TP1 PostgreSQL

Recherche et recommendation

Andrei Arion ,LesFurets.com tp-bigdata@lesfurets.com

1. Recherche exacte / pattern matching

1. Tous les filmes qui ont le mot ***stardust*** dans leur nom.

```
SELECT * FROM MOVIES WHERE title ILIKE '%stardust%' ;
SELECT * FROM MOVIES WHERE title ~~* '%stardust%' ;
```

movie_id	title	genre
611	Stardust	(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0)
1731	Stardust Memories	(0, 0, 0, 5, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

2. Compter tous les filmes dont le titre ne commence pas par le mot ***the***

```
SELECT * FROM MOVIES WHERE title NOT ILIKE 'the%' ;
```

movie_id	title	genre
1	Star Wars	(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0, 0, 0)
2	Forrest Gump	(0, 0, 0, 5, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
3	American Beauty	(0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
4	Citizen Kane	(0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
7	Apocalypse Now	(5, 0, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0)
8	Unforgiven	(0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5)
9	Twelve Monkeys	(0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 7, 0, 7, 0)
10	Absolute Power	(0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 5, 0)

3. Tous les filmes qui ont le mot ***war*** dans le titre mais pas en dernière position

```
 SELECT * FROM MOVIES WHERE title  LIKE '%war_' ;
```

movie_id	title	genre
1723	Hardware	(0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 5, 0, 0, 0)
2235	The Swarm	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0)
2702	Quentin Durward	(0, 5, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0)

(3 rows)

4. Afficher le plan d'exécution de la dernière requête(EXPLAIN).

```
EXPLAIN SELECT * FROM MOVIES WHERE title  LIKE '%war_' ;
```

```
                        QUERY PLAN
-----------------------------------------------------------
 Seq Scan on movies     (cost 0.00..160.76 rows 1 width 315)
   Filter       (title ~~ '%war_'       text)
(2 rows)
```

5. Rajouter un index full text

```
CREATE INDEX movies_title_pattern ON movies (lower(title) text_pattern_ops);
```

6. Plan d'execution:

```
EXPLAIN SELECT * FROM MOVIES WHERE title  LIKE '%war_' ;
```

2. Distance Levenshtein

7. La distance levenshtein entre les mots ***execution*** et ***intention***

```
SELECT levenshtein('execution','intention');
```

levenshtein
5

(1 row)

8. Tous les films qui sont a une distance *levenshtein* inférieure a 9 de la chaine suivante: ***a hard day nght***

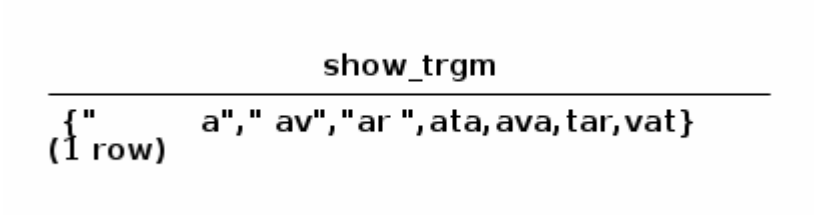
```
SELECT * FROM movies WHERE levenshtein(title,'hard day nght')<9;
```

movie_id	title	genre
245	A Hard Day's Night	(0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0)
561	Hard Target	(5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0)
873	Dead Bang	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0)
899	No Way Out	(0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0)
1274	Hard to Kill	(5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
2033	Last Night	(0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
2276	Bad Timing	(0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0)
2348	Fandango	(0, 5, 0, 5, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
2620	Dead of Night	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0)
2644	Hard Country	(0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

3. N-gram

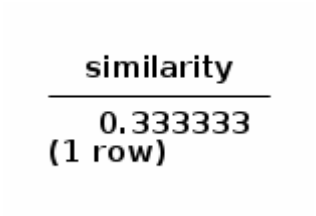
9. Tous les tri-grammes du mot ***Avatar***

```
SELECT show_trgm('Avatar');
```



10. La similarité entre ***VOTKA*** et ***VODKA***

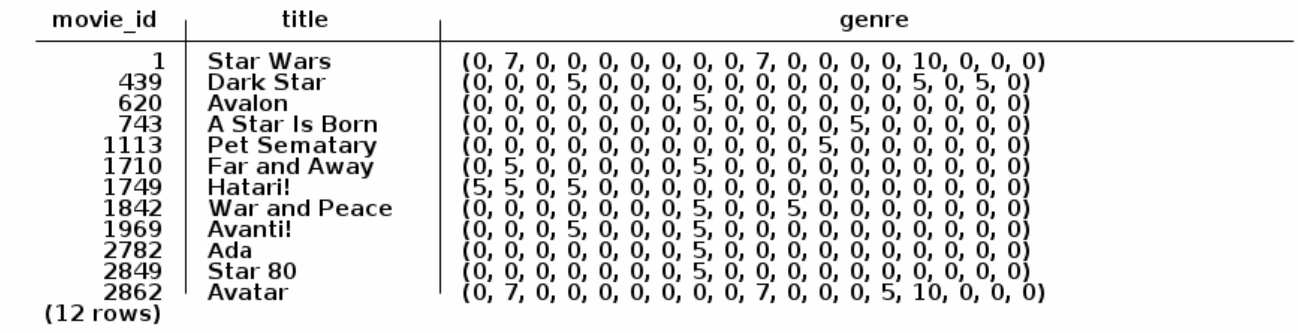
```
SELECT similarity('VOTKA','VODKA');
```



11. Tous les filmes dont le titre est similaire a plus de 0.1% du titre ***Avatar*** .

```
SELECT * FROM movies WHERE similarity(title , 'Avatar')>=0.1;

-- Alternative
SELECT set_limit(0.1);
SELECT * FROM movies WHERE title % 'Avatar';
```



4. Full text search

12. Trouver les filmes qui contiennent les formes grammaticales des mots 'night' et 'day':

```
SELECT to_tsvector('A Hard Day''s Night'), to_tsquery('english', 'night & day');

SELECT title
FROM movies
WHERE to_tsvector(title) @@ to_tsquery('english', 'night & day');

-- Alternative
SELECT title
FROM movies
WHERE title @@ 'night & day';
```

to_tsvector	to_tsquery
'day' 3 'hard' 2 'night' 5	'night' & 'day' & 'dai'

(1 row)

title
A Hard Day's Night
Six Days Seven Nights
Long Day's Journey Into Night

(3 rows)

5. Recherche phonetique

13. Trouver les filmes qui ont des acteurs dont les noms se prononcent pareil.

```
SELECT a1.name,a2.name FROM actors a1, actors a2
WHERE a1.name<>a2.name
AND soundex(a1.name)=soundex(a2.name)
AND metaphone(a1.name,30)=metaphone(a2.name,30)
AND a1.actor_id<a2.actor_id;
```

name	name
Alan Alda	Alan Ladd
Anna Karen	Anna Karina
Annette O'Toole	Annette O`Toole
Bill Pullman	Bill Pulmann
Bill Duck	Bill Duke
Barry Newman	Byron Mann
Cesare Danova	Cesare Doneva
Claude Atkins	Claude Atkuns
David Bowe	David Bowie
David Gale	David Kelly

14. Trouver les acteurs avec un nom similaire a ***Robin Wiliams***, triés par similarité (combiner %, metaphone et levenshtein):

```
-- Exercice
```

6 Recherche multi-dimensionnelle

15. Afficher les notes du film ***Star Wars***

```
SELECT genre FROM movies WHERE title='Star Wars';

SELECT m.title, g.name, cube_ll_coord(m.genre,g.position)
FROM movies m, genres g
WHERE m.title='Star Wars';
```

genre
(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0, 0, 0)

(1 row)

title	name	cube_ll_coord
Star Wars	Action	0
Star Wars	Adventure	7
Star Wars	Animation	0
Star Wars	Comedy	0
Star Wars	Crime	0
Star Wars	Disaster	0
Star Wars	Documentary	0
Star Wars	Drama	0
Star Wars	Eastern	0
Star Wars	Fantasy	7
Star Wars	History	0
Star Wars	Horror	0
Star Wars	Musical	0
Star Wars	Romance	0
Star Wars	SciFi	10
Star Wars	Sport	0
Star Wars	Thriller	0
Star Wars	Western	0

(18 rows)

16. Quelle est la note du film ***Star Wars*** dans la catégorie 'Animation'

```
SELECT m.title, g.name, cube_ll_coord(m.genre,g.position)
FROM movies m, genres g
WHERE m.title='Star Wars' AND g.name='Animation';
```

title	name	cube_ll_coord
Star Wars	Animation	0

(1 row)

17. Afficher les films avec les meilleurs notes dans la catégorie SciFi

```
SELECT m.title, g.name, cube_ll_coord(m.genre,g.position) AS note
FROM movies m, genres g
WHERE g.name='SciFi'
ORDER BY note DESC;
```

title	name	note
The Terminator	SciFi	12
The Island	SciFi	12
Mars Attacks!	SciFi	12
Terminator 2 Judgment Day	SciFi	10
Brazil	SciFi	10
Star Wars	SciFi	10
Alien	SciFi	10
Star Wars Episode V – The Empire Strikes Back	SciFi	10
Star Wars Episode I – The Phantom Menace	SciFi	10
Aliens	SciFi	10
Avatar	SciFi	10
Back to the Future Part II	SciFi	10
Spaceballs	SciFi	10
Star Trek – First Contact	SciFi	10
Blade Runner	SciFi	10

18. Afficher les filmes similaire (**cube_distance**) a ***Star Wars*** (vecteur = (0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0, 0, 0)) du plus similaire au moins similaire

```
SELECT title, cube_distance(genre,'(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0,
0, 0)') as dist
FROM movies
ORDER BY dist DESC;
```

title	dist
The Blair Witch Project	23.7065391822594
The Shining	23.2808934536456
Starship Troopers	22.248595461287
Casino Royale	21.6101827849743
L.A. Confidential	21.6101827849743
Pulp Fiction	21.6101827849743
Casablanca	21.142374511866
Apollo 13	21.142374511866
Goldfinger	21.0237960416286

19. Écrivez une requête pour trouver les filmes qui sont a moins de 5 points de différence sur chaque dimension (utiliser *cube_enlarge* et @>).

```
SELECT title, cube_distance(genre,'(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0,
0, 0)') as dist
FROM movies
WHERE cube_enlarge('(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0, 0, 0)',5,18) @>
genre
ORDER BY dist ASC;
```

title	dist
Star Wars	0
Star Wars Episode V – The Empire Strikes Back	2
Avatar	5
Explorers	5.744562646538 03
Krull	6.480740698407 86
E.T. The Extra Terrestrial	7.615773105863 91

(6 rows)