Customized Wikipedia search engine

Information Retrieval 2021-2-F1801Q110

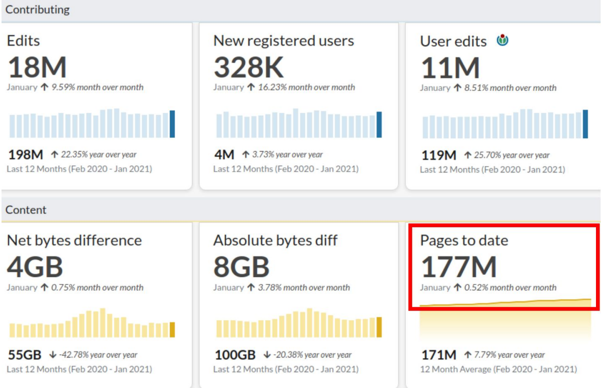
Renzo Arturo Alva Principe
746799

Introduction

Wikipedia is a big project:

- multilingual and open-collaborative
- 177 millions of pages
- pages organized by category

→ need a search engine to retrieve the articles



Contributing

Edits
18M
January ↑ 9.59% month over month
198M ↑ 22.35% year over year
Last 12 Months (Feb 2020 - Jan 2021)

New registered users
328K
January ↑ 16.23% month over month
4M ↑ 3.73% year over year
Last 12 Months (Feb 2020 - Jan 2021)

User edits
11M
January ↑ 8.51% month over month
119M ↑ 25.70% year over year
Last 12 Months (Feb 2020 - Jan 2021)

Content

Net bytes difference
4GB
January ↑ 0.75% month over month
55GB ↓ -42.78% year over year
Last 12 Months (Feb 2020 - Jan 2021)

Absolute bytes diff
8GB
January ↑ 3.78% month over month
100GB ↓ -20.38% year over year
Last 12 Months (Feb 2020 - Jan 2021)

Pages to date
177M
January ↑ 0.52% month over month
171M ↑ 7.79% year over year
12 Month Average (Feb 2020 - Jan 2021)

https://stats.wikimedia.org/#/all-wikipedia-projects

Dataset

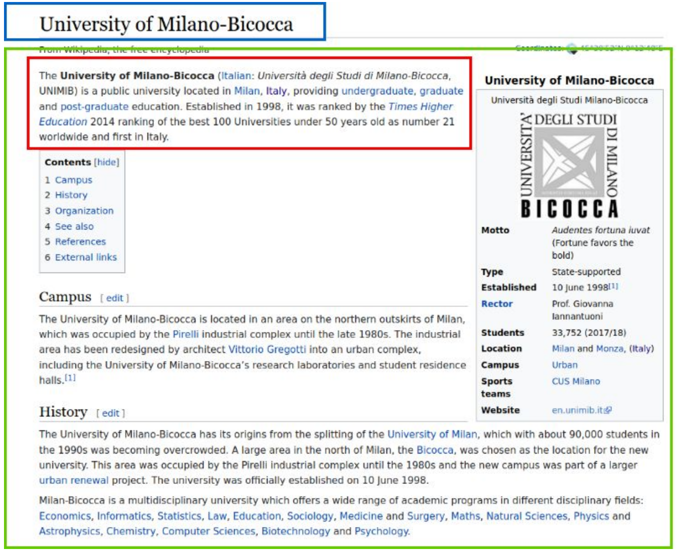
Considered fields:

- **title:** it can be seen as an identifier of the page (I do not consider multi-lingual versions of pages)
- **abstract:** is a concise description of the article which help the reader quickly ascertain the page's content
- **text:** it contains the full description of the page. In this case i incorporated the abstract into the test for indexing purposes
- **url:** the URL to the wikipedia page
- **citations:** number of pages citing the page
- **citations_norm:** number of citations normalized

3 categories selected:

1) Golden Globe Award-winning producers
1) Guitar manufacturing companies of the United States
2) Grammy Lifetime Achievement Award winners

100 pages per category
~17k words per page



University of Milano-Bicocca

From Wikipedia, the free encyclopedia

Coordinates: [illegible]

The **University of Milano-Bicocca** (Italian: *Università degli Studi di Milano-Bicocca*, UNIMIB) is a public university located in Milan, Italy, providing undergraduate, graduate and post-graduate education. Established in 1998, it was ranked by the *Times Higher Education* 2014 ranking of the best 100 Universities under 50 years old as number 21 worldwide and first in Italy.

University of Milano-Bicocca

Università degli Studi Milano-Bicocca

Motto	*Audentes fortuna iuvat* (Fortune favors the bold)
Type	State-supported
Established	10 June 1998[1]
Rector	Prof. Giovanna Iannantuoni
Students	33,752 (2017/18)
Location	Milan and Monza, (Italy)
Campus	Urban
Sports teams	CUS Milano
Website	en.unimib.it

Campus [edit]

The University of Milano-Bicocca is located in an area on the northern outskirts of Milan, which was occupied by the Pirelli industrial complex until the late 1980s. The industrial area has been redesigned by architect Vittorio Gregotti into an urban complex, including the University of Milano-Bicocca's research laboratories and student residence halls.[1]

History [edit]

The University of Milano-Bicocca has its origins from the splitting of the University of Milan, which with about 90,000 students in the 1990s was becoming overcrowded. A large area in the north of Milan, the Bicocca, was chosen as the location for the new university. This area was occupied by the Pirelli industrial complex until the 1980s and the new campus was part of a larger urban renewal project. The university was officially established on 10 June 1998.

Milan-Bicocca is a multidisciplinary university which offers a wide range of academic programs in different disciplinary fields: Economics, Informatics, Statistics, Law, Education, Sociology, Medicine and Surgery, Maths, Natural Sciences, Physics and Astrophysics, Chemistry, Computer Sciences, Biotechnology and Psychology.

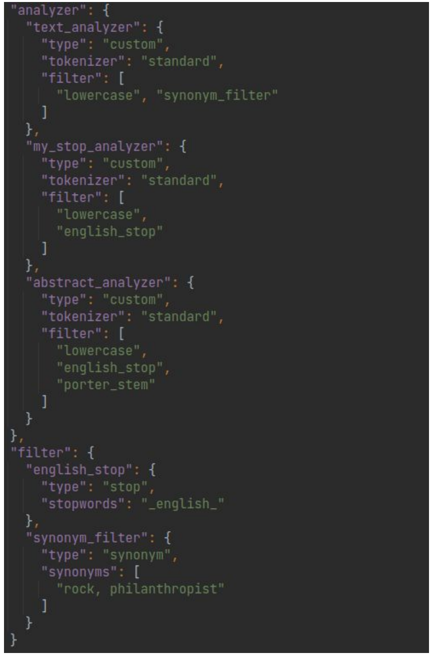
Search Engine - Analyzers and mappings

title: is used for exact matching since i assume that the user who searches in this field have a clear idea of what “entity” is looking for.

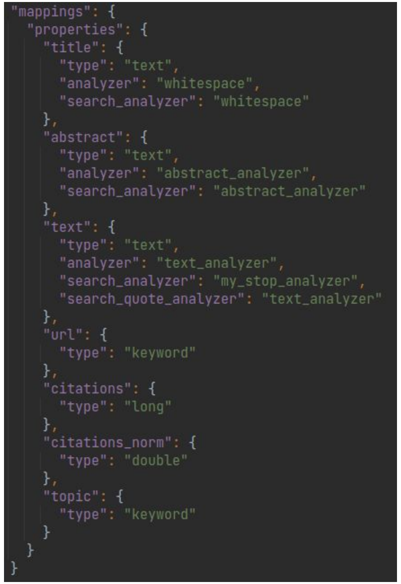
abstract: is meant for exploratory research since it contains most of the keywords in the document

text: is meant for phrase searching. Note that text field includes the abstract content

elasticsearch

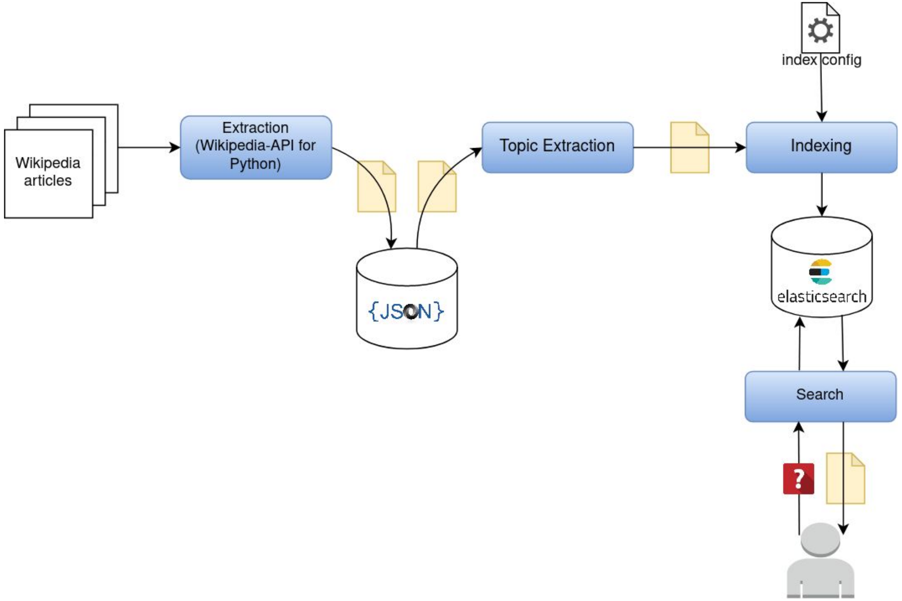


```
"analyzer": {
  "text_analyzer": {
    "type": "custom",
    "tokenizer": "standard",
    "filter": [
      "lowercase", "synonym_filter"
    ]
  },
  "my_stop_analyzer": {
    "type": "custom",
    "tokenizer": "standard",
    "filter": [
      "lowercase",
      "english_stop"
    ]
  },
  "abstract_analyzer": {
    "type": "custom",
    "tokenizer": "standard",
    "filter": [
      "lowercase",
      "english_stop",
      "porter_stem"
    ]
  }
},
"filter": {
  "english_stop": {
    "type": "stop",
    "stopwords": "_english_"
  },
  "synonym_filter": {
    "type": "synonym",
    "synonyms": [
      "rock, philanthropist"
    ]
  }
}
```



```
"mappings": {
  "properties": {
    "title": {
      "type": "text",
      "analyzer": "whitespace",
      "search_analyzer": "whitespace"
    },
    "abstract": {
      "type": "text",
      "analyzer": "abstract_analyzer",
      "search_analyzer": "abstract_analyzer"
    },
    "text": {
      "type": "text",
      "analyzer": "text_analyzer",
      "search_analyzer": "my_stop_analyzer",
      "search_quote_analyzer": "text_analyzer"
    },
    "url": {
      "type": "keyword"
    },
    "citations": {
      "type": "long"
    },
    "citations_norm": {
      "type": "double"
    },
    "topic": {
      "type": "keyword"
    }
  }
}
```

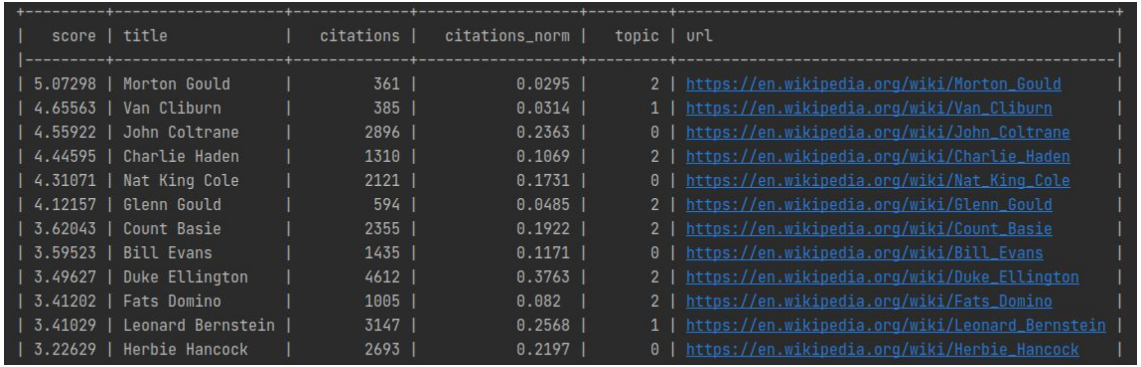
Search Engine - Data flow



Textual search on a specific field (1)

query: {query: {match: {abstract:is an american pianist}}}

notes: it returns both alive and dead pianists (is/was) due to the analyzer.

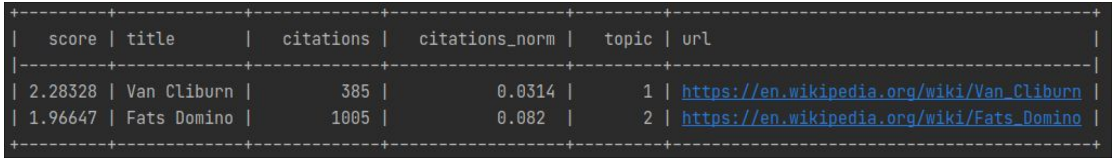


```
+---------+-------------------+-------------+-----------------+---------+------------------------------------------------+
|   score | title             |   citations |  citations_norm |   topic | url                                            |
|---------+-------------------+-------------+-----------------+---------+------------------------------------------------|
| 5.07298 | Morton Gould      |         361 |          0.0295 |       2 | https://en.wikipedia.org/wiki/Morton_Gould     |
| 4.65563 | Van Cliburn       |         385 |          0.0314 |       1 | https://en.wikipedia.org/wiki/Van_Cliburn      |
| 4.55922 | John Coltrane     |        2896 |          0.2363 |       0 | https://en.wikipedia.org/wiki/John_Coltrane    |
| 4.44595 | Charlie Haden     |        1310 |          0.1069 |       2 | https://en.wikipedia.org/wiki/Charlie_Haden    |
| 4.31071 | Nat King Cole     |        2121 |          0.1731 |       0 | https://en.wikipedia.org/wiki/Nat_King_Cole    |
| 4.12157 | Glenn Gould       |         594 |          0.0485 |       2 | https://en.wikipedia.org/wiki/Glenn_Gould      |
| 3.62043 | Count Basie       |        2355 |          0.1922 |       2 | https://en.wikipedia.org/wiki/Count_Basie      |
| 3.59523 | Bill Evans        |        1435 |          0.1171 |       0 | https://en.wikipedia.org/wiki/Bill_Evans       |
| 3.49627 | Duke Ellington    |        4612 |          0.3763 |       2 | https://en.wikipedia.org/wiki/Duke_Ellington   |
| 3.41202 | Fats Domino       |        1005 |          0.082  |       2 | https://en.wikipedia.org/wiki/Fats_Domino      |
| 3.41029 | Leonard Bernstein |        3147 |          0.2568 |       1 | https://en.wikipedia.org/wiki/Leonard_Bernstein |
| 3.22629 | Herbie Hancock    |        2693 |          0.2197 |       0 | https://en.wikipedia.org/wiki/Herbie_Hancock   |
```

Textual search on a specific field (2)

query: {query: {match_phrase: {text:was an american pianist}}}

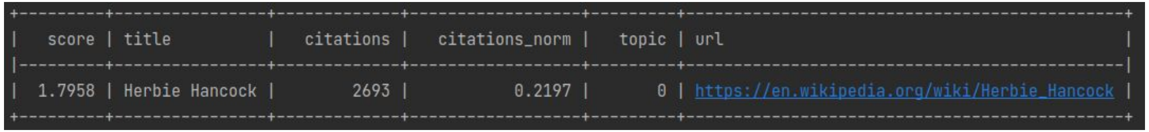
notes: it returns only dead pianists



```
+---------+-------------+-------------+------------------+---------+-------------------------------------------+
|   score | title       |   citations |   citations_norm |   topic | url                                       |
|---------+-------------+-------------+------------------+---------+-------------------------------------------|
| 2.28328 | Van Cliburn |         385 |           0.0314 |       1 | https://en.wikipedia.org/wiki/Van_Cliburn |
| 1.96647 | Fats Domino |        1005 |           0.082  |       2 | https://en.wikipedia.org/wiki/Fats_Domino |
+---------+-------------+-------------+------------------+---------+-------------------------------------------+
```

query: {query: {match_phrase: {text:is an american pianist}}}

notes: it returns only alive pianists



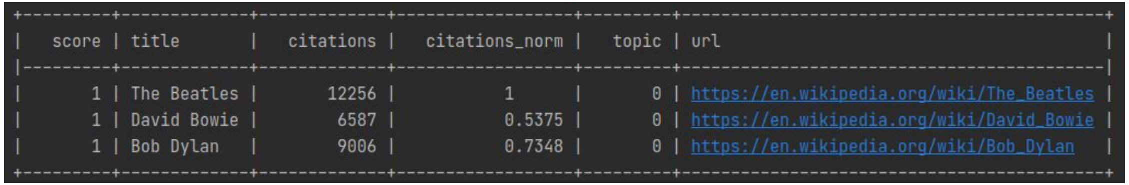
```
+---------+----------------+-------------+------------------+---------+----------------------------------------------+
|   score | title          |   citations |   citations_norm |   topic | url                                          |
|---------+----------------+-------------+------------------+---------+----------------------------------------------|
|  1.7958 | Herbie Hancock |        2693 |           0.2197 |       0 | https://en.wikipedia.org/wiki/Herbie_Hancock |
+---------+----------------+-------------+------------------+---------+----------------------------------------------+
```

Textual search on a combination of fields (1)

query:

```
{query: {bool: {
                must: {match: {abstract: guitarist}},
                must_not: [{match: {abstract: company}}, {match: {abstract: manufacturer}}],
                must: {range: {citations_norm: {gt: 0.500}}}}
}}
```

notes: it returns artists related to guitarists in the abstract of pages with many citations on Wikipedia



```
+---------+-------------+------------+----------------+---------+-------------------------------------------+
|   score | title       |  citations | citations_norm |   topic | url                                       |
|---------+-------------+------------+----------------+---------+-------------------------------------------|
|       1 | The Beatles |      12256 |         1      |       0 | https://en.wikipedia.org/wiki/The_Beatles |
|       1 | David Bowie |       6587 |         0.5375 |       0 | https://en.wikipedia.org/wiki/David_Bowie |
|       1 | Bob Dylan   |       9006 |         0.7348 |       0 | https://en.wikipedia.org/wiki/Bob_Dylan   |
+---------+-------------+------------+----------------+---------+-------------------------------------------+
```

Textual search on a combination of fields (2)

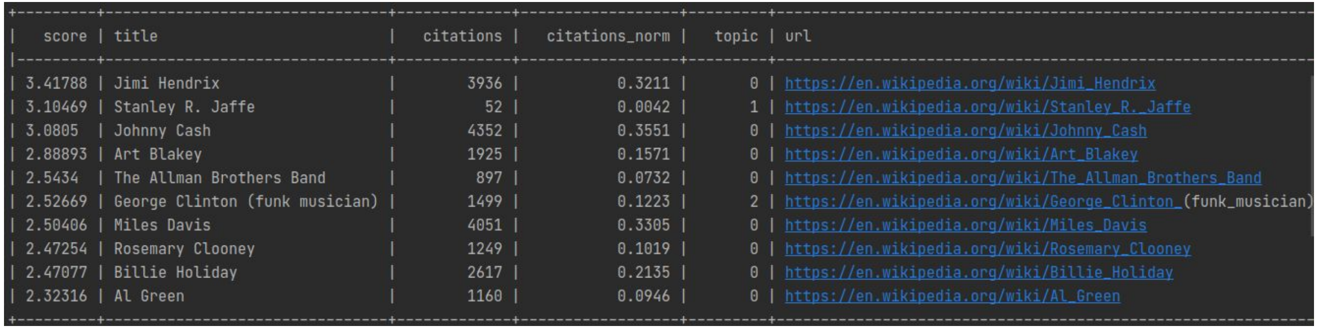
query: {query: {bool: {

must: {match: {abstract: guitarist}},

must: {match: {text: drugs}}}

}}

notes: it returns all the guitarists that have a relation with drugs

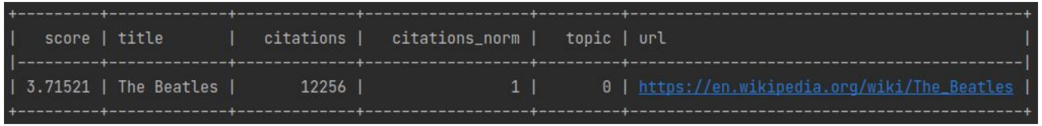


```
+---------+--------------------------------+-------------+------------------+---------+----------------------------------------------------------------------
|   score | title                          |   citations |   citations_norm |   topic | url
|---------+--------------------------------+-------------+------------------+---------+----------------------------------------------------------------------
| 3.41788 | Jimi Hendrix                   |        3936 |           0.3211 |       0 | https://en.wikipedia.org/wiki/Jimi_Hendrix
| 3.10469 | Stanley R. Jaffe               |          52 |           0.0042 |       1 | https://en.wikipedia.org/wiki/Stanley_R._Jaffe
| 3.0805  | Johnny Cash                    |        4352 |           0.3551 |       0 | https://en.wikipedia.org/wiki/Johnny_Cash
| 2.88893 | Art Blakey                     |        1925 |           0.1571 |       0 | https://en.wikipedia.org/wiki/Art_Blakey
| 2.5434  | The Allman Brothers Band       |         897 |           0.0732 |       0 | https://en.wikipedia.org/wiki/The_Allman_Brothers_Band
| 2.52669 | George Clinton (funk musician) |        1499 |           0.1223 |       2 | https://en.wikipedia.org/wiki/George_Clinton_(funk_musician)
| 2.50406 | Miles Davis                    |        4051 |           0.3305 |       0 | https://en.wikipedia.org/wiki/Miles_Davis
| 2.47254 | Rosemary Clooney               |        1249 |           0.1019 |       0 | https://en.wikipedia.org/wiki/Rosemary_Clooney
| 2.47077 | Billie Holiday                 |        2617 |           0.2135 |       0 | https://en.wikipedia.org/wiki/Billie_Holiday
| 2.32316 | Al Green                       |        1160 |           0.0946 |       0 | https://en.wikipedia.org/wiki/Al_Green
+---------+--------------------------------+-------------+------------------+---------+----------------------------------------------------------------------
```

Textual search using fuzzy functionality

query: {query: {fuzzy: {title: {value: batles}}}}

notes: it returns "The Beatles" despite the misspelling



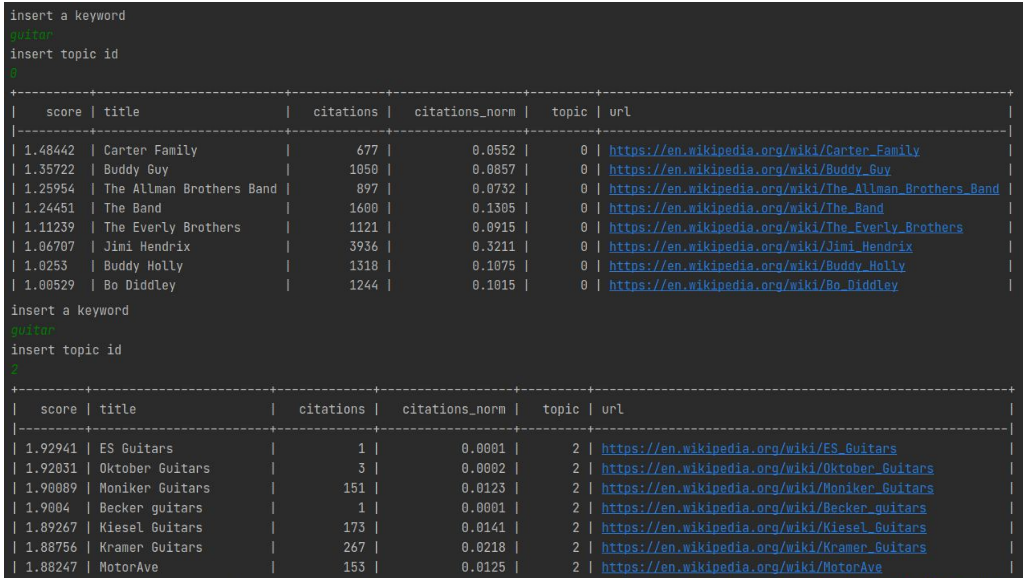
```
+---------+-------------+-------------+------------------+---------+-------------------------------------------+
|   score | title       |   citations |   citations_norm |   topic | url                                       |
|---------+-------------+-------------+------------------+---------+-------------------------------------------|
| 3.71521 | The Beatles |       12256 |                1 |       0 | https://en.wikipedia.org/wiki/The_Beatles |
+---------+-------------+-------------+------------------+---------+-------------------------------------------+
```

Textual search taking into account topics (1)

Topic 0: music, record, album, song, film, award, jazz, fame, includ, hall

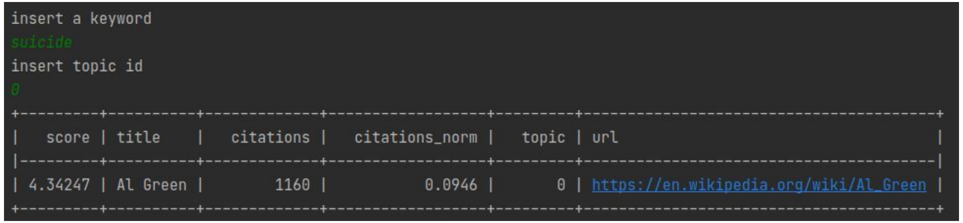
Topic 1: film, award, best, academi, pictur, bear, music, director, nomin, includ

Topic 2: guitar, music, compani, instrument, manufactur, band, record, bass, electr, includ

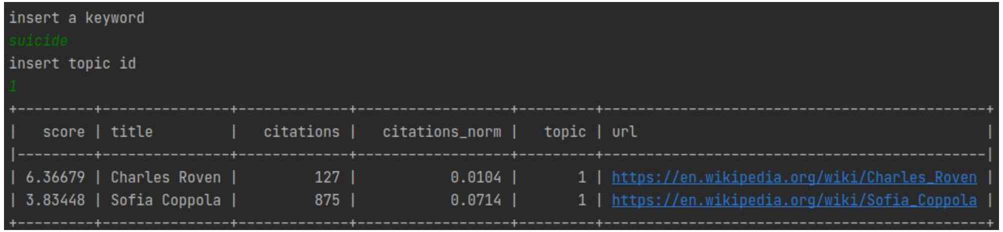


```
insert a keyword
guitar
insert topic id
0
+---------+--------------------------+------------+-----------------+---------+------------------------------------------------------+
|   score | title                    |  citations |  citations_norm |   topic | url                                                  |
|---------+--------------------------+------------+-----------------+---------+------------------------------------------------------|
| 1.48442 | Carter Family            |        677 |          0.0552 |       0 | https://en.wikipedia.org/wiki/Carter_Family          |
| 1.35722 | Buddy Guy                |       1050 |          0.0857 |       0 | https://en.wikipedia.org/wiki/Buddy_Guy              |
| 1.25954 | The Allman Brothers Band |        897 |          0.0732 |       0 | https://en.wikipedia.org/wiki/The_Allman_Brothers_Band |
| 1.24451 | The Band                 |       1600 |          0.1305 |       0 | https://en.wikipedia.org/wiki/The_Band               |
| 1.11239 | The Everly Brothers      |       1121 |          0.0915 |       0 | https://en.wikipedia.org/wiki/The_Everly_Brothers    |
| 1.06707 | Jimi Hendrix             |       3936 |          0.3211 |       0 | https://en.wikipedia.org/wiki/Jimi_Hendrix           |
| 1.0253  | Buddy Holly              |       1318 |          0.1075 |       0 | https://en.wikipedia.org/wiki/Buddy_Holly            |
| 1.00529 | Bo Diddley               |       1244 |          0.1015 |       0 | https://en.wikipedia.org/wiki/Bo_Diddley             |
insert a keyword
guitar
insert topic id
2
+---------+----------------------+-----------+-----------------+---------+------------------------------------------------------+
|   score | title                |  citations |  citations_norm |   topic | url                                                 |
|---------+----------------------+-----------+-----------------+---------+------------------------------------------------------|
| 1.92941 | ES Guitars           |         1 |          0.0001 |       2 | https://en.wikipedia.org/wiki/ES_Guitars             |
| 1.92031 | Oktober Guitars      |         3 |          0.0002 |       2 | https://en.wikipedia.org/wiki/Oktober_Guitars        |
| 1.90089 | Moniker Guitars      |       151 |          0.0123 |       2 | https://en.wikipedia.org/wiki/Moniker_Guitars        |
| 1.9004  | Becker guitars       |         1 |          0.0001 |       2 | https://en.wikipedia.org/wiki/Becker_guitars         |
| 1.89267 | Kiesel Guitars       |       173 |          0.0141 |       2 | https://en.wikipedia.org/wiki/Kiesel_Guitars         |
| 1.88756 | Kramer Guitars       |       267 |          0.0218 |       2 | https://en.wikipedia.org/wiki/Kramer_Guitars         |
| 1.88247 | MotorAve             |       153 |          0.0125 |       2 | https://en.wikipedia.org/wiki/MotorAve               |
```

Textual search taking into account topics (2)



```
insert a keyword
suicide
insert topic id
0
+---------+----------+-------------+------------------+---------+----------------------------------------+
|   score | title    |   citations |   citations_norm |   topic | url                                    |
|---------+----------+-------------+------------------+---------+----------------------------------------|
| 4.34247 | Al Green |        1160 |           0.0946 |       0 | https://en.wikipedia.org/wiki/Al_Green |
+---------+----------+-------------+------------------+---------+----------------------------------------+
```

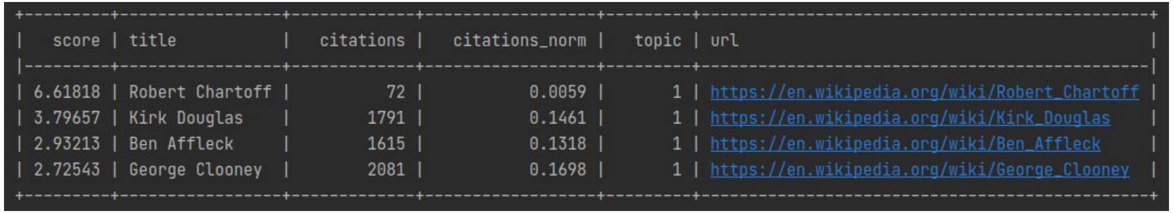


```
insert a keyword
suicide
insert topic id
1
+---------+---------------+-------------+------------------+---------+---------------------------------------------+
|   score | title         |   citations |   citations_norm |   topic | url                                         |
|---------+---------------+-------------+------------------+---------+---------------------------------------------|
| 6.36679 | Charles Roven |         127 |           0.0104 |       1 | https://en.wikipedia.org/wiki/Charles_Roven |
| 3.83448 | Sofia Coppola |         875 |           0.0714 |       1 | https://en.wikipedia.org/wiki/Sofia_Coppola |
+---------+---------------+-------------+------------------+---------+---------------------------------------------+
```

Textual search using synonyms

query: { query: {match: {abstract: philanthropist}}}

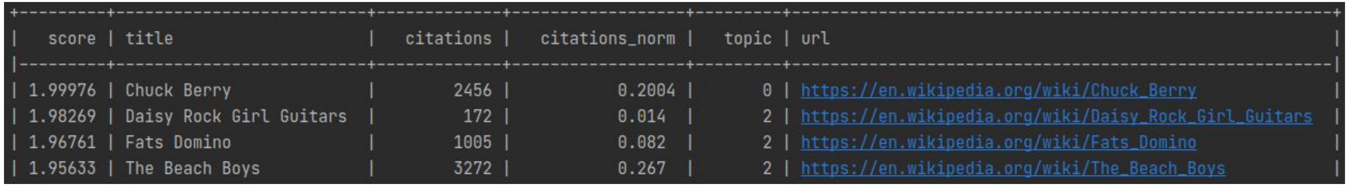
notes: it returns al the philanthropist from the corpus which are producers



```
+---------+-----------------+-------------+------------------+---------+-----------------------------------------------+
|   score | title           |   citations |   citations_norm |   topic | url                                           |
|---------+-----------------+-------------+------------------+---------+-----------------------------------------------|
| 6.61818 | Robert Chartoff |          72 |           0.0059 |       1 | https://en.wikipedia.org/wiki/Robert_Chartoff |
| 3.79657 | Kirk Douglas    |        1791 |           0.1461 |       1 | https://en.wikipedia.org/wiki/Kirk_Douglas    |
| 2.93213 | Ben Affleck     |        1615 |           0.1318 |       1 | https://en.wikipedia.org/wiki/Ben_Affleck     |
| 2.72543 | George Clooney  |        2081 |           0.1698 |       1 | https://en.wikipedia.org/wiki/George_Clooney  |
+---------+-----------------+-------------+------------------+---------+-----------------------------------------------+
```

query: {query: {match_phrase: {text: philanthropist}}}

notes: since i intentionally declare "philanthropist" as synonym of "rock" in the text_analyzer filter, this query returns rocks stars



```
+---------+---------------------------+-------------+------------------+---------+---------------------------------------------------------+
|   score | title                     |   citations |   citations_norm |   topic | url                                                     |
|---------+---------------------------+-------------+------------------+---------+---------------------------------------------------------|
| 1.99976 | Chuck Berry               |        2456 |           0.2004 |       0 | https://en.wikipedia.org/wiki/Chuck_Berry               |
| 1.98269 | Daisy Rock Girl Guitars   |         172 |           0.014  |       2 | https://en.wikipedia.org/wiki/Daisy_Rock_Girl_Guitars   |
| 1.96761 | Fats Domino               |        1005 |           0.082  |       2 | https://en.wikipedia.org/wiki/Fats_Domino               |
| 1.95633 | The Beach Boys            |        3272 |           0.267  |       2 | https://en.wikipedia.org/wiki/The_Beach_Boys            |
```

Conclusions

Elasticsearch:

- flexible
- indexing high customizable
- boolean operators, vectors similarity
- support synonym, fuzzy queries
- etc ..

Topic modeling: automatic and in some cases helps to filter the results

→ high quality results