An Introduction to Data Mining with Open-Source Technologies

Blake Galbreath
Washington State University, blake.galbreath@wsu.edu

Follow this and additional works at: https://pdxscholar.library.pdx.edu/onlinenorthwest

Let us know how access to this document benefits you.

https://pdxscholar.library.pdx.edu/onlinenorthwest/2018/presentations/8

This Presentation is brought to you for free and open access. It has been accepted for inclusion in Online Northwest by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Blake L. Galbreath
Core Services Librarian, WSU
An Introduction to Data Mining with Open-Source Technologies

OpenRefine · RapidMiner · Voyant Tools
Learning Outcomes

Learn how to grab relevant data from a website with OpenRefine

Learn how to use text processing operators in RapidMiner

Learn how to use various tools in Voyant to display data analysis
What is Data Mining?

A step in the process of knowledge discovery from data (KDD)

'(Semi) automated discovery of trends and patterns across very large datasets, usually for the purpose of decision making'

A proactive process that automatically searches data for new relationships and anomalies on which to base business decisions in order to gain competitive advantage

--Attempt to stay ahead of your competition by having more complete information to proactively make better-informed decisions.'
What is Text Mining?

A branch or a sibling of data mining; Also called ‘text data mining’

'Text mining is all about extracting patterns and associations previously unknown from large text databases' (Thuraisingham 1999:167).

Text mining is 'a way to examine a collection of documents and discover information not contained in any individual document in the collection' (Lucas 1999/2000:1).

Text mining 'performs various searching functions, linguistic analysis and categorizations' (Chen 2001:5,9).
OpenRefine
What is OpenRefine?

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Also: see my presentation from last year: Galbreath, Blake, "Using OpenRefine to Standardize and Augment Your Data" (2017). Online Northwest. 8.

http://pdxscholar.library.pdx.edu/onlinenorthwest/2017/schedule/8
Pulling Data from Website with OpenRefine

- Construct URLs
- Fetch Data
- Parse Data
Idea Exchange

- Primo Ideas

**Browses: add See Also and Usage info from authority headings**

Primo supports display of See references (not same as See Also) in browse headings. If those cross-references are in the bibliographic data ingested by Primo. There is no capacity to display See Also and Usage info from authority records, which are important so that users understand how and why certain terms, especially subjects, may be used.

We propose that See Also and Usage Info display in the Browse headings.

Example headings and See Also and Usage info that does NOT currently display in Primo:

- Blacks

  Search also under: subdivision Blacks under individual wars, e.g. World War, 1939-1945--Blacks; and headings beginning... more

3 comments - User Interface

**Add Urbadoc database to Primo Central index**

Urbadoc bibliographic database indexes European architectural journals and can enrich PCI with unique references.

It would be very useful to have this database searchable in Primo Central Index. Urbadoc is an OAI-PMH data provider. More info at: http://www.cnba.it/2017/09/28/novita-urbadoc/

Urbadoc URL: http://www.urbadoc.com

27 votes
Gather URLs

- Create Project
- Clipboard
- Paste data from clipboard here
Fetch Data

- Edit Column
- Add Column by Fetching URLs

<table>
<thead>
<tr>
<th>18 records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show as: rows records</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Facet</td>
</tr>
<tr>
<td>2. Text filter</td>
</tr>
<tr>
<td>3. Edit cells</td>
</tr>
<tr>
<td>4. Edit column</td>
</tr>
<tr>
<td>5. Transpose</td>
</tr>
<tr>
<td>6. Sort...</td>
</tr>
<tr>
<td>7. View</td>
</tr>
</tbody>
</table>

Add column based on this column...
Add column by fetching URLs...
Add columns from reconciled values...
Rename this column
Remove this column
Move column to beginning
Move column to end
Move column left
Move column right

Web page URLs:
- [Page 6](http://com/forums/308176-primo?page=6)
- [Page 7](http://com/forums/308176-primo?page=7)
- [Page 8](http://com/forums/308176-primo?page=8)
Parse Data

- Undo/Redo Tab
- Apply
- Paste
- Extracted JSON

```
[
  {
    "op": "core/column-addition-by-fetching-urls",
    "description": "Create column Fetch Html at index 1 by fetching URLs based on column Column 1 using:
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "newColumnName": "Fetch Html",
    "columnInsertIndex": 1,
    "baseColumnName": "Column 1",
    "urlExpression": "grel:,value",
    "onError": "set-to-blank",
    "delay": 500,
    "cacheResponses": true
  }
]
```
Admire Data

- And compare against original website data
What is RapidMiner?

A lightning fast unified data science platform.

A software platform for data science teams that unites data prep, machine learning, and predictive model deployment.

Is it really open-source? Actually it has moved to “business source.”
Analyze Sentiment: Document

Synopsis
Analyzes Sentiment of text.

Description
Extracting sentiment from a piece of text such as a tweet, a review or an article can provide us with valuable insight about the author's emotions and perspective: whether the tone is positive, neutral or negative, and whether the text is subjective (meaning it's reflecting the author's opinion) or objective (meaning it's expressing a fact).
Processing Operators

- Read Document
- Tokenize
- Sentiment Analysis
Analysis

- Polarity = Positive
- Polarity_confidence = .907
Social Media: Search Twitter

Synopsis
This operator searches for Twitter statuses.

Description
With the Search Twitter operator, you can specify a query and get Twitter statuses containing this query. The list of statuses contains additional data with context of the statuses. In the expert mode, you can specify additional search restrictions.
Processing Operators

- Search Twitter
- Analyze Sentiment
- Query = “trump -rt -http” (-retweets) (-links)
Analysis

- Sentiment = Positive
- Confidence = .855

@jeneps And don’t forget Trump’s broad shoulders that Pence loves so much.
Process Documents: WordList

Synopsis
Generates word vectors from a text object.

Description
This operator uses one single TextObject as input for generating a term vector. The resulting example set will hence consist of only one single example. This makes this operator especially useful for applying a model on one single text.
Processing Operators

- Retrieve
- Select Attributes
- Data to Documents
- Process Documents
- WordList to Data
- Store
Processing Operators (Sub-routine)

- Extract Content
- Tokenize
- Transform Cases
- Stopwords
- n-Grams
- Tokens
- Stem
Analysis

- Journal
- Effect
- Histori
- Educ
- Studi
- State
- ...

ExampleSet (39160 examples, 0 special attributes, 3 regular attributes)

<table>
<thead>
<tr>
<th>Row No.</th>
<th>word</th>
<th>total</th>
<th>in documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>19094</td>
<td>journal</td>
<td>1093</td>
<td>1029</td>
</tr>
<tr>
<td>10992</td>
<td>effect</td>
<td>700</td>
<td>695</td>
</tr>
<tr>
<td>15312</td>
<td>histori</td>
<td>612</td>
<td>582</td>
</tr>
<tr>
<td>10826</td>
<td>educ</td>
<td>598</td>
<td>545</td>
</tr>
<tr>
<td>33997</td>
<td>studi</td>
<td>581</td>
<td>572</td>
</tr>
<tr>
<td>33417</td>
<td>state</td>
<td>536</td>
<td>509</td>
</tr>
<tr>
<td>38403</td>
<td>women</td>
<td>493</td>
<td>464</td>
</tr>
<tr>
<td>30795</td>
<td>us</td>
<td>474</td>
<td>462</td>
</tr>
<tr>
<td>32552</td>
<td>social</td>
<td>462</td>
<td>442</td>
</tr>
<tr>
<td>29969</td>
<td>research</td>
<td>428</td>
<td>413</td>
</tr>
<tr>
<td>15840</td>
<td>health</td>
<td>427</td>
<td>387</td>
</tr>
<tr>
<td>6856</td>
<td>commun</td>
<td>363</td>
<td>333</td>
</tr>
<tr>
<td>21735</td>
<td>manag</td>
<td>363</td>
<td>336</td>
</tr>
<tr>
<td>33932</td>
<td>student</td>
<td>363</td>
<td>347</td>
</tr>
</tbody>
</table>
Voyant Tools
What is Voyant Tools?

A web-based text reading and analysis environment. It is a scholarly project that is designed to facilitate reading and interpretive practices for digital humanities students and scholars as well as for the general public. Possibilities:

- Study texts that you find on the web or texts that you have carefully edited and have on your computer.
- Add functionality to your online collections, journals, blogs or websites so others can see through your texts with analytical tools.
- Learn how computers-assisted analysis works.
Interface

- Drop in text, URL, upload file
- Reveal!
Multiple Visualizations

- Open
- Science
- Danielle
- Portland
Links: Keywords and Collocates

- Science:
  - fellowship
  - mozilla day
  - datarescue
  - hack
References


http://openrefine.org/

http://voyant-tools.org/docs/#!/guide/about

https://rapidminer.com/

https://rapidminer.com/blog/the-core-of-rapidminer-is-open-source/
Questions
Other Experiments

Retrieve Primo and Alma ideas from Ex Libris Idea Exchange

Determine the sentiment of these ideas

Analyze sentiment of ideas to see if there was any correlation between polarity and number of votes received
Retrieve Ideas from Idea Exchange

- Construct URLs
- Fetch Data
- Parse Data
Sentiment Analysis, Count: Primo

- Positive = 20
- Negative = 160
- Neutral = 175
Sentiment Analysis, Count: Alma

- Positive = 13
- Negative = 418
- Neutral = 568
Sentiment Analysis, Scatterplot: Primo

- Shows some separation in the Neutral and Negative categories
Sentiment Analysis, Scatterplot: Alma

- Shows some separation in the Negative and Positive categories
Correlation Matrix for + and - Polarities: Primo

- $|r| = .113$ for Polarity = Positive
- $|r| = .115$ for Polarity = Negative
Correlation Matrix for + and - Polarities: Alma

- $|r| = 0.283$ for Polarity = Positive
- $|r| = 0.094$ for Polarity = Negative
Voyant Tools: Cirrus

- Cirrus is a wordcloud of the most frequently occurring words in the corpus.
Voyant Tools: Links

- Collocates graph shows a network graph of higher frequency terms that appear in proximity.