

# 1. projekt – tražilica dokumenata

Izraditi (web) aplikaciju koja će omogućiti

1. Unos tekstualnog sadržaja
2. Pretragu u bazi podataka pohranjenog sadržaja - tražilica dokumenta koji zadovoljavaju postavljeni uvjet
3. Analizu najčešće postavljenih upita u zadanom vremenskom periodu



## FULL & FUZZY TEXT SEARCH

- Menu
  - ◊ [Search](#)
  - ◊ Add

Search

Copyright © ZPR FER 2013

# 1. projekt – unos tekstualnog sadržaja

## 1. Unos tekstualnog sadržaja

- Temu odabrati proizvoljno
- Sadržaj spremati u PostgreSQL bazu podataka (dovoljna je jedna relacija s 2 atributa: ključ i tekstualni podatak). Koristiti tekstove na engleskom jeziku zbog nepostojanja potpune podrške za FTS za hrvatski jezik.



## FULL & FUZZY TEXT SEARCH

◆ Menu

- ◇ Search
- ◇ [Add](#)

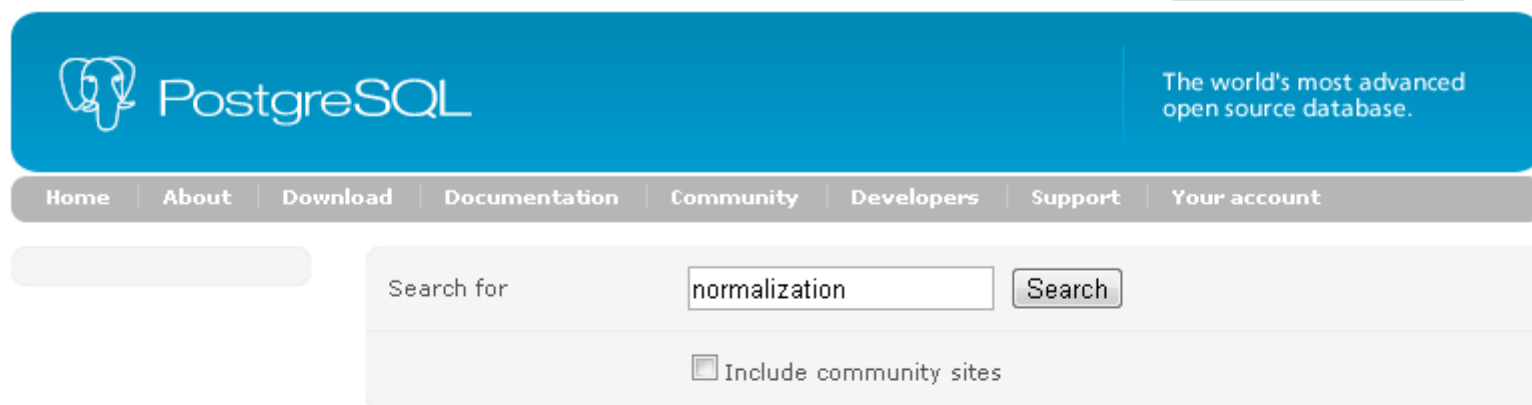
Silver Linings Playbook

Add

Row added

Copyright © ZPR FER 2013

# 1. projekt – tražilica dokumenta



PostgreSQL

The world's most advanced open source database.

Home | About | Download | Documentation | Community | Developers | Support | Your account

Search for

☐ Include community sites

Results 1-20 of more than 1000.

Result pages: 1 [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [Next](#)

1. [PostgreSQL: Documentation: 9.1: Dictionaries](#) [0.58]  
...**normalized** word is called a lexeme . Aside from improving search quality, **normalization** and removal of stop...  
<http://www.postgresql.org/docs/9.1/static/textsearch-dictionaries.html>
2. [PostgreSQL: Documentation: 9.1: Hot Standby](#) [0.54]  
...**Normal** Large Home Documentation Manuals PostgreSQL 9.1 This page in other versions: 9.2 9.1 9.0 Unsupported...  
<http://www.postgresql.org/docs/9.1/static/hot-standby.html>
3. [PostgreSQL: Documentation: 9.1: Continuous Archiving and Point-in-Time Recovery \(PITR\)](#) [0.54]  
...**Normal** Large Home Documentation Manuals PostgreSQL 9.1 This page in other versions: 9.2 9.1 9.0 8.4 Unsupported...  
<http://www.postgresql.org/docs/9.1/static/continuous-archiving.html>

# 1. projekt – tražilica dokumenta

## 2. Pretragu u bazi podataka pohranjenog sadržaja

- Potrebno je realizirati tri vrste pretrage:
  - a. Dohvat dokumenata koji sadrže traženi uzorak (uzorke) u neizmijenjenom obliku - **Exact string matching**
  - b. Dohvat dokumenata koji sadrže traženi uzorak (uzorke) u normaliziranom obliku - **Use dictionaries**
  - c. Dohvat dokumenata koji sadrže približno jednak uzorak (uzorke) - **Fuzzy string matching**
- Podržati povezivanje zadanih uzoraka pretrage logičkim operatorom
  1. AND
  2. OR
- Prikazati sadržaj SQL upita kojim će se dohvatiti dokumenti za zadani uzorak (SQL string)

The screenshot shows a web-based search interface. On the left is a sidebar menu with 'Menu', 'Search', and 'Add'. The main area contains a search bar, logical operator buttons (AND/OR), matching method buttons (Exact string matching, Use dictionaries, Fuzzy string matching), and a large text area for the SQL string. Annotations with blue arrows point to specific elements: '1' points to the search bar, '2' points to the AND/OR buttons, 'a' points to the 'Exact string matching' button, 'b' points to the 'Use dictionaries' button, and 'c' points to the 'Fuzzy string matching' button.

1 2 a b c

Menu  
◊ Search  
◊ Add

AND OR Exact string matching Use dictionaries Fuzzy string matching

SQL string:

# 1. projekt – tražilica dokumenta

2. Pretraga u bazi podataka pohranjenog sadržaja
  - a) u predviđeno područje se unosi uzorak (niz riječi)
  - b) obzirom na odabrane opcije "gradi se" SQL upit i prikazuje u predviđenom području
  - c) korisniku se prikazuju informativni podaci o dohvaćenim dokumentima s podebljanim riječima temeljem kojih je dokument kvalificiran kao rezultat, i rang dokumenta
  - d) odabirom konkretnog rezultata dohvata prikazuje se sadržaj dokumenta.

The screenshot shows a web-based search interface. At the top, there is a search input field containing the text "Legend of Tarzan" "Lord of" and a "Search" button. Below the input field, there are two groups of radio buttons. The first group has "AND" selected. The second group has "Exact string matching" selected, with "Use dictionaries" and "Fuzzy string matching" as options. Below these options, the text "SQL string:" is followed by a text area containing a SQL query. The query is: 

```
SELECT movie_id,
       ts_headline(title, to_tsquery(' (Legend & of & Tarzan) & (Lord & of) ')),
       title
, ts_rank(title_tsvector, to_tsquery (' (Legend & of & Tarzan) & (Lord & of) ')) rank
FROM movies
WHERE title LIKE '%Legend of Tarzan%'
AND title LIKE '%Lord of%'
ORDER BY rank DESC
```

 Below the SQL query, it says "Number of documents retrieved: 1". Below that, there is a blue hyperlink: "Greystoke: The Legend of Tarzan, Lord of the Apes [0,2669128]". To the right of the search results, there is a snippet of the document: "Greystoke: The Legend of Tarzan, Lord o". Annotations: A blue arrow labeled 'a' points to the search input field. A blue arrow labeled 'b' points to the SQL query text area. A blue arrow labeled 'c' points to the search results, specifically the hyperlink. A blue arrow labeled 'd' points to the document snippet.

"Legend of Tarzan" "Lord of" Search

☒ AND ☐ OR ☒ Exact string matching ☐ Use dictionaries ☐ Fuzzy string matching

SQL string:

```
SELECT movie_id,
       ts_headline(title, to_tsquery(' (Legend & of & Tarzan) & (Lord & of) ')),
       title
, ts_rank(title_tsvector, to_tsquery (' (Legend & of & Tarzan) & (Lord & of) ')) rank
FROM movies
WHERE title LIKE '%Legend of Tarzan%'
AND title LIKE '%Lord of%'
ORDER BY rank DESC
```

Number of documents retrieved: 1

[Greystoke: The Legend of Tarzan, Lord of the Apes \[0,2669128\]](#)

Greystoke: The Legend of Tarzan, Lord o

# 1. projekt – tražilica dokumenta

Za označenu opciju:

- |                                 |  |
|---------------------------------|--|
| <b>a. Exact string matching</b> | koristiti neki od operatora: LIKE, SIMILAR TO, ~ |
| <b>b. Use dictionaries</b>      | koristiti operator @@                            |
| <b>c. Fuzzy string matching</b> | koristiti operator %                             |
- 
- Podržati traženje fraza (znakovni niz naveden unutar navodnika) i "jednostavnih" riječi kombiniranih logičkim operatorima AND i OR
  - Za prikazivanje sažetih informacija o dohvaćenim dokumentima s podebljanim ključnim riječima koristiti funkciju *ts\_headline*
  - Za rangiranje dokumenata koristiti funkciju *ts\_rank* ili *ts\_rank\_cd*

# 1. projekt – Exact string matching

## Primjer:

Želimo pronaći dokumente koji sadrže fraze “[Legend of Tarzan](#)” ili “[Lord of](#)” ili riječ [Dance](#)

☐ AND ☒ OR

☒ Exact string matching ☐ Use dictionaries ☐ Fuzzy string matching

SQL string:

```
SELECT movie_id,  
       ts_headline(title, to_tsquery('Dance | (Legend & of & Tarzan) | (Lord &  
of)')), title  
, ts_rank(title_tsvector, to_tsquery ('Dance | (Legend & of & Tarzan) |  
(Lord & of)')) rank  
FROM movies  
WHERE title LIKE '%Legend of Tarzan%'  
OR title LIKE '%Lord of%'  
OR title LIKE '%Dance%'  
ORDER BY rank DESC
```

Number of documents retrieved: 4

[Greystoke: The Legend of Tarzan, Lord of the Apes \[0,03647562\]](#)

[Lord of Illusions \[0,01215854\]](#)

[Shall We Dance? \[0\]](#)

[Dances with Wolves \[0\]](#)

# 1. projekt – Use dictionaries

**Primjer:** Želimo pronaći dokumente koji u normaliziranom obliku sadrže normalizirane fraze “[Legend of Tarzan](#)” ili “[Lord of](#)” ili riječ [Dance](#)

Za razliku od opcije Exact string matching, ovdje treba dohvatiti i dokumente koji sadrže neki drugi oblik riječi iz teksta pretrage tj. potrebno je normalizirati riječi, ukloniti stop riječi i sl.

Modelirajte bazu podataka tako da koliko god možete ubrzate pretragu: npr. dodatna pohrana normaliziranog teksta, kreiranje specijalnih indeksa.

☐ AND ☒ OR

☐ Exact string matching ☒ Use dictionaries ☐ Fuzzy string matching

SQL string:

```
SELECT movie_id,  
       ts_headline(title, to_tsquery('Dance | (Legend & of & Tarzan) | (Lord &  
of)')), title  
  , ts_rank(title_tsvector, to_tsquery ('Dance | (Legend & of & Tarzan) |  
(Lord & of)')) rank  
FROM movies  
WHERE title_tsvector @@ to_tsquery('english','Dance')  
OR title_tsvector @@ to_tsquery('english','Legend & of & Tarzan')  
OR title_tsvector @@ to_tsquery('english','Lord & of')  
ORDER BY rank DESC
```

Number of documents retrieved: 11

[Greystoke: The Legend of Tarzan, Lord of the Apes \[0.03647562\]](#)

[Lord Of The Flies \[0.01215854\]](#)

[Lord Jim \[0.01215854\]](#)

[Lord of Illusions \[0.01215854\]](#)



# 1. projekt – Fuzzy string matching

**Primjer:** Želimo pronaći dokumente koji sadrže dijelove nalik frazama “[Legend of Tarzan](#)” ili “[Lord of](#)” ili dijelove nalik riječi [Dance](#)

Modelirajte bazu podataka tako da koliko god možete ubrzate pretragu (specijalni indeksi).

</div>
<div>
<input type="button" value="Search" />
</div>
<div>
<input type="radio" /> AND
<input checked="" type="radio" /> OR
<input type="radio" /> Exact string matching
<input type="radio" /> Use dictionaries
<input checked="" type="radio" /> Fuzzy string matching
</div>
<div>
SQL string:
<pre>
SELECT movie\_id,
 ts\_headline(title, to\_tsquery('Dance | (Legend & of & Tarzan) | (Lord & of)')), title
, ts\_rank(title\_tsvector, to\_tsquery ('Dance | (Legend & of & Tarzan) | (Lord & of)')) rank
FROM movies
WHERE title % 'Legend of Tarzan'
OR title % 'Lord of'
OR title % 'Dance'
ORDER BY rank DESC
</pre>
</div>
</div>
<div>
Number of documents retrieved: 12
<div>
<a href="#">Greystoke: The Legend of Tarzan, Lord of the Apes [0.03647562]
<a href="#">Lord of Illusions [0.01215854]
<a href="#">The Legend of Billie Jean [0.01215854]
<a href="#">Urban Legend [0.01215854]
<a href="#">Lord Of The Flies [0.01215854]
<a href="#">Lord Jim [0.01215854]
<a href="#">Legends of the Fall [0.01215854]
</div>
</div>
</div>

© ZPR-FER - Zagreb

Napredni modeli i baze podataka 2014/2015

47

# 1. projekt - Analiza najčešće postavljanih upita u zadanom vremenskom periodu

1. Zadati vremenski period u kojem treba provesti analizu:  
datumOd – datum do (npr. 10/10/2014 -13/10/2014)
2. Odabrati granulaciju analize:
  - dan ili
  - sat
3. Korištenjem naprednih mogućnosti SQL-a (pivotiranje) izraditi izvještaj s pregledom broja postavljanja konkretnog upita za zadani period:

Po danima:

querystring character(200)	d10102014 integer	d11102014 integer	d12102014 integer	d13102014 integer
'Dance' & 'Legend of Tarzan' & 'Lord'	4	3	2	
'Lord' & 'Dance'	3	2	2	

Ili po satima:

	querystring character(200)	s00_01 integer	s01_02 integer	s02_03 integer	s03_04 integer	s04_05 integer	s05_06 integer	s06_07 integer	s07_08 integer	s08_09 integer	s09_10 integer	s10_11 integer	s11_12 integer	s12_13 integer
1	'Dance' & 'Legend of Tarzan' & 'Lord'								1		3	3	1	
2	'Lord' & 'Dance'										3	3	1	

Da bi analiza bila moguća potrebno je bilježiti upite koje korisnici postavljaju.

Gornje slike su ilustracija rješenja (pgAdmin screenshot) koje je potrebno implementirati u okviru (web) aplikacije.

# 1. projekt

- Da biste mogli koristiti funkcije za Full Text Search i Fuzzy Text Search morate uključiti module `fuzzystrmatch` i `pg_trgm`, odnosno `tablefunc` za korištenje funkcija za pivotiranje. Module je potrebno uključiti u svakoj bazi podataka u kojoj ih namjeravate koristiti. „Registriraju“ se izvođenjem sljedećih naredbi:

```
CREATE EXTENSION fuzzystrmatch;      -- (soundex, levenshtein, metaphone)
CREATE EXTENSION pg_trgm;             -- (similarity , show_Trgm,..., %, <->)
CREATE EXTENSION tablefunc;           -- (crosstab)
```

- 20 bodova.
- Rješenje (dokumentaciju i datoteku sa shemom baze podataka) postaviti u vlastiti NMIBP\P1 direktorij na FTP serveru
- Objasniti ulogu eventualno kreiranih indeksa.
- **Rok za predaju: 27.10.2014 u 10:00.**