

Sistemi informatici avanzati

Information Retrieval Project

A.A. 2015/2016



Index

- Database selection and cleaning;
- solrconfig.xml and managed-schema.xml config;
- Synonyms and Stopwords;
- Spellcheck;
- Loading data on Solr platform;
- Visualization;
- Demo.

Index

- **Database selection and cleaning;**
- solrconfig.xml and managed-schema.xml config;
- Synonyms and Stopwords;
- Spellcheck;
- Loading data on Solr platform;
- Visualization;
- Demo.

The Italian schools database (released by MIUR) has been chosen from the Italian website of the Open Data.

The dataset, composed by 15 fields and 72356 tuples, was downloaded in CSV format and prepared for the cleaning phase.

The screenshot shows the DatiOpen.it website interface. The browser address bar displays the URL: www.datiopen.it/it/opendata/Anagrafe_strutture_scolastiche?t=Scarica. The website header includes the logo 'DatiOpen.it' and the tagline 'Il portale Italiano dell'Open Data'. A navigation bar contains links: Home, Esplora Catalogo, Ricerca Open Data, Cosa sono gli Open Data, Il progetto, Blog, FAQ, Links, and Cerca nel sito. The main content area is titled 'ANAGRAFE DELLE SCUOLE ITALIANE' and includes the following information:

- FONTE:** MIUR (Ministero Università e Ricerca Scientifica)
- AGGIORNATO AL:** 02/Apr/2013
- DESCRIZIONE:** Elenco di tutte le scuole italiane, completo di Codice scuola, Denominazione, Descrizione tipo scuola, Indirizzo, Comune, Provincia, Regione, Codice postale, Telefono, Fax, Posta elettronica, Posta elettronica certificata, Sito web, Codice istituto principale, tipologia (Statale/Paritaria).
- DATA CREAZIONE:** 24/Set/2012
- AGGIORNATO AL:** 02/Apr/2013
- TEMATICA:** Istruzione
- PAROLE CHIAVE:** Istruzione, Scuola, Scuola elementare, Scuola media, Scuola superiore
- ALTRI LINK:** Visualizza su MIUR (Ministero Università e Ricerca Scientifica), Vai alla fonte del dato

On the right side, there is a section for 'PUBBLICATO DA' (Dati Open.it), 'LICENZE' (Creative Commons BY), 'BANCA DATI' (MIUR - La scuola in chiaro), and 'STATISTICHE' (Visualizzato: 31090, Ricercato: 31098, Scaricato: 7346). At the bottom, there is a list of 'OPENDATA SIMILI' including 'Mappa delle scuole materne in Italia', 'Mappa delle scuole in Italia', 'Mappa delle università in Italia', 'Regione Veneto - Laureati per università, facoltà e anno', and 'Regione Lazio - Posti alloggio per studenti universitari'.

1 Database selection and cleaning

COMUNE	PROVINCIA	REGIONE	CODICE SCUOLA	DENOMINAZIONE DELL'ISTITUTO	TIPOLOGIA ISTITUTO - DENOMINAZIONE	INDIRIZZO (VIA + NUMERO)	CODICE POSTALE	TELEFONO	FAX	EMAIL	EMAIL - PEC	SITO WEB	ISTITUTO PRINCIPALE	STATALE/PARITARIA
Aglie'	TORINO	Piemonte	TO1A16900R	REGINA M.CRISTINA	SCUOLA DELL'INFANZIA	V.BALUARDI 1	10011	012433352	012433352	asliomc@interfree.it			TO1A16900R	PARITARIA
Aglie'	TORINO	Piemonte	TOEE091108	D.D.CASTELLAMONTE-AGLIE'	SCUOLA PRIMARIA	PIAZZA MARTIRI LIBERTA'	10011	012433497		TOEE09100R@istruzione.it			TOEE09100R	STATALE
Aglie'	TORINO	Piemonte	TOMM230013	AGLIE' - OLIVETTI	SCUOLA SECONDARIA DI PRIMO GRADO	PIAZZA MARTIRI LIBERTA'	10011	0124330239	0124330239	TOMM230002@istruzione.it		http://www.isafaccio.it	TOMM230002	STATALE
Airasca	TORINO	Piemonte	TOAA835006	I.C. - AIRASCA	SCUOLA DELL'INFANZIA	VIA STAZIONE	10060	0119908555	0119909475	TOIC83500A@istruzione.it			TOIC83500A	STATALE
Airasca	TORINO	Piemonte	TOAA835017	I.C. AIRASCA - VIA DEL PALAZZO	SCUOLA DELL'INFANZIA	VIA DEL PALAZZO 13	10060	0119909673		TOIC83500A@istruzione.it		nuke.icalrasca.it	TOIC83500A	STATALE
Airasca	TORINO	Piemonte	TOEE83501C	I.C. AIRASCA - CAP.	SCUOLA PRIMARIA	VIA STAZIONE 26	10060	0119909497		TOIC83500A@istruzione.it			TOIC83500A	STATALE
Airasca	TORINO	Piemonte	TOIC83500A	I.C. - AIRASCA	SCUOLA SECONDARIA DI SECONDO GRADO	VIA STAZIONE	10060	0119908555	0119909475	TOIC83500A@istruzione.it	toic83500a@pec.istruzione.it		TOIC83500A	STATALE
Airasca	TORINO	Piemonte	TOMM835018	I.C. AIRASCA - VIA STAZIONE	SCUOLA SECONDARIA DI PRIMO GRADO	VIA STAZIONE	10060	0119908555	0119909475	TOIC83500A@istruzione.it			TOIC83500A	STATALE
Ala di Stura	TORINO	Piemonte	TO1A281008	PARROCCHIA SAN NICOLA	SCUOLA DELL'INFANZIA	V.PIAN DEL TETTO 9	10070	012355152	0123551818	scuolamaterna.treves@alice.it			TO1A281008	PARITARIA
Ala di Stura	TORINO	Piemonte	TOEE809021	I.C.CERES-ALA DI STURA	SCUOLA PRIMARIA	PIAZZA CENTRALE 22	10070	012355228		TOIC80900T@istruzione.it		http://www.icmuriardo.it	TOIC80900T	STATALE
Albiano d'Ivrea	TORINO	Piemonte	TOAA894079	I.C. AZEGLIO - ALBIANO	SCUOLA DELL'INFANZIA	VIA RICCARDI 2	10010	012559524		TOIC894006@istruzione.it		http://digilander.libero.it/c	TOIC894006	STATALE
Albiano d'Ivrea	TORINO	Piemonte	TOEE89405C	I.C. AZEGLIO-ALBIANO D'IVREA	SCUOLA PRIMARIA	VIA RICCARDI 17	10010	012559702		TOIC894006@istruzione.it		http://digilander.libero.it/c	TOIC894006	STATALE
Almese	TORINO	Piemonte	TO1A23600A	RIVA ROCCI	SCUOLA DELL'INFANZIA	V.VIGLIANIS 16	10040	011 9351468	011	ASILOINFANTILRIVAROCCI@istruzione.it			TO1A23600A	PARITARIA
Almese	TORINO	Piemonte	TOAA821008	I.C. - ALMESE	SCUOLA DELL'INFANZIA	PIAZZA DELLA FIERA	10040	0119350258	0119350258	TOIC82100C@istruzione.it			TOIC82100C	STATALE

1

di 1.448

50

Visualizzati 1 - 50 di 72.356

Cleaning

The dataset has been cleaned by using Kettle. Here some examples:

- URLs correction;
 - Ex. *http://, hptt://, htt://, hhttp://, http://, ww, etc.*
- Whitespace removal;
 - Ex. *Via Roma 7*
- Others.
 - Ex. hypens in empty fields, [at] instead of @, etc.

Index

- Database selection and cleaning;
- **solrconfig.xml and managed-schema.xml config;**
- Synonyms and Stopwords;
- Spellcheck;
- Loading data on Solr platform;
- Visualization;
- Demo.

schema.xml

- **Types:**

Two new types have been created *text_custom* e *text_not*. Moreover, the already existing type *text_it* has been modified for searching and faceting purpose.

- **text_custom**: uses a Classic Tokenizer where whitespaces and punctuation are delimiters. It recognizes internet domains, email addresses and numbers with hypens, that are considered single tokens. It also uses a case-insensitive filter and a stopwords filter.

```
<fieldType name="text_custom" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.ClassicTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.StopFilterFactory" format="snowball" words="lang/stopwords_it.txt" ignoreCase="true"/>
  </analyzer>
</fieldType>
```


schema.xml

- **Types:**

Two new types have been created *text_custom* e *text_not*. Moreover, the already existing type *text_it* has been modified for searching and faceting purpose.

- **text_not**: uses a Keyword Tokenizer. It takes the entire text as a single token. Then a case-insensitive filter and a synonyms filter are used.

```
<fieldType name="text_not" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.KeywordTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.SynonymFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
  </analyzer>
</fieldType>
```

schema.xml

- **Types:**

Two new types have been created *text_custom* e *text_not*. Moreover, the already existing type *text_it* has been modified for searching and faceting purpose.

- **text_it:** uses a Standard Tokenizer where whitespaces and punctuation are delimiters. Then elision filter, case-insensitive filter, stopwords filter, stemming filter and synonyms filter are used.

```
<fieldType name="text_it" class="solr.TextField" positionIncrementGap="100">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.ElisionFilterFactory" articles="lang/contractions_it.txt" ignoreCase="true"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.SynonymFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
    <filter class="solr.StopFilterFactory" format="snowball" words="lang/stopwords_it.txt" ignoreCase="true"/>
    <filter class="solr.ItalianLightStemFilterFactory"/>
  </analyzer>
</fieldType>
```

schema.xml

- **Fields:**

Three new fields have been created as copies from already existing ones in order to separate searching from faceting.

```
<field name="ISTAT_COM" type="tlongs"/>
<field name="ISTAT_PROV" type="tlongs"/>
<field name="ISTAT_REG" type="tlongs"/>
<field name="ISTAT_REG_7C" type="tlongs"/>
<field name="cap" type="text_not"/>
<field name="cod_istituto_principale" type="text_not"/>
<field name="codice_scuola" type="text_not"/>
<field name="comune" type="text_not"/>
<field name="denominazione" type="text_it"/>
<field name="tipo_scuola" type="text_not"/>
<field name="email" type="text_custom"/>
<field name="fax" type="text_not"/>
<field name="indirizzo" type="text_custom"/>
<field name="pec" type="text_custom"/>
<field name="sito_web" type="text_custom"/>
<field name="statale" type="text_not"/>
<field name="telefono" type="text_not"/>
<field name="comune_tok" type="text_custom"/>
<field name="tipo_scuola_tok" type="text_it"/>
<field name="denominazione_tok" type="text_custom"/>
```

solrconfig.xml

```
<requestHandler name="/browse" class="solr.SearchHandler">
  <lst name="defaults">
    <str name="echoParams">explicit</str>

    <str name="wt">velocity</str>
    <str name="v.template">browse</str>
    <str name="v.layout">layout</str>
    <str name="title">scuole</str>

    <str name="defType">edismax</str>
    <str name="qf">
      ISTAT_COM^1.0
      ISTAT_PROV^1.0
      ISTAT_REG^1.0
      ISTAT_REG_7C^1.0
      cap^4.0
      cod_istituto_principale^10.0
      cod_scuola^10.0
      comune_tok^8.0
      denominazione^7.0
      tipo_scuola_tok^6.0
      email^1.0
      fax^2.0
      indirizzo^3.0
      pec^1.0
      sito_web^1.0
      statale^5.0
      telefono^2.0
    </str>
    <str name="q.alt">*:*</str>
    <str name="rows">10</str>
    <str name="fl">*</str>
```

- The following entries in the **browse** requestHandler have been modified:
 - **wt** (Writer Type): the response is given by using the *velocity* response writer;
 - **defType**: edismax (ExtendedDisMax) is set as query parser;
 - **qf** (Query Fields): fields and weights on which perform the query are selected.

solrconfig.xml

- Faceting shows a list of preset query based on indexed terms that determine categories. These terms come from a field or a query. Two faceting options have been added:
 - **Facet Field:** A field determines a category. *tipo_scuola*, *statale* and *comune* were chosen;
 - **Facet Query:** A query determines a category. Seven queries were specified for seven different kinds of schools (*Istituto*);

```
<str name="facet">on</str>
<str name="facet.missing">true</str>

<str name="facet.field">tipo_scuola</str>
<str name="facet.field">statale</str>
<str name="facet.field">comune</str>

<str name="facet.query">LS</str>
<str name="facet.query">LC</str>
<str name="facet.query">IA</str>
<str name="facet.query">ITC</str>
<str name="facet.query">ITI</str>
<str name="facet.query">ITA</str>
<str name="facet.query">IPSAR</str>

<str name="facet.mincount">1</str>
```

Index

- Database selection and cleaning;
- solrconfig.xml and managed-schema.xml config;
- **Synonyms and Stopwords;**
- Spellcheck;
- Loading data on Solr platform;
- Visualization;
- Demo.

Synonyms

- The Solr file *synonyms.txt* has been modified to fit the context.

Es:

infanzia, materna, asilo
elementare, primaria
media, primo grado, sm, s.m.
secondo grado, superiore
classico, lc, l.c.
scientifico, ls, l.s., scienze applicate, l.s.a., lsa, scientifico tecnologico, lst, l.s.t.
tecnico commerciale, itc, i.t.c., ragioneria, ipc, i.p.c., ipsct
tecnico industriale, iti, i.t.i., itis, i.t.i.s., industriale
alberghiero, ipsar, ipsa, i.p.s.a.r., i.p.s.a. , ristorazione, ipa, i.p.a.
artistico, arte, ia, i.a.
privata, paritaria
...

Stopwords

- Solr file *stopwords_it.txt* already contains the most common italian stopwords.

Ex: articles, pronouns, prepositions etc.

Index

- Database selection and cleaning;
- solrconfig.xml and managed-schema.xml config;
- Synonyms and Stopwords;
- **Spellcheck;**
- Loading data on Solr platform;
- Visualization;
- Demo.

Spellcheck

- **spellcheck.collate**: provided by Lucene library has been set as spellcheck. It finds the best suggestions (if any) from each word of the inserted phrase.
The field *denominazione* was chosen as source of suggestions for it.
- Spellcheck works in two different phases:
 - **Autocomplete**;
 - **Did you mean?** .

```
#foreach($t in $response.response.terms.denominazione_tok)
  #if($foreach.count > 7)
    #break
  #end
  $t.key
#end
```

Index

- Database selection and cleaning;
- solrconfig.xml and managed-schema.xml config;
- Synonyms and Stopwords;
- Spellcheck;
- **Loading data on Solr platform;**
- Visualization;
- Demo.

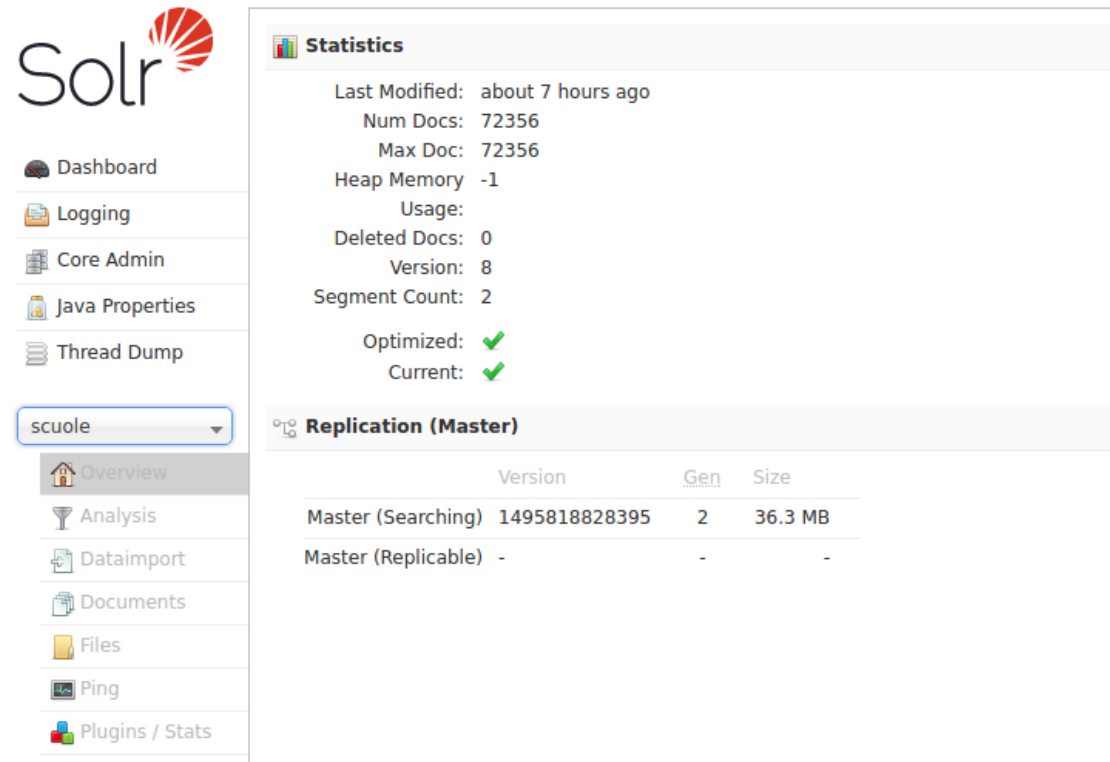
Index Creation

- Dataset has been loaded on Solr using command line:

```
pc@ubuntu:~/solr-6.5.0$ bin/solr start  
pc@ubuntu:~/solr-6.5.0$ bin/solr create -c scuole  
pc@ubuntu:~/solr-6.5.0$ bin/post -c scuole scuole.csv
```

Index Creation

- Dataset has been loaded on Solr using command line:



The screenshot displays the Solr Admin UI for the 'scuole' index. The left sidebar contains navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a dropdown menu currently showing 'scuole'. Below the dropdown are links for Overview, Analysis, Dataimport, Documents, Files, Ping, and Plugins / Stats. The main content area is divided into two sections: 'Statistics' and 'Replication (Master)'. The 'Statistics' section shows the following data:

Statistic	Value
Last Modified:	about 7 hours ago
Num Docs:	72356
Max Doc:	72356
Heap Memory	-1
Usage:	
Deleted Docs:	0
Version:	8
Segment Count:	2
Optimized:	✓
Current:	✓



The 'Replication (Master)' section shows a table with the following data:

	Version	Gen	Size
Master (Searching)	1495818828395	2	36.3 MB
Master (Replicable)	-	-	-

Index

- Database selection and cleaning;
- solrconfig.xml and managed-schema.xml config;
- Synonyms and Stopwords;
- Spellcheck;
- Loading data on Solr platform;
- **Visualization;**
- Demo.

Visualization



Cerca:

Submit Reset

Field Facets

Ciclo Istruzione

[scuola dell'infanzia](#) (29267)

[scuola primaria](#) (20125)

[scuola secondaria di second...](#) (14892)

[scuola secondaria di primo ...](#) (8072)

Tipologia

[statale](#) (58392)

[paritaria](#) (13964)

Comune

[roma](#) (2282)

[napoli](#) (1142)

[milano](#) (938)

[palermo](#) (842)

[torino](#) (636)

[genova](#) (532)

[catania](#) (479)

[messina](#) (358)

[firenze](#) (351)

[bologna](#) (330)

[bari](#) (317)

72356 risultati trovati in 21 ms. Pagina 1 di 7236 [Successiva](#)

I.P.C. G. BRUNO - SEZ. ASSOCIATA SERALE
Tipo Scuola: SCUOLA SECONDARIA DI SECONDO GRADO
Comune: MOLINELLA
Indirizzo: VIA MAZZINI 371
Email: bois00300a@istruzione.it



Via Giuseppe Mazzini, 371
[View larger map](#)

CASSIANO DA IMOLA
Tipo Scuola: SCUOLA SECONDARIA DI SECONDO GRADO
Comune: IMOLA
Indirizzo: VIALE DANTE N.1/A
CAP: 40026
Numero telefonico: 054225751
Fax: 054229841
Sito Web: [www.nonlineacassiano.it](#)



Viale Dante Alighieri, 1
[View larger map](#)

Index

- Database selection and cleaning;
- solrconfig.xml and managed-schema.xml config;
- Synonyms and Stopwords;
- Spellcheck;
- Loading data on Solr platform;
- Visualization;
- **Demo.**

Demo

Please see *demo.mp4*.

Thanks.