



Lazy song

progetto per il corso di Gestione dell'Informazione
dell'università di Modena e Reggio Emilia anno 2020/2021

[Lorenzo Stigliano e Alessandra Bonaccorso]

Indice

1. idea
2. raccolta documenti
3. preprocessing e algoritmo di indexing
4. search engine e gui
5. esempi di query di ricerca
6. benchmark (grafici e valutazioni)
7. conclusioni
8. fine

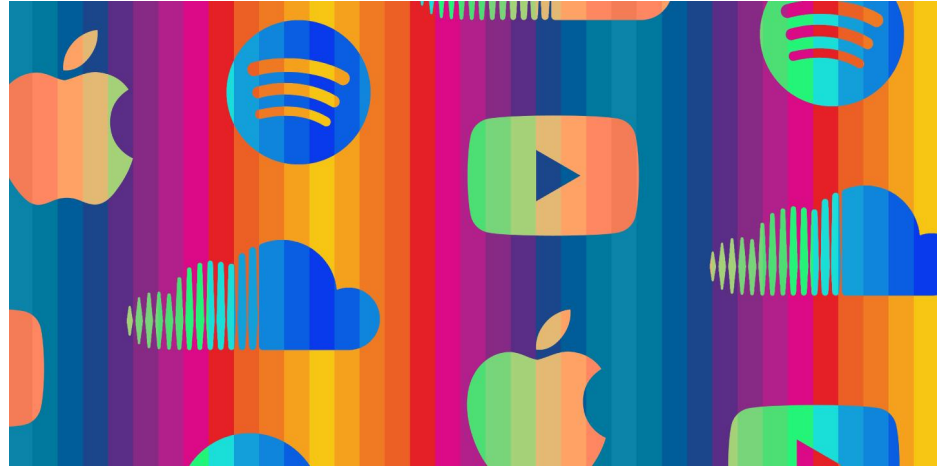


Idea di base

Come abbiamo appreso dal corso di gestione dell'informazione, i dati intorno a noi si materializzano in varie forme ed in modi differenti.

Per questo progetto ci è stato chiesto di realizzare un search engine tematico. Abbiamo scelto la musica, perché è qualcosa che accomuna tutti e che ci circonda ogni giorno.

Nel nostro search engine sono presenti i testi di vari artisti ed è possibile ricercare un termine ed avere come risultato canzoni che contengono quel termine (o simili).



Il vero programmatore/informatico è Lazy! Dunque non ci sembrava corretto scaricare tramite copia ed incolla ogni canzone manualmente.

Una rapida ricerca su internet ci ha permesso di scoprire il pacchetto per python GeniusLyrics, che dato in pasto il nome di un artista scarica i testi delle sue canzoni dal sito più famoso di lyrics, Genius.com.

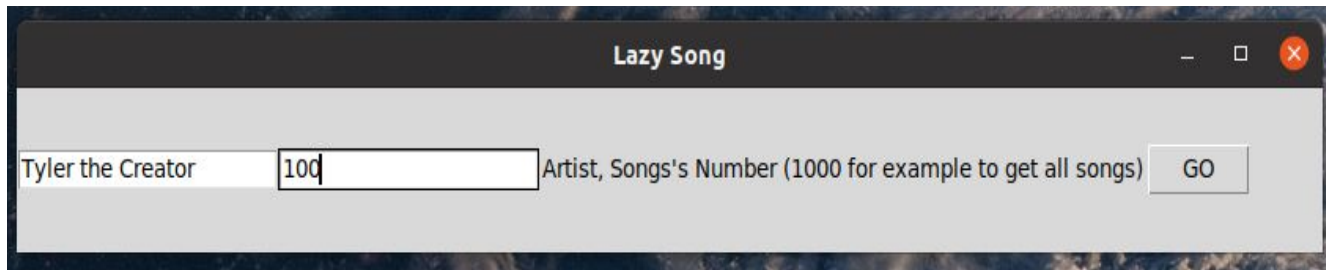
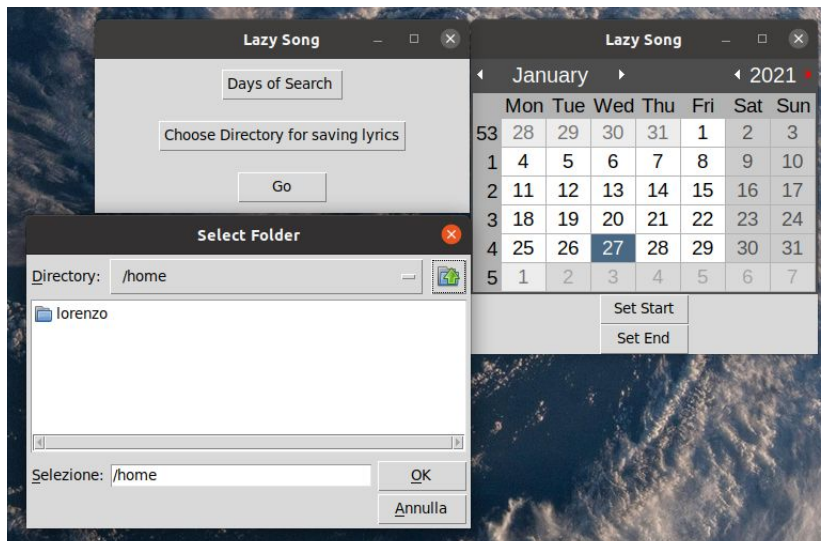
GENIUS

Raccolta del materiale

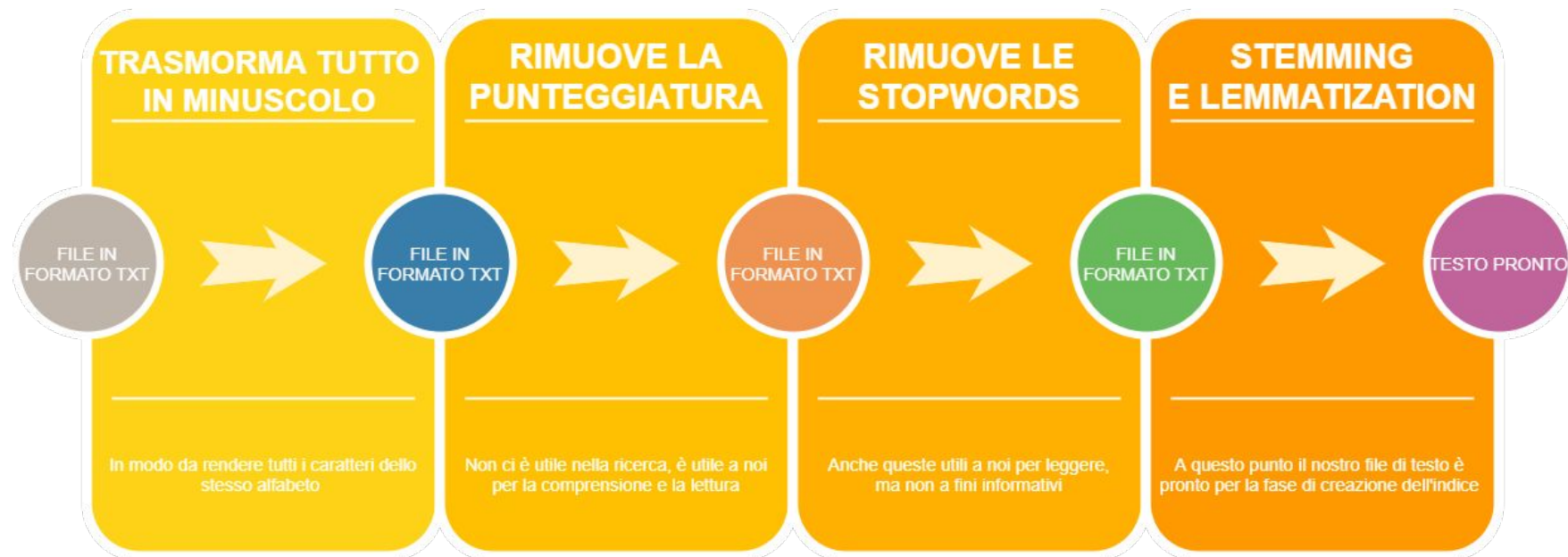
Per scegliere gli artisti abbiamo selezionato quelli che hanno raggiunto le classifiche negli ultimi anni. Di ognuno abbiamo selezionato la top ten delle sue canzoni, in modo da avere una selezione più omogenea.



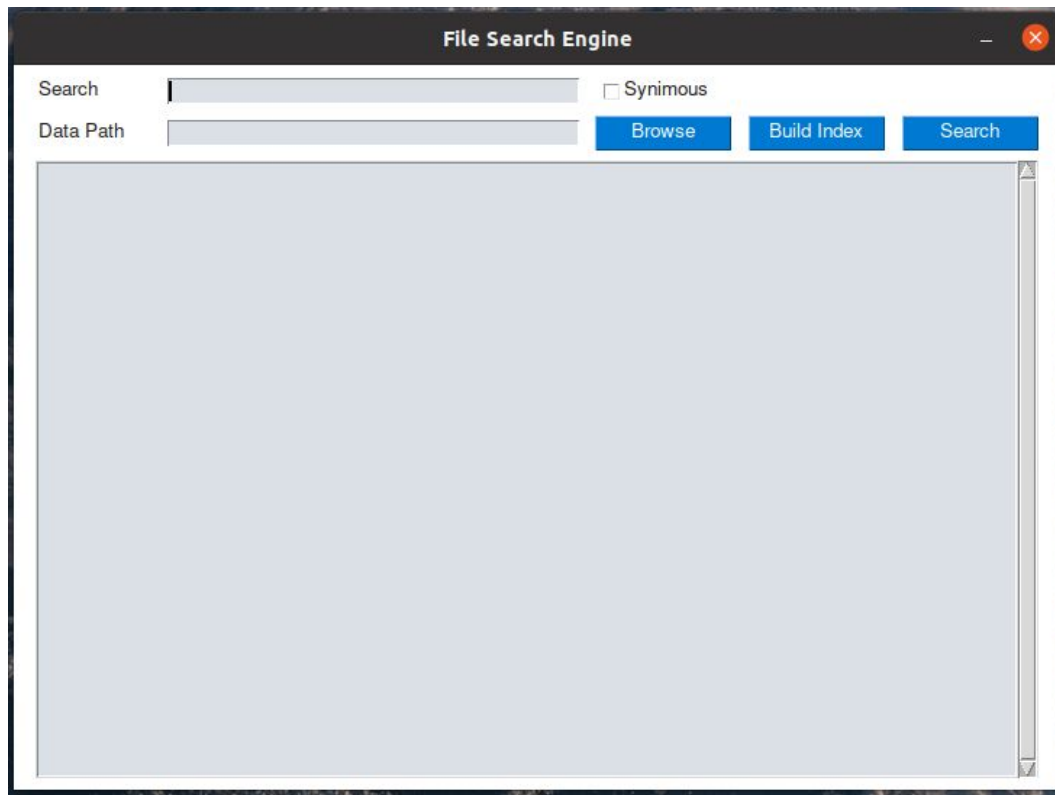
Screenshoots Dump



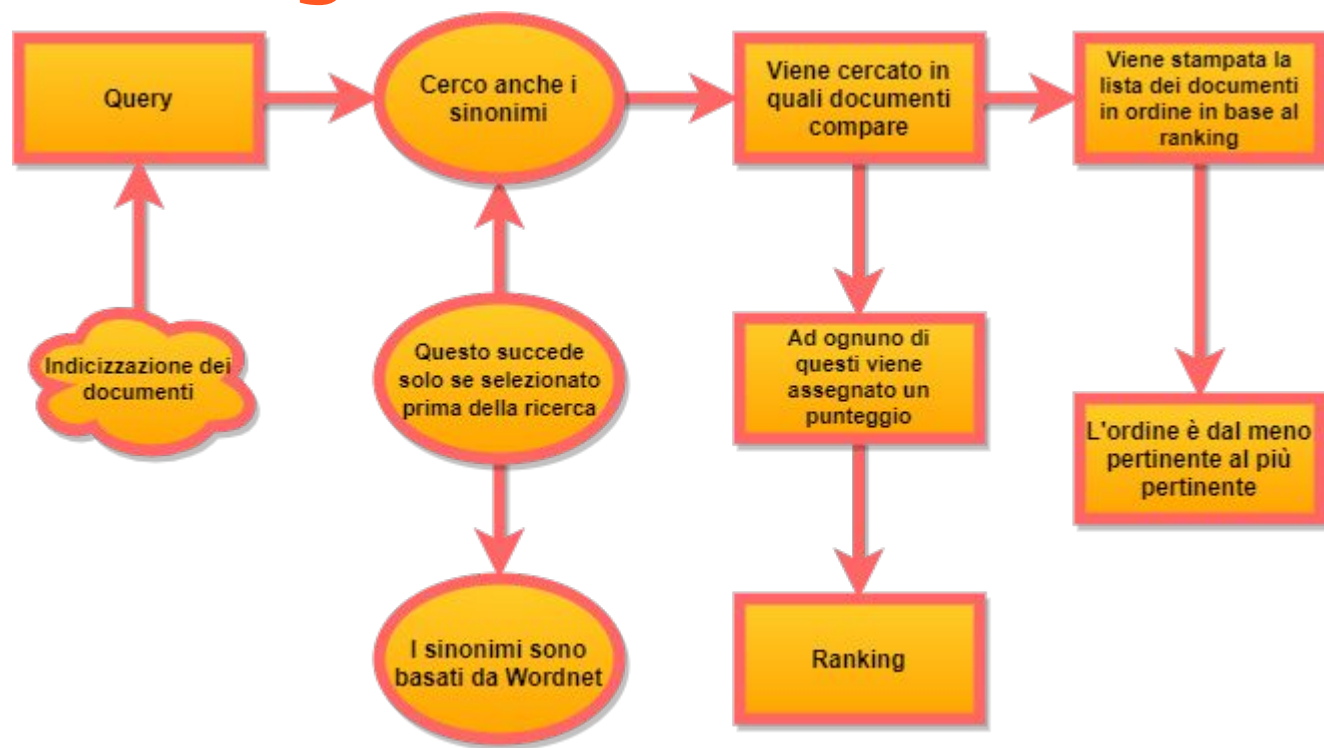
Pre-processing



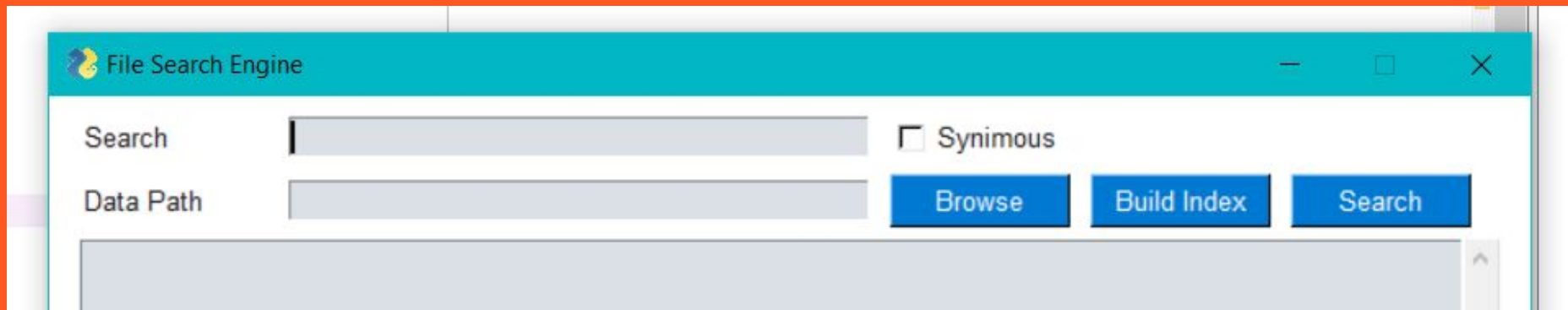
Search Engine



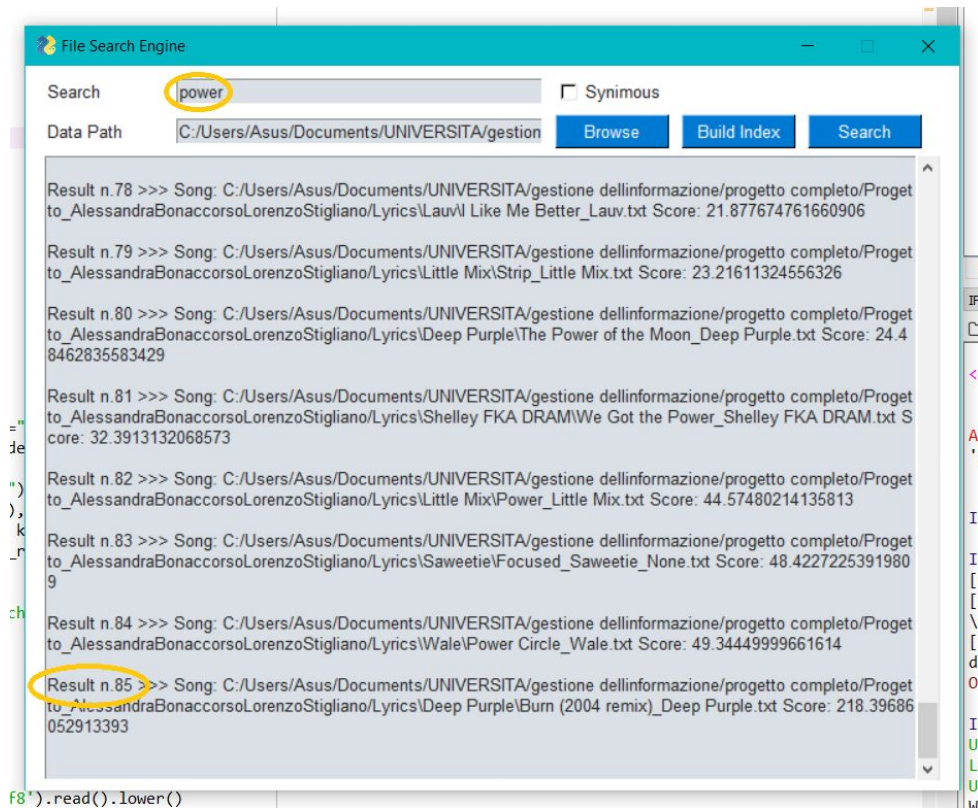
Search engine



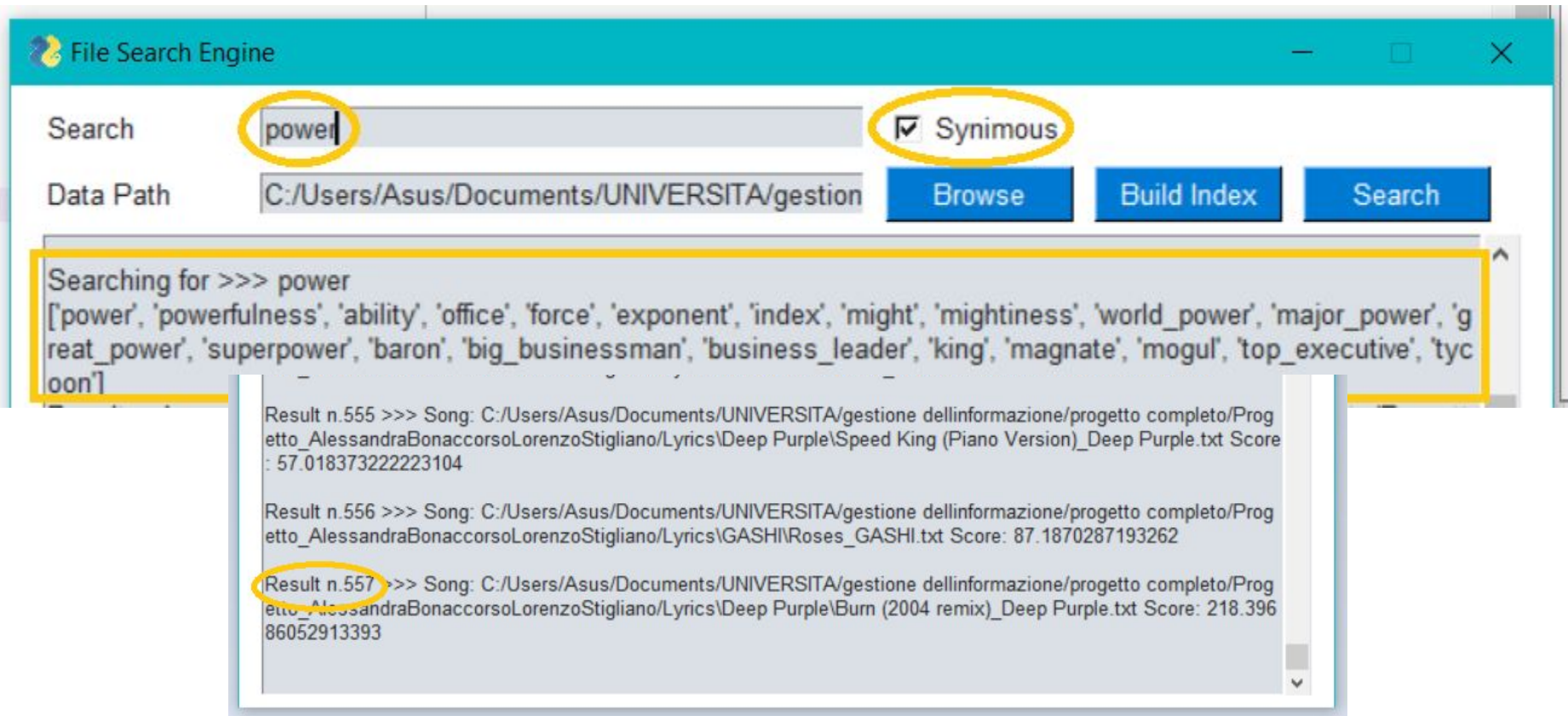
Esempi di query



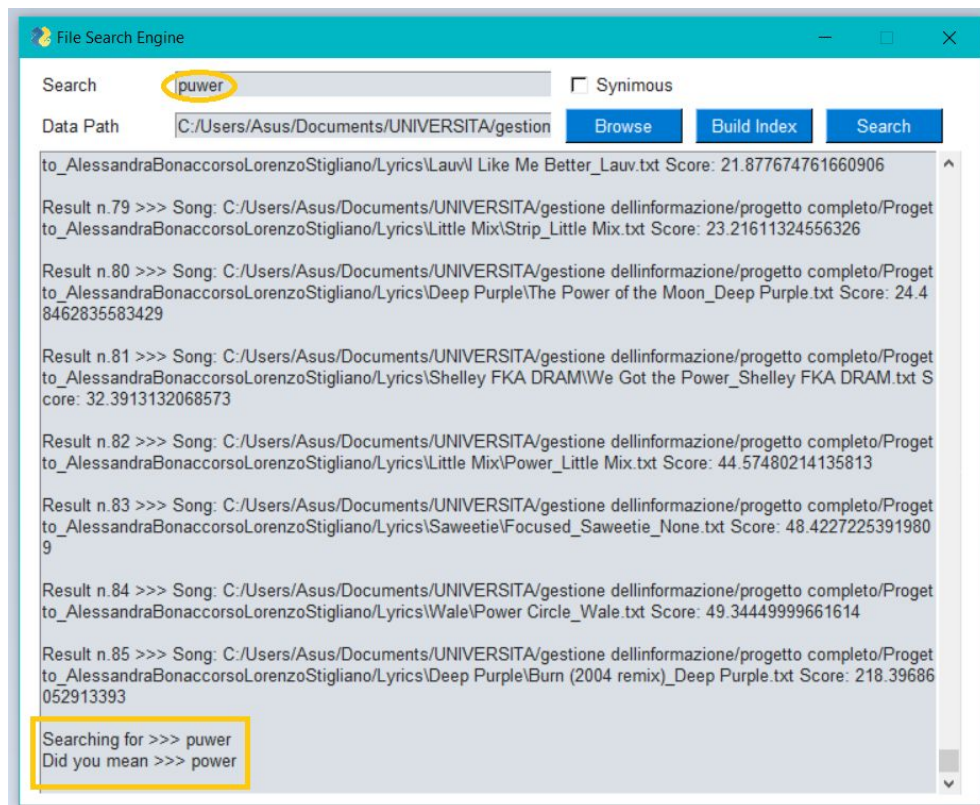
Ricerca della parola "power"



Ricerca con sinonimi di “power”



In caso di errore di spelling



Screenshoot benchmark

```
Terminale
0.614044189453125
{'results': <Top 10 Results for Term('content', 'love') runtime=0.00090
39319993462414>, 'total': 84, 'pagecount': 9, 'pagenum': 1, 'offset': 0
, 'pagelen': 10}

Process returned 0 (0x0)      execution time : 0.711 s
Press [ENTER] to continue...
```

File Search Engine

Search ☐ Synimous

Data Path

Building the Inverted Index >>> ...
0.010276079177856445
Done!

Searching for >>> love

Result n.1 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/R.I.P. (Rock in Peace)_AC_DC_Dirty Deeds Done Dirt Cheap [Australian Edition].txt Score: 0.595519803721176

Result n.2 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Demon Fire_AC_DC_POWER UP.txt Score: 0.606701029371403

Result n.3 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Rock or Bust_AC_DC_Rock or Bust.txt Score: 0.6506819660893806

Result n.4 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Have a Drink On Me_AC_DC_Back in Black.txt Score: 0.6709479325865876

Result n.5 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Let's Make It_A_C_DC_The Razors Edge.txt Score: 0.7333819631187235

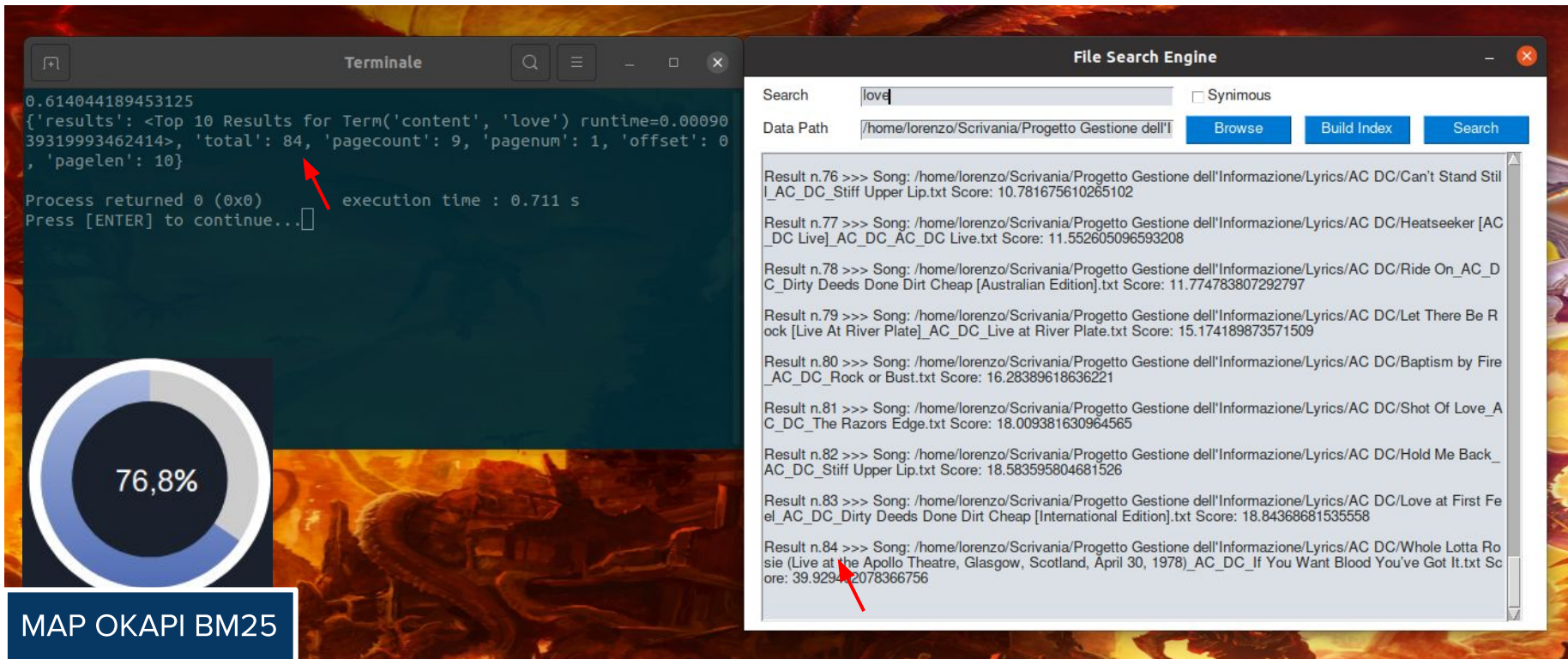
Result n.6 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Give It Up_AC_DC_Stiff Upper Lip.txt Score: 0.7469440574289467

Result n.7 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Through the Mist s of Time_AC_DC_POWER UP.txt Score: 0.7504133192594962

Result n.8 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Whole Lotta Rose [AC_DC Live]_AC_DC_AC_DC Live.txt Score: 0.764618734315693

Result n.9 >>> Song: /home/lorenzo/Scrivania/Progetto Gestione dell'Informazione/Lyrics/AC DC/Got Some Rock

Screenshoot Benchmark



Conclusioni

Soluzioni alternative implementate:

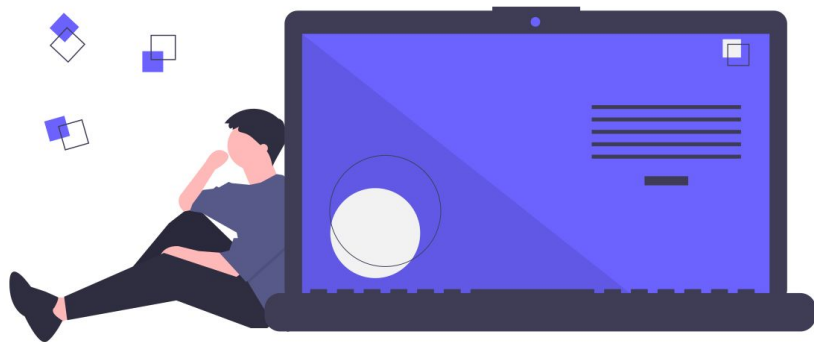
1. Modello booleano
2. Indexing (non utilizzando le librerie di whoosh e lucene)

Cosa avremmo voluto implementare:

- ricerca per emozioni/mood
- ricerca per generi

Maggiori difficoltà:

- download dump
- problemi di encoding a seconda del sistema operativo
- didattica a distanza
- lavoro di gruppo a distanza



Fine

