

# Clustering and visualization of Solr search results

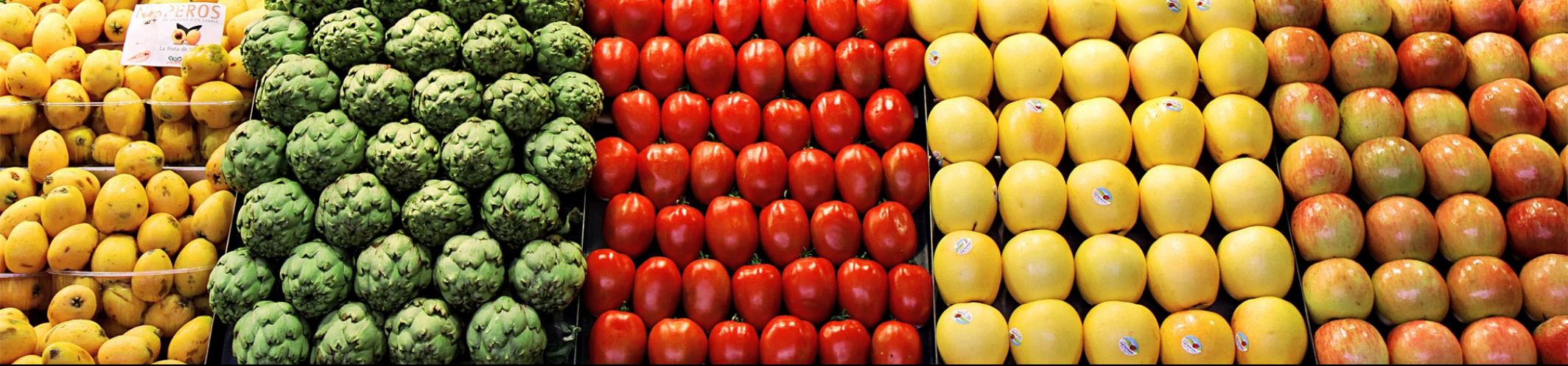
B E R L I N  
**BUZZ**  
W O R D S

Stanislaw Osinski

[stanislaw.osinski@carrotsearch.com](mailto:stanislaw.osinski@carrotsearch.com)







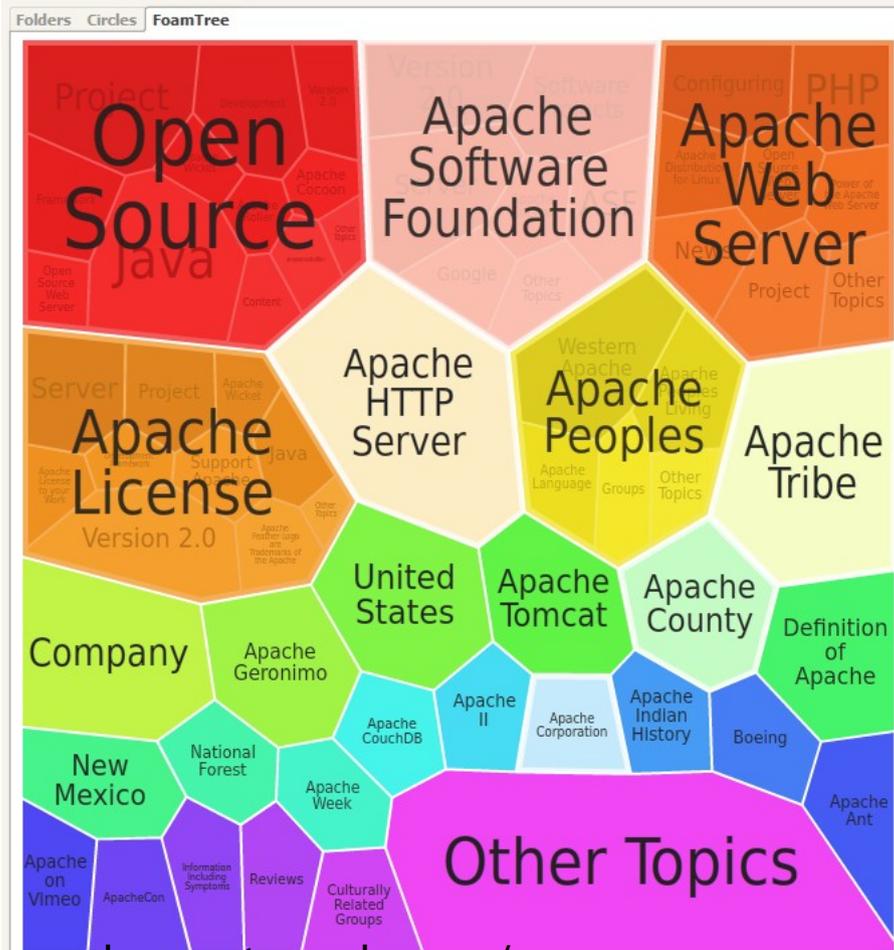
Put similar objects together,  
separate dissimilar ones



Put similar objects together,  
separate dissimilar ones







- Cluster with 17 documents
- 1 [Welcome to The Apache Software Foundation!](#)  
Supports the development of a number of open-source software projects, including the Apache web server. Includes license information, latest news, ...  
<http://www.apache.org/>
  - 34 [Apache Software Founda](#)  
The Apache Software Founda (Apache Software Foundation) to support Apache HTTP Server .  
[http://en.wikipedia.org/wiki/Apache\\_HTTP\\_Server](http://en.wikipedia.org/wiki/Apache_HTTP_Server) [Wikipedia]
  - 99 [Apache Subversion](#)  
... License, Version 2.0 . Apache feather logo are trademarks of the Apache Software Foundation.  
<http://subversion.apache.org/>
  - 100 [Apache Pizza - Pizza restaurants](#)  
Irish Pizza Restaurants, pizza Takeout, Eat in or At Home  
<http://www.apache.ie/>
  - 101 [Western Apache language](#)  
The Western Apache language spoken by over 1 million peoples living primarily in the United States.  
[http://en.wikipedia.org/wiki/Western\\_Apache\\_language](http://en.wikipedia.org/wiki/Western_Apache_language) [Wikipedia]
  - 104 [Apache Lenya - Open Source \(Java/XML\)](#)  
Apache Lenya | Java based publishing system designed for Apache Cocoon from the Apache Software Foundation based on open standards  
<http://lenya.apache.org/>
  - 109 [Apache](#)

- ▶ [Welcome to The Apache Software Foundation!](#)  
Supports the development of a number of open-source software projects, including the Apache web server. Includes license information, latest news, ...  
[www.apache.org/](http://www.apache.org/) - Cached - Similar
- [Apache Web Server Project](#)    [FAQs](#)  
[Mirrors](#)    [Cocoon](#)  
[Tomcat](#)    [Felix](#)  
[Projects](#)    [POI](#)
- [More results from apache.org »](#)
- [Welcome! - The Apache HTTP Server Project](#)  
The Apache HTTP Server Project is an effort to develop and maintain an open source web server software project.  
[httpd.apache.org/](http://httpd.apache.org/) - Cached - Similar
- [Apache HTTP Server - Wikipedia, the free encyclopedia](#)  
The Apache HTTP Server, commonly referred to as Apache is web server software notable for playing a key role in the initial growth of the World Wide Web.  
[en.wikipedia.org/wiki/Apache\\_HTTP\\_Server](http://en.wikipedia.org/wiki/Apache_HTTP_Server) - Cached - Similar
- [Apache - Wikipedia, the free encyclopedia](#)  
Apache is the collective term for several culturally related groups of people in the United States, Canada, Egypt, the United Kingdom North Sea, ...  
[en.wikipedia.org/wiki/Apache](http://en.wikipedia.org/wiki/Apache) - Cached - Similar
- [Apache Corporation : Home Page](#)  
Apache Corporation is an oil and gas exploration and production company with operations in the United States, Canada, Egypt, the United Kingdom North Sea, ...  
[www.apachecorp.com/](http://www.apachecorp.com/) - Cached - Similar
- [Apache](#)  
The best Apache web server resource for helpful tips, advice, and configuration. Get help with configuring the Apache server for your use.  
[www.apache.com/](http://www.apache.com/) - Cached - Similar
- [Open Source Initiative OSI - Apache License, Version 2.0: Licensing ...](#)  
To apply the Apache License to your work, attach the following boilerplate notice, with the fields enclosed by brackets "[ ]" replaced with your own ...  
[www.opensource.org/licenses/apache2.0.php](http://www.opensource.org/licenses/apache2.0.php) - Cached - Similar
- [Apache County Website](#)  
Apache County is unique among all counties in the United States in many ways. Particularly because it is the longest county in the country, 211 miles from ...  
[www.co.apache.ar.us/](http://www.co.apache.ar.us/) - Cached - Similar
- [APACHE TRIBE](#)  
The word "apache" comes from the Yuma word for "fighting-men" and from the Zuni word meaning "enemy." The Apache tribe consists of six subtribes: the ...  
[www.greatdreams.com/apache/apache-tribe.htm](http://www.greatdreams.com/apache/apache-tribe.htm) - Cached - Similar

search.carrotsearch.com /  
search.carrot2.org

Search results clustering  
has been there for years

google.com

Document clustering

off-line process

global clusters

no query-time cost

Search results clustering

on-line process

query-specific clusters

no maintenance cost

cluster labels needed

# Enabling search results clustering in

# Solr

Performance impact and tuning

Cluster tuning

Rapid prototyping: cluster

visualization

```
<searchComponent name="clustering"  
    class="solr.clustering.ClusteringComponent">  
  <lst name="engine">  
    <str name="name">default</str>  
    <str name="carrot.algorithm">  
      org.carrot2.clustering.lingo.LingoClusteringAlgorithm  
<!-- org.carrot2.clustering.stc.STCClusteringAlgorithm  
    org.carrot2.clustering.kmeans.BisectingKMeansClusteringAlgorithm -->  
    </str>  
    <str name="MultilingualClustering.defaultLanguage">ENGLISH</str>  
  </lst>  
</searchComponent>
```

# Add clustering component

```
<requestHandler name="/ clustering" class="solr.SearchHandler">
  <lst name="defaults">
    ...
    <bool name="clustering">on</ bool>
    <bool name="clustering.results">>true</ bool>
    <str name="clustering.engine">default</ str>

    <!-- Fields to cluster on, must be stored -->
    <str name="carrot.title">name</ str>
    <str name="carrot.snippet">features</ str> <!-- optional -->
    <str name="carrot.url">url</ str> <!-- optional -->

    <!-- Set other parameters if needed -->
  </ lst>

  <arr name="last-components">
    ...
    <str>clustering</ str>
  </ arr>
</ requestHandler>
```

# Add clustering to your handler

# Test if it works

[http://localhost/solr/clustering?q=\\*:\\*&rows=100](http://localhost/solr/clustering?q=*:*&rows=100)

```
<response>
  <lst name="responseHeader">...</lst>

  ...

  <arr name="clusters">
    <lst>
      <arr name="labels"><str>Data Analysis</str></arr>
      <double name="score">56.481022171223046</double>
      <arr name="docs">
        <str>c2dm0</str> <!-- document id -->
        <str>c2dm4</str>
        <str>c2dm6</str>

        ...
      </arr>
    </lst>
    <lst>
      <arr name="labels"><str>Data Mining Software</str></arr>

      ...
    </lst>

    ...

    <lst>
      <arr name="labels">
        <str>Other Topics</str> <!-- unclustered documents -->
      </arr>
      <double name="score">0.0</double>
      <bool name="other-topics">true</bool>
      <arr name="docs">
        <str>c2dm9</str>
        <str>c2dm11</str>
        <str>c2dm12</str>
        <str>c2dm14</str>

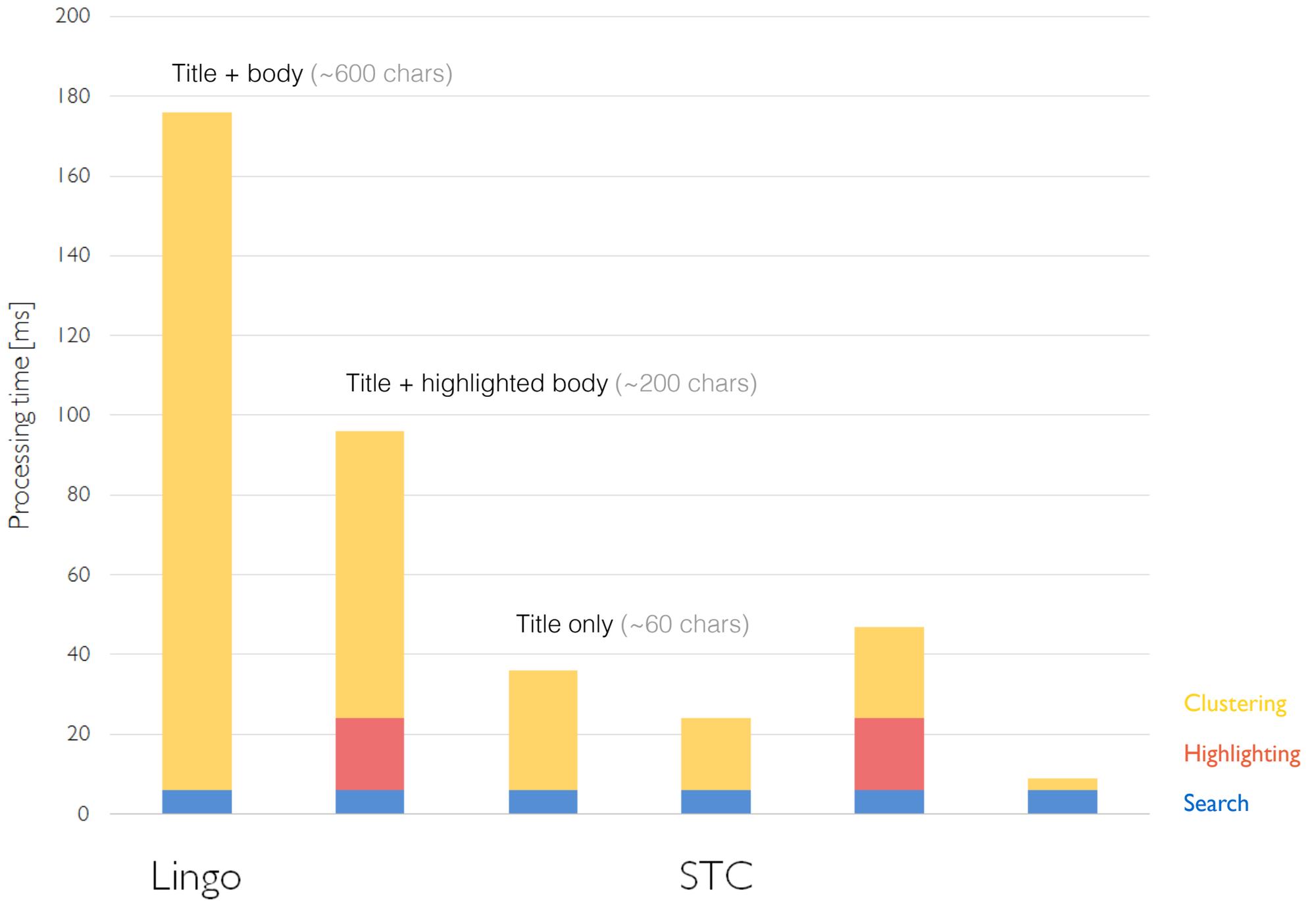
        ...
      </arr>
    </lst>
  </arr>
</response>
```

Performance impact and tuning

Cluster tuning

Rapid prototyping: cluster

visualization



For better  
clustering  
performance

Limit the length of clustered content or use highlighting

Cluster only document titles

Use STC instead of Lingo

Tune algorithms' settings

[download.carrot2.org/stable/manual/#section.advanced-topics](https://download.carrot2.org/stable/manual/#section.advanced-topics)

Cluster tuning

Rapid prototyping: cluster

visualization

# Carrot<sup>2</sup> Clustering Workbench

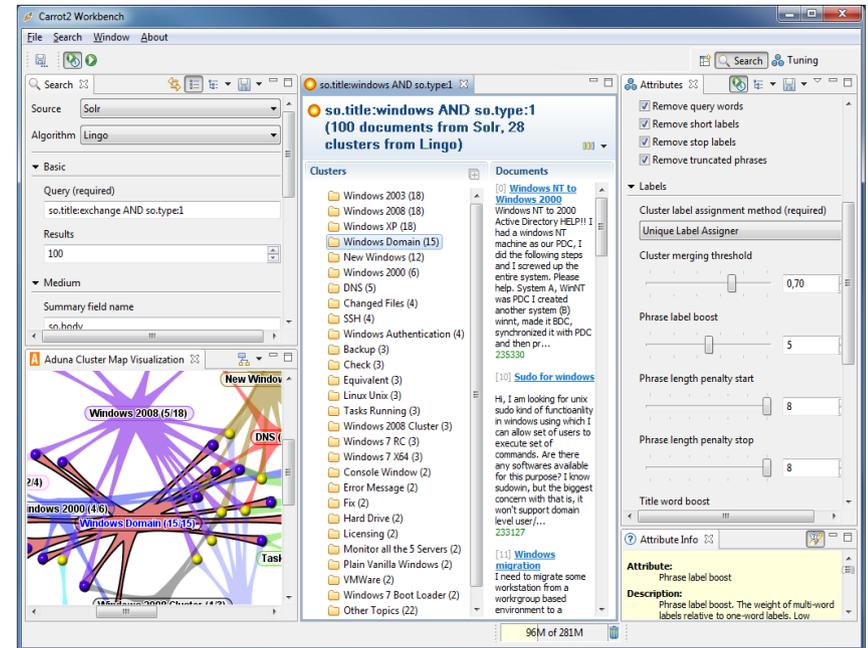
get from [download.carrot2.org](http://download.carrot2.org)  
provide your Solr URL, field names and query

## Clustering settings

tune number of clusters,  
size of clusters, label lengths, ...

## Lexical resources

edit stop words, stop labels



Save modified settings as XML,  
apply XSLT\* and paste to `solrconfig.xml`

Copy modified stop words and stop labels to `solr-home/conf/clustering`

! Solr3.2/4.0

\* see [wiki.apache.org/solr/ClusteringComponent](http://wiki.apache.org/solr/ClusteringComponent)

Rapid prototyping: cluster  
visualization



All documents (250) > Holland Casino Rotterdam (8)

[APS: Vrouw wordt miljonair op speelautomaat](#)  
 PERSBERICHT Rotterdam, 17 april 2011 - Een 28-jarige vrouw uit Zuid-Holland is gisteravond miljonair geworden op een speelautomaat in Holland Casino Rotterdam. Om zeven uur viel de Mega Millions Jackpot, waarmee de vrouw met een inzet van 3,75 euro het belastingvrije bedrag van 1.049.286,78 euro won. De vrouw had pas door dat ze de miljoenenjackpot had gewonnen toen andere gasten van het casino haar daar op zezen.  
 2011-04-17, ANP Pers Support

[Vrouw wordt miljonair in casino](#)  
 ROTTERDAM - Door in te zetten op een speelautomaat in Holland Casino in Rotterdam is een 28-jarige vrouw van het weekend miljonair geworden. Om zeven uur viel de Mega Millions Jackpot, waarmee de vrouw met een inzet van 3,75 euro ruim een miljoen belastingvrij won.  
 2011-04-22, De Havenloods Noord

[Miljonair](#)  
 Een 28-jarige vrouw uit Zuid-Holland is zaterdagavond miljonair geworden op een speelautomaat in de vestiging van Holland Casino in Rotterdam. Om 19.00 uur viel de Mega Millions Jackpot waardoor de vrouw met een inzet van EUR 3,75 het belastingvrije bedrag van EUR 1.049.286,79 won. Het is de 3e keer in iets meer dan een maand dat de Mega Millions Jackpot is gevallen. Op 11 maart won een vrouw een bedrag van zo'n EUR 1,07 miljoen in Holland Casino Valkenburg.  
 2011-04-17, Omroep Vlaardingen

[First Friday Borrel - Netwerkborrel voor ondernemers uit regio Rotterdam & Haaglanden - 04-03-2011](#)  
 Flevum organiseert in samenwerking met Holland Casino Rotterdam de 'Netwerkborrel voor ondernemers uit regio Rotterdam & Haaglanden' Ons doel is dat er 60 deelnemers tot 100 (leden en introducés) elkaar ontmoeten op deze borrel! De borrel is zo georganiseerd dat u geniet van een drankje en een hapje gedurende de spits. De borrel begint om 16:30 en eindigt rond 19.00.

We'll need

\*

\*

\*

```

<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
  <xsl:output indent="yes" omit-xml-declaration="no"
    media-type="application/xml; charset=UTF-8" encoding="utf-8" />

  <xsl:variable name="title.field"><xsl:value-of select="//str[@name = 'carrot.title']" /></xsl:variable>
  <xsl:variable name="snippet.field"><xsl:value-of select="//str[@name = 'carrot.snippet']" /></xsl:variable>
  <xsl:variable name="url.field"><xsl:value-of select="//str[@name = 'carrot.url']" /></xsl:variable>

  <xsl:template match="/">
    <searchresult>
      <xsl:apply-templates select="response/result/doc" />
      <xsl:apply-templates select="response/arr[@name = 'clusters']/lst" />
    </searchresult>
  </xsl:template>

  <xsl:template match="doc">
    <document id="{str[@name = 'id']}">
      <title><xsl:value-of select="str[@name = $title.field]" /></title>
      <snippet><xsl:value-of select="str[@name = $snippet.field]" /></snippet>
      <url><xsl:value-of select="str[@name = $url.field]" /></url>
    </document>
  </xsl:template>

  <xsl:template match="arr[@name = 'clusters']/lst">
    <group id="{generate-id(.)}" score="{double[@name = 'score']}">
      <title><xsl:apply-templates select="arr[@name = 'labels']/str" /></title>
      <xsl:apply-templates select="arr[@name = 'docs']/str" />
      <xsl:apply-templates select="arr[@name = 'clusters']/lst" />
    </group>
  </xsl:template>

  <xsl:template match="arr[@name = 'labels']/str">
    <phrase><xsl:apply-templates /></phrase>
  </xsl:template>

  <xsl:template match="arr[@name = 'docs']/str">
    <document refid="{.}" />
  </xsl:template>
</xsl:stylesheet>

```

XSLT: Solr XML → FoamTree XML

```

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
  <head>
    <title>Simple visualization of Solr search results</title>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
    <link rel="stylesheet" type="text/css" href="css/layout.css"></link>
  </head>

  <body>
    <input type="text" name="url" />
    <button name="visualize">Visualize</button>

    <div id="swfcontainer"><div id="swf"></div></div>

    <script type="text/javascript" src="js/swfobject.js"></script>
    <script type="text/javascript" src="js/carrotsearch.foamtree.min.js"></script>
    <script type="text/javascript" src="js/jquery-1.4.4.min.js"></script>

    <script type="text/javascript">
      $(document).ready(function() {
        var foamtree = new CarrotSearchFoamTree({
          visualizationSwfLocation: "swf/com.carrotsearch.visualizations.foamtree.swf",
          flashPlayerInstallerSwfLocation: "swf/playerProductInstall.swf",
          id: "swf"
        });

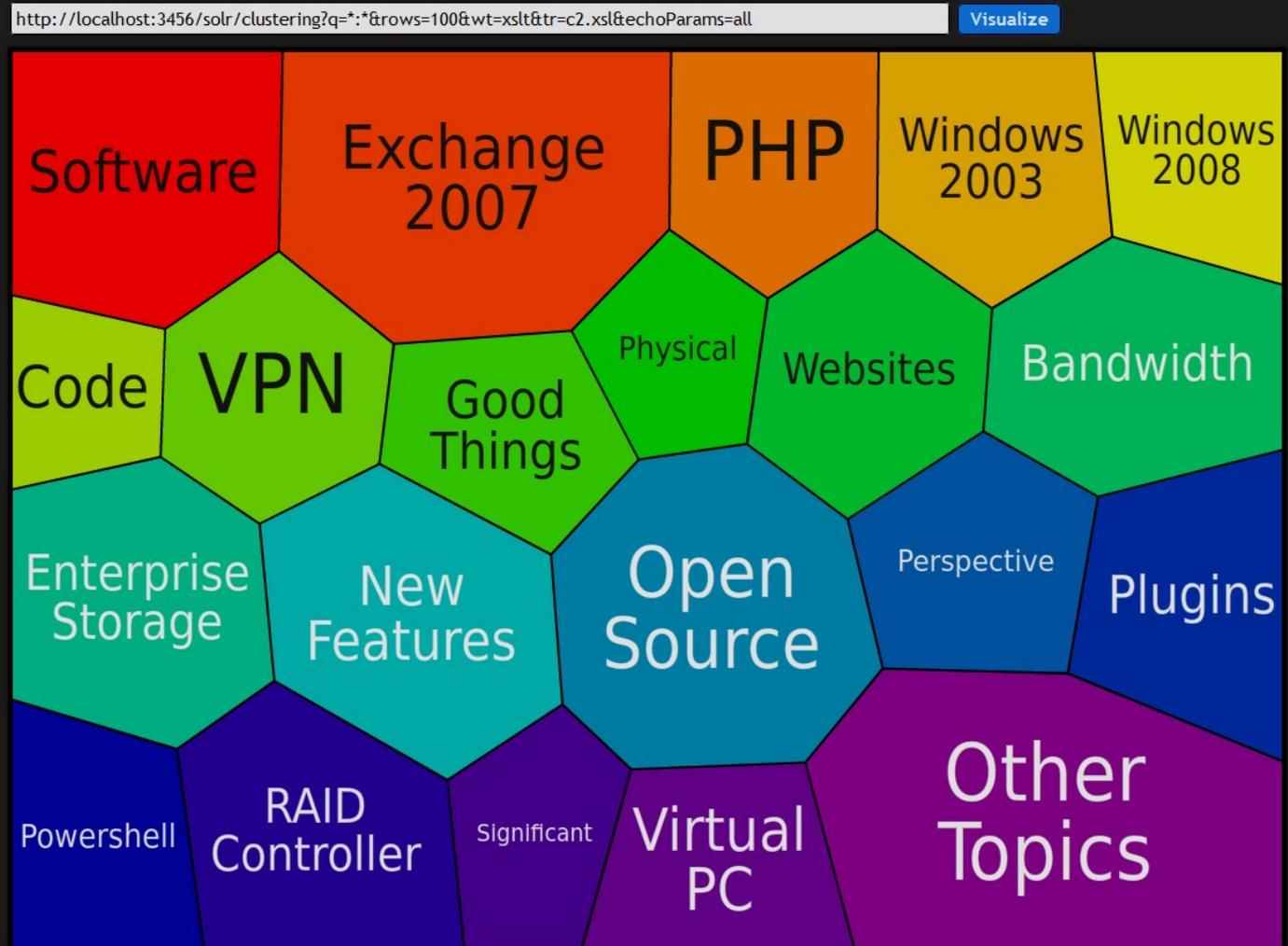
        $("button[name = 'visualize']").click(function() {
          foamtree.set("dataUrl", $("input[name = 'url']").val());
        });
      });
    ]&gt;&lt;/script&gt;
  &lt;/body&gt;
&lt;/html&gt;
</pre>
</div>
<div data-bbox="48 863 681 938" data-label="Section-Header">
<h1>HTML / JavaScript magic</h1>
</div>
```

XSLT

Solr URL

Save to `${solr.home}/conf/xslt/c2.xsl`

`http://localhost:3456/solr/clustering?q=*&rows=100&wt=xslt`

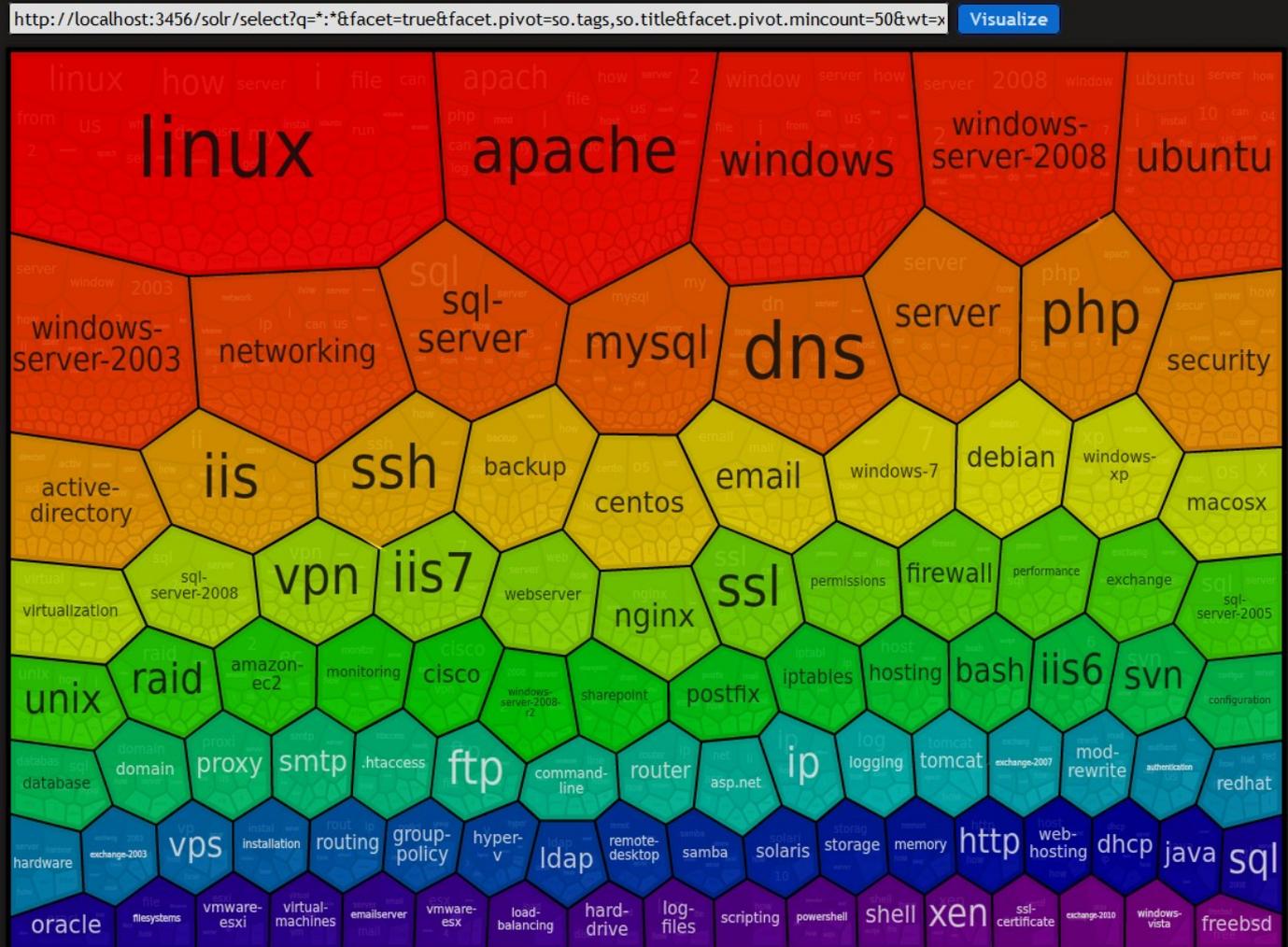


XSLT

Solr URL

Save to `${solr.home}/conf/xslt/c2-pivot`

`/solr/select?q=*&facet=true&facet.pivot=so.tags,so.title&wt=`



Configuration,  
performance  
tuning, cluster  
tuning

[wiki.apache.org/solr/ClusteringC](http://wiki.apache.org/solr/ClusteringC)  
clustering component reference & extra mater

[project.carrot2.org](http://project.carrot2.org)  
Carrot<sup>2</sup> downloads, documentation, publication

Visualization

[github.com/carrotsearch/solr-vis](https://github.com/carrotsearch/solr-vis)  
cluster & facet visualization, XSLTs,  
HTML/JS

Thanks for  
listening!