Portland State University PDXScholar

Online Northwest

Online Northwest 2018

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

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.

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	Browses: add See Also and Usage info from authority headings
votes	Primo supports display of See references (not same as See Also) in browse headings, if those cross-reference are in the bibliographic data ingested by Primo. There is no capacity to display See Also and Usage info from
Vote	authority records, which are important so that users understand how and why certain terms, especially subject 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, 1939-1945Blacks; and heading: beginning more
	3 comments · User Interface
27	Add Urbadoc database to Primo Central index
Z/ votes	Urbadoc bibliographic database indexes european architectural journals and can enrich PCI with unique
VOICS	references.
Vote	It would be very useful to have this database searchable in Primo Central Index.
	Urbadoc is an OAI-PMH data provider. More info at: <u>http://www.cnba.it/2017/09/28/novita-urbadoc/</u>

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

Gather URLs

- Create
 Project
- Clipboard

Paste
 data from
 clipboard
 here

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

Fetch Data

EditColumn

Add
 Column
 by
 Fetching
 URLs

[De sere e lise la		Add column based on this column	n Expo
Ľ	Permalink		Add column by fetching URLs	
L	18 records		Add columns from reconciled values	
	Show as: rows	records Sho	Rename this column	ious 1 - 18
Н		Column 1	Remove this column	
ł	☆ 🖓 1. 🛛 Fa	acet 🕨	Move column to beginning	
L		ext filter	Move column to end	
H	☆ 덕 3. ☆ 덕 4.	dit cells	Move column left	
L		dit column 🔹 🕨	Move column right	
L		anspose 🕨 🕨	.com/forums/308176-primo?page=6	
L		ort	.com/forums/308176-primo?page=7	
L	🛣 ק 8. Vi	ew 🕨	.com/forums/308176-primo?page=8	

Parse Data

 Undo/ Redo Tab

• Apply

Paste
 Extracted
 JSON

```
Apply Operation History
Paste an extracted JSON history of operations to perform:
     "op": "core/column-addition-by-fetching-urls",
     "description": "Create column Fetch Html at index 1 by fetching URLs based on column Column 1 usi
     "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

355 record	ds		Extensions
Show as: rows	s records Show: 5 10 25 50 reco	ords	« first < previous 1 - 25 next > last
VoteCount	TitleUrl	💌 Title	Description
55	http://ideas.exlibrisgroup.com/forums/308176- primo/suggestions/33137959-browses-add- see-also-and-usage-info-from-authorit	Browses: add See Also and Usage info from authority headings	Primo supports display of See references (not same as See Also) in br those cross-references are in the bibliographic data ingested by Primo. capacity to display See Also and Usage info from authority records, wh so that users understand how and why certain terms, especially subjec We propose that See Also and Usage info display in the Browse headin headings and See Also and Usage info that does NOT currently display Blacks: Search also under: subdivision Blacks under individual wars, e. 1939-1945Blacks; and headings beginning with the word Black Here a works on blacks as an element in the population. Theoretical works dis race from an anthropological point of view are entered under Black race this enhancement could be achieved through application of an API that Library of Congress systems. Display of narrow/broader terms would at valuable. Primo supports display See references (not same as See A headings, if those cross-references are in the bibliographic data ingeste There is no capacity to display See Also and Usage info from authority are important so that users understand how and why certain terms, esp may be used. We propose that See Also and Usage info display in the headings. Example headings and See Also and Usage info that does N display in Primo: Blacks: Search also under: subdivision Blacks under i e.g. World War, 1939-1945Blacks; and headings beginning
63	http://ideas.exlibrisgroup.com/forums/308176- primo/suggestions/32978050-search-slot- scope-profile-usage-in-primo-analytics	Search Slot/Scope/Profile Usage in Primo	We would like to be able to report the total number of searches that hav performed in each search/slot/scope/profile that we have configured in the only report in Analytics that has Search Scope search information is Depute Deput. This information will be required instruction be

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

Process X Process >				
Process Read	d Document	Tokenize	Analyze Sentiment	

Analysis

- Polarity = Positive
- Polarity_confidence
 - = .907

🗙 🧥 IOObjectCollection (Analyze Sentiment (2))

 \wedge The Master and Margarita Mikhail Bulgakov Translated from the russian by Michael Glenny Published by Collins and Harvill file \bulgako... Press London BOOK ONE Never Talk to Strangers Pontius Pilate The Seventh Proof The Pursuit The Affair at Griboyedov file_type txt Schizophrenia The Haunted Flat A Duel between Professor and Poet Korovievis Tricks News from Yalta The Two Ivans Black Magic Revealed Enter the Hero Saved by Cock Crow The Dream of Nikanor Ivanovich The Execution A Day of Anxiety path C:\Users Unwelcome Visitors BOOK TWO Margarita Azazello s Cream The Flight By Candlelight Satan s Rout The Master is Released date Feb 23. .. How the Procurator Tried to Save Judas of Karioth The Burial The Last of Flat No The Final Adventure of Koroviev and Behemoth The Fate of the Master and Margarita is Decided Time to Go On Sparrow Hills Absolution and Eternal Refuge Epilogue Sav at 806232 size last who art thou That Power I serve Which wills forever evil Yet does forever good Goethe Faust BOOK ONE Never Talk to polarity positive Strangers At the sunset hour of one warm spring day two men were to be seen at Patriarch's Ponds The first of them aged about forty dressed in a grevish summer suit was short dark haired well fed and bald He carried his decorous pork pie hat by subjectivity unknown the brim and his neatly shaven face was embellished by black hornrimmed spectacles of preternatural dimensions The other a 0.907 broad shouldered young man with curly reddish hair and a check cap pushed back to the nape of his neck was wearing a tartan polarity_.. shirt chewed white trousers and black sneakers The first was none other than Mikhail Alexandrovich Berlioz editor of a subjectiv... 0 highbrow literary magazine and chairman of the management cofnmittee of one of the biggest Moscow literary clubs known by its abbreviation as massolit his young companion was the poet Ivan Nikolavich Poniryov who wrote under the pseudonym of Bezdomny Reaching the shade of the budding lime trees the two writers went straight to a gaily painted kiosk labelled Beer and Minerals There was an oddness about that terrible day in May which is worth recording not only at the kiosk but along the whole avenue parallel to Malaya Bronnaya Street there was not a person to be seen it was the hour of the day when people feel too exhausted to breathe when Moscow glows in a dry haze as the sun disappears behind the Sadovava Boulevard yet no one had come out for a walk under the limes no one was sitting on a bench the avenue was empty A glass of lemonade please said Berlioz There isn't any replied the woman in the kiosk For some reason the request seemed to offend her Got any beer enquired Bezdomny in a hoarse voice Beer's being delivered later this evening said the woman Well what have you got asked Berlioz Apricot juice only it s warm was the answer All right let s have some The apricot juice produced a rich yellow froth making the air smell like a hairdresser's After drinking it the two writers immediately began to hiccup They paid and sat down on a bench facing the pond their backs to Bronnaya Street Then occurred the second oddness which affected Berlioz alone He suddenly stopped hiccuping his heart thumped and for a moment vanished then returned but with a blunt needle sticking into it In addition Berlioz was seized by a fear that was groundless but so powerful that he had an immediate impulse to run away

X

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)

	Para	rameters X	
175% 🔑 🔑 🛃 🥁 💣 🖯	3	Search Twitter	
	^ conn	nection	TweetTweet
Secreb Twitter Analyze Continent	quer	ry	trump -rt -http
Search Twitter Analyze Sentiment	resul	ılt type 💙	recent or popular
	limit	t	100
	since	e id	
	max	c id	

Analysis

- Sentiment
 = Positive
- Confidence

= .855

ExampleSet (100	examples, 5 speci	al attributes, 11 reg	gular attributes)		
Row No.	polarity \downarrow	polarity_con	subjectivity	Text	From-User
15	positive	0.657	subjective	Congratulations, Mr. Clint. You will be a great asset to the Trump train https://t.co/6h5n34x0CY	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

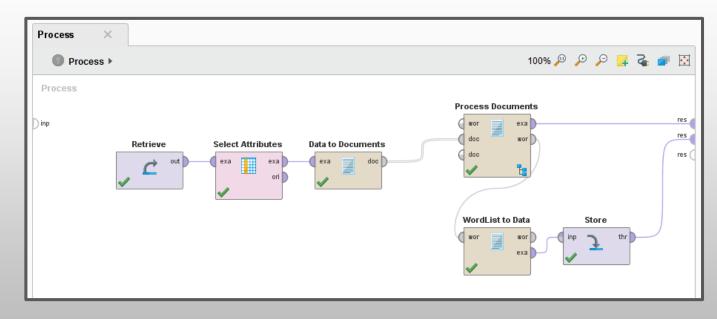
Generates word vectors from a text object.

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.

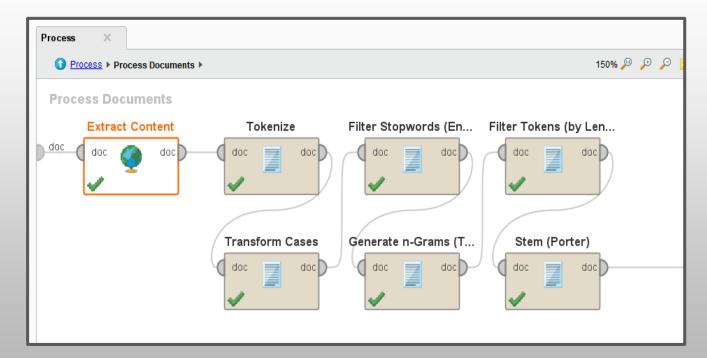
Processing Operators

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



Processing Operators (Sub-routine)

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



Analysis

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

• ...

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/	
Den Upload	Reveal
Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (@ 2018) Privacy v. 2.4 (M5pr3)	

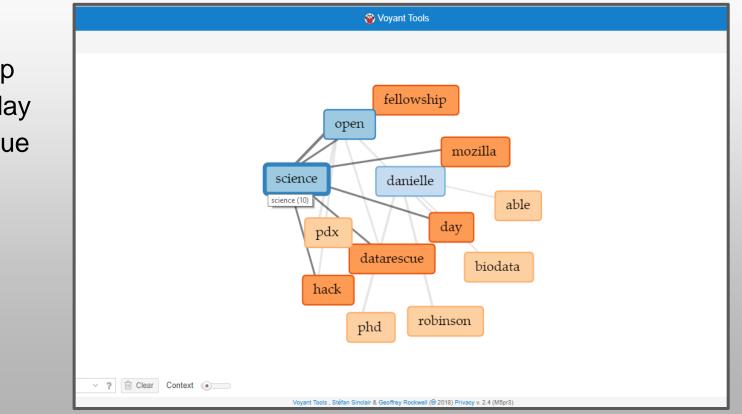
Multiple Visualizations

- Open
- Science
- Danielle
- Portland

Cirrus III Terms 🔩 Links ?	Reader O TermsBerry			?	🛃 Trends 🗰 D	ocument	t Terms
including today working working brogram working brogram projects projects base base base base base base base bas	Online Northwest 2018 Keynote Spec A conference on libraries, technology, and cul Toggle navigation Home About Registration Proposals Travel Pri Registration is now open for Online Northwest We are delighted to announce that registration Please Join us this year on March 30th, 2018 a that includes presentations on privacy. OERs, instruction. Register today! Registration is \$135. It includes morning and a The full program and schedule are available a	ture 2018! is nov it Porti user to	rs Archive v open for Online Northwest 2018 A State University for a program esting, and tools for effective	Ť	Bave Leducation	• sci	ence danielle portland library
Terms: ()						~ ?	Frequencies V Segments:
Summary Documents III Phrases	?		Contexts @ Bubblelines	⊞ (Correlations		
This corpus has 1 document with 379 total words and 200 unique word forms. Created now.			Document		Left	Term	Right
Vocabulary Density: 0.528					Archive Registration is now	open	for Online Northwest 2018! We
Average Words Per Sentence: 19.9				annour	nce that registration is now	open	for Online Northwest 2018! Please
Most frequent words in the corpus: open (11); science (10); danielle (5); portland (5); library (4)			1) Onlin		C. Robinson, PhD, is an	open	
and a second state and a second state and s			1) Onlin		in the power of inclusive	open	
		(H)	1) Onlin	sh	e continues to develop the	open	science community at OHSU, in

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

Retrieve Ideas from Idea Exchange

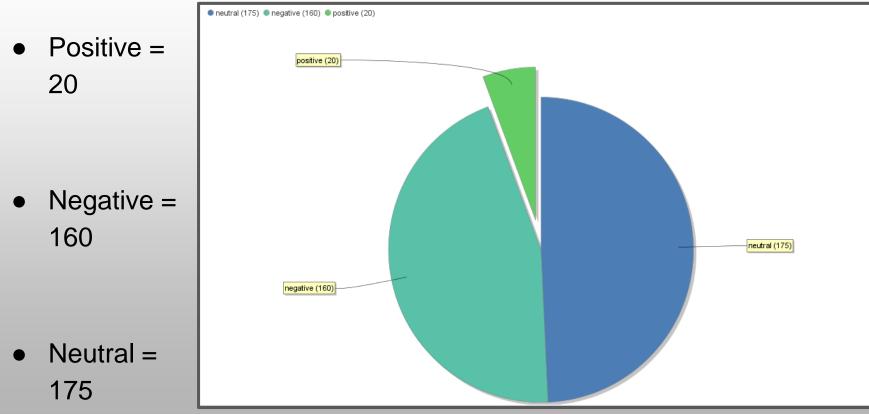
 Construct URLs

Fetch
 Data

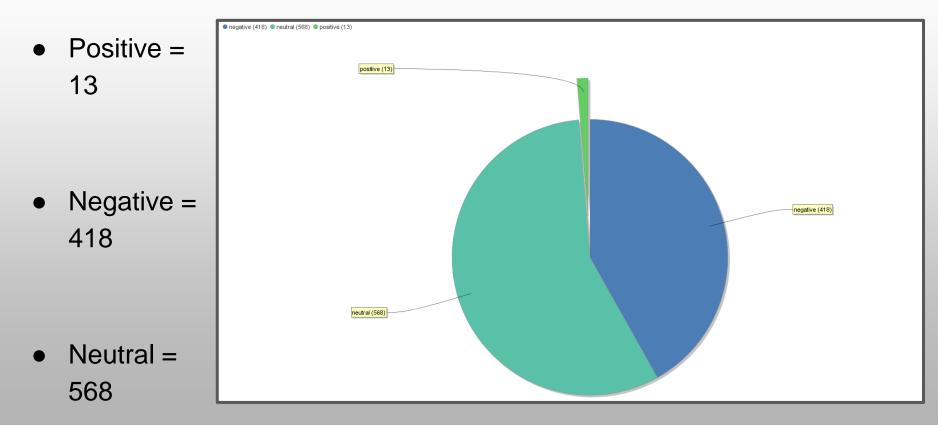
Parse Data

355 record	ds		Extensions
Show as: rows	s records Show: 5 10 25 50 reco	rds	« first < previous 1 - 25 next > last
VoteCount	▼ TitleUrl	Title	Description
55	http://ideas.exlibrisgroup.com/forums/308176- primo/suggestions/33137959-browses-add- see-also-and-usage-info-from-authorit	Browses: add See Also and Usage info from authority headings	Primo supports display of See references (not same as See Also) in brithose cross-references are in the bibliographic data ingested by Primo. capacity to display See Also and Usage info from authority records, whis of that users understand how and why certain terms, especially subject. We propose that See Also and Usage info display in the Browse headin headings and See Also and Usage info that does NOT currently display Blacks: Search also under: subdivision Blacks under individual wars, e. 1939-1945Blacks; and headings beginning with the word Black Here a works on blacks as an element in the population. Theoretical works dis race from an anthropological point of view are entered under Black race this enhancement could be achieved through application of an API that Library of Congress systems. Display of See references (not same as See A headings, if those cross-references are in the bibliographic data ingeste There is no capacity to display See Also and Usage info from authority are important so that users understand how and why certain terms, esp. may be used. We propose that See Also and Usage info that does N display in Primo: Blacks: Search also under: subdivision Blacks under i e.g. World War, 1939-1945Blacks; and headings beginning
63	http://ideas.exlibrisgroup.com/forums/308176- primo/suggestions/32978050-search-slot- scope-profile-usage-in-primo-analytics	Search Slot/Scope/Profile Usage in Primo	We would like to be able to report the total number of searches that hav performed in each search/slot/scope/profile that we have configured in the only report in Analytics that has Search Scope search information is Deputie Deput. This information will provide involveble information as

Sentiment Analysis, Count: Primo

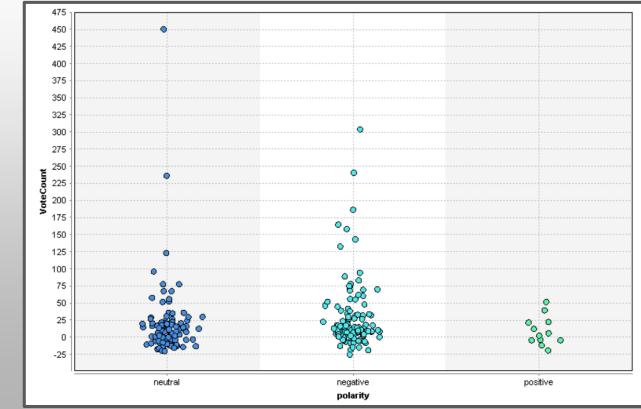


Sentiment Analysis, Count: Alma



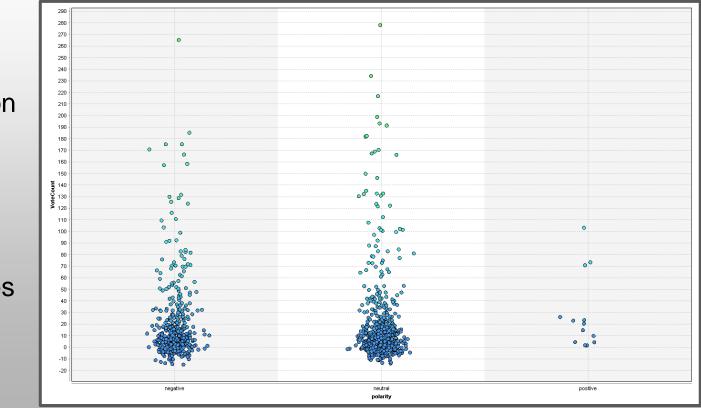
Sentiment Analysis, Scatterplot: Primo

• 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		Page	Url	TitleL	Irl	Title		Descriptio	n	VoteCou	nt	polar	rity	subjectiv	/ity	polarity_confidence	subjectivity_confidence
	PageUrl TitleUrl Title		1		?		?		?		?		?		?		?	?
			?		1		?		?		?		?		?		?	?
			?		?		1		?		?		?		?		?	?
	Description		?		?		?		1		?		?		?		?	?
	VoteCount		?		?		?		?		1	?			?		0.113	0.121
	polarity		?		?		?		?		?		1		?		?	?
	subjectivity		?		?	?			?		?		?		1		?	?
ttril	polarity_confidence		nce ? ?		?		?		?		0.113 ?		?	?			1	0.240
age	subjectivity_confide	ence	??			?		?		0.121		?		?		0.240	1	
itleUrl	1	?		1		?		?		?		?		?		?		?
itle		?		?		1		?		?		?		?		?		?
escri	ption	?		?		?		1		?		?	? ?		?			?
oteCo	ount	?		?		?		?		1		?		?		0.11	5	0.014
olarity	plarity ?			?		?		?		?		1		?		?		?
subjectivity		?		?		?		?		?		?		1		?		?
olarity	_confidence	?		?		?		?		0.11	15	?	?		1			-0.027
uhied	tivity_confidence	?		?		?		?		0.01	14	?		?	-0.027		7	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 Title Description		1	?		?		?		?	???			?		?	
			?	1		?		?		?	?	?		?		?	
			?	?		1		?		?	?		?		?		
	VoteCount		?	?		?		1		??		?	-0.283			?	
	polarity		?	?		?		?		1	1 ?			?		?	
	subjectivity polarity_confidence subjectivity_confidence		?	?		? ?		? -0.283		? ?		1 ?		? 1 ?		? ? 1	
Attri			?	?													
Title			?	?		?		?		? ?		?					
Title		?	1		?		?		?		?		?		?		
Description		?	?		1		?		?		?		?		?		
/oteCount		?	?		?	1		?		?		0.094		1	0.035		
polarity		?	?	?		?			1		?	?			?		
subjectivity		?	?		?		?		?		1	?			?		
polarity_confidence		?	?		?		0.094		?		?		1		0.034		
subjectivity_confidence		?	?		?		0.035		?	?		0.034		1	1		

Voyant Tools: Cirrus

• Cirrus is a wordcloud of the most frequently occuring words in the corpus



Voyant Tools: Links

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

