MeTA: A Unifying Toolkit for the Management and Analysis of Text Data

Chase Geigle and ChengXiang ("Cheng") Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

gingle1@Illinois.edu; czhai@Illinois.edu

https://meta-toolkit.org/
Text data cover all kinds of topics

**Topics:**
- People
- Events
- Products
- Services, ...

**Sources:**
- Blogs
- Microblogs
- Forums
- Reviews, ...

- 45M reviews
- 53M blogs
- 1307M posts
- 65M msgs/day
- 115M users
- 10M groups

...
Humans as **Subjective & Intelligent “Sensors”**

Real World → Sense → Sensor → Report → Data

- **Weather**
  - Thermometer → 3℃, 15°F, ...

- **Locations**
  - Geo Sensor → 41°N and 120°W, ...

- **Networks**
  - Network Sensor → 01000100011100

Perceive → Express

“Human Sensor”
Main Techniques for Harnessing Big Text Data: Text Retrieval + Text Analysis

Big Text Data

Text Retrieval

Small Relevant Data

Text Analysis

Knowledge

Many Applications
This tutorial: MeTA Toolkit

https://meta-toolkit.org/

Big Text Data

Small Relevant Data

Knowledge

Many Applications

Text Retrieval

Text Analysis

Many Applications
Overview of MeTA

• **Founders & Key contributors:** Chase Geigle, Sean Massung

• **Advisor:** ChengXiang Zhai

• **Design philosophy**
  – Unified toolkit for full support of text data retrieval and analysis
  – Facilitate education, research, and application development (open source)
  – Highly efficient and extensible (C++ with Python binding)
  – Continuously updated to reflect research progress (hope you can help!)

• **Current uses**
  – Mostly education (MOOCs on Coursera, courses at UIUC)
### MeTA vs. Related Toolkits

| Feature                  | Indri | Lucene | MALLET | LIBLINEAR | SVM\(^{MULT}\) | scikit | CoreNLP | MeTA
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature generation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Search</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Classification</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regression</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>POS tagging</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Parsing</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Topic models</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>n-gram LM</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Word embeddings</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Graph algorithms</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multithreading</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

MeTA Pipeline

Before the break

After the break
Anticipated Schedule

• 8:30am-8:40am: Introduction and overview
• 8:40am-8:50am: MeTA setup
• 8:50am-9:20am: Raw text processing
• 9:20am-9:30am: Indexing
• 9:30am-10:00am: BREAK
• 10:00am-10:45am: Information retrieval (Search engine competition)
• 10:45am-11:30am: Text classification (competition if time permitting)
• 11:30am-11:55am: Clustering
• 11:55am-12:00noon: Wrapup and feedback