Discussion Session
Week 6

INF 141: Information Retrieval
Winter 2008

Yasser Ganjisaffar
yganjisa@ics.uci.edu
Assignment 03

• Longest known (to me) Palindrome in Wikipedia:

<table>
<thead>
<tr>
<th>Discography</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Singles</strong></td>
</tr>
<tr>
<td>1966 &quot;They're Coming To Take Me Away, Ha-Haaa!&quot; / &quot;IaaHaH ,yawA eMekaT oT gnimoC er'yeht&quot; Warner Bros.</td>
</tr>
<tr>
<td>1966 &quot;I'm In Love With My Little Red Tricycle&quot; / &quot;Doin' The Napoleon&quot; Warner Bros.</td>
</tr>
<tr>
<td>1973 &quot;They're Coming To Take Me Away, Ha-Haaa!&quot; / &quot;IaaHaH ,yawA eMekaT oT gnimoC er'yeht&quot; Warner Bros.</td>
</tr>
</tbody>
</table>


• Longest reported Palindrome:
  - s. tropeR trebloC ehT si sihT." ("This is The Colbert Report" s
Assignment 03

• Longest known (to me) Rhopalic in Wikipedia (17 words):

• Longest reported Rhopalic:
  – of ABN Amro Hoare Govett, Tourism Training Australia
  – and hair salon, called "Dalohan," becoming customers themselves
## Assignment 03 – Your Results

<table>
<thead>
<tr>
<th>Pages</th>
<th>Links</th>
<th>Duration</th>
<th>Text</th>
<th>Content</th>
<th>Rhopalic</th>
<th>Lipogram</th>
<th>Palindrome</th>
<th>Obama</th>
<th>Link/Page</th>
<th>HTML/Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>912,500</td>
<td>137,338,042</td>
<td>2.5d</td>
<td>9.5</td>
<td>41.3</td>
<td>7</td>
<td>OK-vandalism</td>
<td>OK-vandalism</td>
<td>Not Found</td>
<td>150.5074433</td>
<td>4.347368421</td>
</tr>
<tr>
<td>233,921</td>
<td>79,486,017</td>
<td>2.5d</td>
<td>3.3</td>
<td>13.4</td>
<td>7</td>
<td>Mid</td>
<td>Short - Not vandalism</td>
<td>4846</td>
<td>339.7985516</td>
<td>4.060606061</td>
</tr>
<tr>
<td>100,000</td>
<td>23,164,593</td>
<td>6.4h</td>
<td>1.6</td>
<td>6.4</td>
<td>6+ bug</td>
<td>OK-Source?</td>
<td></td>
<td></td>
<td>231.64953</td>
<td></td>
</tr>
<tr>
<td>235,712</td>
<td>60,394,129</td>
<td>?</td>
<td>4.4</td>
<td>7.8</td>
<td>5</td>
<td>Too Short-Not vanda</td>
<td>Very Short - Not vandalism</td>
<td>1</td>
<td>256.2200015</td>
<td>1.772727273</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>321,000</td>
<td>59,358,861</td>
<td>1d</td>
<td>4.2</td>
<td></td>
<td>6</td>
<td>OK</td>
<td>OK-vandalism</td>
<td>3340</td>
<td>184.9185701</td>
<td></td>
</tr>
<tr>
<td>17,000</td>
<td>4,926,699</td>
<td>4h</td>
<td>0.4</td>
<td></td>
<td>5</td>
<td>Mid</td>
<td>Mid</td>
<td>3486</td>
<td>289.8058235</td>
<td></td>
</tr>
<tr>
<td>800,000</td>
<td>123,857,058</td>
<td>2.5d</td>
<td>8.6</td>
<td>37.5</td>
<td>8</td>
<td>OK-vandalism</td>
<td>OK-vandalism</td>
<td>47</td>
<td>154.8213225</td>
<td>4.360465116</td>
</tr>
<tr>
<td>445,950</td>
<td>74,909,376</td>
<td>3d</td>
<td>5.4</td>
<td>22.5</td>
<td>7</td>
<td>OK</td>
<td>OK (Peel's foe not a sw)</td>
<td>2120</td>
<td>167.9770737</td>
<td>4.166666667</td>
</tr>
<tr>
<td>360,000</td>
<td>61,636,580</td>
<td>2d</td>
<td>4.6</td>
<td>18.7</td>
<td>5</td>
<td>Mid+</td>
<td>OK-non sense</td>
<td>40772</td>
<td>171.2127222</td>
<td>4.065217391</td>
</tr>
<tr>
<td>137,500</td>
<td>27,755,110</td>
<td>2d (6h?)</td>
<td>2.1</td>
<td>8.3</td>
<td>6</td>
<td>OK</td>
<td>---</td>
<td>2 (seeded)</td>
<td>201.8553455</td>
<td>3.952380952</td>
</tr>
<tr>
<td>1,012,148</td>
<td>147,904,209</td>
<td>2d</td>
<td>10.7</td>
<td>46.6</td>
<td>8</td>
<td>OK-vandalism</td>
<td>OK - All a -vandalism</td>
<td>253</td>
<td>146.1290335</td>
<td>4.355140187</td>
</tr>
<tr>
<td>845,200</td>
<td>1.27×10^4???</td>
<td>12h (24h?)</td>
<td></td>
<td></td>
<td>7</td>
<td>OK</td>
<td>OK</td>
<td>3340</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37,273</td>
<td>8,353,413</td>
<td>12h</td>
<td>0.6</td>
<td>2.4</td>
<td>7+ bug</td>
<td>OK</td>
<td>OK (Peel's foe not a sw)</td>
<td>2</td>
<td>224.1143187</td>
<td>4</td>
</tr>
<tr>
<td>/3 /</td>
<td>218,646</td>
<td>11h</td>
<td>---</td>
<td></td>
<td>---</td>
<td>Bad</td>
<td>---</td>
<td></td>
<td>296.6 jU2849</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>259,550</td>
<td>46,478,416</td>
<td>3d</td>
<td>3.3</td>
<td>13.6</td>
<td>8</td>
<td>OK</td>
<td>OK (s. tropeR trebloC)</td>
<td>17213</td>
<td>179.0730726</td>
<td>4.121212121</td>
</tr>
<tr>
<td>202,789</td>
<td>1,770,923</td>
<td>6d</td>
<td>2.2</td>
<td>10</td>
<td>8</td>
<td>OK</td>
<td>OK</td>
<td>25480</td>
<td>8.732835607</td>
<td>4.545454545</td>
</tr>
</tbody>
</table>
What is Hadoop

Flexible infrastructure for large scale computation and data processing on a network of commodity hardware.
Current State of Hadoop Project

• Top level Apache Foundation project
• In production at Yahoo!, Facebook, IBM, Amazon, Fox, NY Times, ...
• Very active and strong development team

• Last year, Yahoo! Launched a Hadoop application that runs on a more than 10,000 core Linux Cluster and produces data that is now used in every Yahoo! Web search query.
What is Hadoop?

The Linux of Distributed Processing.
Very Large Storage Requirements

10,000,000 GB

No single storage can handle this amount of data.

You need a large set of nodes each storing part of the data.
HDFS

1. filename, index

2. Datanodes, Blockid

3. Read Data

Client

Namenode

Datanodes
Distributed Word Count

Very big data

Split data → count → count → merge → merged count

Split data → count → count

Split data → count → count

Split data → count → count
Programming Model: Map/Reduce

• Very simple programming model:
  – Map(anything) -> key, value
  – Sort, partition on key
  – Reduce(key, value) -> key, value

• No parallel processing / message passing semantics
Counting Words by Map/Reduce

Hello World
Bye World
Hello Hadoop
Goodbye Hadoop

Split

Hello World
Bye World

Hello Hadoop
Goodbye Hadoop
Counting Words by Map/Reduce

Node 1

Hello World  
Bye World  

Mapper 1

<Hello, 1>  
<World, 1>  
<Bye, 1>  
<World, 1>

Combiner 1

<Bye, 1>  
<Hello, 1>  
<World, 11>

Node 2

Hello Hadoop  
Goodbye Hadoop  

Mapper 2

<Hello, 1>  
<Hadoop, 1>  
<Goodbye, 1>  
<Hadoop, 1>

Combiner 2

<Goodbye, 1>  
<Hadoop, 2>  
<Hello, 1>
Counting Words by Map/Reduce

 Merge & Sort By key

<Bye, 1>
<Hello, 1>
<World, 2>

<Bye, 1>
<Goodbye, 1>
<Hadoop, 2>
<Hello, 1, 1>
<World, 2>

Split

<Bye, 1>
<Goodbye, 1>
<Hadoop, 2>

<Hello, 1, 1>
<World, 2>

<Bye, 1>
<Hello, 1>
<Goodbye, 1>
<Hadoop, 2>
Counting Words by Map/Reduce

Node 1

<Bye, 1>
<Goodbye, 1>
<Hadoop, 2>

Reducer 1

<Bye, 1>
<Goodbye, 1>
<Hadoop, 2>

Node 2

<Hello, 1, 1>
<World, 2>

Reducer 2

<Hello, 2>
<World, 2>
Installation Requirements

- Linux
- Java 1.6.x
- ssh

Links