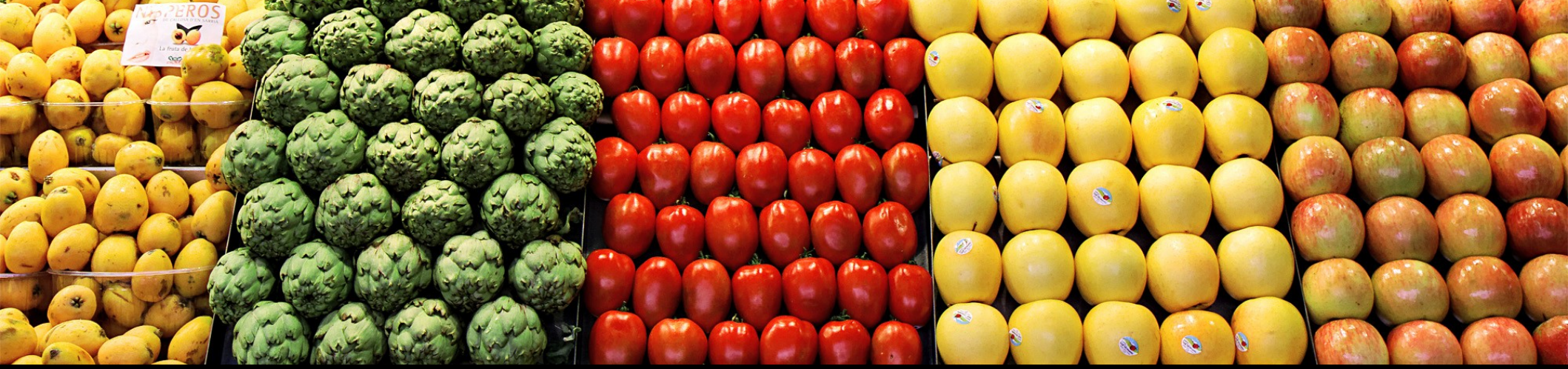
Clustering and visualization of Solr search results

BERLIN BUZZ WORDS

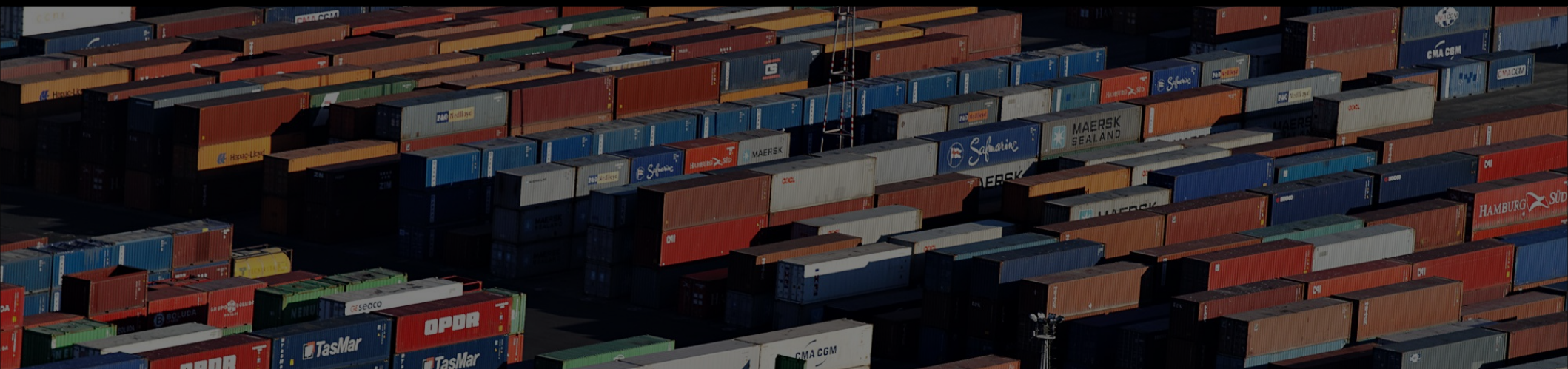
Stanislaw Osinski
stanislaw.osinski@carrotsearch.com

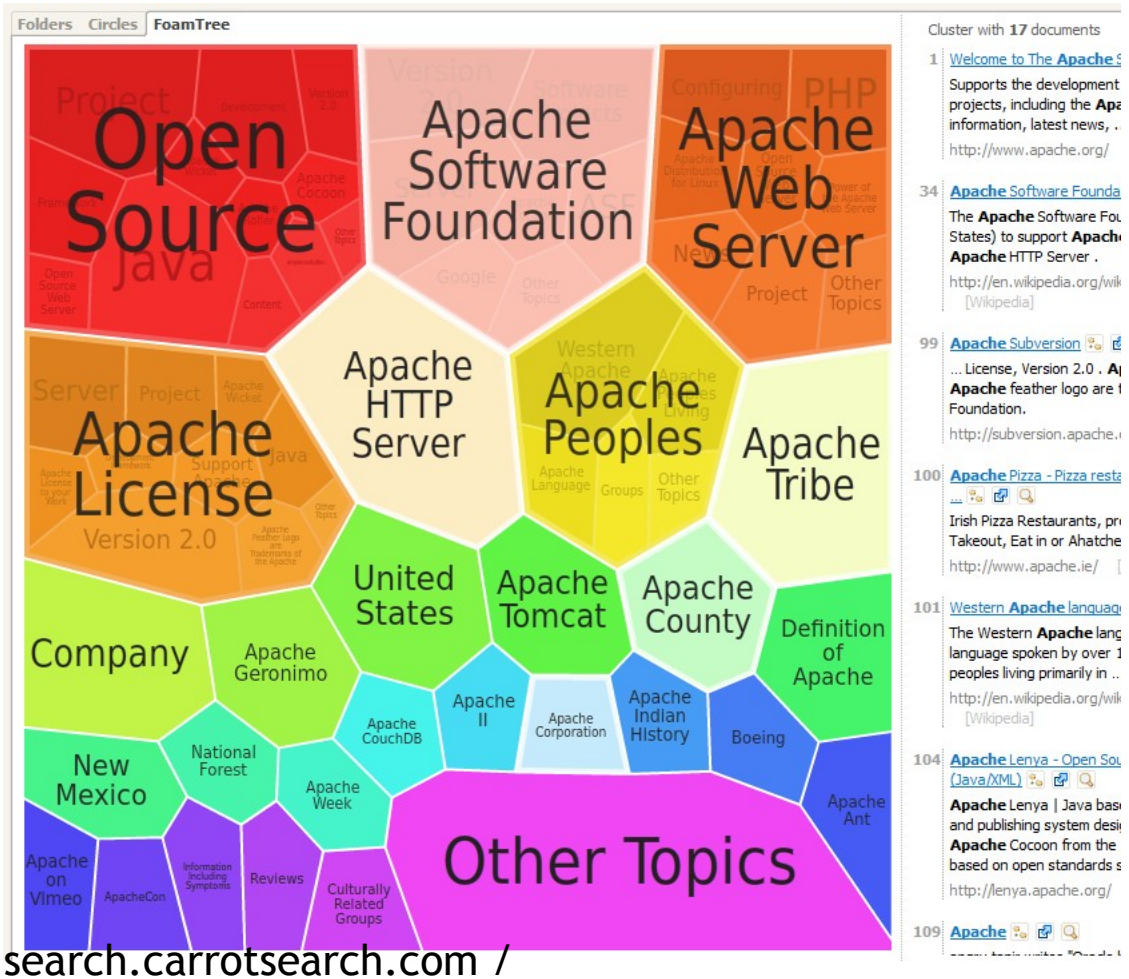
CARROT search



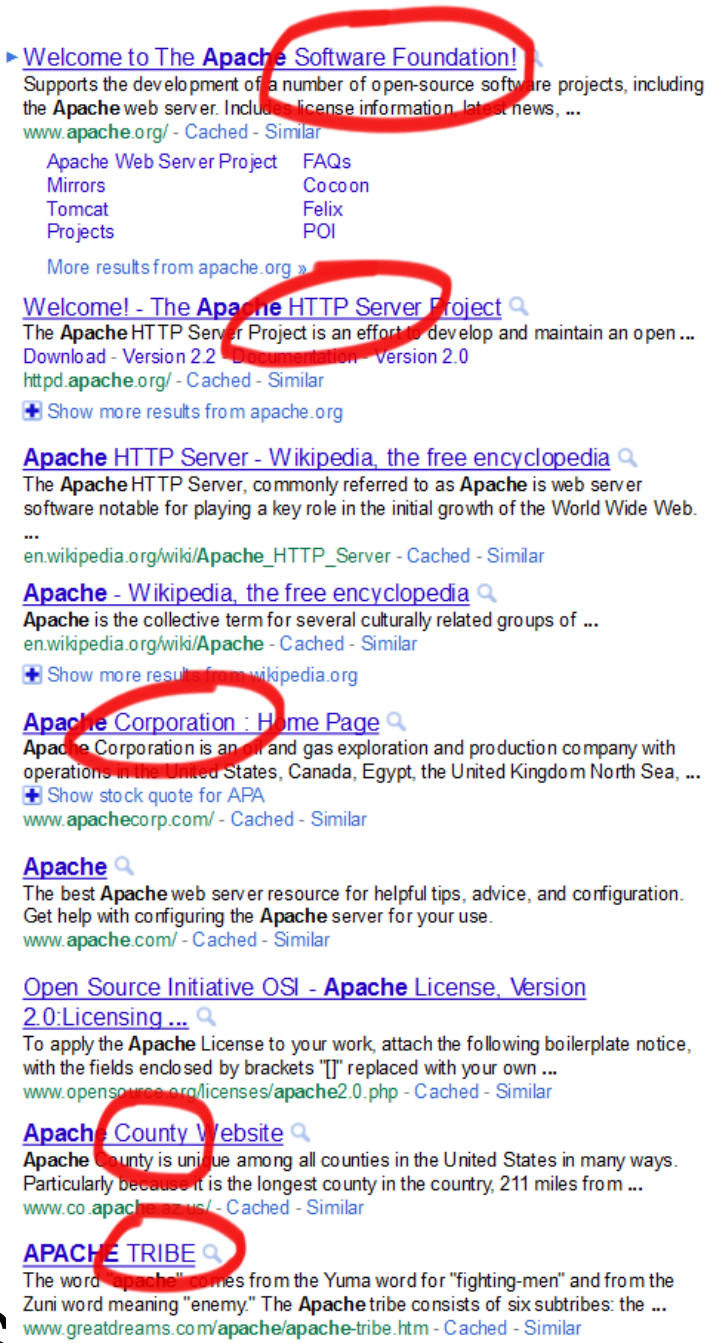
Put **similar** objects together, separate **dissimilar** ones

Put **similar** objects together, separate **dissimilar** ones


MAERSK
SEALAND
HAMBURG SÜD
CMA CGM
TasMar
OPDR
seaco



search.carrotsearch.com / search.carrot2.org



google.com

Search results clustering has been there for years

Document clustering	Search results clustering
off-line process	on-line process
global clusters	query-specific clusters
no query-time cost	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

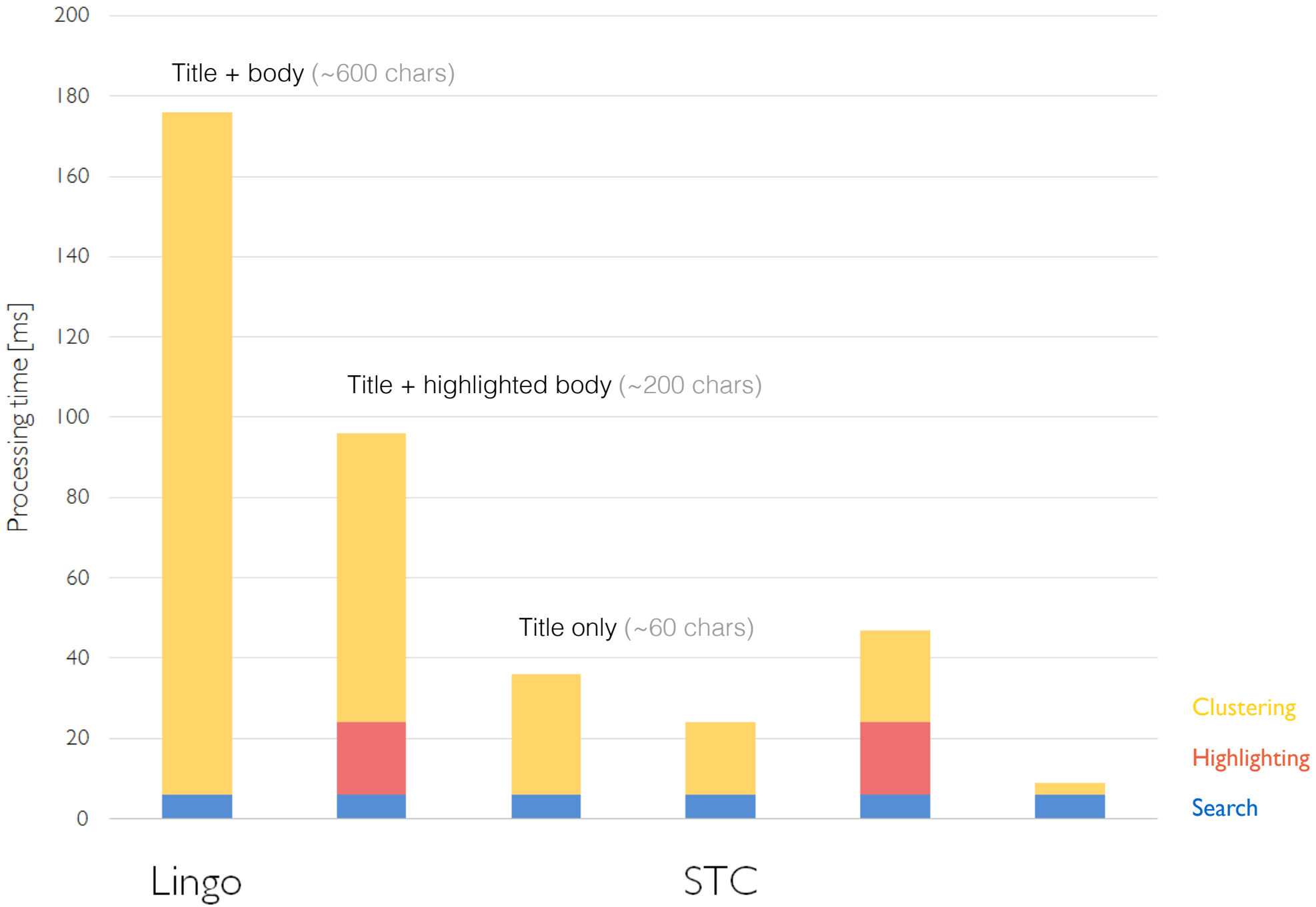
http://localhost/solr/clustering
?q=*:*&rows=100

```
<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 documensts -->
   </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



Solr 4.0 (trunk) | single-term query | **100 rows** | 1 thread | average of 50 runs after JVM warm-up | Core2 3 GHz, Intel X25-M

For better clustering performanc

- Limit the length of clustered content or use highlighting
- Cluster only document titles
- Use STC instead of Lingo
- Tune algorithms' settings

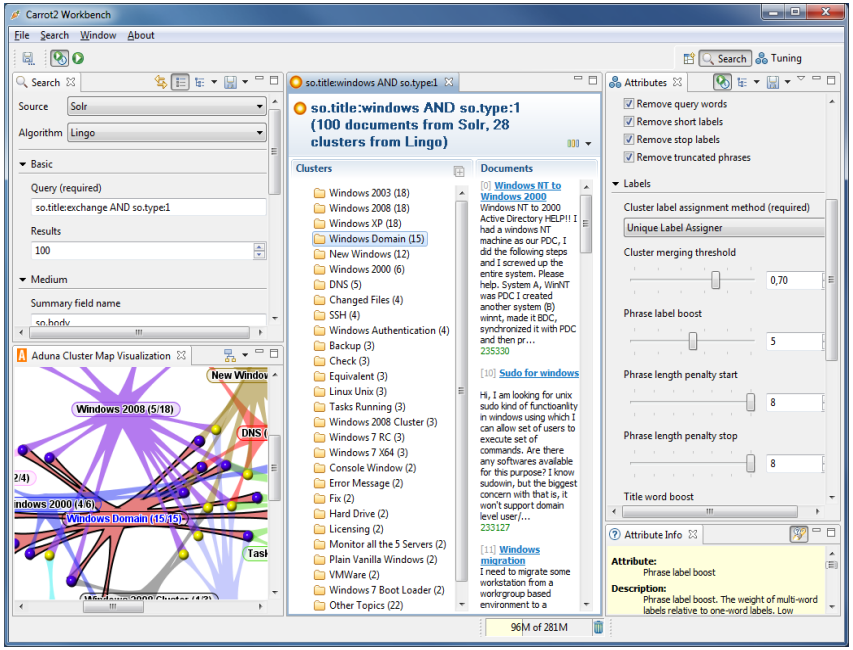
download.carrot2.org/stable/manual/#section.advanced-topi

Cluster tuning

Rapid prototyping: cluster visualization

Carrot² Clustering Workbench

get from **download.carrot2.org**
provide your Solr URL, field names and query



Clustering settings

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

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

Lexical resources

edit stop words, stop labels

Copy modified stop words and stop labels to **${solr-home}/conf/clustering**

⚠ Solr3.2/4.0

* see **wiki.apache.org/solr/ClusteringComponent**

Rapid prototyping: cluster visualization

"holland casino" 250 results from the last year Search × link absolute default colors print

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 wezen.

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

*

*

*

* available on **github.com/carrotsearch/solr-visualization**

```
<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">//<![CDATA[
      $(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());
        });
      });
    //]]></script>
  </body>
</html>
```

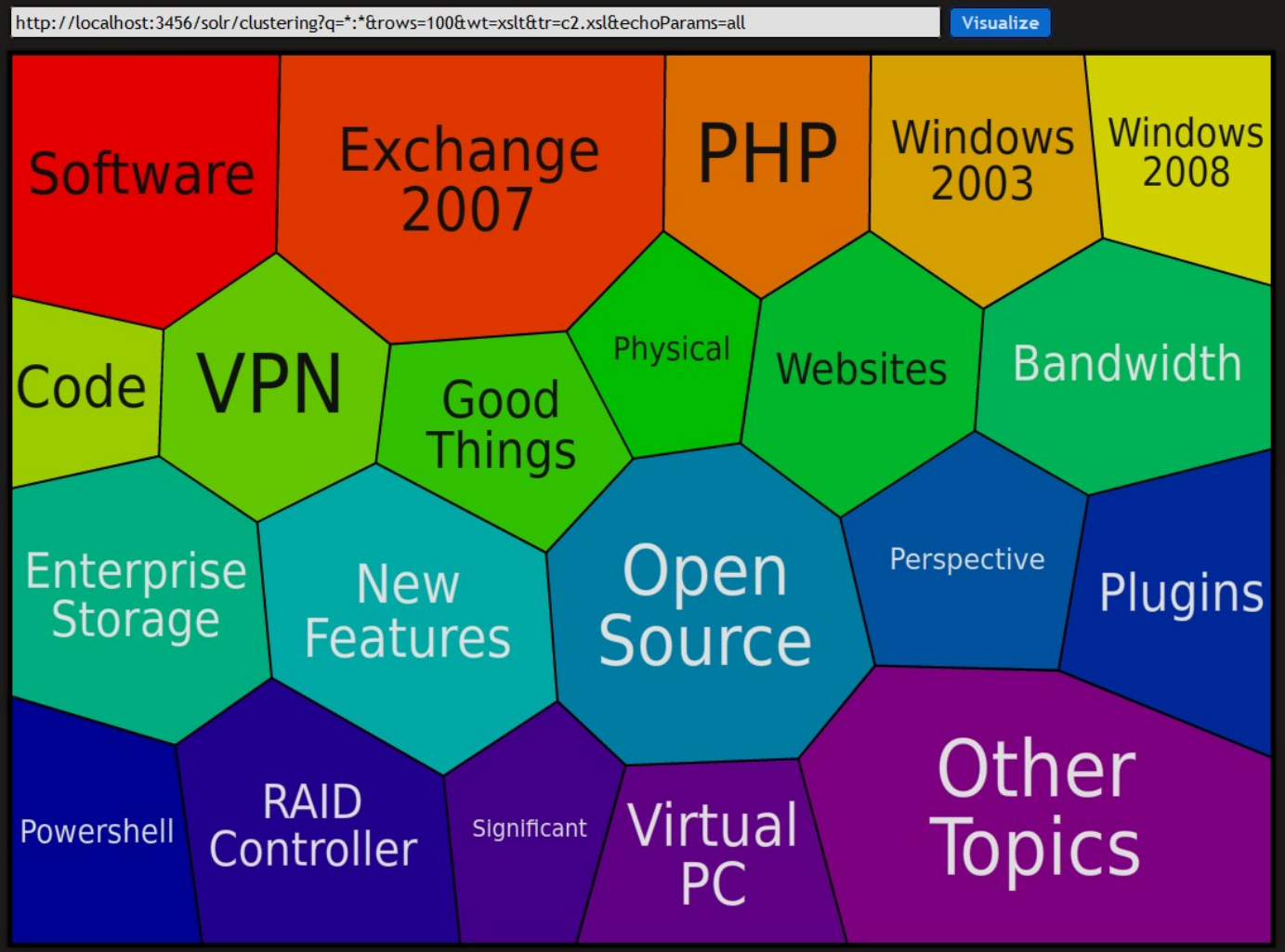
HTML / JavaScript magic

XSLT

Solr URL

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

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



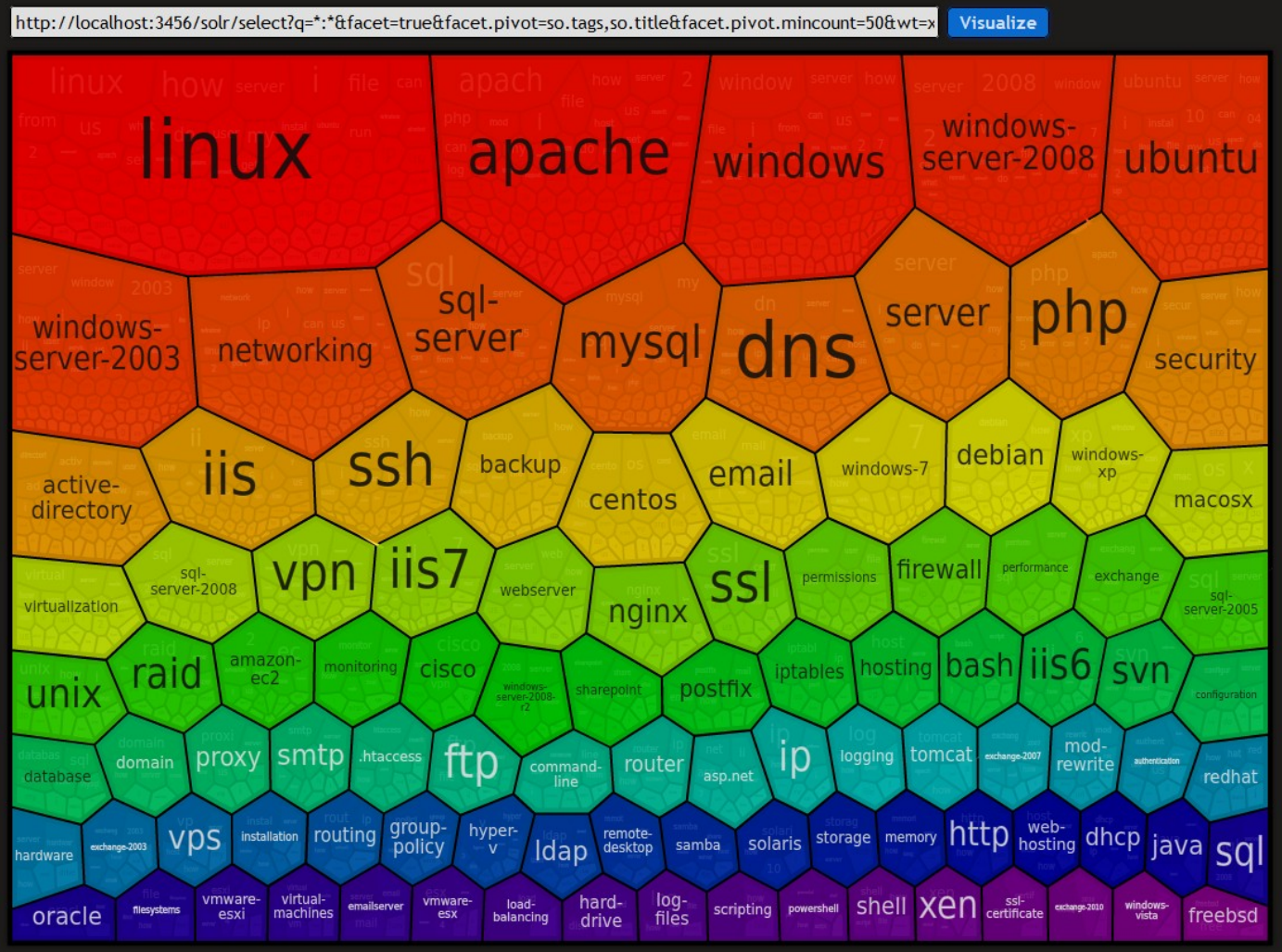
data from: serverfault

XSLT

Solr URL

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

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



data from: serverfault

Configuration, performance tuning, cluster tuning

wiki.apache.org/solr/ClusteringC
clustering component reference & extra materi

project.carrot2.org
Carrot^2 downloads, documentation, publicatior

Visualization

github.com/carrotsearch/solr-vis
cluster & facet visualization, XSLTs, HTML/JS

Thanks for listening!