Developing an Analytics Dashboard for Coursera MOOC Discussion Forums

CNI Fall 2014 Membership Meeting

Bill Parod
Northwestern University Information Technology
• Private / Big Ten
• Campuses in Evanston, Chicago, Qatar
• 12 Schools and Colleges

• 19,000 Students
• 2,500 Full-time Faculty
• $1.9B Budget
• $7B Endowment
Northwestern Coursera MOOCs
### Coursera Analytics Dashboard

<table>
<thead>
<tr>
<th>Reach</th>
<th>Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>27,034 total learners joined</td>
<td>244 with Signature Track</td>
</tr>
<tr>
<td>15,365 have visited all time</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity</th>
<th>This week</th>
<th>All time</th>
</tr>
</thead>
<tbody>
<tr>
<td>visited the course</td>
<td>630</td>
<td>15,365</td>
</tr>
<tr>
<td>watched a lecture</td>
<td>226</td>
<td>10,340</td>
</tr>
<tr>
<td>submitted an exercise</td>
<td>15</td>
<td>2,772</td>
</tr>
<tr>
<td>browsed the forums</td>
<td>99</td>
<td>3,346</td>
</tr>
</tbody>
</table>
Coursera Data Exports

- 48 Tables:
- Structure & Content
- Users
- Permissions
- Assessments
- Forums
- Quiz Submissions
- Lecture Views

ER Diagram from University of Michigan
Initial Analysis Efforts

• Past Project Experience
  – Humanities Computing
  – Corpus Linguistics
  – Repository Applications

• Discussion Forum Analysis and Visualization
  – Topic Analysis
  – Sentiment Analysis
  – Geo-location
  – Data Model Simplification
  – Distinguishing Group Statistics
  – Visualization Techniques
Top 18 Topics and their Keywords

1. news articles sources local newspaper story newspapers source times topic
2. google data users gmail privacy services email maps service tools
3. http www youtube video watch videos html user blog watching
4. world things life thing human make lives live change thinking
5. knowledge learn learning research students great taking education experience mooc
6. information people public access power internet agree making fact government
7. week assignment points work lectures assignments quiz peer students written
8. content free pay advertising money paid model market revenue paying
9. time work people day great things job school home lot
10. google good thing makes searching start bad lot find mind
11. ads ad page click sites site view website relevant pages
12. google http en wikipedia org world filter bubble wiki project
13. google search results engine searches web result bing page seo
14. agree long problem time issues issue words check bit type
15. media internet tv online music create age today impact audience
16. question answer questions give understand provide find doesn answers wrong
17. google company companies business products big product services service customers
18. facebook social people friends twitter personal media don privacy share

Topic Modeling
UMass Mallet
TOPIC: news articles sources local newspaper story newspapers ...

<table>
<thead>
<tr>
<th>rank</th>
<th>words</th>
<th>docID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>187</td>
<td>post630</td>
</tr>
<tr>
<td>3</td>
<td>131</td>
<td>post4748</td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>post6717</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>post4215</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>post6550</td>
</tr>
<tr>
<td>7</td>
<td>54</td>
<td>post4300</td>
</tr>
<tr>
<td>8</td>
<td>53</td>
<td>post5118</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
<td>post6507</td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>post5210</td>
</tr>
<tr>
<td>11</td>
<td>49</td>
<td>post4724</td>
</tr>
<tr>
<td>12</td>
<td>49</td>
<td>post464</td>
</tr>
<tr>
<td>13</td>
<td>49</td>
<td>comment179</td>
</tr>
<tr>
<td>14</td>
<td>47</td>
<td>post4086</td>
</tr>
<tr>
<td>15</td>
<td>45</td>
<td>post4226</td>
</tr>
<tr>
<td>16</td>
<td>43</td>
<td>post4864</td>
</tr>
</tbody>
</table>

Topic Modeling
UMass Mallet
Topics / Posts Network Graph

Gephi
Topics / Posts Network Graph

Gephi
<table>
<thead>
<tr>
<th>Question</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>13A. Communications from the course staff and instructor (in course-wide emails and announcements, or in the discussion forums)</td>
<td>Positive (0.132069)</td>
</tr>
<tr>
<td>13B. Video lectures and interviews</td>
<td>Positive (0.0994045)</td>
</tr>
<tr>
<td>13C. Assessments (quizzes and peer-graded homework)</td>
<td>Negative (-0.195772)</td>
</tr>
<tr>
<td>13D. Coursera learning platform</td>
<td>Positive (0.130757)</td>
</tr>
<tr>
<td>13E. Final comments and suggestions</td>
<td>Positive (0.0543005)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Relevance</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>course</td>
<td>0.999076</td>
<td>Positive (0.0457994)</td>
</tr>
<tr>
<td>staff</td>
<td>0.454871</td>
<td>Positive (0.00740671)</td>
</tr>
<tr>
<td>forums</td>
<td>0.420342</td>
<td>Negative (-0.0817278)</td>
</tr>
<tr>
<td>Youngman</td>
<td>0.419881</td>
<td>Positive (0.104288)</td>
</tr>
<tr>
<td>class</td>
<td>0.408228</td>
<td>Positive (6.12065E-4)</td>
</tr>
<tr>
<td>time</td>
<td>0.407215</td>
<td>Negative (-0.00467706)</td>
</tr>
<tr>
<td>communication</td>
<td>0.391934</td>
<td>Positive (0.235323)</td>
</tr>
<tr>
<td>course staff</td>
<td>0.390933</td>
<td>Negative (-0.0220035)</td>
</tr>
<tr>
<td>instructor</td>
<td>0.388226</td>
<td>Positive (0.255751)</td>
</tr>
<tr>
<td>Prof. Youngman</td>
<td>0.38083</td>
<td>Positive (0.101429)</td>
</tr>
</tbody>
</table>

Sentiment Analysis
AlchemyAPI
IP-Address Geolocation

*GeoLite2 + KML + Google Earth*
MOOC Statistical Analysis

Two-sample t-test on postVotes Grouped by AllHWsDone vs Alternative = 'not equal'

<table>
<thead>
<tr>
<th>GROUP</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>54,334</td>
<td>0.102</td>
<td>4.239</td>
</tr>
<tr>
<td>1</td>
<td>1,190</td>
<td>7.339</td>
<td>13.129</td>
</tr>
</tbody>
</table>

Separate Variance

Difference in Means : -7.238
95.00% Confidence Interval : -7.985 to -6.490
\( t \) : -18.995
\( df \) : 1,194.436
p-value : 0.000
Dunn-Sidak Adjusted p-value : 0.000

Distinguishing Groups
All Homeworks Done / Post Votes
Bubble Graph of MOOC Components

Assignments
Forums
Quizzes
Video Lectures
... but what are people saying in the Forums? Well, let’s search them!
Full-text search with faceted browsing
Faceted Browse Application Framework

- Drupal Interface Components
- Drupal and LDAP Identity and Access Management (IAM)
- Solr Indexing and Discovery
- Google Charts and D3 Visualization

- ... but can we offer more facets?
Metadata Enhancement Services

• Named Entity Recognition (Stanford NLPG NER)
  – People
  – Place Names
  – Organizations
• Keyword Extraction (topia.termextract)
• Sentiment Analysis (local impl. of Narayanan et al)
• Topical Analysis (DMOZ “Open Directory Project”)
• IP Address to Geographic Location (MaxMind)
• Place Name to Geographic Location (GeoNames)
Enterprise SOA Meets Humanities Computing

• Enterprise Service Bus and Registry
  – Portfolio of services
  – Reuse locally developed software
  – Leverage growing abundance of external services

• Flexible Service Orchestration
  – Compose and orchestrate services for specific needs
  – Obtain low incremental cost for new projects
  – Used for Metadata Enhancement & Indexing Pipeline
  – Leverage Apache Camel / Fuse / JMS Messaging
Service Orchestration

Metadata Enhancement Pipeline

Event Aggregation  Entity Extraction  Keyword Extraction  Geo Location  Topic Analysis  Sentiment Analysis

Metadata Enhancement Services
Visualization Graphics for Result Sets
Streamgraph of Learner Country

United States
Streamgraph of *Organizations Mentioned*
Latitude / Longitude Heatmap
Geographic Coverage by Country
Pie Chart for any Facet

Learner City Pivot on Learner Country (India)
Word Cloud on Keyword Facet
Word Cloud on Keyword Facet
Word Cloud on *Places Mentioned* Facet
Remove, Pivot, and Inspect Posts
3 Ideas...

- Routinize Analysis Services with SOA
  - Text analysis services
  - Geographic lookup services
  - Sentiment analysis

- Leverage Analysis Services with Orchestration Layer
  - Metadata enhancement
  - Metadata and full text indexing

- Application Framework with Visualization Graphics
  - Full Text and Faceted Discovery
  - Result Set Visualization

next steps...
Next Steps

• Faculty Feedback
  – More demographic fields (skill level, education, time expectation, etc…)
  – Train classifiers for various types of interaction
  – Alert systems for “On Demand” courses?
  – Geolocation for Place Names
  – … gather more feedback – We just recently launched

• Application Framework
  – Evolve application platform (instead of Drupal)
  – Evaluate other “noSQL” indexing solutions (instead of Solr)

• Learning Analytics
  – LTI Apps for Instructure Canvas Transition Reporting
  – Early Discussions at NU and CIC about Learning Analytics
Thank You
Questions?

bill-parod@northwestern.edu
Northwestern University Information Technology