Český fulltext a sdílené slovníky

Prague PostgreSQL Developers Day 2012

Tomáš Vondra (tv@fuzzy.cz)

lightning talk

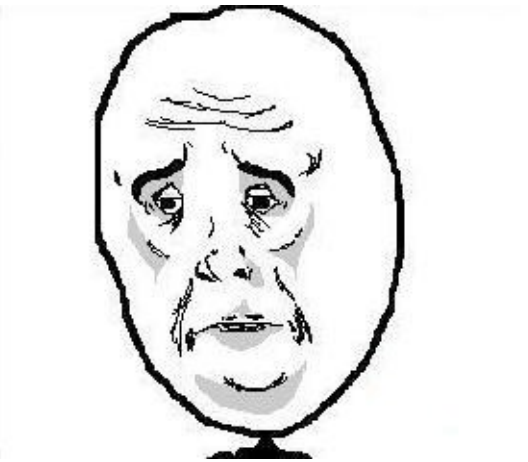
Kdo používá zabudovaný fulltext?

```
SELECT id, nadpis
  FROM dokumenty
 WHERE fts_idx @@ to_tsquery('cs','hledám');
```

snowball

- algoritmická normalizace slov
  - slovesa na infinitiv
  - podstatná jména na 1. pád
  - ...
- super věc
  - málo paměti
  - rychlé odezvy

Bohužel nefunguje pro češtinu :-(



Okay

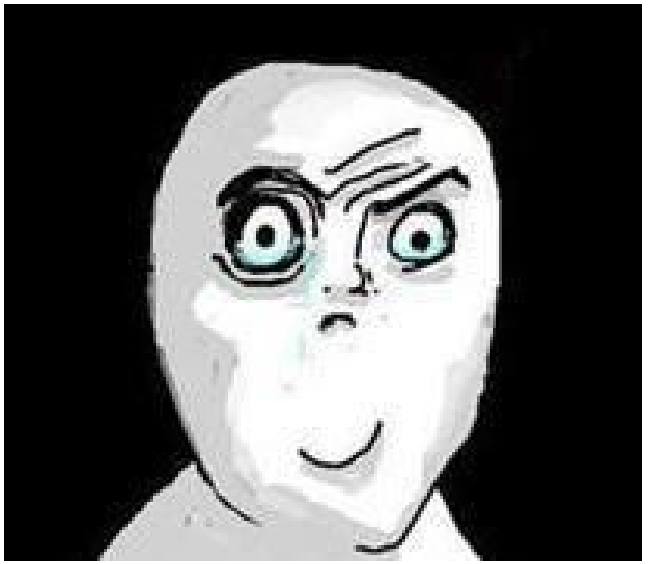
ispell

- založeno na slovníku
  - sada pravidel (skloňování, časování)
  - slovník (slovo + jak skloňovat)

ispell

- založeno na slovníku
  - sada pravidel (skloňování, časování)
  - slovník (slovo + jak skloňovat)

- tři problémy
  - per connection
  - inicializace
  - zabírá paměť
  - …



Actually...It's Not Okay

inicializace

- slovník se musí načíst a naparsovat
  - CPU, I/O
  - ~ přes vteřinu, podle systému
  - … nic moc :-(
- standardní řešení
  - connection pool
  - … ale nic není tak růžové

CPU vs. RAM

- krátká vs. dlouhá spojení
  - oslíkovo dilema
  - krátká spojení – opakovaná inicializace (CPU)
  - dlouhá spojení – kopie statických dat (RAM)
- 25 connections + 2 slovníky (cs + cs_ascii)

  500 MB
- standardně víc connections (100 ?)

ispell vs. RAM

- virtualizovaná prostředí
- VPS
- sdílené servery

Tam všude je (často) nedostatek RAM.

co s tím?

Challenge Accepted

shared_ispell

- načítá slovníky do sdílené paměti
  - jednorázová inicializace
  - každý slovník uložen jen jednou
- používá se stejně jako obyčejný ispell
- paměť je nutno vyhradit předem
- ne všechny slovníky

shared_ispell API

- info o paměti
  - shared_ispell_mem_used()
  - shared_ispell_mem_available()

- slovníky / stop words
  - shared_ispell_dicts()
  - shared_ispell_stoplists()
  - shared_ispell_reset()

shared_ispell

http://github.com/tvondra/shared_ispell

http://pgxn.org/dist/shared_ispell/

ispell slovník (UTF-8)

http://github.com/tvondra/ispell_czech