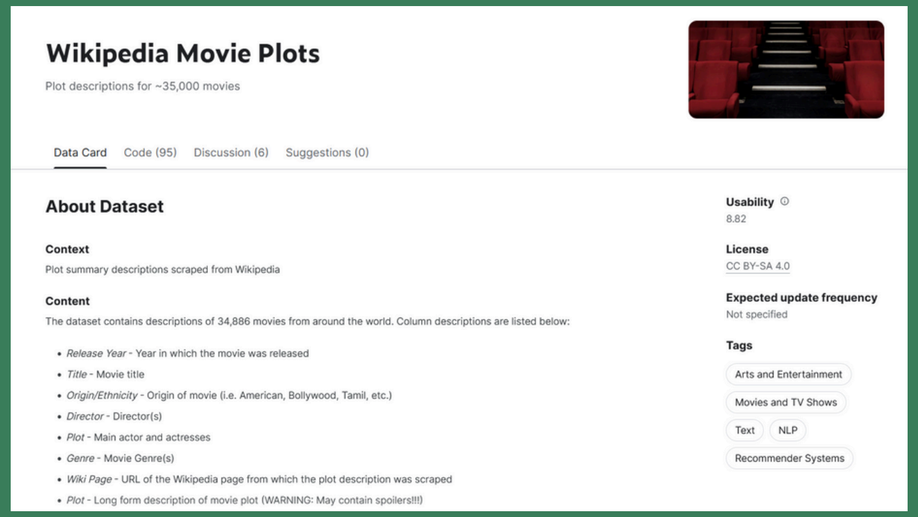
19 June 2024

Overview

Project goals:

- Creation of an indexing function to efficiently store documents
- Creation of a retrieving function to collect all documents matching given terms

Dataset



Wikipedia Movie Plots

Plot descriptions for ~35,000 movies

Data Card | Code (95) | Discussion (6) | Suggestions (0)

About Dataset

Context

Plot summary descriptions scraped from Wikipedia

Content

The dataset contains descriptions of 34,886 movies from around the world. Column descriptions are listed below:

- *Release Year* - Year in which the movie was released
- *Title* - Movie title
- *Origin/Ethnicity* - Origin of movie (i.e. American, Bollywood, Tamil, etc.)
- *Director* - Director(s)
- *Plot* - Main actor and actresses
- *Genre* - Movie Genre(s)
- *Wiki Page* - URL of the Wikipedia page from which the plot description was scraped
- *Plot* - Long form description of movie plot (WARNING: May contain spoilers!!!)

Usability
8.82

License
CC BY-SA 4.0

Expected update frequency
Not specified

Tags

Arts and Entertainment
Movies and TV Shows
Text
NLP
Recommender Systems

- *Number of documents: 35.000*
- *Metadata: Release Year, Director, Genre, Plot, ...*
- *Original file format: CSV file*
- *Dimension: 82MB*
- *Column of interest: Plot*

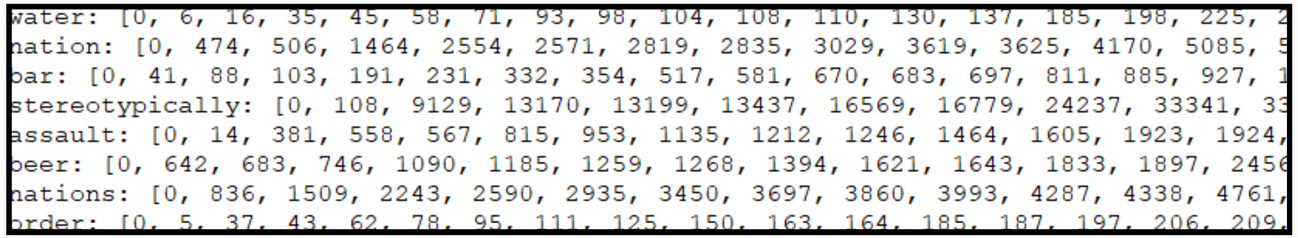
Inverted Index

Creation of an inverted index on column “Plot”

Steps for text tokenization:

- removal of punctuation
- removal of what is not a word
- removal of stopwords based on a list

Creation of inverted index based on a Single-Pass In-Memory Indexing algorithm and storage into a txt file



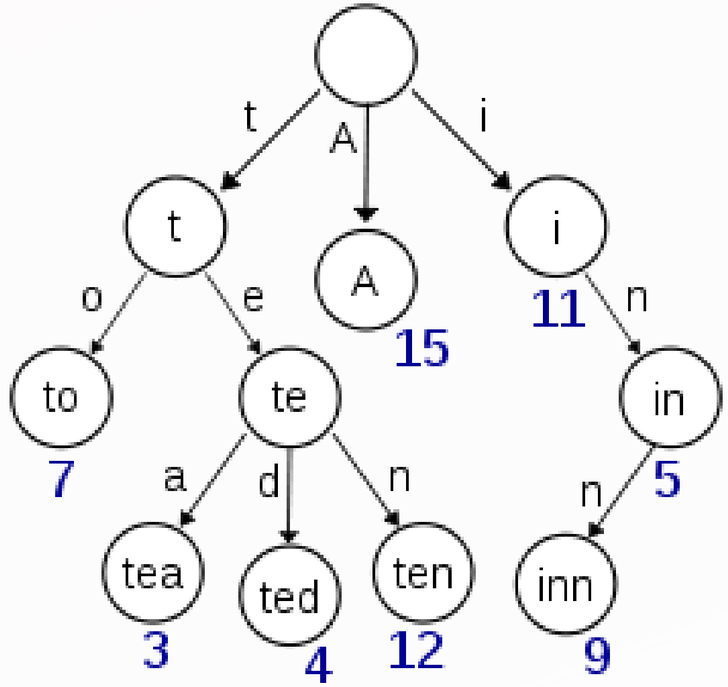
B-Tree construction

Construction of Search Tree and Reversed Search Tree based on B-Tree
(later used for single term search queries)

Steps:

- Reading of txt file containing the inverted index
- Distribution of tokens and posting lists in nodes and leafs

B-Trees have a height of five levels and are stored inside a python pickle file



Queries and Search Results

- *Single word query: word*
- *Conjunctive query: word1 AND word2*
- *Disjunctive query: word1 OR word2*
- *Prefix query: prefix**
- *Suffix query: *suffix*
- *Wildcard query: prefix*suffix*

Output: metadata of all documents retrieved

```
PS C:\Users\andre\Documents\Università\Digital Content Retrieval Mod.B\DCRB-Part2> python .\retrieving.py
Possible queries:
        - Single word query: word
        - Conjunctive query: word1 AND word2
        - Disjunctive query: word1 OR word2
        - Prefix query: prefix*
        - Suffix query: *suffix
        - Wildcard query: prefix*suffix

Enter a query or press 'Ctrl+C' to quit: abstraction
Number of documents retrieved for 'abstraction': 1

Documents matching the IDs:
Document ID: 20571
{
    "Release Year": 1992,
    "Title": "Orlando",
    "Origin/Ethnicity": "British",
    "Director": "Sally Potter",
    "Cast": "Tilda Swinton, Billy Zane",
    "Genre": "fantasy",
    "Wiki Page": "https://en.wikipedia.org/wiki/Orlando_(film)",
```

```
Enter a query or press 'Ctrl+C' to quit: woolf*
Number of documents retrieved for 'woolf*': 12
Words starting with the prefix:
woolf
woolfolk
woolfs

Documents matching the IDs:
Document ID: 14432
{
    "Release Year": 2004,
    "Title": "Closer",
    "Origin/Ethnicity": "American",
    "Director": "Mike Nichols",
    "Cast": "Julia Roberts, Jude Law, Natalie Portman, Clive Owen",
    "Genre": "drama",
    "Wiki Page": "https://en.wikipedia.org/wiki/Closer_(film)",
    "Plot": "In the opening scene, 24-year-old \"Alice Ayres\" (Portman) and Dan Woolf (Law)
```

Conclusions

01 Efficient text processing and indexing

02 Scalable and balanced data storage based on B-Tree

03 Versatile query handling on single terms searches

04 Persistent and quick data retrieval