Publikováno z *NANO Solution s.r.o.* (https://www.na-no.cz)

Jak na vyhledávání v češtině - Apache Solr, Drupal 7

Keywords: Apache Solr, searchapi, Tomcat, Drupal 7, czech language, schema.xml, server.xml, views, more like this, autocomplete.
Klíčová slova: čeština, diakritika, vyhledávání, našeptávač.

Reference:
http://en.wikipedia.org/wiki/Solr
http://lucidworks.lucidimagination.com/display/solr/Apache+Solr+Referenc...

Anotace: Nastavení vyhledávání v češtině, tedy s diakritikou, za pomocí Apache Solr na Drupal 7.

Základní vybavení:

- Apache Tomcat, Apache Solr na strane serveru
- Drupal 7
- modul - Apache Solr Search - umožňuje použít vyhledávání solr v Drupal, more like this (http://drupal.org/project/apachesolr)

Vychytávky:

- modul - Search API - umožnuje vyhledávání, formátování výsledků ve views, facceting... (http://drupal.org/project/search_api)
- modul - Views - prezentace dat v drupal
- modul - Search API page - vytváří jednoduchou vyhledávací stránku (http://drupal.org/project/search_api_page)
- modul - Search API autocomplete - našeptávač slov při zadávání dotazu (http://drupal.org/project/search_api_autocomplete)
- modul - Search API live results - našeptávač, kompletuje při psaní textu (http://drupal.org/project/search_api_live_results)

Nastavení Apache Tomcat:
(http://wiki.apache.org/solr/SolrTomcat#URI_Charset_Config)

V souboru "server.xml" je třeba přidat atribut **URIEncoding="UTF-8"** do elementu "Connector"
Pokud se toto nastavení neprovede indexování obsahu stránek bude správně fungovat, ale vyhledávání nebude fungovat s diakritikiou i přesto, že na straně serveru a při testování přes administraci solr bude vše v pořádku.

URIEncoding="UTF-8"
connectionTimeout="*****"
redirectPort="****" />

Jestli se vyskytuje "useBodyEncodingForUR" v elementu "Connector" , pak vyjmout.

Nastavení "schema.xml" searchapi

Pro správné indexování obsahu a vyhledávání v něm i s diakritikou je třeba upravit soubor "schema.xml" Soubor je umístěn v adresáři modulu search_api nebo apachesolr dle toho který je využíván. Tento se po úpravě spolu se souborem "solrconfig.xml" nakopíruje do konfiguračního adresáře serveru Apache Tomcat. Po úpravě souboru je třeba restartovat server a reindexovat obsah!

Pro vyhledávání veškerého fulltextového obsahu je v souboru "schema.xml" pole **"fieldType name="text" class="solr.TextField..."** . Toto je pole, které nás zajímá.
Obsah prochází při indexaci i vyhledávání postupně filtry definovanými v tomto poli. Popis některých filtrů:

- **Whitespace Tokenizer Factory** --- rozděluje na jednotlivá slova: "Být či nebýt" ---> "Být","či","nebýt"
- **Stop Filter Factory** --- vyhazuje slova definovana v souboru "stopwords.txt", např. "a", "nebo"
- **HTML Strip Char Filter Factory** --- odstraňuje html kód a vrátí čistý text
- **Word Delimiter Filter Factory** --- odděluje/spojuje slova dělená oddělovači: "Big-Bang","BigBang" ---> "Big","Bang"
- **Lower Case Filter Factory**--- změna na malá písmena: "Ahoj" ---> "ahoj"
- **Edge N-Gram Filter Factory** --- dělí slova na znaky, umožní vyhledat část slova!
- **Remove Duplicates Token Filter Factory** --- vyhodí duplikace

Příklad nastavení:

generateWordParts="1"
generateNumberParts="1"
catenateWords="1"
catenateNumbers="1"
catenateAll="0"
splitOnCaseChange="1"
preserveOriginal="1"/>

Článek je pahýl a upravuje se....

Jazyk Česky

URL zdroje:https://www.na-no.cz/jak-na-vyhledavani-v-cestine-apache-solr-drupal-7