Outline

• System Idea/Problem Statement
• Method/Approach
• System Evaluation
• Future Work
System Idea/Problem Statement

- Aggregated sports articles
- Social media
- Fast-paced updates
Method/Approach (1/2)

- **Crawler**
  - Wget crawler
  - Python based document parser

- **Twitter Integration**
  - Search API (versus Streaming API)
  - Twitter4J

- **Building Inverted Index**
  - Document parsing
  - Lucene
Method/Approach (2/2)

- **Back-end**
  - Solr & Lucene index compatibility
  - Configuration - schema.xml
  - Querying

- **Front-end**
  - HTML & CSS (Bootstrap)
  - Website Design/Structure
  - Javascript & AJAX search box and connection to back-end

- **Automation**
System Evaluation

- Precision@k evaluation

<table>
<thead>
<tr>
<th>Query</th>
<th>Precision@10 for News Articles</th>
<th>Precision@10 for Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeBron James</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Christiano Ronaldo (note: incorrect spelling)</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Football</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Baseball</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Soccer</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>TOTAL</td>
<td>0.96</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Future Work

- Crawling more content providers
- Addition of Topic Modeling
- Front-end integration of top sports news summary (categorized by topics)
- Sentiment Analysis
- Incremental improvement of back-end/automation
Questions?
Brian: 1 - 5,8 (last 2)
Chris: 6,7,8 (first 3)
topic modeling: go by sports, connect to front-end, LDA (Latent_Dirichlet_allocation)