

ANNIC

The art and craft of JAPE rules

- You know by now how to write some not so simple JAPE rules
- The question is: how do you design them? How do you find patterns which are frequent in your test corpus?
- Given a dataset of tweets, how can you be sure that the JAPE LHS pattern you are about to implement doesn't do more harm than good?

ANNIC: Annotations in Context

- Motivation
 - Need for a corpus analysis tool
 - Useful for authoring of IE patterns for rules
- ANNIC is an IR engine that can search over:
 - Document Content (words)
 - Metadata (Annotation types, features and values)
 - for example: `Person.gender=="male"`

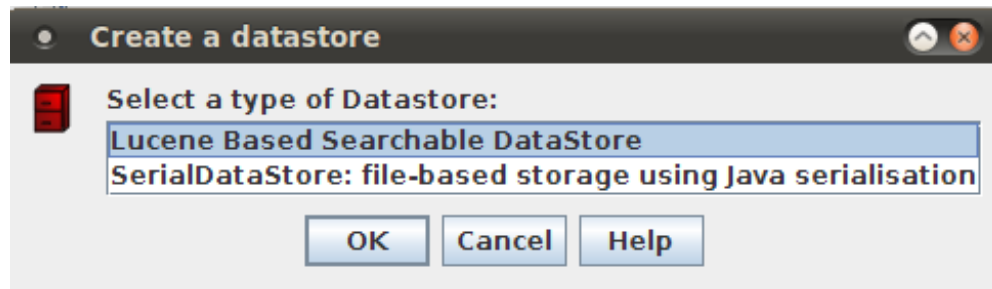
- is based on Apache Lucene technology.
- can index any document supported by GATE
- is integrated in GATE as a Searchable Datastore
- has an advanced GUI that provides:
 - view of annotation mark-ups over the matched patterns
 - interactive way of developing new patterns
 - annotation statistics

How does it work?

- Initialization
 - where to store
 - what to index and what to exclude
 - context boundary (e.g. restricted within sentence or paragraph boundaries)
- Index actions linked with Datastore actions
 - when document is saved, index or re-index if already indexed
 - when document is deleted, delete it from the index

Hands-on 1: creating a datastore

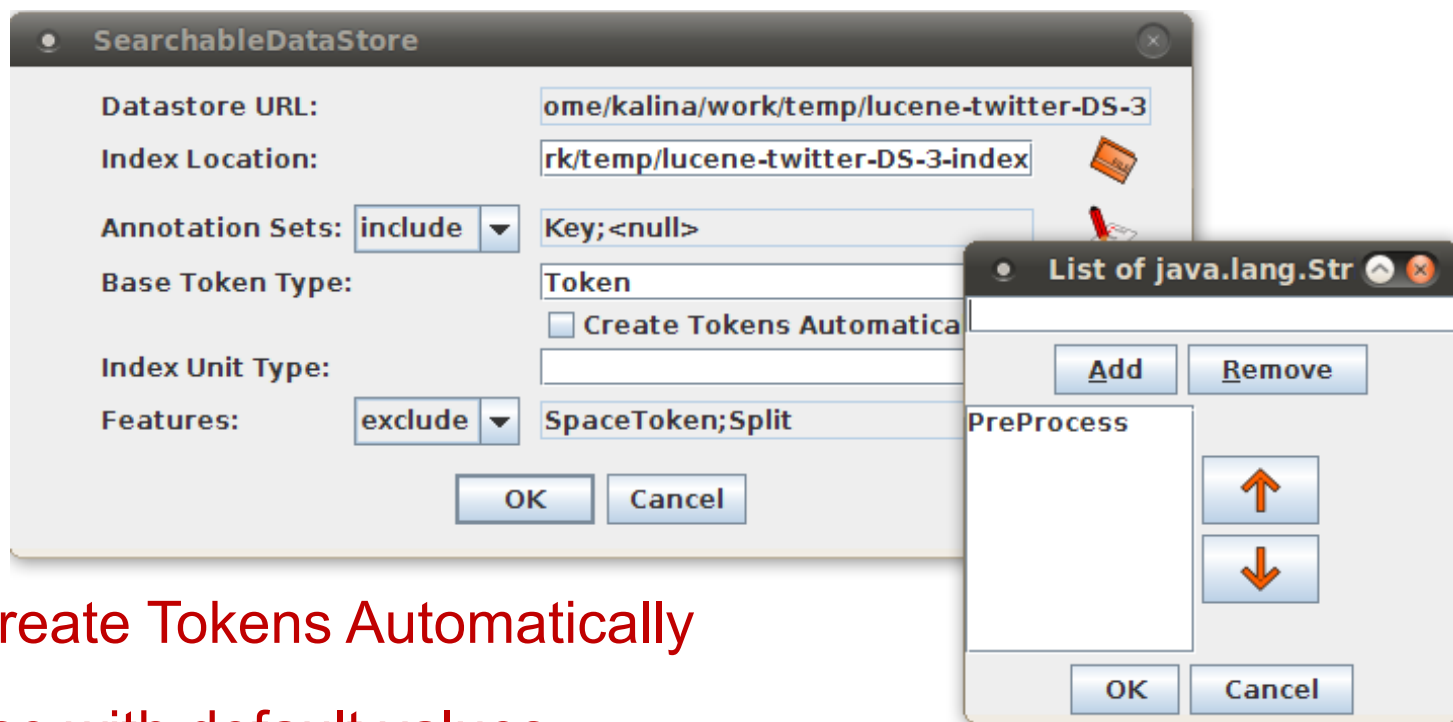
- In GATE, right click on Datastores, then Create Datastore



- Specify a new empty directory for the index
- By default, the annotation sets to be indexed are the default set (<null>) and the Key set (where by convention we put gold-standard annotations)
- We want to index only the PreProcess annotation set
- This needs to be specified at index creation time – we cannot change it later

Create Lucene Datastore (2)

- Click on the pencil button opposite Annotation Sets
- In the list box, delete the default values, type PreProcess and press the Add button



- Uncheck "Create Tokens Automatically"
- Leave all else with default values
- Click OK, the new datastore is now ready to use

ANNIC: The Query Language

- JAPE –like LHS Pattern syntax
 - String within quotes or without quotes
 - e.g. “ubuntu”
 - {AnnotationType}
 - e.g. {Person}
 - {AnnotationType == string}
 - e.g. {Organization == “University of Sheffield”}
 - {AT.featureName==value}
 - e.g. {Person.gender == male}
 - {AT.feature==value, AT.feature==value}
 - e.g. {Token.orth == “upperInitial”, Token.length == “3”}

ANNIC: The Query Language (2)

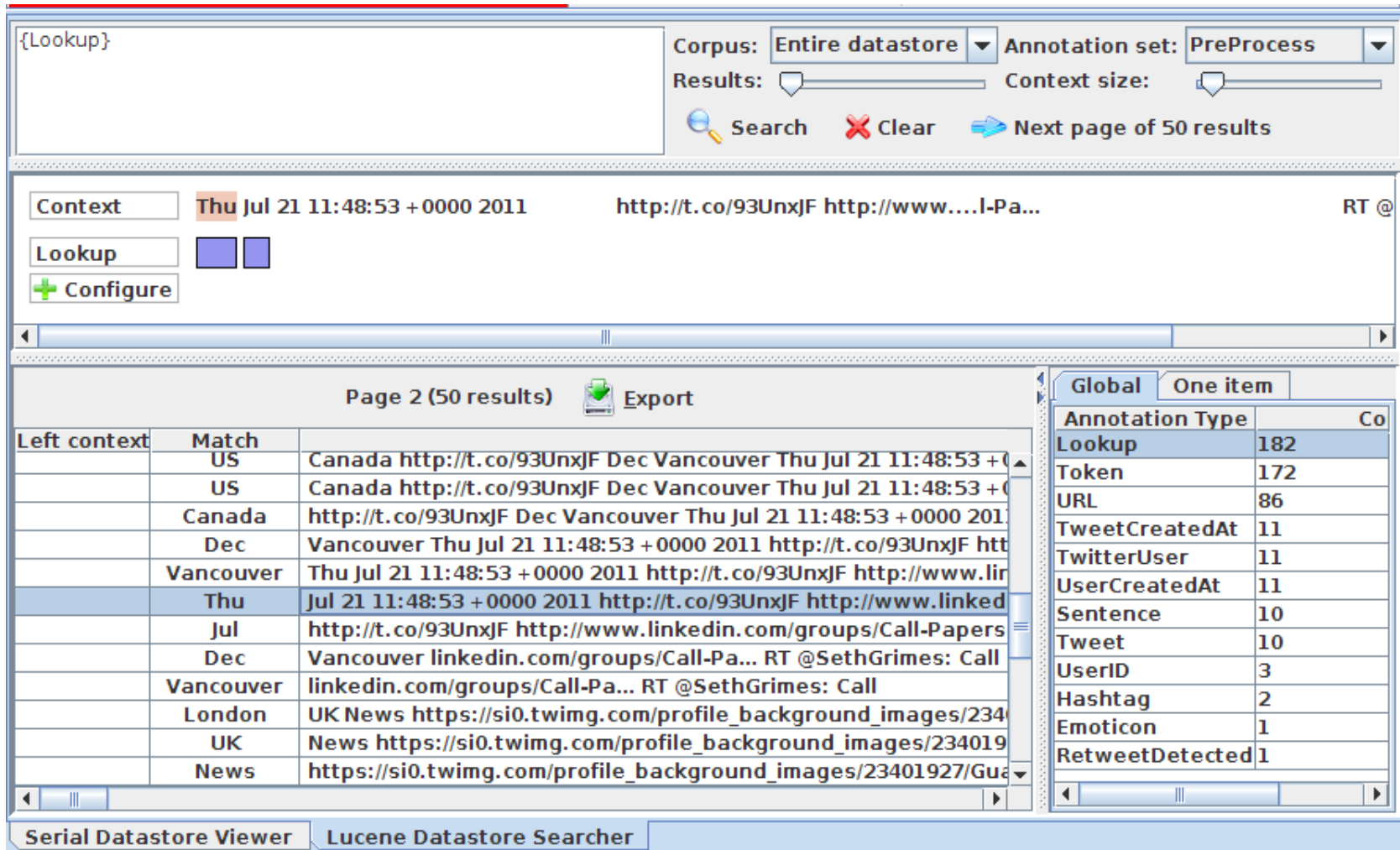
- Klene Operator + and * but they need to be quantified
 - $\{Person\}\{Token\}^*3\{Organization\}$ – find all Person and Organization annotations **within up to 3 tokens** of each other
- Logical | (OR) operator
 - $\{A\}(\{B\} | \{C\})$
- Order of query terms is very important

Hands-on: ANNIC Pattern Searches

- Create a corpus and save it to the newly created Lucene Datastore
- Populate the corpus from the **annic-documents** directory
- Note: if you populate the corpus before saving it, all documents are loaded in memory – this might be bad if you have a lot!
- Double click on the datastore
- Click on the “Lucene Datastore Searcher” tab at the bottom
- Choose over which annotation set you wish to search (top right). Enter a test ANNIC query (e.g. {Lookup} or {Hashtag}) in the big search field, then press Search

Example: Building a Date pattern

- Let us first start by checking the {Lookup} annotations in the PreProcess set and the context in which they appear



The screenshot shows the GATE Serial Datastore Viewer interface. At the top, the search criteria are set to {Lookup} in the Serial Datastore, with the corpus set to 'Entire datastore' and the annotation set to 'PreProcess'. The results are displayed on 'Page 2 (50 results)'. A table shows the left context, the match, and the corresponding text snippet. A summary table on the right shows the count of annotations for various types.

Context: **Thu Jul 21 11:48:53 +0000 2011** <http://t.co/93UnxjF> <http://www....l-Pa...> RT @

Lookup:


Page 2 (50 results)

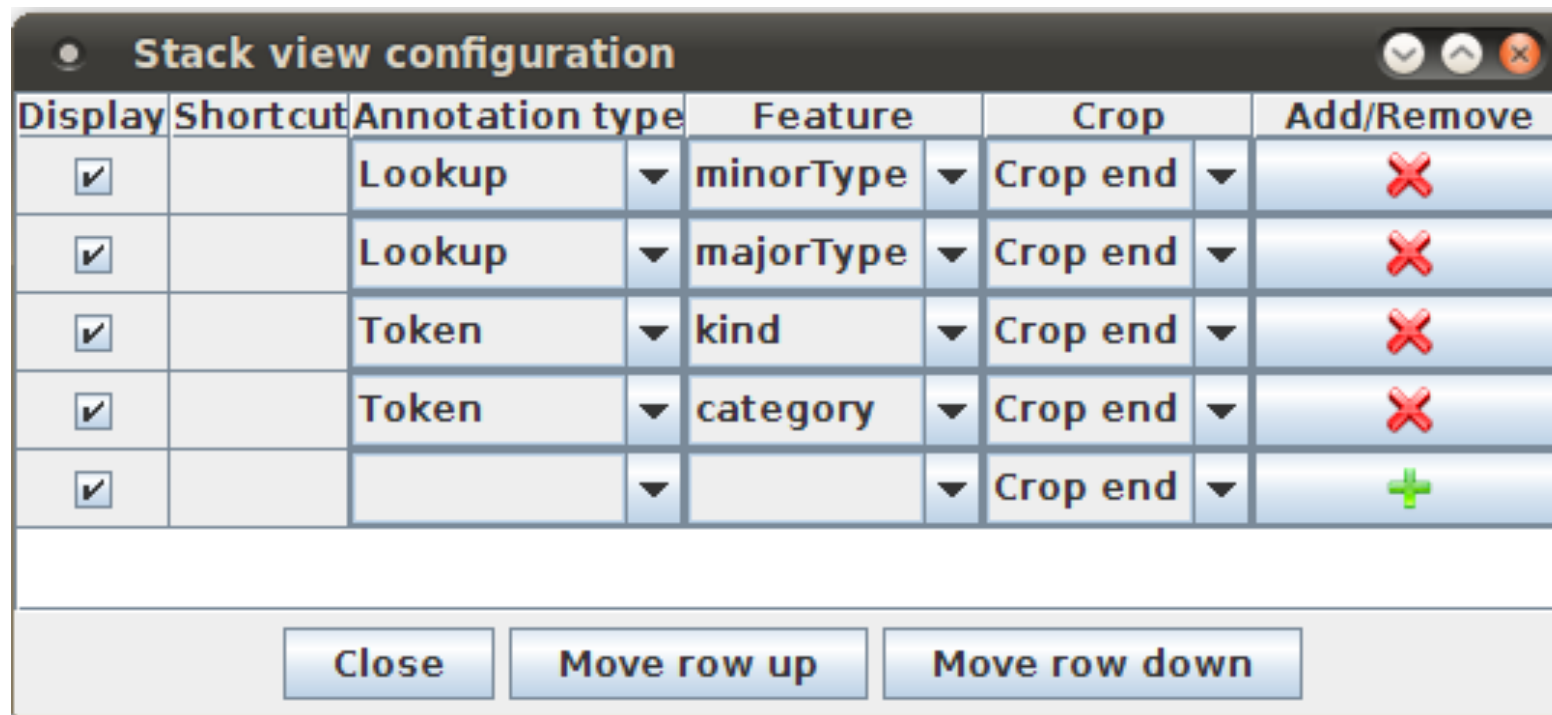
Left context	Match	Text
	US	Canada http://t.co/93UnxjF Dec Vancouver Thu Jul 21 11:48:53 +0000 2011
	US	Canada http://t.co/93UnxjF Dec Vancouver Thu Jul 21 11:48:53 +0000 2011
	Canada	http://t.co/93UnxjF Dec Vancouver Thu Jul 21 11:48:53 +0000 2011
	Dec	Vancouver Thu Jul 21 11:48:53 +0000 2011 http://t.co/93UnxjF http://www.lir
	Vancouver	Thu Jul 21 11:48:53 +0000 2011 http://t.co/93UnxjF http://www.linkedin.com/groups/Call-Papers
	Thu	Jul 21 11:48:53 +0000 2011 http://t.co/93UnxjF http://www.linkedin.com/groups/Call-Pa... RT @SethGrimes: Call
	Jul	http://t.co/93UnxjF http://www.linkedin.com/groups/Call-Pa... RT @SethGrimes: Call
	Dec	Vancouver http://t.co/93UnxjF http://www.linkedin.com/groups/Call-Pa... RT @SethGrimes: Call
	Vancouver	http://t.co/93UnxjF http://www.linkedin.com/groups/Call-Pa... RT @SethGrimes: Call
	London	UK News https://si0.twimg.com/profile_background_images/23401927/Gua
	UK	News https://si0.twimg.com/profile_background_images/23401927/Gua
	News	https://si0.twimg.com/profile_background_images/23401927/Gua

Annotation Type	Count
Lookup	182
Token	172
URL	86
TweetCreatedAt	11
TwitterUser	11
UserCreatedAt	11
Sentence	10
Tweet	10
UserID	3
Hashtag	2
Emoticon	1
RetweetDetected	1

Serial Datastore Viewer | Lucene Datastore Searcher

Seeing More Context

- Click the Configure button 
- In the dialog box, keep adding rows for the annotation types (and optionally features) that you'd like displayed in the viewer
- A good set for our example is this:



Display	Shortcut	Annotation type	Feature	Crop	Add/Remove
<input checked="" type="checkbox"/>		Lookup	minorType	Crop end	X
<input checked="" type="checkbox"/>		Lookup	majorType	Crop end	X
<input checked="" type="checkbox"/>		Token	kind	Crop end	X
<input checked="" type="checkbox"/>		Token	category	Crop end	X
<input checked="" type="checkbox"/>				Crop end	+

Close Move row up Move row down

Seeing More Context (2)



{Lookup}

Corpus: Entire datastore Annotation set: All sets

Results: Context size:

Search Clear Next page of 50 results

Context: traffic : - D . Thu Jul 21 13 : 06

Lookup.minorType:

Lookup.majorType: emoticon date date

Token.kind: word punctuation punctuation word punctuation word word number number punctuation number

+ Configure

Page 1 (50 results) Export

Left context	Match	Right context	Features
cs industry ...	analyst	, consultant, writer -	Lookup.majorType=jobtitl
traffic :-D.	Thu	Jul 21 13:06	Lookup.majorType=date
le driving ~25	%	of traffic :-D	Lookup.majorType=perce
se Eastern ...	US	& Canada) BDDCAD False	Lookup.majorType=locati
se Eastern ...	US	& Canada) BDDCAD False	Lookup.majorType=locati
se Eastern ...	US	& Canada) C0	Lookup.majorType=curren
se Eastern ...	US	& Canada) C0	Lookup.majorType=curren
n Time (US &	Canada) BDDCAD False 3151 False	Lookup.majorType=locati
n Time (US &	Canada) CODEED True	Lookup.majorType=locati
48:43 +0000	2008	False Eastern Time (US	Lookup.majorType=year

Annotation Type	Count
Token	1430
Lookup	182
URL	86
TweetCreatedAt	11
TwitterUser	11
UserCreatedAt	11
Sentence	10
Tweet	10
UserID	3
Hashtag	2
Emoticon	1

Building Up A Date Pattern

- Let's look for dates which contain a day of the week
- We start the query by typing `{Lookup.minorType=="day"}`
- 22 results are returned. The subsequent word is typically a Lookup of type month
- Expand the query:
`{Lookup.minorType=="day"}{Lookup.minorType=="month"}`
- This still returns 22 results, which means we haven't lost anything or introduced noise
- From inspection, we notice that what follows next is a number. These can be recognised from `Token.kind == "number"`
- Final Date LHS pattern:
`{Lookup.minorType=="day"}{Lookup.minorType=="month"}{Token.kind=="number"}`

Example Results



Corpus: Annotation set:
 Results: Context size:

traffic : - D . Thu Jul 21 13 : 06 :

Lookup.minorType		day	month									
Lookup.majorType	emoticon	date	date	emoticon								
Token.kind	word	punctuation	punctuation	word	punctuation	word	word	number	number	punctuation	number	punctuation


Page 1 (22 results)

Left context	Match	Right context	Features
traffic :-D.	Thu Jul 21	13:06:38	Lookup.majorType=dat...d=wo
http://t.co/93UnxjF De ...	Thu Jul 21	13:06:09	Lookup.majorType=dat...d=wo
False False http://ow.ly/5jSoS...	Thu Jul 21	13:01:21	Lookup.majorType=dat...d=wo
): http://ow.ly/5jSoS #somany...	Thu Jul 21	13:01:21	Lookup.majorType=dat...d=wo
000 jobs cuts by 2015	Thu Jul 21	13:02:46	Lookup.majorType=dat...d=wo
deputy Rudolf Hess exhumed...	Thu Jul 21	13:09:07	Lookup.majorType=dat...d=wo
of sales tweets? http://bit.ly/...	Thu Jul 21	13:07:11	Lookup.majorType=dat...d=wo
head its British operations h...	Thu Jul 21	13:07:09	Lookup.majorType=dat...d=wo
False 0 Some Person http://di...	Mon Feb 09	16:33:16	Lookup.majorType=dat...d=wo
False 0 Manu Sporny http://di...	Mon Feb 09	16:33:16	Lookup.majorType=dat...d=wo
False 50 Kalina Bontcheva ht...	Thu Jul 30	12:14:39	Lookup.majorType=dat...d=wo
False 0 Pete Cashmore http://...	Mon Mar 12	01:28:01	Lookup.majorType=dat...d=wo

Annotation Type	Count
Token	1430
Lookup	182
URL	86
TweetCreatedAt	11
TwitterUser	11
UserCreatedAt	11
Sentence	10
Tweet	10
UserID	3
Hashtag	2
Emoticon	1
RetweetDetected	1

Hands-on: Expand to include the time

- Double-click on the datastore, open the ANNIC GUI
- In the ANNIC GUI:
 - Expand the pattern to include the time expressions

		Page 1 (22 results)  Export			
	Match		Right context		Fe
traffic :-D.	Thu Jul 21 13:06:38	+0000	2011 False 36	Lookup.majorType	
://t.co/93UnxJF De ...	Thu Jul 21 13:06:09	+0000	2011 False 3	Lookup.majorType	
manyerrorsitsfunny	Thu Jul 21 13:01:21	+0000	2011 False 93	Lookup.majorType	
manyerrorsitsfunny	Thu Jul 21 13:01:21	+0000	2011 False 93	Lookup.majorType	
00 jobs cuts by 2015	Thu Jul 21 13:02:46	+0000	2011 False 9402955...	Lookup.majorType	
http://bbc.in/q8E6g2	Thu Jul 21 13:09:07	+0000	2011 False 36	Lookup.majorType	
http://bit.ly/oKd8lQ	Thu Jul 21 13:07:11	+0000	2011 False 46	Lookup.majorType	
http://bit.ly/pkqXy0	Thu Jul 21 13:07:09	+0000	2011 False 65	Lookup.majorType	

Converting the Pattern to a JAPE Rule

- You might wish to create several different annotations from this JAPE LHS, e.g. Date, Time, and Offset
- Use different named blocks in the pattern to achieve this
- We leave this as homework, especially if you wish to link the year (which appears at the end) with the rest of the date
- A relevant PR here is the DateNormalizer:
 - <http://gate.ac.uk/userguide/sec:misc-creole:datenormalizer>



Other interesting searches

Messages **annic-ds**

{Hashtag}

Corpus: Entire datastore Annotation set: PreProcess

Results: Context size:

Search Clear Next page of 332 results

Context False 189 False False http://ow.ly/5JSoS #somanyerrorsitsfunny Thu July21 13 01

Token [red boxes]

Hashtag [green box]

+ Configure

Page 1 (2 results) Export

Left context	Match	Right context
False 189 False False http://ow.ly/5JSoS	#somanyerrorsitsfunny	Thu July21
/Microformats): http://ow.ly/5JSoS	#somanyerrorsitsfunny	Thu July21

Annotation Type	Count
Emoticon	8
Percent	5
UserID	4
Hashtag	2
Money	2