|  |  |
| --- | --- |
| http://students.uni-sofia.bg/wp/wp-content/uploads/2010/11/logo-su-s-nadpis.jpg | Софийски Университет „Св. Климент Охридски“  Факултет по Математика и Информатика |

**Курсова работа**

за курса по

**Извличане на информация**

на тема

**„Извличане и търсене на данните от http://aop.bg“**

Тодор Михайлов, M24425

Ина Бунова, M24174

Дата 09.02.2014

# Въведение

В момента в сайта на Агенцията за Обществени Поръчки се публикува информация за обществените поръчки, техните възложители и документи свързани с тях. Сайтът предоставя възможност за търсене върху базата от данни на АОП, но няма основни неща като пълнотекстово търсене или важна информация като размер(стойност) на обществената поръчка и цената с която изпълнителят е спечелил

Нашата идея е да извлечем данните от АОП, да структурираме по някои по-важни полета и да извлечем и да предоставим и проздрачно търсене.

# Цел

Да се извлекат данните за обществените поръчки <http://aop.bg>, да се индексират и да се предостави лесно търсене чрез уеб приложение.

# Реализация

Основните дейности по реализацията на проекта са следните:

* Изтегляне на данните за обществени поръчки – от AOP.bg
* Обработка на документи – Извличане на структурирани полета
* Съхранение на данните и индексиране – Apache Solr
* Потребителски интерфейс – Web приложение за търсене и преглеждане на данните

## Изтегляне и обработка на данните

Първата стъпка на проекта е да бъдат изтеглени данните от сайта за обществени поръчки. Това става като се правят заявки за търсене върху сайта и се взема върнатия резултат като отговор.

Един от проблемите е че за търсенето сайтът изисква регистрация и затова трябва в заявките да се включи cookie за идентификация на потребителя.

Реализирано е приложение AOPSearch.Cralwer, което по зададен брой страници от ресултати, бисквитка(cookie) за идентификация на потребителя и изходен файл да свали структурирани данни за обществените поръчки и да ги запише в изходен файл. Файлът по-късно се използва инициализиране на индекса в Solr.

### Обработка на данните

При извършване на заявката за 1 страница от резултати се връща HTML-а от страницата, който се обработва от специална библиотека, като се взема структурата на документа – извличат се таблиците с резултати и в зависимост от полетата и текстовото им представяне се съхраняват. Използвват се регулярни изрази за откриване на структурирани данни в резултата.

## Индексиране

За индексирането на изтеглените данни се използва Solr. За целта е конфигурирана колекция от тип CaseCollection, който съдържа информация за документи в съответната структура на Case(Обществена поръчка) обект. Значещите полета са в съответния тип text, date, double и др. За текстовите полета са създадени по-сложи типове, върху които са приложени различни филтри.

* Филтри за основните текстови полета
  + StandardTokenizerFactory
  + LowerCaseFilterFactory
  + StopFilterFactory
  + BulgarianStemFilterFactory
* Spellcheck – позволява търсене със сгрешени думи. На базата на данните въведени до момента Solr извлича най-често срещаните думи и ги използва за correction
  + SpellCheckComponent – изграден върху думите от текстове полета

Настройите на полетата и типовете са в приложения schema.xml файл:

<!--Основни полета в индекса на Solr-->

<field name=**"id"** type=**"string"** indexed=**"true"** stored=**"true"** required=**"true"** multiValued=**"false"** />

<field name=**"case\_number"** type=**"text\_en\_splitting\_tight"** indexed=**"true"** stored=**"true"** omitNorms=**"true"**/>

<field name=**"name"** type=**"text\_general"** indexed=**"true"** stored=**"true"**/>

<field name=**"assigner"** type=**"text\_general"** indexed=**"true"** stored=**"true"** omitNorms=**"true"**/>

<field name=**"case\_type"** type=**"string"** indexed=**"true"** stored=**"true"** />

<field name=**"case\_descr"** type=**"text\_general"** indexed=**"true"** stored=**"true"** termVectors=**"true"** termPositions=**"true"** termOffsets=**"true"**/>

<field name=**"max\_amount"** type=**"float"** indexed=**"true"** stored=**"true"**/>

<field name=**"max\_duration"** type=**"int"** indexed=**"true"** stored=**"true"**/>

<field name=**"status"** type=**"string"** indexed=**"true"** stored=**"true"** />

<field name=**"recieved"** type=**"date"** indexed=**"true"** stored=**"true"**/>

<field name=**"deadline"** type=**"date"** indexed=**"true"** stored=**"true"**/>

<field name=**"content"** type=**"text\_general"** indexed=**"false"** stored=**"true"** multiValued=**"true"**/>

<field name=**"text"** type=**"text\_general"** indexed=**"true"** stored=**"false"** multiValued=**"true"**/>

<field name=**"text\_rev"** type=**"text\_general\_rev"** indexed=**"true"** stored=**"false"** multiValued=**"true"**/>

<field name=**"assigner\_exact"** type=**"string"** indexed=**"true"** stored=**"false"**/>

<!--Полетата се копират в Text полето,което служи за търсене-->

<copyField source=**"case\_type"** dest=**"text"**/>

<copyField source=**"name"** dest=**"text"**/>

<copyField source=**"assigner"** dest=**"text"**/>

<copyField source=**"case\_descr"** dest=**"text"**/>

<copyField source=**"assigner"** dest=**"assigner\_exact"**/>

<!--Представяне на сумата и като money-->

<copyField source=**"max\_amount"** dest=**"price\_c"**/>

Използвани филтри за полетата:

<fieldType name=**"text\_general"** class=**"solr.TextField"** positionIncrementGap=**"100"**>

<analyzer>

<tokenizer class=**"solr.StandardTokenizerFactory"**/>

<filter class=**"solr.LowerCaseFilterFactory"**/>

<filter class=**"solr.StopFilterFactory"** ignoreCase=**"true"** words=**"lang/stopwords\_bg.txt"** enablePositionIncrements=**"true"**/>

<filter class=**"solr.BulgarianStemFilterFactory"**/>

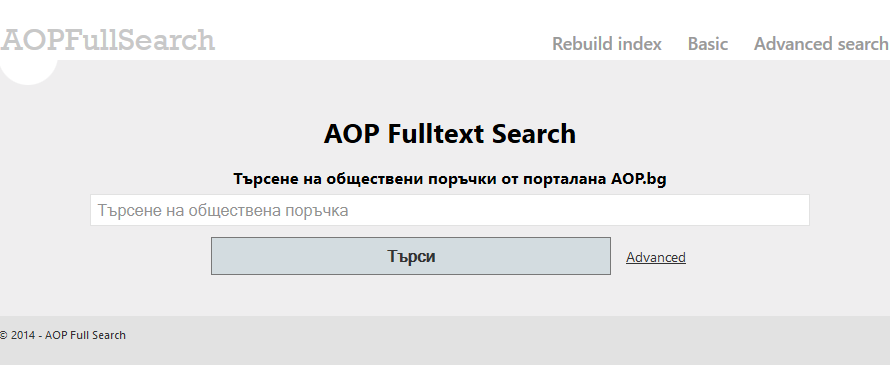
</analyzer>

</fieldType>

## Уеб приложение

Предоставя търсене върху колекцията документи в Solr. На фигурите по-долу са показани части от приложението.

### Търсене

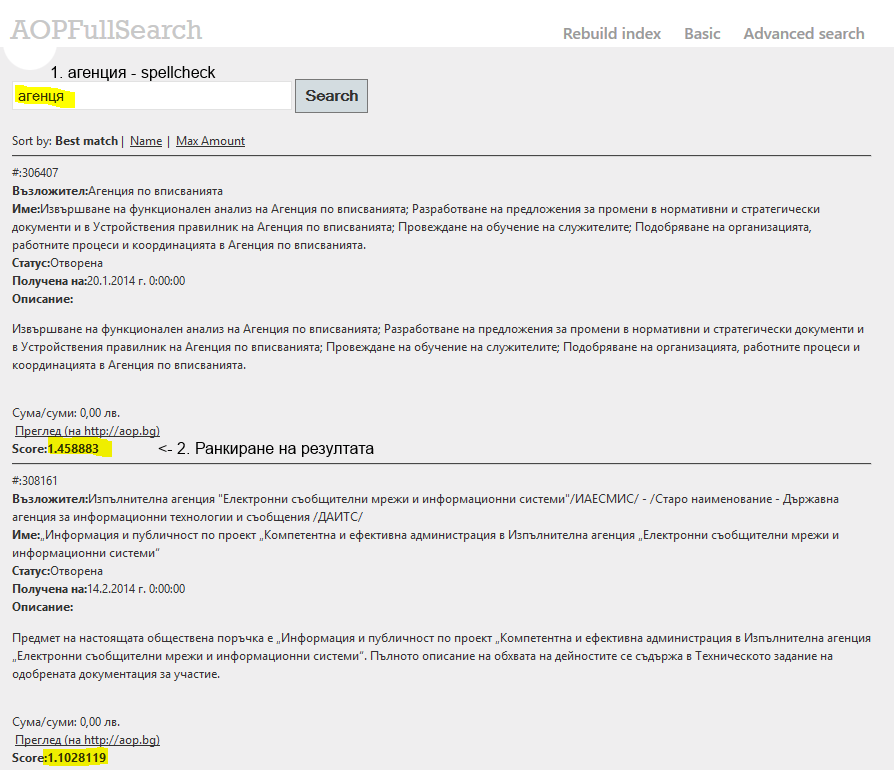


*Основна страница за търсене*

В момента е реализирано търсене само по текстовите полета странициране и сортиране.

### Представяне на резултатите

* Търсенето работи и при грешка в изписването на често срещаните думи
* Включен е Score от търсенето – ранкът, който Solr връща за съответния документ.



Уеб приложението е реализирано с ASP.NET MVC и C#. За връзка със Solr е използвана библиотеката SolrNet и примери от нейната документацията.

# Екип

* Тодор Михайлов, M24425 – ИИОЗ
  + Crawler
  + Уеб приложение за търсене върху Solr идекса
* Ина Бунова, M24174 – РСМТ
  + Конфигуриране на Solr
  + Създаване на индекс и схема ан данните

# Ресурси

* [http://Lucene.apache.org/solr](http://lucene.apache.org/solr) - инсталиране, конфигурация, индексиране
* <https://github.com/mausch/SolrNet/blob/master/Documentation> - документация и примери
* Stackoverflow – помощ и примери