

Mar 30th, 2:15 PM - 3:00 PM

# An Introduction to Data Mining with Open-Source Technologies

Blake Galbreath  
Washington State University, [blake.galbreath@wsu.edu](mailto: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.**

---

Galbreath, Blake, "An Introduction to Data Mining with Open-Source Technologies" (2018). *Online Northwest*. 8.

<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](mailto: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

65

votes

Vote

## 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

27

votes

Vote

## Add Urdadoc database to Primo Central index

Urdadoc 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.

Urdadoc is an OAI-PMH data provider. More info at: <http://www.cnba.it/2017/09/28/novita-urbadoc/>

Urdadoc URL: <http://www.urbadoc.com>

# Gather URLs

- Create Project
- Clipboard
- Paste data from clipboard here

	A
1	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=1">https://ideas.exlibrisgroup.com/forums/308176-primo?page=1</a>
2	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=2">https://ideas.exlibrisgroup.com/forums/308176-primo?page=2</a>
3	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=3">https://ideas.exlibrisgroup.com/forums/308176-primo?page=3</a>
4	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=4">https://ideas.exlibrisgroup.com/forums/308176-primo?page=4</a>
5	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=5">https://ideas.exlibrisgroup.com/forums/308176-primo?page=5</a>
6	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=6">https://ideas.exlibrisgroup.com/forums/308176-primo?page=6</a>
7	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=7">https://ideas.exlibrisgroup.com/forums/308176-primo?page=7</a>
8	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=8">https://ideas.exlibrisgroup.com/forums/308176-primo?page=8</a>
9	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=9">https://ideas.exlibrisgroup.com/forums/308176-primo?page=9</a>
10	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=10">https://ideas.exlibrisgroup.com/forums/308176-primo?page=10</a>
11	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=11">https://ideas.exlibrisgroup.com/forums/308176-primo?page=11</a>
12	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=12">https://ideas.exlibrisgroup.com/forums/308176-primo?page=12</a>
13	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=13">https://ideas.exlibrisgroup.com/forums/308176-primo?page=13</a>
14	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=14">https://ideas.exlibrisgroup.com/forums/308176-primo?page=14</a>
15	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=15">https://ideas.exlibrisgroup.com/forums/308176-primo?page=15</a>
16	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=16">https://ideas.exlibrisgroup.com/forums/308176-primo?page=16</a>
17	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=17">https://ideas.exlibrisgroup.com/forums/308176-primo?page=17</a>
18	<a href="https://ideas.exlibrisgroup.com/forums/308176-primo?page=18">https://ideas.exlibrisgroup.com/forums/308176-primo?page=18</a>

# Fetch Data

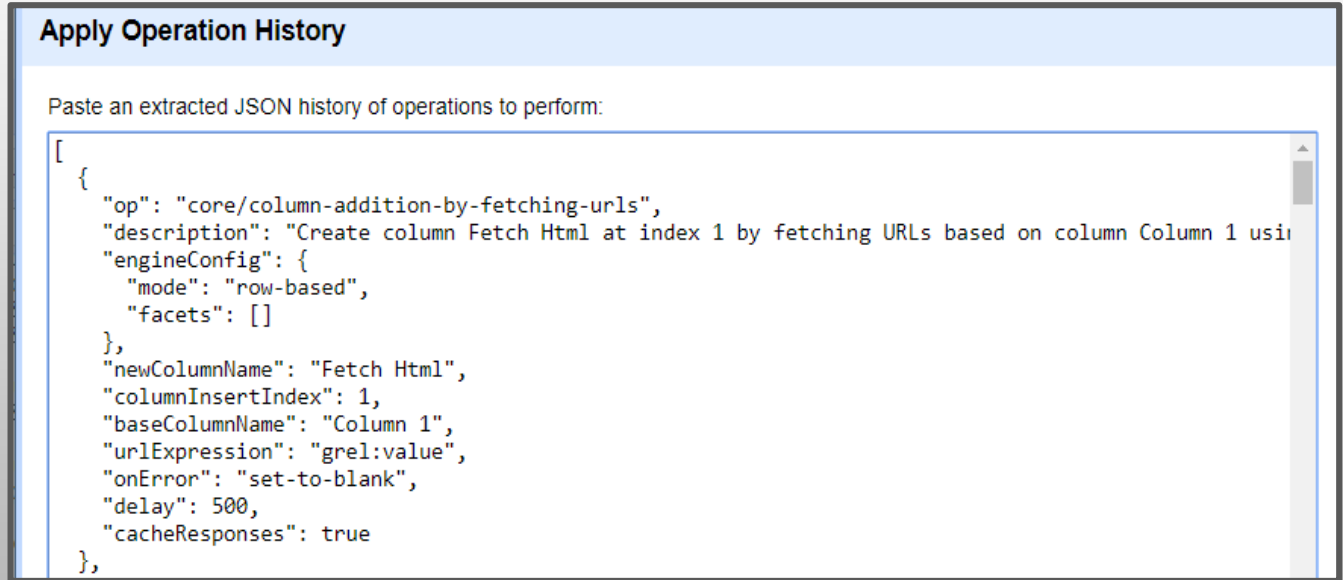
- Edit Column
- Add Column by Fetching URLs

The screenshot shows a data table with 18 records. A context menu is open over the first column, titled 'Column 1'. The menu options include: 'Add column based on this column...', 'Add column by fetching URLs...' (highlighted), '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', 'Facet', 'Text filter', 'Edit cells', 'Edit column' (highlighted), 'Transpose', 'Sort...', and 'View'. The table header shows 'Permalink' and '18 records'. The table body has columns for 'All' and 'Column 1', with rows numbered 1 through 8. The 'Edit column' option is highlighted, and the 'Add column by fetching URLs...' option is also highlighted.

Permalink	
<b>18 records</b>	
Show as: rows records	Show
▼ All	▼ Column 1
☆	1.
☆	2.
☆	3.
☆	4.
☆	5.
☆	6.
☆	7.
☆	8.

# Parse Data

- Undo/  
Redo Tab
- Apply
- Paste  
Extracted  
JSON



The screenshot shows a dialog box titled "Apply Operation History" with a light blue header. Below the header, there is a text area containing the instruction "Paste an extracted JSON history of operations to perform:". Below this instruction is a code editor with a white background and a vertical scrollbar on the right. The code editor contains a JSON array with one object. The object has the following properties: "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" (row-based mode, no facets), "newColumnName" (Fetch Html), "columnInsertIndex" (1), "baseColumnName" (Column 1), "urlExpression" (grel:value), "onError" (set-to-blank), "delay" (500), and "cacheResponses" (true).

```
[
  {
    "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
  },
]
```



**RapidMiner**

# 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

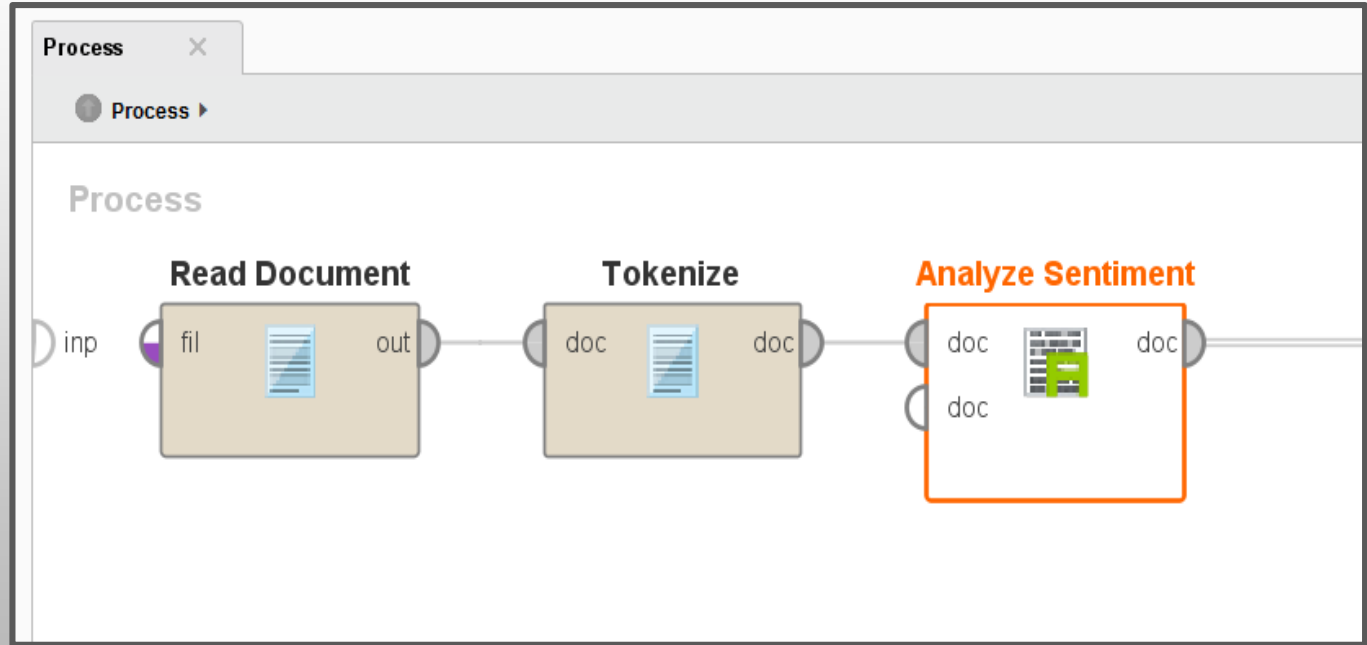
Analyzes Sentiment of text.

## 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



The screenshot shows a window titled "IOObjectCollection (Analyze Sentiment (2))". The main area contains a text document with a paragraph of Russian text. The text is color-coded, with words in red and blue. The text is a translation of a passage from "The Master and Margarita" by Mikhail Bulgakov, describing a scene at a kiosk in Moscow.

file | \bulgako...

file_type	txt
path	C:\Users...
date	Feb 23, ...
size	806232
polarity	positive
subjectivity	unknown
polarity_...	0.907
subjectiv...	0

# Social Media: Search Twitter

## Synopsis

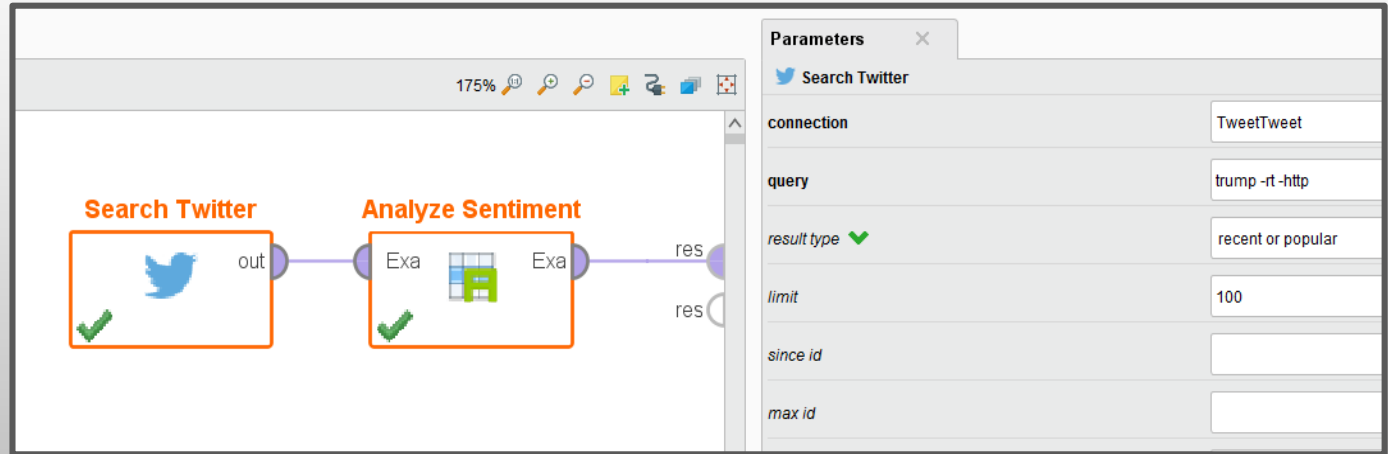
This operator searches for Twitter statuses.

## 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

ExampleSet (100 examples, 5 special attributes, 11 regular attributes)

Row No.	polarity ↓	polarity_con...	subjectivity	Text	From-User
15	positive	0.657	subjective	Congratulations, Mr. Clint. You will be a great asset to the Trump train <a href="https://t.co/6h5n34x0CY">https://t.co/6h5n34x0CY</a>	David Molina
33	positive	0.771	subjective	Does anyone have the Clip of Trump hyping the market asking	Gaslighting A Nation
36	positive	0.908	subjective	@WandaWstewart @TruthMatters13 @UG0TTRUMPED @girl4_trump @busylizzie48 @joesmomlover @Brenderm @Gma...	Rightwingmadman
38	positive	0.737	subjective	@washingtonpost I don't wish these particular voters ill. Trump understands what makes a reality show work: he encourag...	Annie

**@jeneps And don't forget Trump's broad shoulders that Pence loves so much.**

6	neutral	0.585	subjective	Liberals is this what you truly want and believe in? You may be in a position where this won't affect you, but think about your ...	UCan'tHandleTheTruth
9	neutral	0.430	subjective	All you hard core republicans and conservatives like me , we need to elect new fresh blood officials that support our boss D...	Edwin Lopez
10	neutral	0.665	subjective	@MingGao26 What did China do to make this man baby so mad? He acts like his crush is being mean to him...	Lunatic Raven

# Process Documents: WordList

## Synopsis

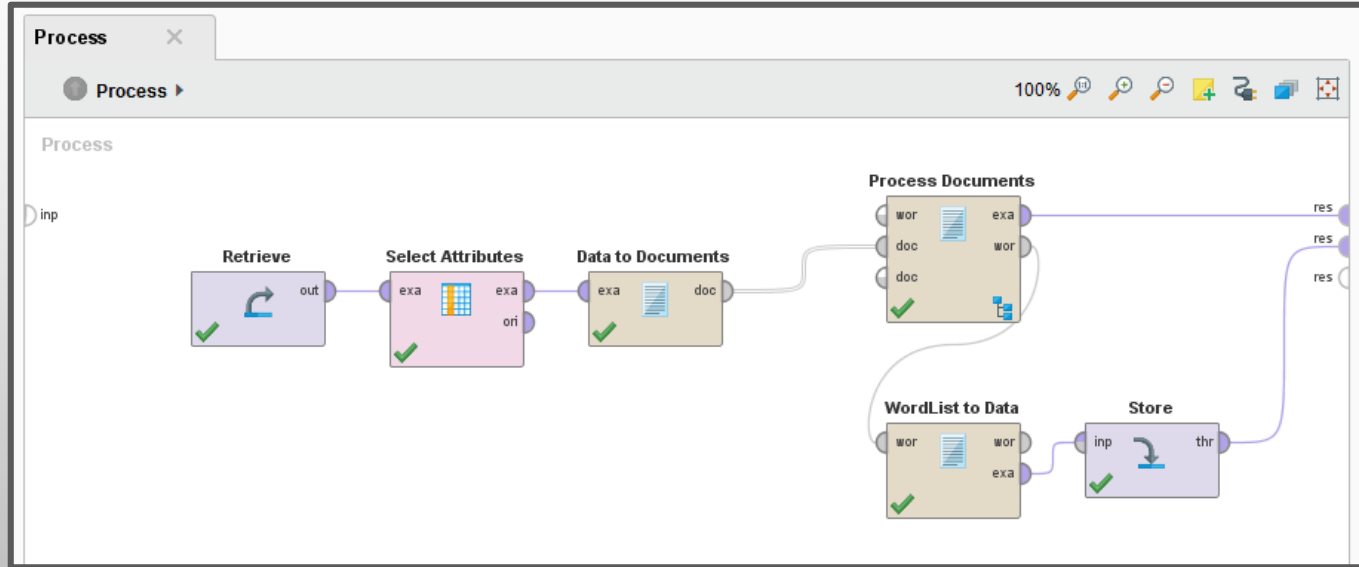
Generates word vectors from a text object.

## Description

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

# Processing Operators

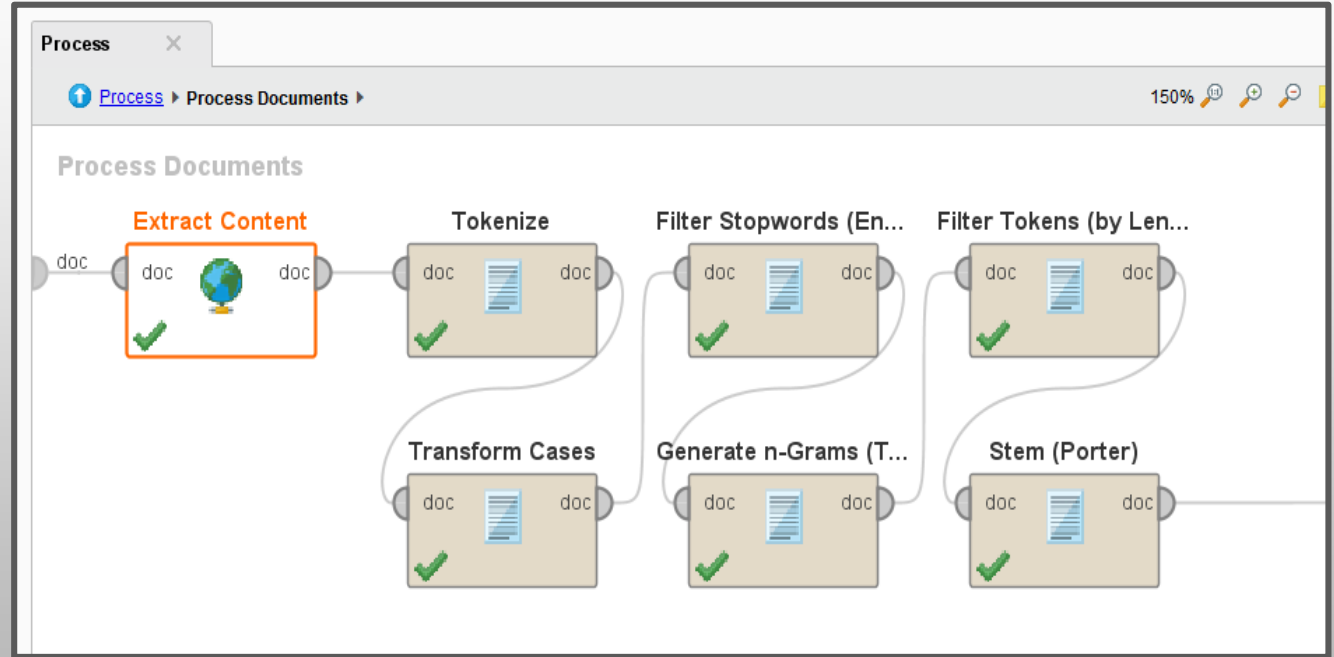
- Retrieve
- Select Attributes
- Data to Documents
- Process Documents
- WordList to Data
- Store





# Processing Operators (Sub-routine)

- Extract Content
- Tokenize
- Filter Stopwords (En...)
- Filter Tokens (by Len...)
- Transform Cases
- Generate n-Grams (T...)
- Stem (Porter)



# Analysis

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

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

Row No.	word	total ↓	in documents
19094	journal	1093	1029
10992	effect	706	695
16312	histori	612	582
10826	educ	598	545
33997	studi	581	572
33417	state	536	509
38403	women	493	464
36795	us	474	462
32552	social	462	442
29969	research	428	413
15849	health	427	387
6856	commun	363	333
21735	manag	363	336
33932	student	363	347

# **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!



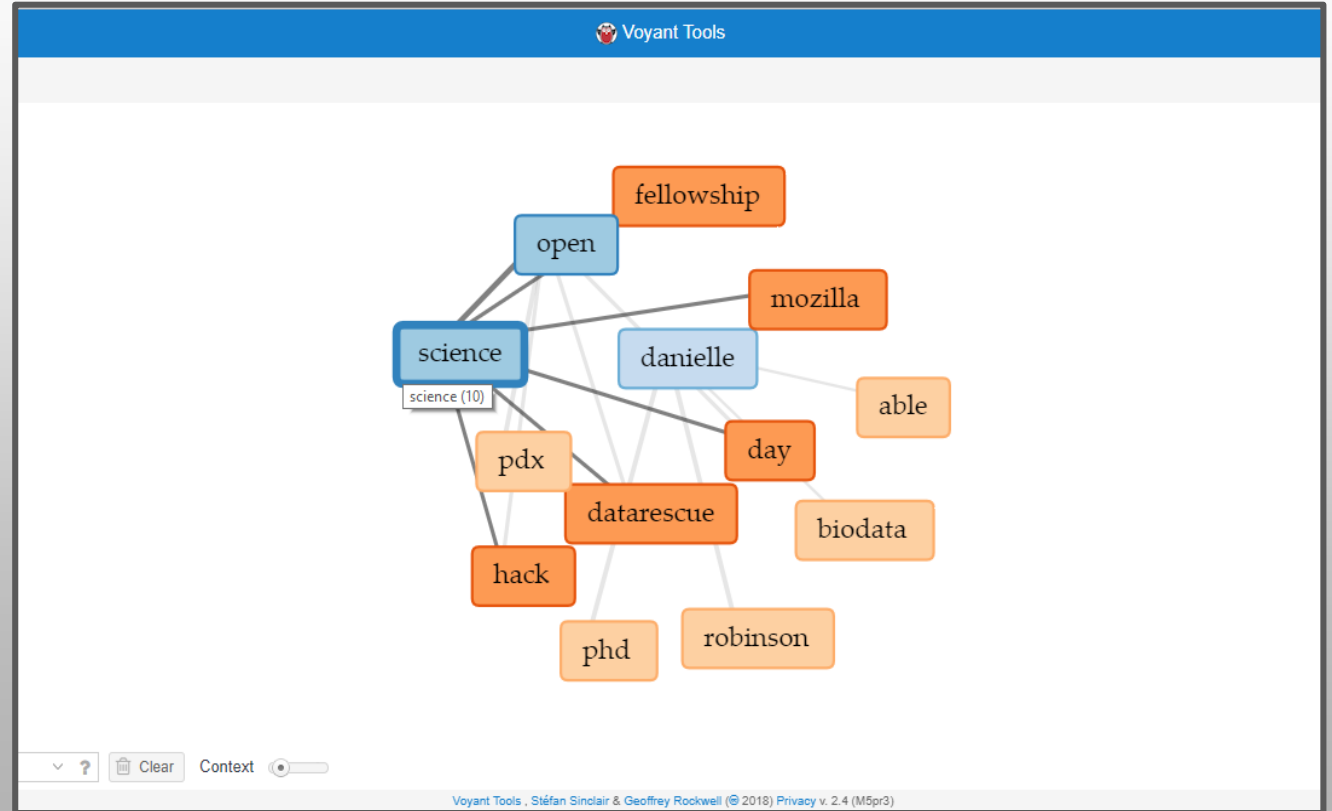
Add Texts 🔍 🌙 ?

<https://onlinenorthwest.org/>



# Links: Keywords and Collocates

- Science:  
fellowship  
mozilla day  
datarescue  
hack



# References

Kroeze, J., Matthee, M., & Bothma, T. (2004). Differentiating between data-mining and text-mining terminology. 6(4), 297-306.

<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

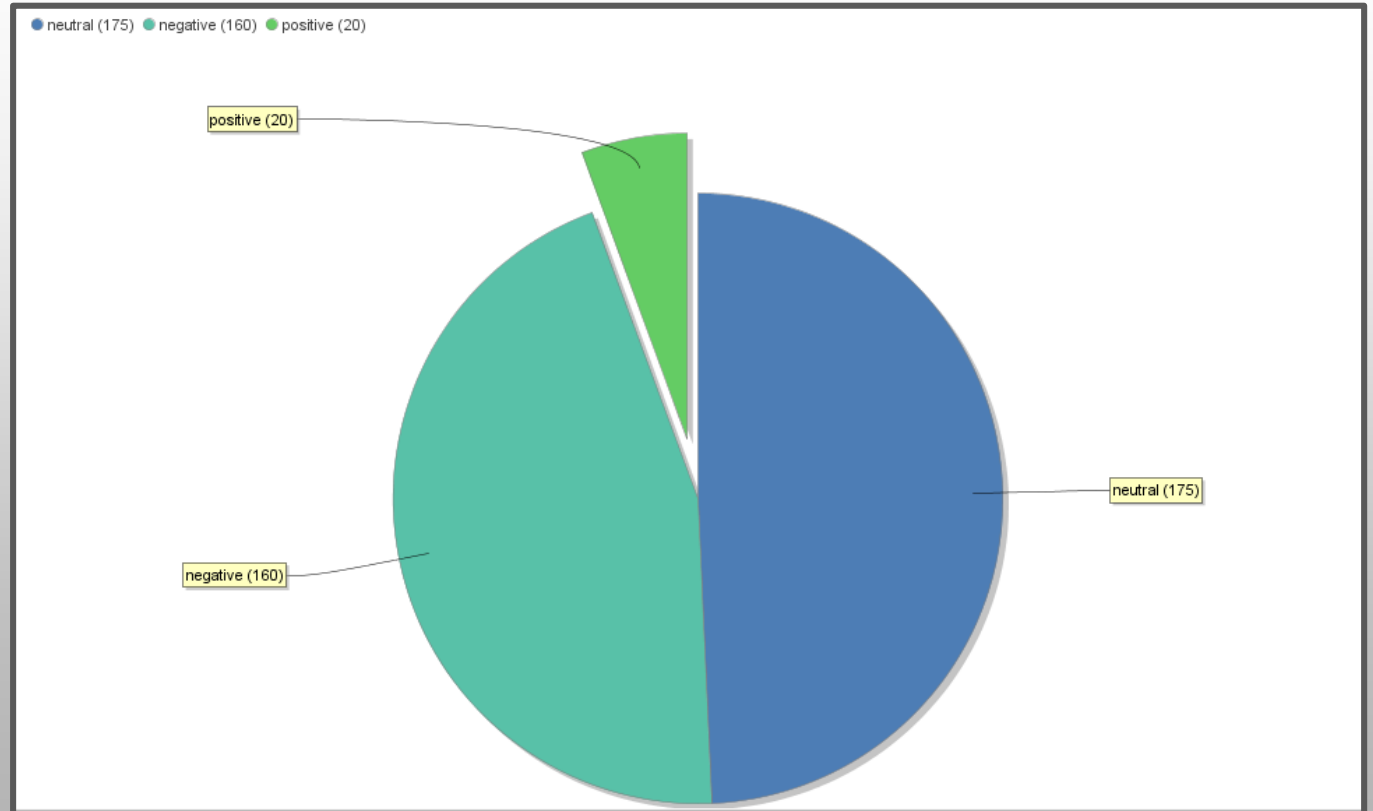


# Sentiment Analysis, Count: Primo

- Positive = 20

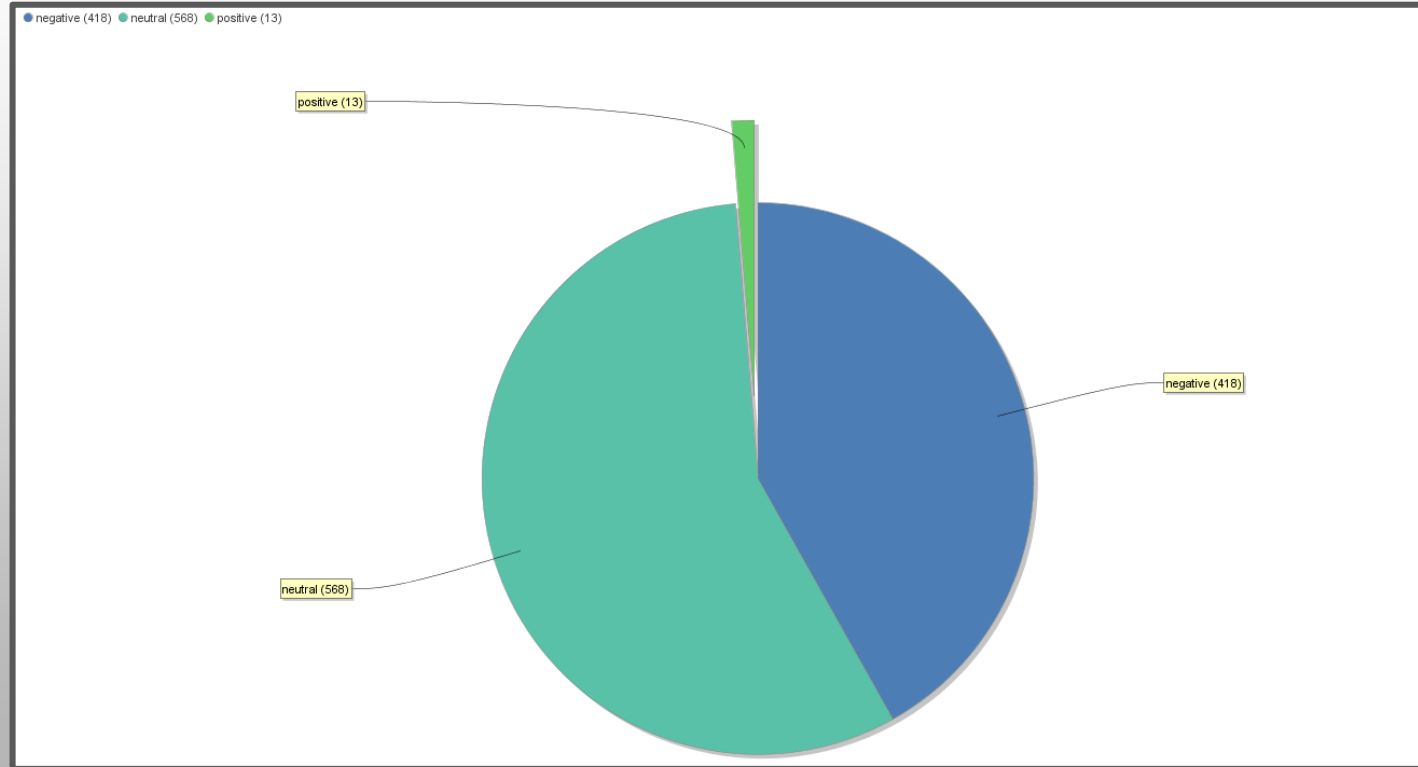
- Negative = 160

- Neutral = 175



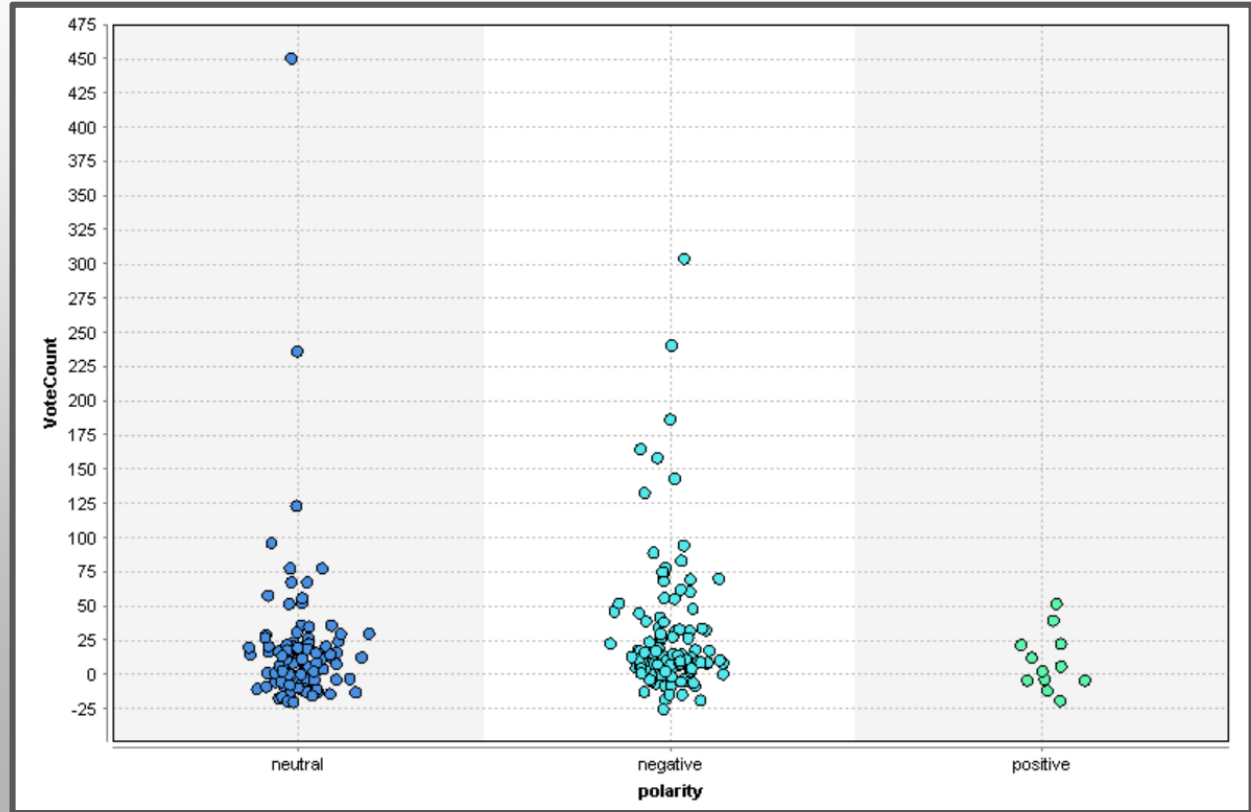
# Sentiment Analysis, Count: Alma

- Positive = 13
- Negative = 418
- Neutral = 568



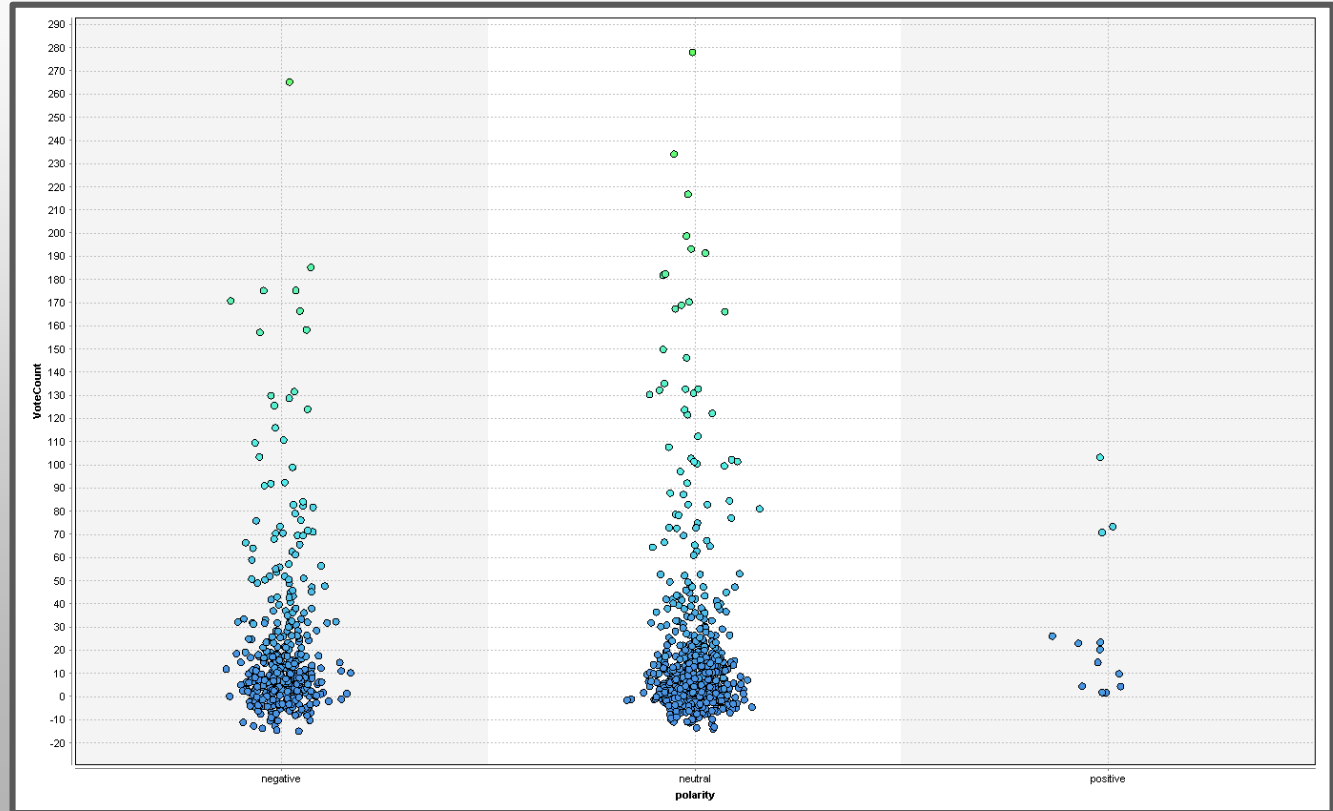
# Sentiment Analysis, Scatterplot: Primo

- 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

Attributes	PageUrl	TitleUrl	Title	Description	VoteCount	polarity	subjectivity	polarity_confidence	subjectivity_confidence
PageUrl	1	?	?	?	?	?	?	?	?
TitleUrl	?	1	?	?	?	?	?	?	?
Title	?	?	1	?	?	?	?	?	?
Description	?	?	?	1	?	?	?	?	?
VoteCount	?	?	?	?	1	?	?	0.113	0.121
polarity	?	?	?	?	?	1	?	?	?
subjectivity	?	?	?	?	?	?	1	?	?
polarity_confidence	?	?	?	?	0.113	?	?	1	0.240
subjectivity_confidence	?	?	?	?	0.121	?	?	0.240	1

Attri	Page	TitleUrl	Title	Description	VoteCount	polarity	subjectivity	polarity_confidence	subjectivity_confidence
TitleUrl	?	1	?	?	?	?	?	?	?
Title	?	?	1	?	?	?	?	?	?
Description	?	?	?	1	?	?	?	?	?
VoteCount	?	?	?	?	1	?	?	0.115	0.014
polarity	?	?	?	?	?	1	?	?	?
subjectivity	?	?	?	?	?	?	1	?	?
polarity_confidence	?	?	?	?	0.115	?	?	1	-0.027
subjectivity_confidence	?	?	?	?	0.014	?	?	-0.027	1



# Correlation Matrix for + and - Polarities: Alma

- $|r| = .283$   
for  
Polarity =  
Positive
- $|r| = .094$   
for  
Polarity =  
Negative

Attributes	TitleUrl	Title	Description	VoteCount	polarity	subjectivity	polarity_confidence	subjectivity_confidence
TitleUrl	1	?	?	?	?	?	?	?
Title	?	1	?	?	?	?	?	?
Description	?	?	1	?	?	?	?	?
VoteCount	?	?	?	1	?	?	-0.283	?
polarity	?	?	?	?	1	?	?	?
subjectivity	?	?	?	?	?	1	?	?
polarity_confidence	?	?	?	-0.283	?	?	1	?
subjectivity_confidence	?	?	?	?	?	?	?	1

Attributes	TitleUrl	Title	Description	VoteCount	polarity	subjectivity	polarity_confidence	subjectivity_confidence
TitleUrl	1	?	?	?	?	?	?	?
Title	?	1	?	?	?	?	?	?
Description	?	?	1	?	?	?	?	?
VoteCount	?	?	?	1	?	?	0.094	0.035
polarity	?	?	?	?	1	?	?	?
subjectivity	?	?	?	?	?	1	?	?
polarity_confidence	?	?	?	0.094	?	?	1	0.034
subjectivity_confidence	?	?	?	0.035	?	?	0.034	1



