HATHI TRUST
RESEARCH CENTER

Demonstration of Capability

Stacy Kowalczyk

Beth Plale

Loretta Auvil

Topics to be Covered

- Brief overview of HTRC web applications
- Results of experimental HPC applications and the HTRC data
- SEASR Analytics for HTRC

Web App Demo

RESEARCH CENTER

Experiment: Large Scale Data Analysis on XSEDE

RESEARCH CENTER

Experimental Environment and Results

- Dataset

  2,592,210 volumes, in total 2.1 TB, divided into 1024 partitions of 2GB each

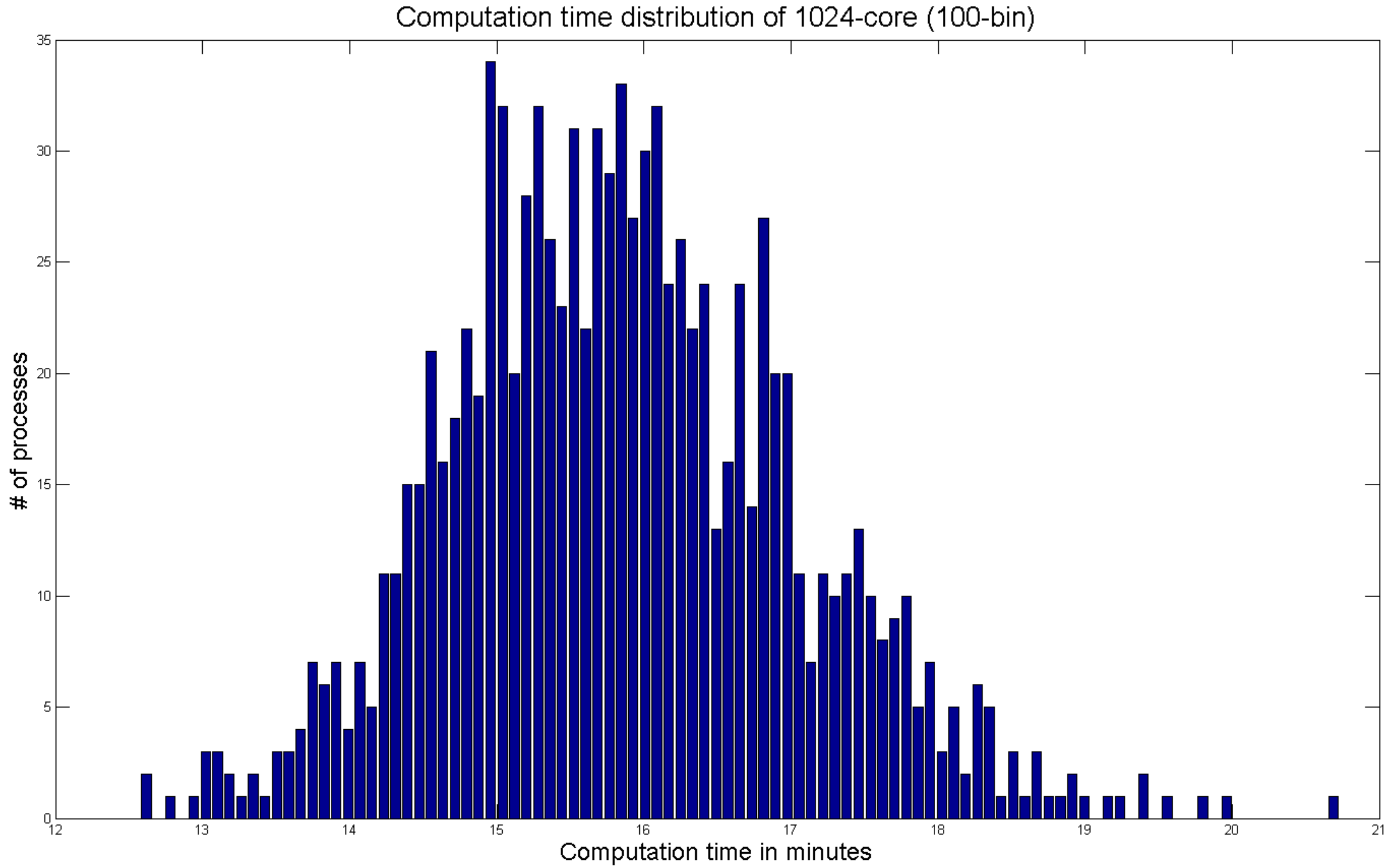
- Computation platform

  XSEDE Blacklight, 1024-core of each 2.27 GHz, 8192 GB memory. Each core processes one partition

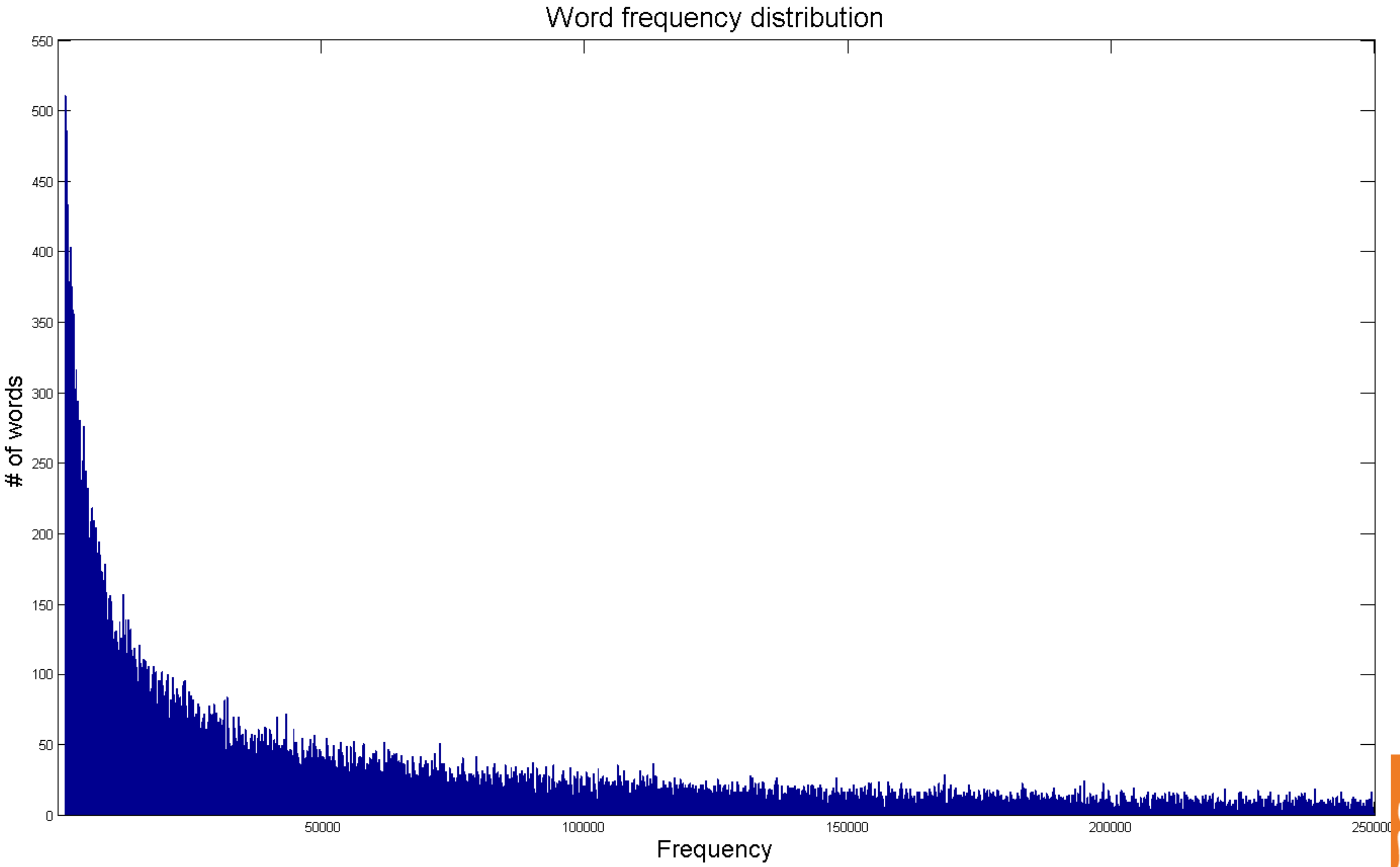
- Results

  Whole corpus word count finished in 1,454 seconds or 24.23 minutes

Computation Time Distribution



Word Frequency Distribution



RESEARCH CENTER

HATHI TRUST
RESEARCH CENTER

SEASR Analytics for HTRC

Loretta Auvil
University of Illinois

What is SEASR?

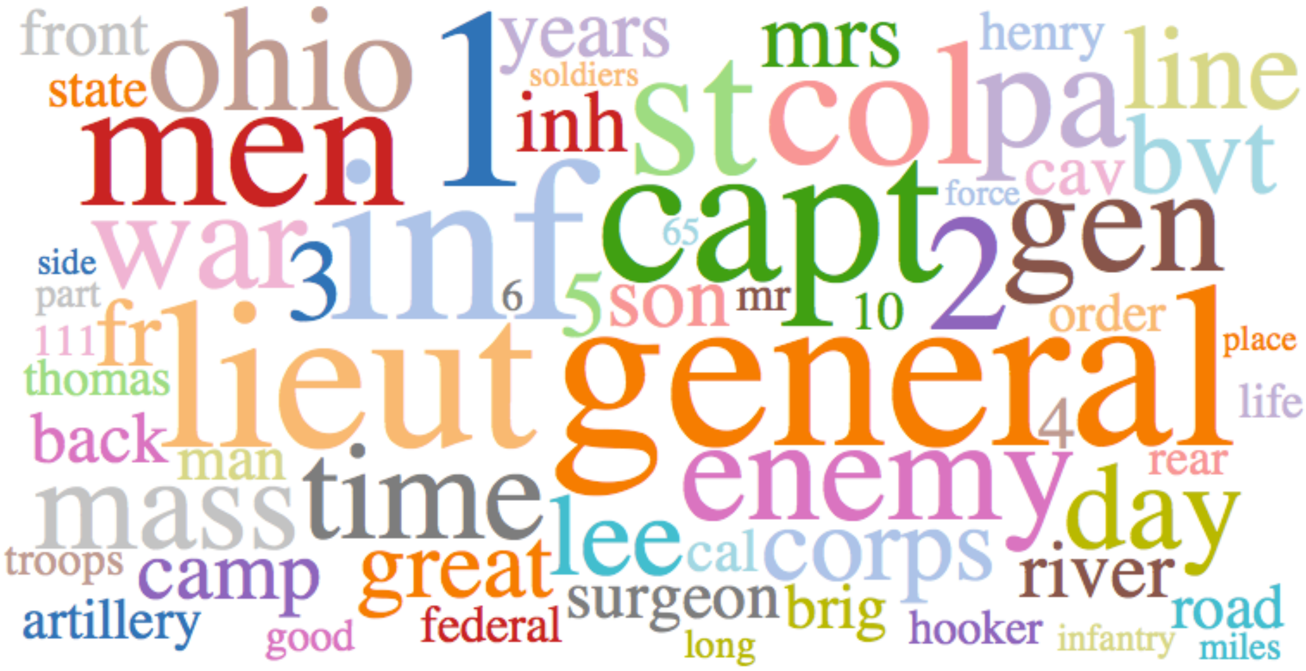
This project focus on

- developing,
- integrating,
- deploying, and
- sustaining

a set of reusable and expandable software components and a supporting framework,

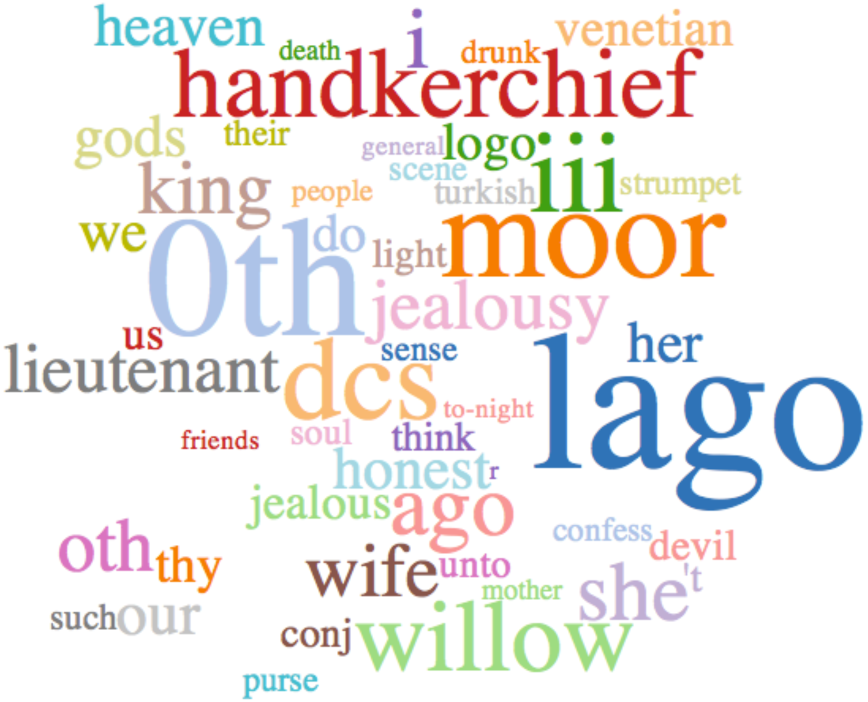
to benefit a broad set of data mining applications for scholars in humanities.

Tag Cloud Analysis

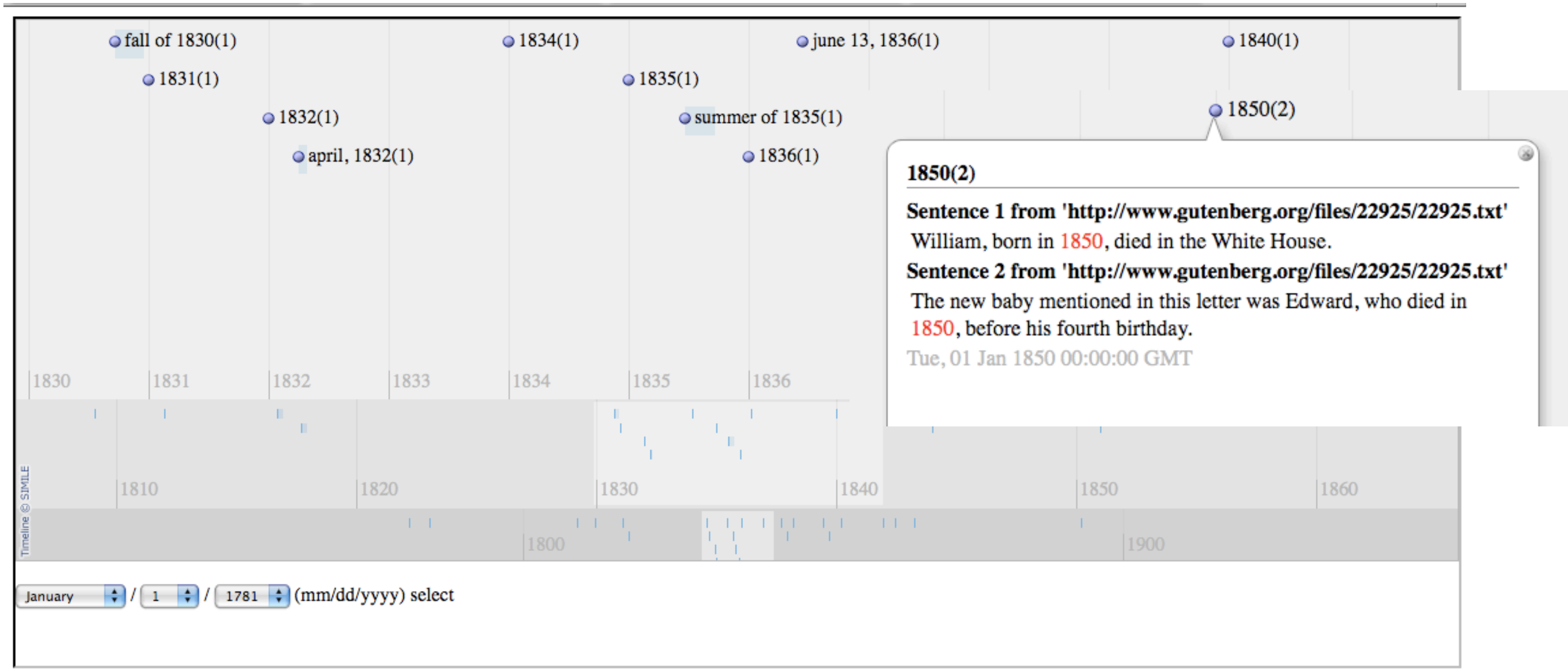


Dunning Loglikelihood

- Words more frequent in Othello than Shakespeare's other tragedies



Date Extraction to Simile Timeline



1850(2)

Sentence 1 from 'http://www.gutenberg.org/files/22925/22925.txt'

William, born in 1850, died in the White House.

Sentence 2 from 'http://www.gutenberg.org/files/22925/22925.txt'

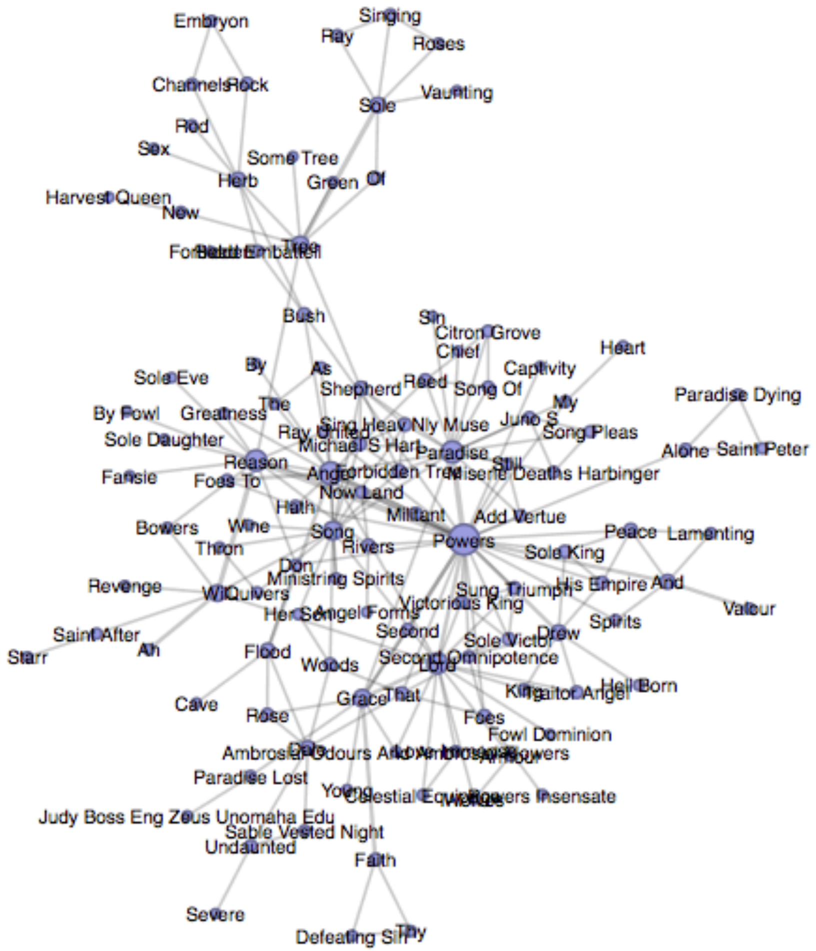
The new baby mentioned in this letter was Edward, who died in 1850, before his fourth birthday.

Tue, 01 Jan 1850 00:00:00 GMT

January / 1 / 1781 (mm/dd/yyyy) select

RESEARCH CENTER

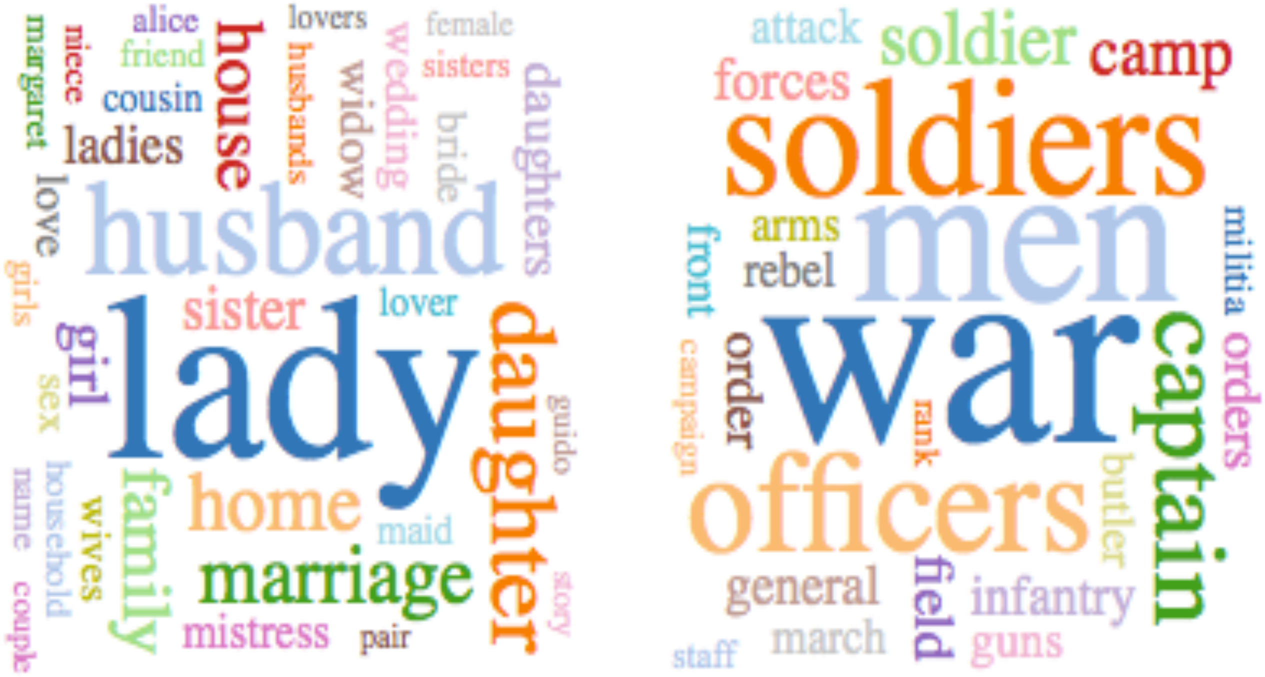
Entity Extraction for Network Analysis



RESEARCH CENTER

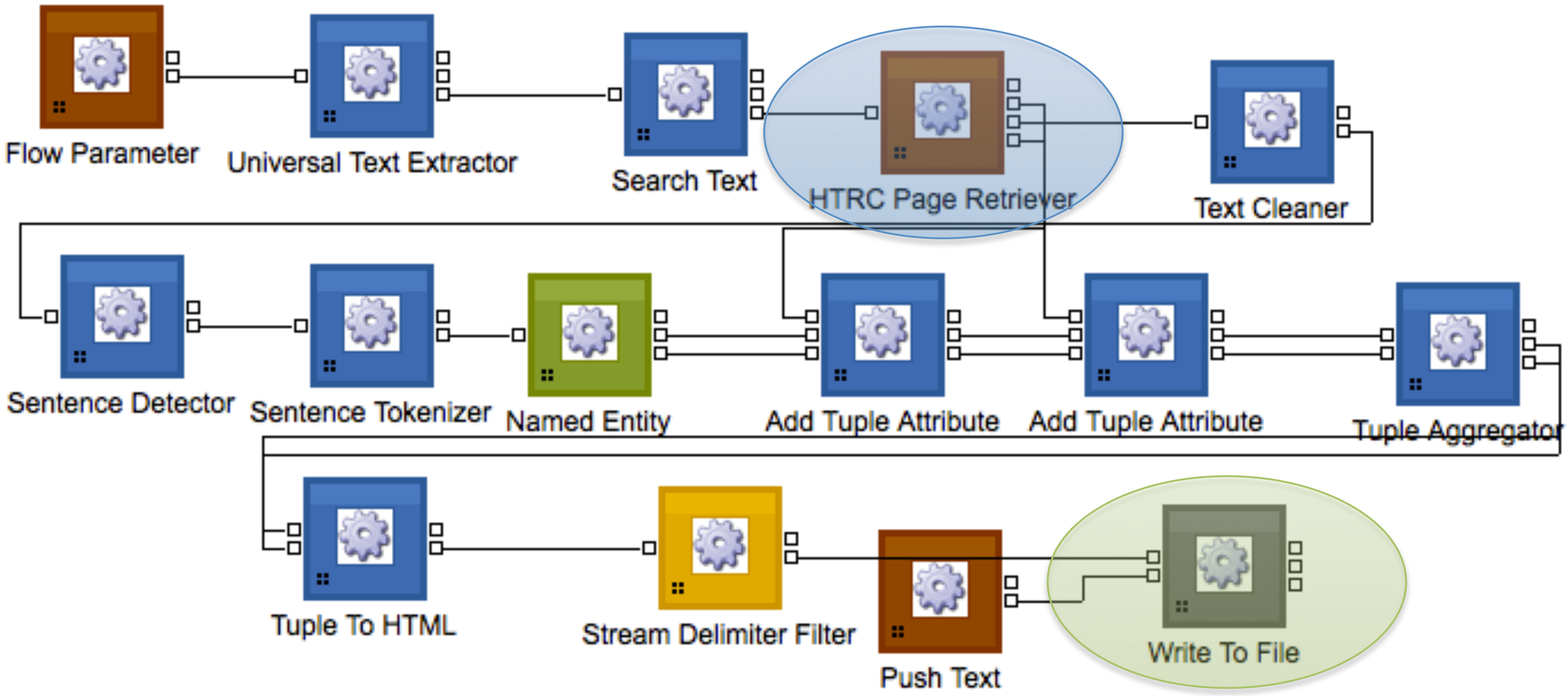
Topic Modeling

Two topics from Charles Dicken’s as author



Meandre Flow

Encapsulation and integration environment for tools and algorithms



RESEARCH CENTER

Algorithm Info

```
<algorithm>
 <info>
  <name>Meandre_Topic_Modeling</name>
  <description></description>
  <authors></authors>
<parameters>
   <param
    name="input_collection"
    type="collection"
    required="true">
    <label>Please select a collection for analysis</label>
    <description>The collection containing the volume ids to be used for analysis.</
description>
   </param>
  </parameters>
 </info>
```

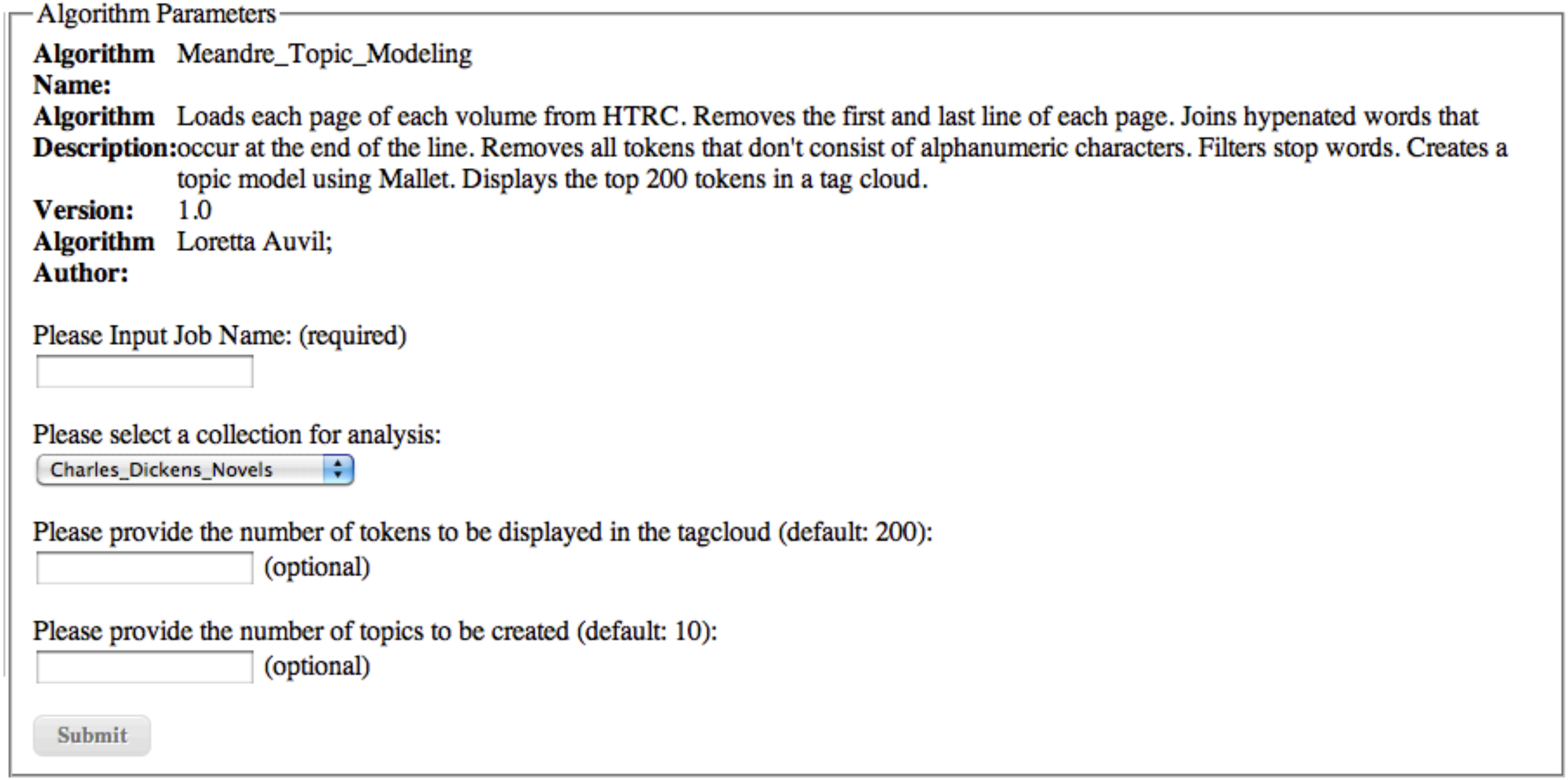
Algorithm Execution

```
<run_script>run_HTRC_Meandre_Topic_Modeling.sh</run_script>
 <properties_file_name>HTRC_Meandre_Topic_Modeling.properties
 </properties_file_name>

 <dependencies>
   <dependency name="run_HTRC_Meandre_Topic_Modeling.sh" path="htrc/agent/dependencies/meandre/run_HTRC_Meandre_Topic_Modeling.sh"/>
 </dependencies>

 <system_properties>
   <e key="volume_id">$input_collection</e>
 </system_properties>
```

Algorithm Results

```
<results>
  <result type="text/html" name="topic_tagclouds.html"/>
  <result type="text/xml" name="topic_top_words.xml"/>
</results>
```

HTRC Algorithm UI



Algorithm Parameters

Algorithm Name: Meandre_Topic_Modeling

Algorithm Description: Loads each page of each volume from HTRC. Removes the first and last line of each page. Joins hypenated words that occur at the end of the line. Removes all tokens that don't consist of alphanumeric characters. Filters stop words. Creates a topic model using Mallet. Displays the top 200 tokens in a tag cloud.

Version: 1.0

Algorithm Author: Loretta Auvil;

Please Input Job Name: (required)

Please select a collection for analysis:

Charles_Dickens_Novels

Please provide the number of tokens to be displayed in the tagcloud (default: 200):

(optional)

Please provide the number of topics to be created (default: 10):

(optional)

Examples on the Wall