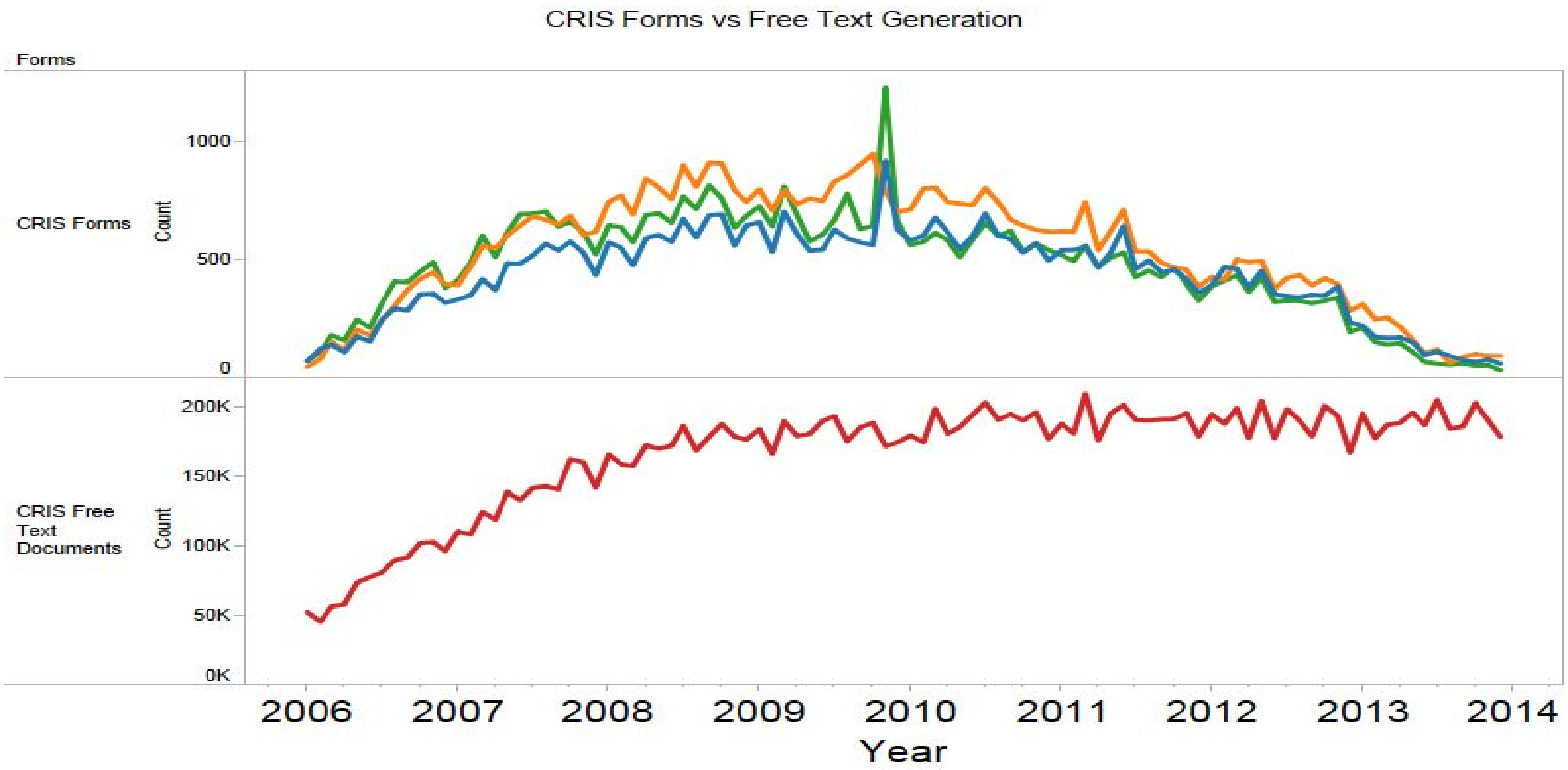
Clinical Natural Language Processing - An Introduction

Richard Jackson, King’s College London

What is Natural Language Processing?

Subdomains

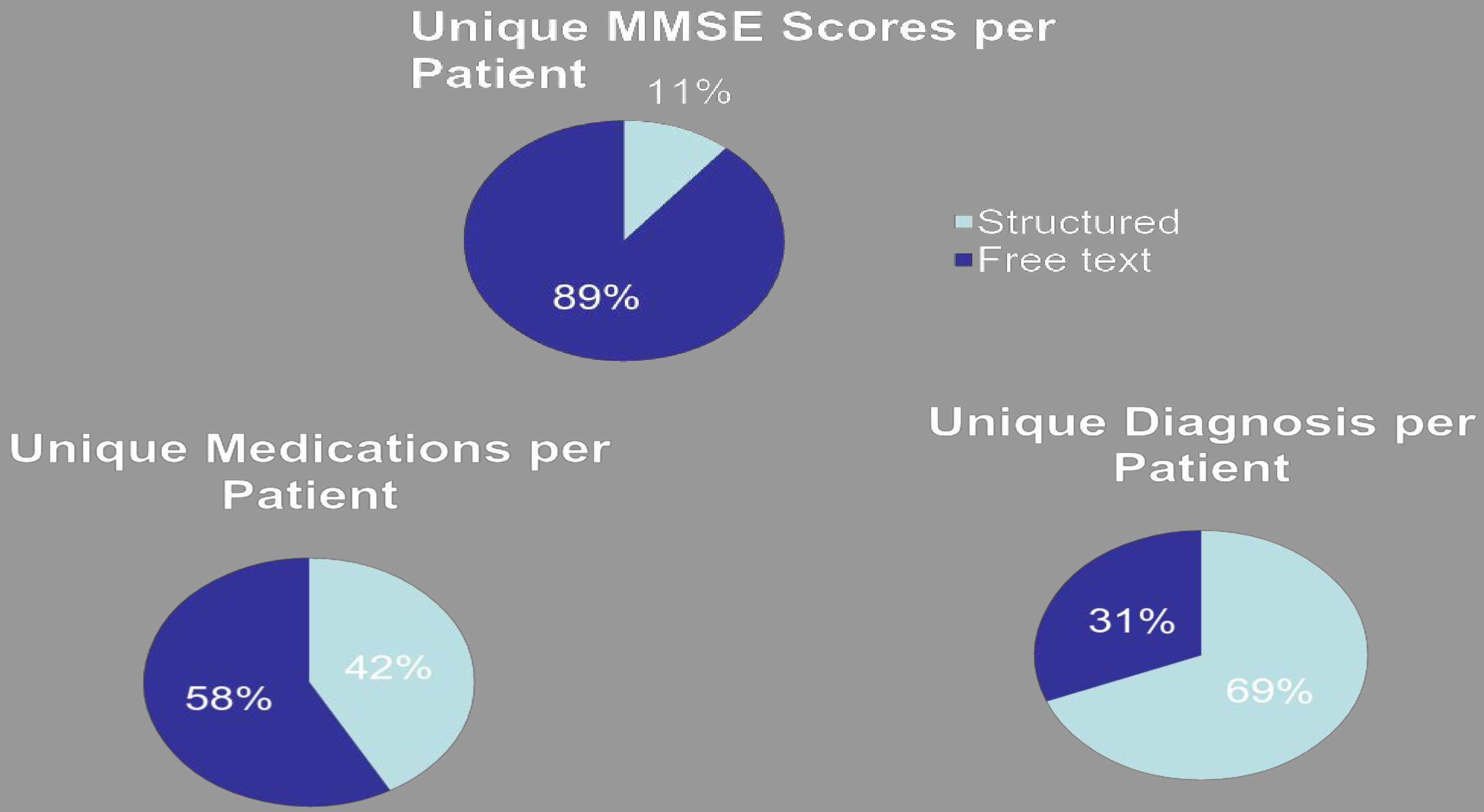
- **Information Retrieval,**
- **Information Extraction,**
- Translation,
- Summarisation,
- Object Character Recognition,
- Document Classification,
- Co-reference Resolution
- etc. etc



The trend of count of CN_Doc_ID for date Month broken down by Forms. Color shows details about event_category. The data is filtered on date, which includes the last 8 years relative to 31/12/2013. The filter associated with this field ranges from 01/01/2006 to 31/12/2013.

event_category

- History
- Mental State Formulation
- Presenting Circumstances
- Text_Document



Unique MMSE Scores per Patient
11%
89%
Structured
Free text
Unique Medications per Patient
42%
58%
Unique Diagnosis per Patient
31%
69%

Information Retrieval

Inf ble

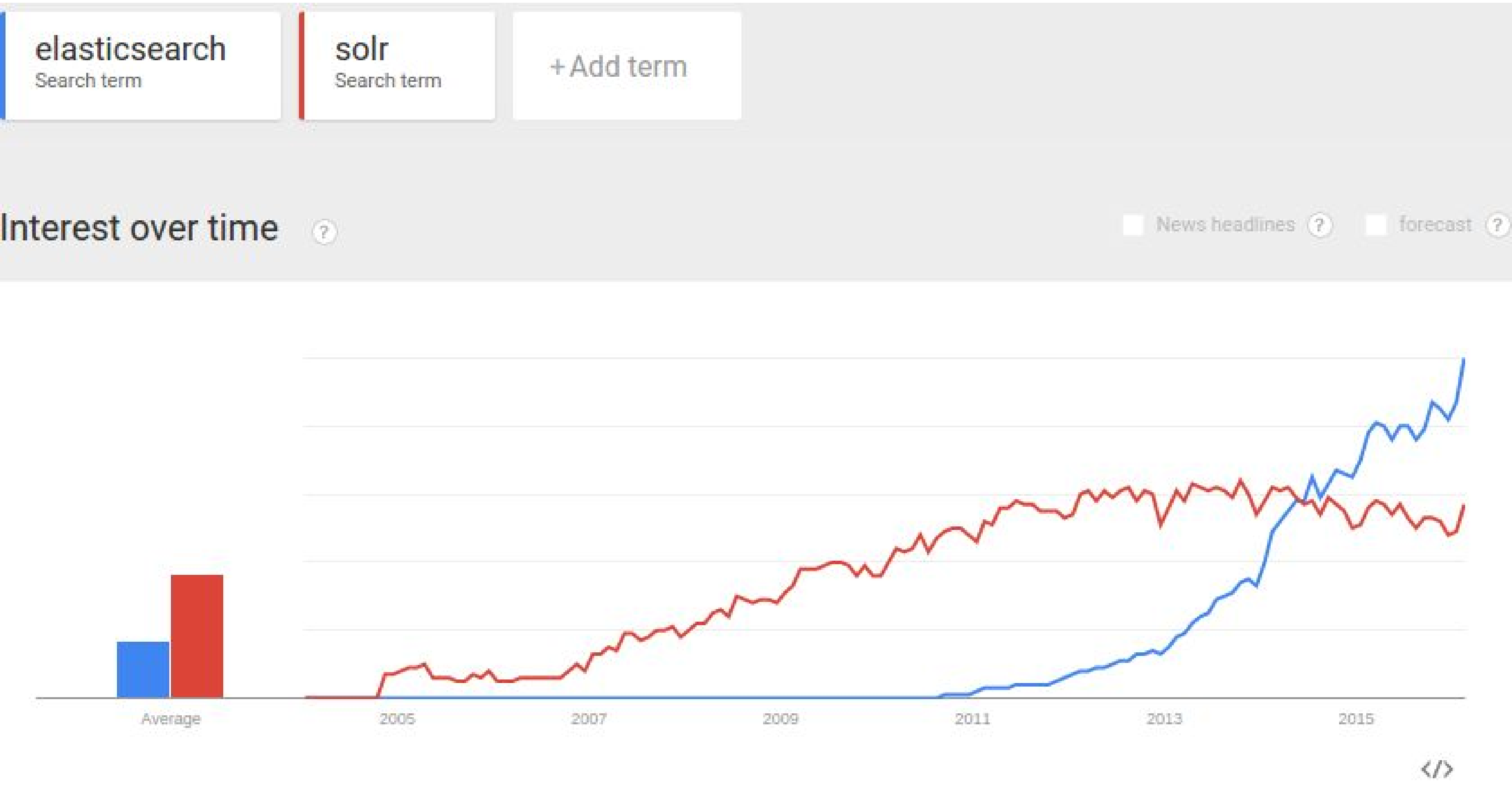

Google!
Search the web using Google!
10 results
Google Search
I'm feeling lucky
Index contains ~25 million pages (soon to be much bigger)
About Google!
Stanford Search Linux Search
Get Google! updates monthly!
your e-mail
Subscribe
Archive
Copyright ©1997-8 Stanford University

Where is the field of Information Retrieval today?

- Google effectively ended the commercial provision of public web search services (with some exceptions)
- However, private, customisable information retrieval remains an in-demand capability for many business concerns
  - Enterprise search
  - data management
  - systems monitoring
  - e-commerce

Open Source search engines

Today's private search arena is dominated by two similar products


elasticsearch
Search term
solr
Search term
+ Add term
Interest over time
News headlines
forecast
Average
2005
2007
2009
2011
2013
2015

Principal differences are mainly at the API level

Solr

Around since 2006

Largest community by far

Many plugins of variable quality

Suffers from open source bloat

Massive community response since arrival of elasticsearch

elasticsearch

Around since 2010

The ‘new kid on the block’

Cleaner design, easier to use due to tightly controlled development by elastic.co

rapidly overtaking Solr as search endige of choice

high commercial focus ($100m VC, open core)

Movement into analytics market

How do they work?

‘Inverted index’

At ingestion time, document is tokenised, and an index is created

An index is a mapping of which tokens belong to what documents, and is very efficient to compute over.

What sorts of queries are possible?

Simple keyword “schizophrenia”

Wildcard “schizophren*”

Stemming “schizophrenic”

Fuzzy “schisophrenia”~0.4

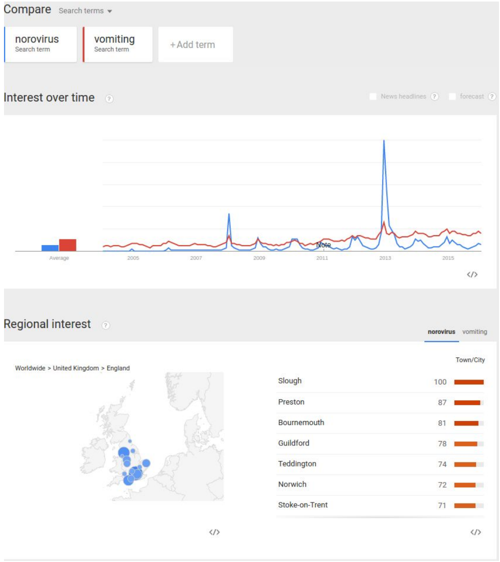
Span “diagnosis schizophrenia”~4

Custom.....

Kibana Demo

Kibana analytics: Norovirus seasonality

http://www.google.com/trends/explore#geo=GB-ENG&cmpt=q&q=norovirus,+vomiting



Norovirus in the local community

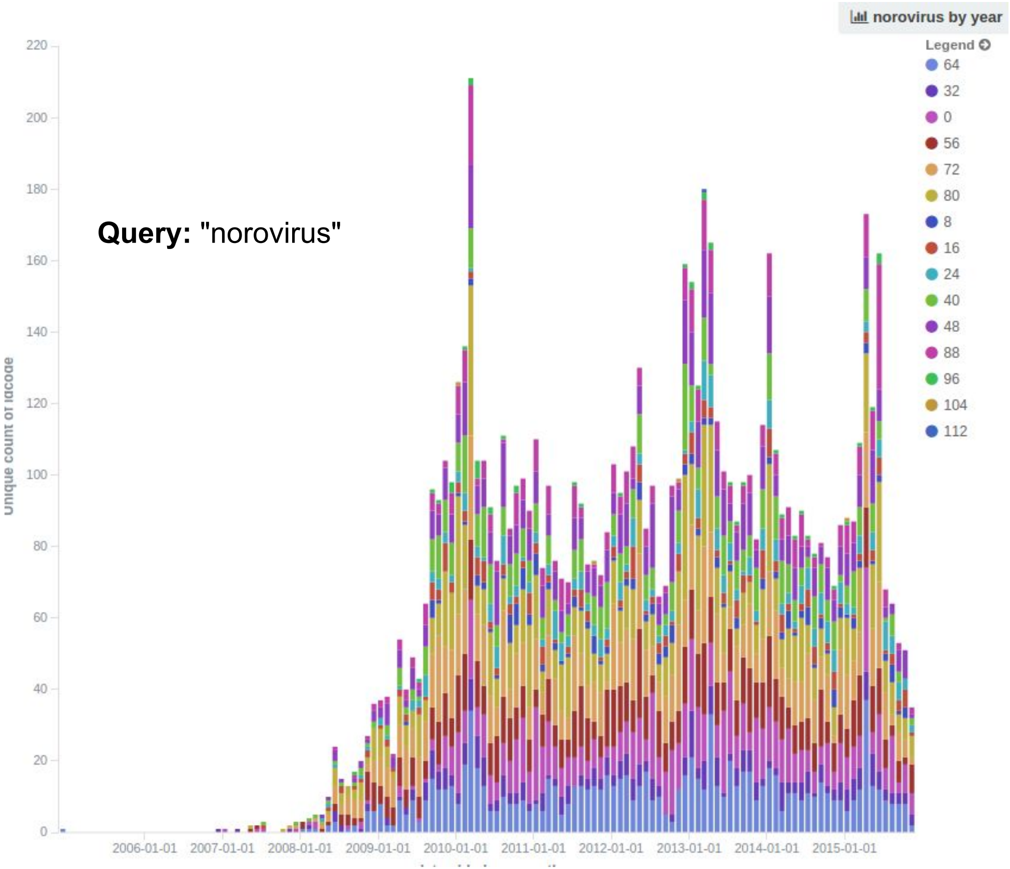
26 January 2016 - Help us stop the spread of infection at King's

There is currently an increase of Norovirus in the local community. The virus, also known as the winter vomiting bug, causes diarrhoea and vomiting.

Norovirus symptoms usually last for between 24 and 48 hours, and most people recover very quickly.

If you have symptoms, it is best to stay at home, rest and drink plenty of fluids. If you are still unwell after four or five days, you should speak to your GP or call the NHS helpline on 111.

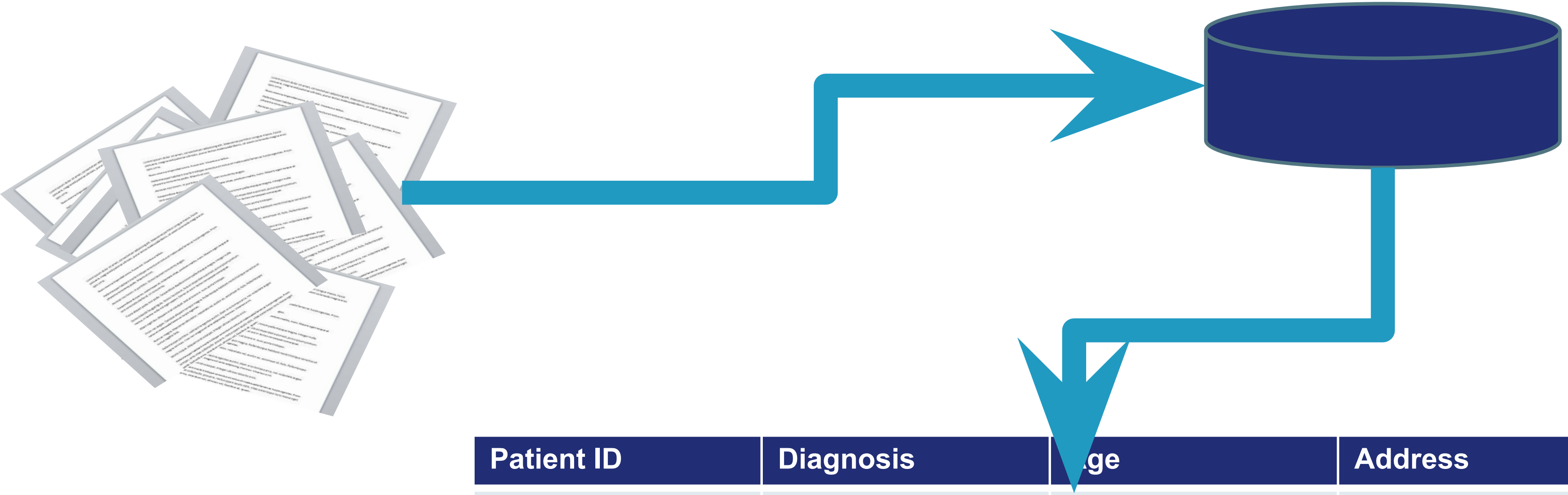
https://www.kch.nhs.uk/news/public/news/view/19193




kibana
Discover
Visualize
Dashboard
Settings
epub_date:[2008/01/01 TO 2014/12/31] AND body:"kinase inhibitor"~5
pmc_corpus
Data
Options
Y-Axis
Aggregation
Unique Count
Field
Analyzed Field
pmid
Advanced
+ Add metrics
buckets
Split Bars
Aggregation
Significant Terms
Field
Analyzed Field
body
Size
30
Advanced
X-Axis
Sub Aggregation
Date Histogram
Field
epub_date
Interval
Yearly
Advanced
JSON Input
Any JSON formatted properties you add here will be merged with the elasticsearch aggregation definition for this section. For example shard_size on a terms aggregation
24,000
22,000
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Unique count of pmid
2009
2010
2011
2012
2013
2014
epub_date per year
Legend
kinase
inhibitor
cyclin
kinases
tyrosine
phosphorylation
akt
imatinib
erk
pi3k
inhibitors
signaling
phosphorylated
mapk
tki
sunitinib
phospho
inhibition
apoptosis
mek
p21
proliferation
mitogen
pathway
activated
inhibits
inhibited
dasatinib
activation
sorafenib

Information Extraction

Information Extraction (IE)



Patient ID	Diagnosis	Age	Address
1	Depression	31	Flat 1, XYZ road
2	Cancer	67	2 Another Lane
3	Heart Attack	58	78 A place

Many tools and frameworks

How to build an information extraction algorithm

1. Define problem
2. Produce annotation guidelines
3. Create gold standard and training sets
4. Chose a NLP method (Rules, Machine Learning, Hybrid)
5. Validate model against gold standard and roduce performance statistics

1. Define Problem

- Is an NLP approach appropriate? Signal/noise ratio
- Data sufficiency considerations? How many features need to be extracted for a concept to be complete?
- Subject matter expert driven

2. Produce annotation guidelines

- Required to ensure consistent rules are applied when producing training data
- Often starts as a simple process, but rapidly becomes unwieldy if not properly managed
- Feeds back into assessing feasibility of task
- Some guidelines (THYME ML) are huge! >50 pages

3. Create gold standard and training data

- Uses annotation guidelines to describe how a human user should annotate a corpus of documents
- Often a boring, unpopular task, but completely necessary for the building of a model
- May involve teams of annotators double annotating to ensure consistency
- The availability of annotated data is often the bottleneck in improving the performance of an algorithm

4. Choose an NLP method

Rules

Deterministic method that describes a clear grammatical logic

Generally work very well for simple problems

Can become complicated quickly for complex tasks

Requires a large amount of communication between a language engineer and a subject matter expert

examples: JAPE (GATE), RUTA (UIMA)

Machine Learning

Do not require a language engineer to understand the domain, but may require large amounts of training data to be effective

Often ‘Black Box’ and therefore hard to determine best way to improve performance

Can be combined with rules, to generate richer features for the ML algo to work with

examples: Support Vector Machines, Naive Bayes, Conditional random fields and Maximum Entropy

5. Validate model

Most common performance statistics are precision, recall and F1

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Predicted condition negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$	

GATE Demo

TEXT HUNTER - CONCEPT EXTRACTION SYSTEM

NEGATIVE SYMPTOMS CASE STUDY

SLAM Clinical records

~250 000 patient records

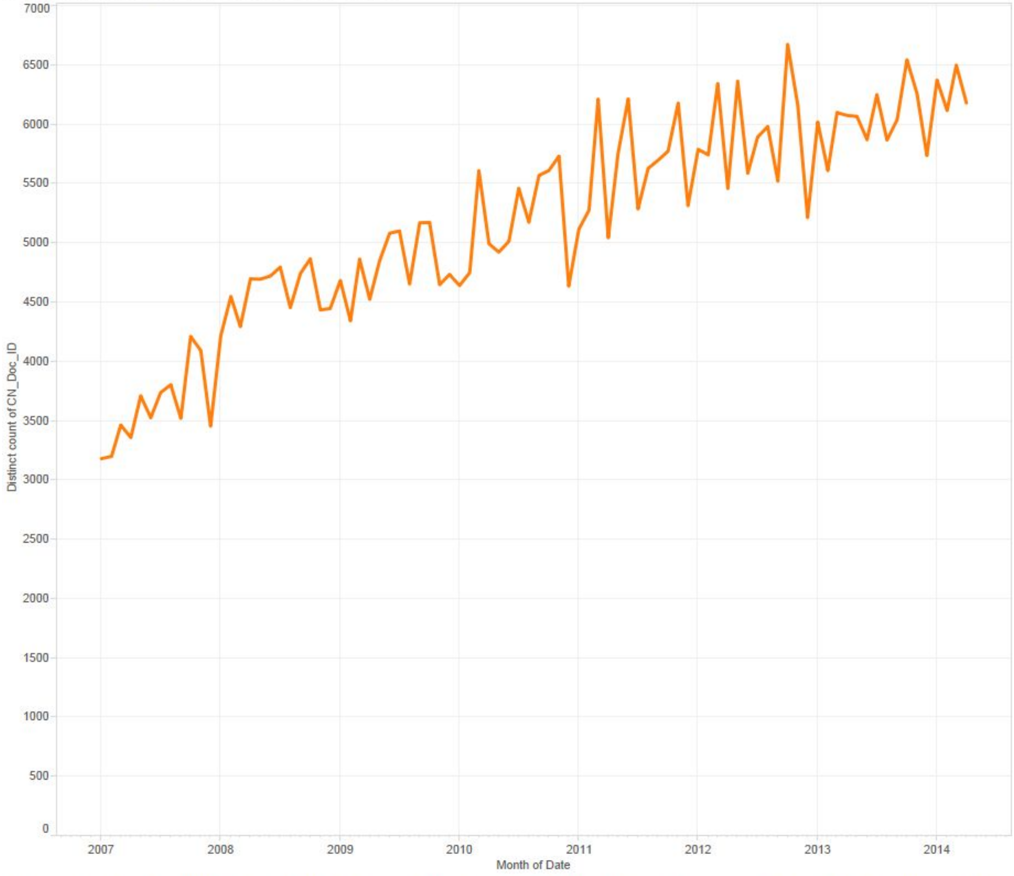
18 million free text documents

Available for research via the CRIS project

Negative symptoms of psychosis

- Deficits of normal emotional behaviour
  - Social withdrawal
  - Anhedonia (inability to experience pleasure)
  - Poverty of speech
  - Etc.

Count of Mentions of "eye contact" 2007 - 2014



The trend of distinct count of CN_Doc_ID for Date Month. Color shows details about Word. The data is filtered on Date and Date Month. The Date filter includes the last 8 years relative to 15/05/2014. The filter associated with this field ranges from 01/01/2007 to 31/12/2014. The Date Month filter excludes May 2014, June 2014, August 2014, October 2014 and December 2014. The view is filtered on Word and Exclusions (Word,MONTH(Date)). The Word filter excludes "rapport". The Exclusions (Word,MONTH(Date)) filter keeps 487 members.

Word

"eye contact"

Example sentences

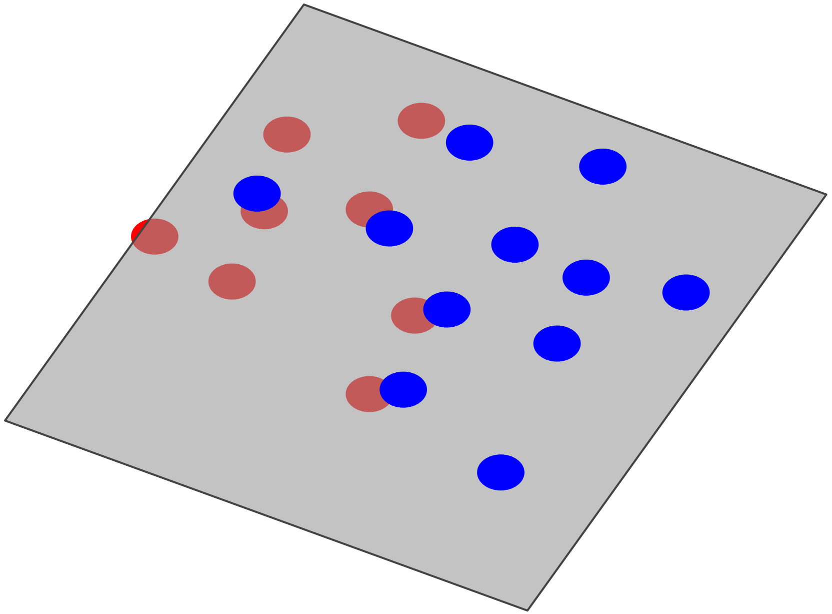
- ‘Patient X has poor eye contact’
- ‘I assessed the patient on 01/03/12. I noted that eye contact was poor’
- ‘Saw patient X yesterday. Eye contact was bad, even worse than before’

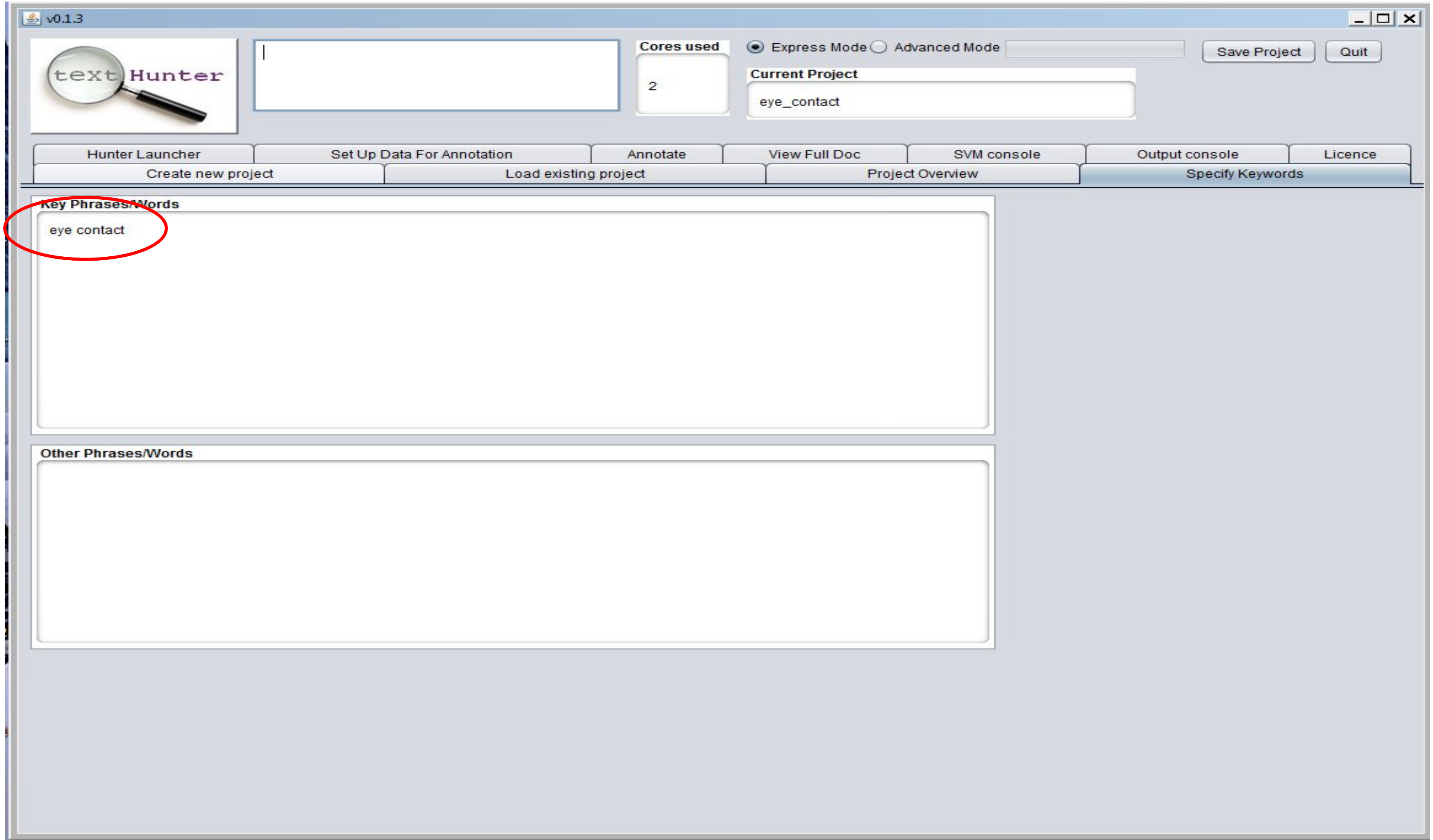
- ‘I spoke to patient X over the telephone, and was thus unable to assess eye contact’
- ‘Patient X presented with the same level of eye contact as on our last meeting’
- Patient had good eye contact

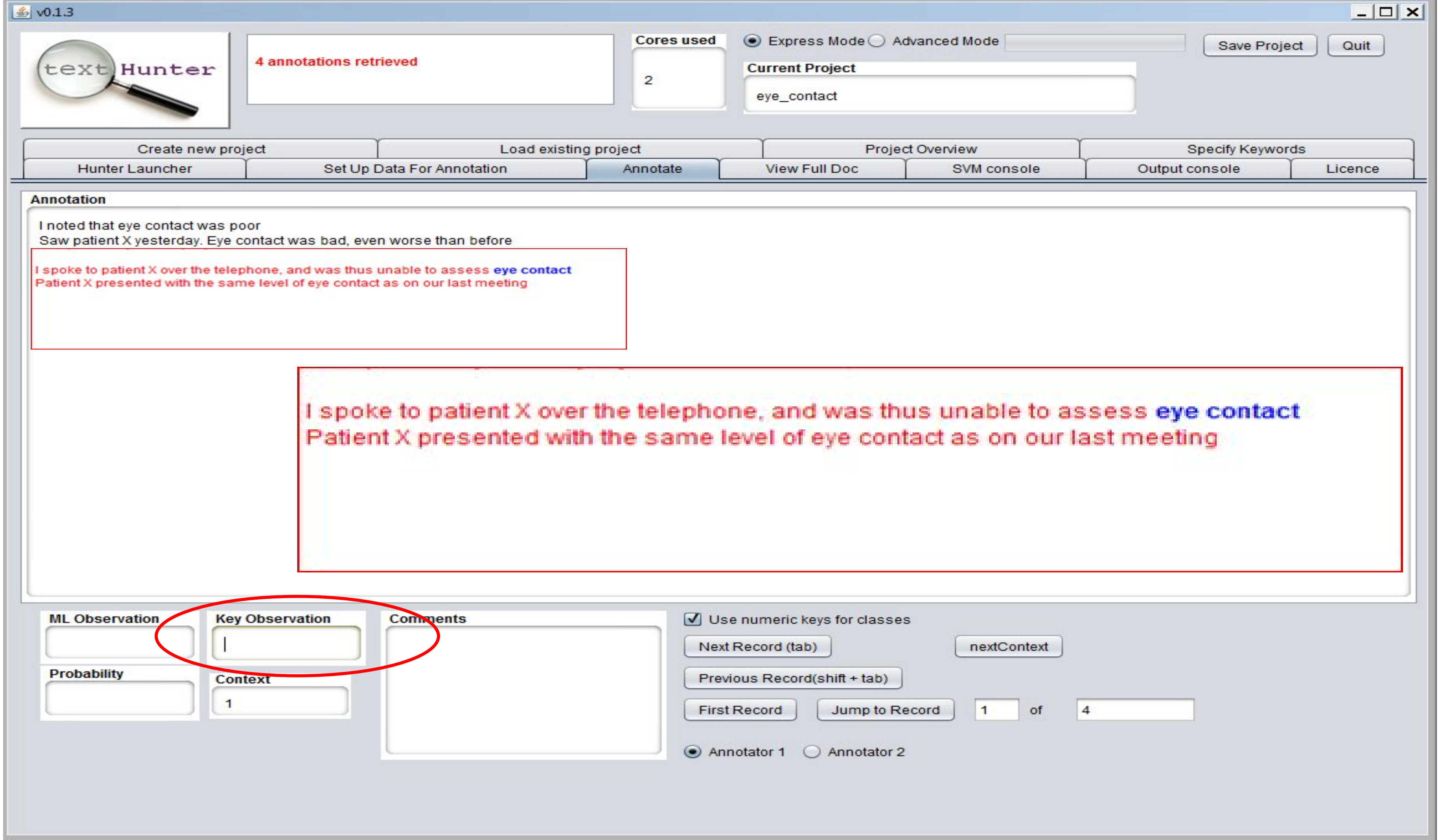
Support Vector Machines

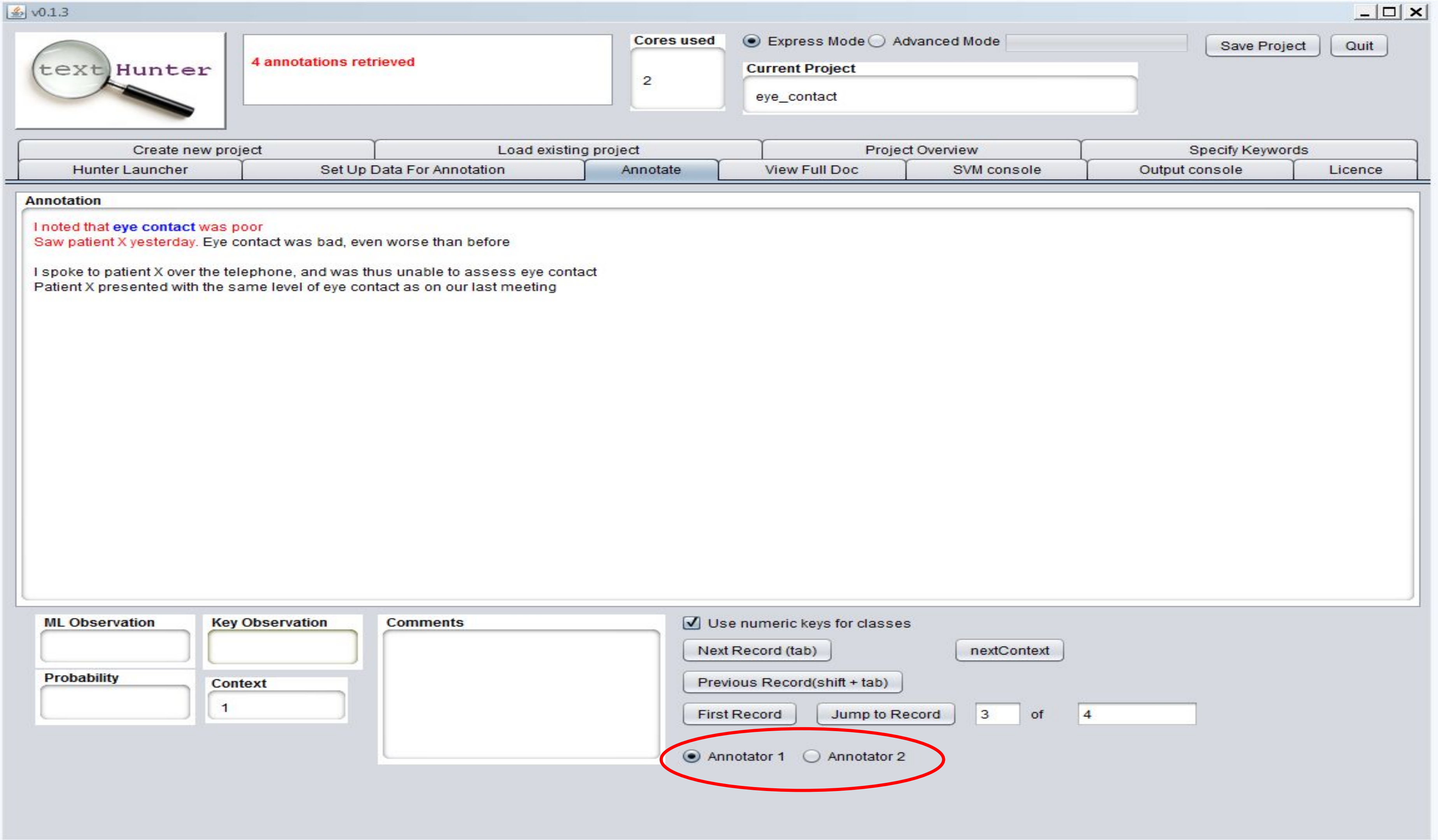
	patient	eye	contact	poor	bad	worse	unable	assess	good
Sentence 1	1	1	1	1	0	0	0	0	0
Sentence 2	1	1	1	0	1	1	0	0	0
Sentence 3	1	1	1	1	0	0	0	0	0
Sentence 4	1	1	1	0	0	0	1	1	0
Sentence 5	1	1	1	0	0	0	0	0	0
Sentence 6	1	1	1	0	0	0	0	0	1

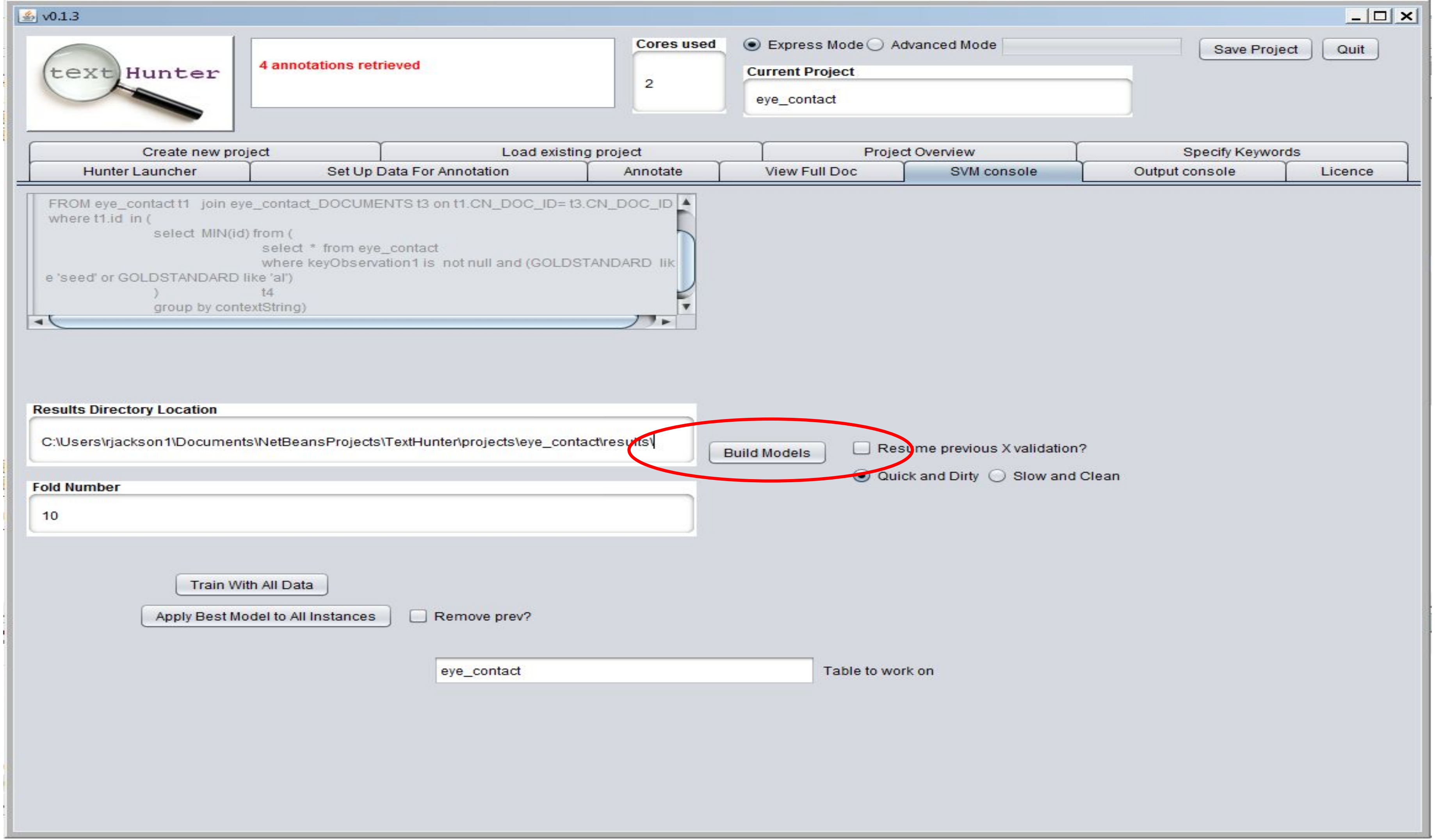
Support Vector Machines



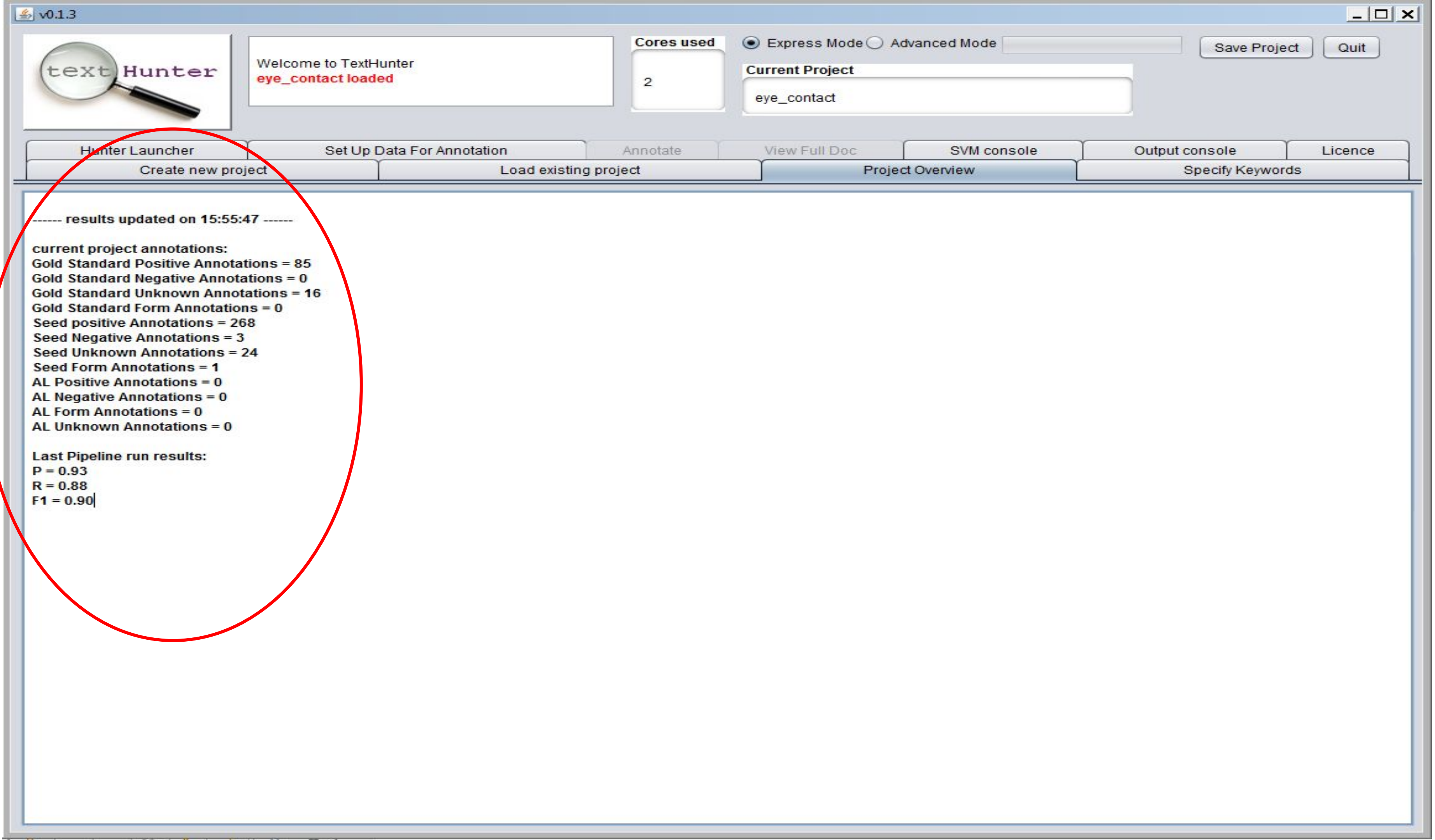

v0.1.3
textHunter
Cores used
2
Express Mode
Advanced Mode
Save Project
Quit
Current Project
eye_contact
Hunter Launcher
Set Up Data For Annotation
Annotate
View Full Doc
SVM console
Output console
Licence
Create new project
Load existing project
Project Overview
Specify Keywords
Key Phrases/Words
eye contact
Other Phrases/Words


v0.1.3
text Hunter
4 annotations retrieved
Cores used
2
Express Mode
Advanced Mode
Save Project
Quit
Current Project
eye_contact
Create new project
Load existing project
Project Overview
Specify Keywords
Hunter Launcher
Set Up Data For Annotation
Annotate
View Full Doc
SVM console
Output console
Licence
Annotation
I noted that eye contact was poor
Saw patient X yesterday. Eye contact was bad, even worse than before
I spoke to patient X over the telephone, and was thus unable to assess eye contact
Patient X presented with the same level of eye contact as on our last meeting
I spoke to patient X over the telephone, and was thus unable to assess eye contact
Patient X presented with the same level of eye contact as on our last meeting
ML Observation
Key Observation
Comments
Probability
Context
1
Use numeric keys for classes
Next Record (tab)
nextContext
Previous Record(shift + tab)
First Record
Jump to Record
1
of
4
Annotator 1
Annotator 2


v0.1.3
text Hunter
4 annotations retrieved
Cores used
2
Express Mode
Advanced Mode
Save Project
Quit
Current Project
eye_contact
Create new project
Load existing project
Project Overview
Specify Keywords
Hunter Launcher
Set Up Data For Annotation
Annotate
View Full Doc
SVM console
Output console
Licence
Annotation
I noted that eye contact was poor
Saw patient X yesterday. Eye contact was bad, even worse than before
I spoke to patient X over the telephone, and was thus unable to assess eye contact
Patient X presented with the same level of eye contact as on our last meeting
ML Observation
Key Observation
Comments
Use numeric keys for classes
Next Record (tab)
nextContext
Previous Record(shift + tab)
Probability
Context
1
First Record
Jump to Record
3
of
4
Annotator 1
Annotator 2

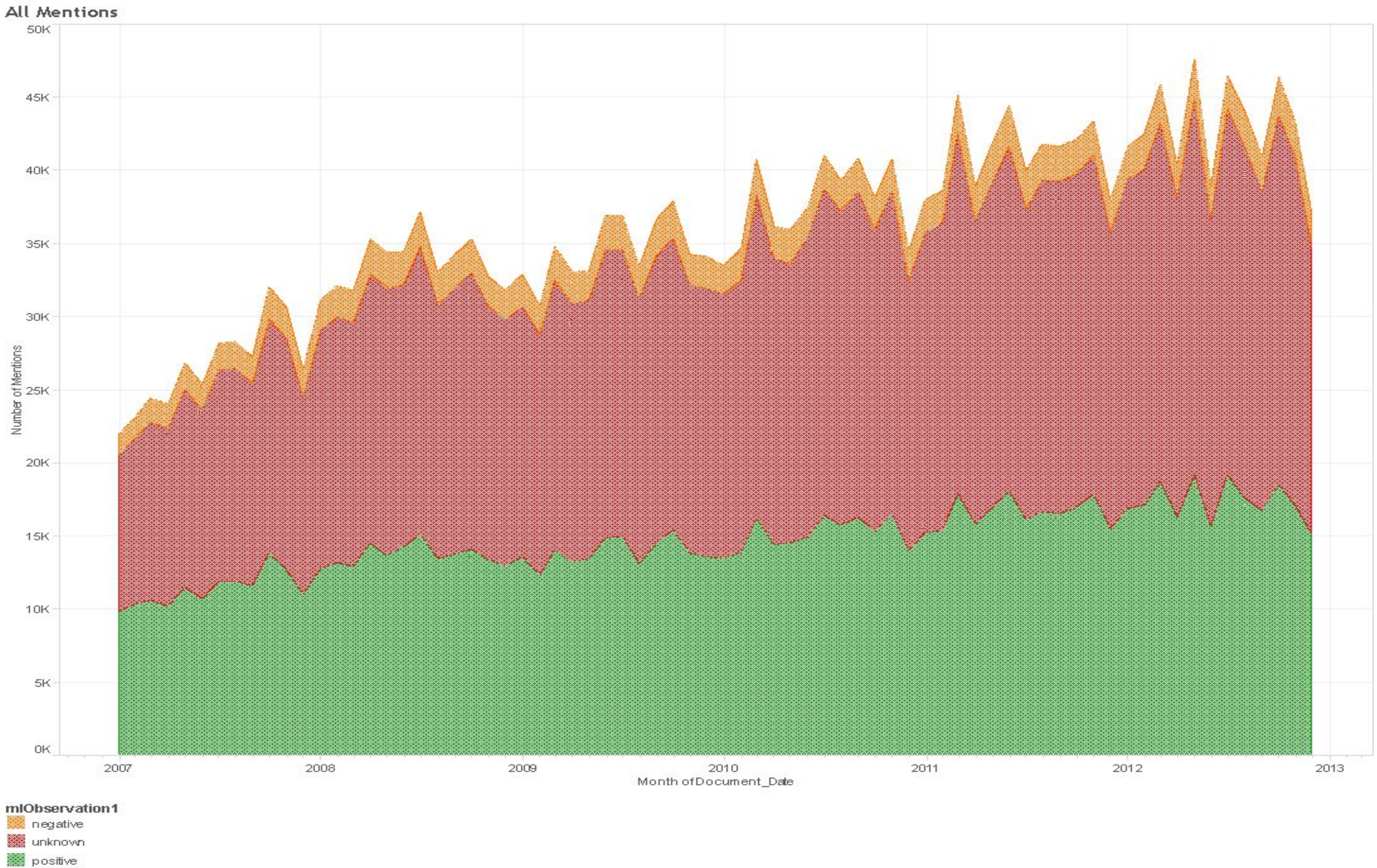


v0.1.3
text Hunter
4 annotations retrieved
Cores used
2
Express Mode
Advanced Mode
Save Project
Quit
Current Project
eye_contact
Create new project
Load existing project
Project Overview
Specify Keywords
Hunter Launcher
Set Up Data For Annotation
Annotate
View Full Doc
SVM console
Output console
Licence
FROM eye_contact t1 join eye_contact_DOCUMENTS t3 on t1.CN_DOC_ID= t3.CN_DOC_ID
where t1.id in (
select MIN(id) from (
select * from eye_contact
where keyObservation1 is not null and (GOLDSTANDARD lik
e 'seed' or GOLDSTANDARD like 'al')
) t4
group by contextString)
Results Directory Location
C:\Users\rjackson1\Documents\NetBeansProjects\TextHunter\projects\eye_contact\results\
Build Models
Resume previous X validation?
Quick and Dirty
Slow and Clean
Fold Number
10
Train With All Data
Apply Best Model to All Instances
Remove prev?
eye_contact
Table to work on


v0.1.3
text Hunter
Welcome to TextHunter
eye_contact loaded
Cores used
2
Express Mode
Advanced Mode
Save Project
Quit
Current Project
eye_contact
Hunter Launcher
Set Up Data For Annotation
Annotate
View Full Doc
SVM console
Output console
Licence
Create new project
Load existing project
Project Overview
Specify Keywords
------ results updated on 15:55:47 ------
current project annotations:
Gold Standard Positive Annotations = 85
Gold Standard Negative Annotations = 0
Gold Standard Unknown Annotations = 16
Gold Standard Form Annotations = 0
Seed positive Annotations = 268
Seed Negative Annotations = 3
Seed Unknown Annotations = 24
Seed Form Annotations = 1
AL Positive Annotations = 0
AL Negative Annotations = 0
AL Form Annotations = 0
AL Unknown Annotations = 0
Last Pipeline run results:
P = 0.93
R = 0.88
F1 = 0.90

Psychosis Symptomatology

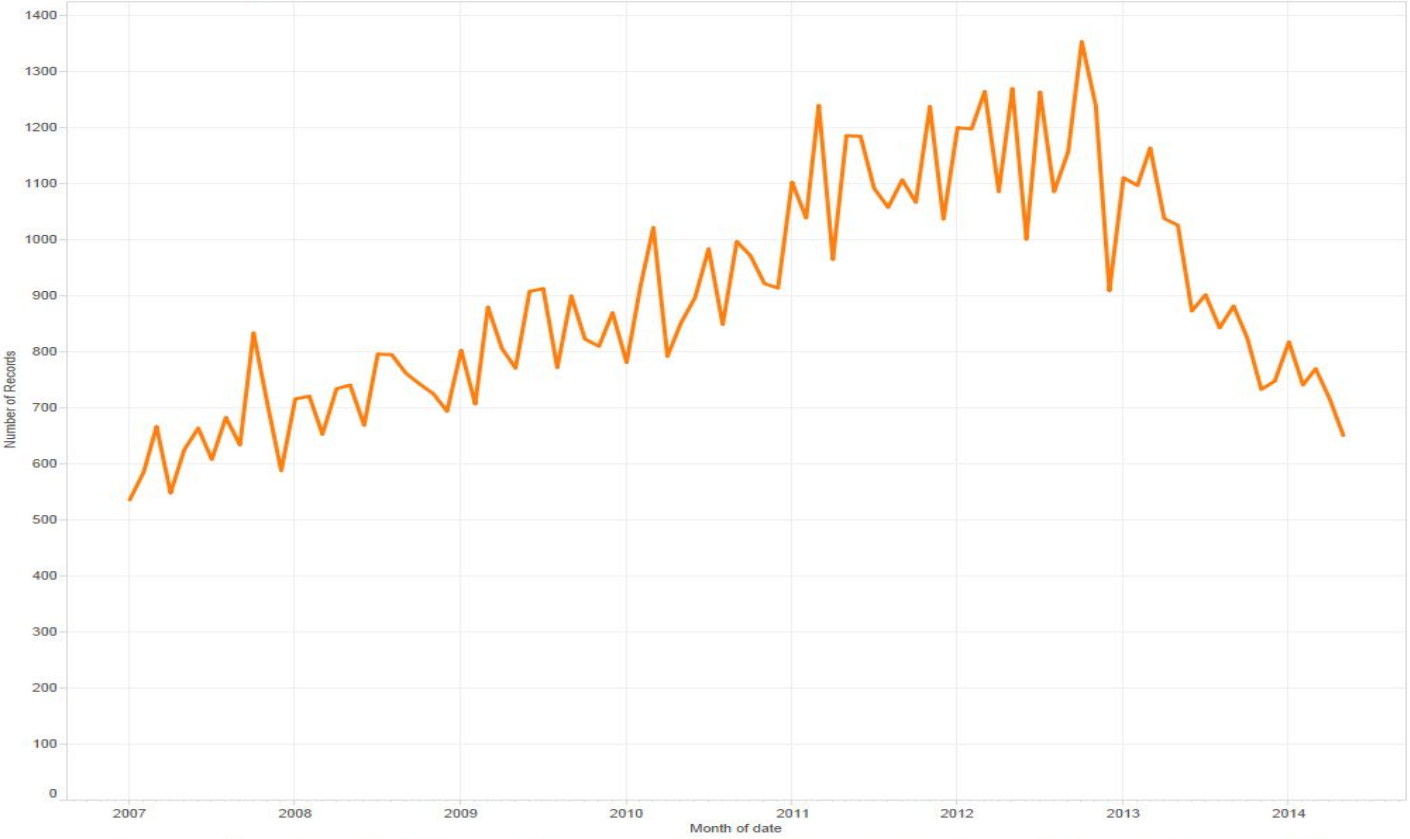
app	P	R	F1
Apathy	0.85	1	0.93
Blunted/Flat affect	1	0.74	0.84
Concrete thinking	0.97	0.6	0.74
Emotional withdrawal	0.78	0.76	0.77
Motivation	0.75	0.63	0.68
Poverty of speech	0.81	0.87	0.84
Rapport	0.85	1	0.91
Social withdrawal	0.9	1	
Anhedonia	0.96	0.83	0.89
Associations	1	0.87	0.94
Circumstantial	0.9	1	0.94
Coherence	0.85	0.98	0.91
Delusions	0.91	1	0.95
Derailment	0.91	0.96	0.94
Flight of ideas	0.93	0.97	0.94
Hallucinations	0.85	0.98	0.91
Incoherence	0.82	0.99	0.9
Poverty of thought	0.92	0.96	0.94
Tangential	0.92	1	0.95



All Mentions
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0K
Number of Mentions
2007
2008
2009
2010
2011
2012
2013
Month of Document_Date
mlObservation1
negative
unknown
positive

The MMSE story

MMSE score information extraction rates 2007 - 2014

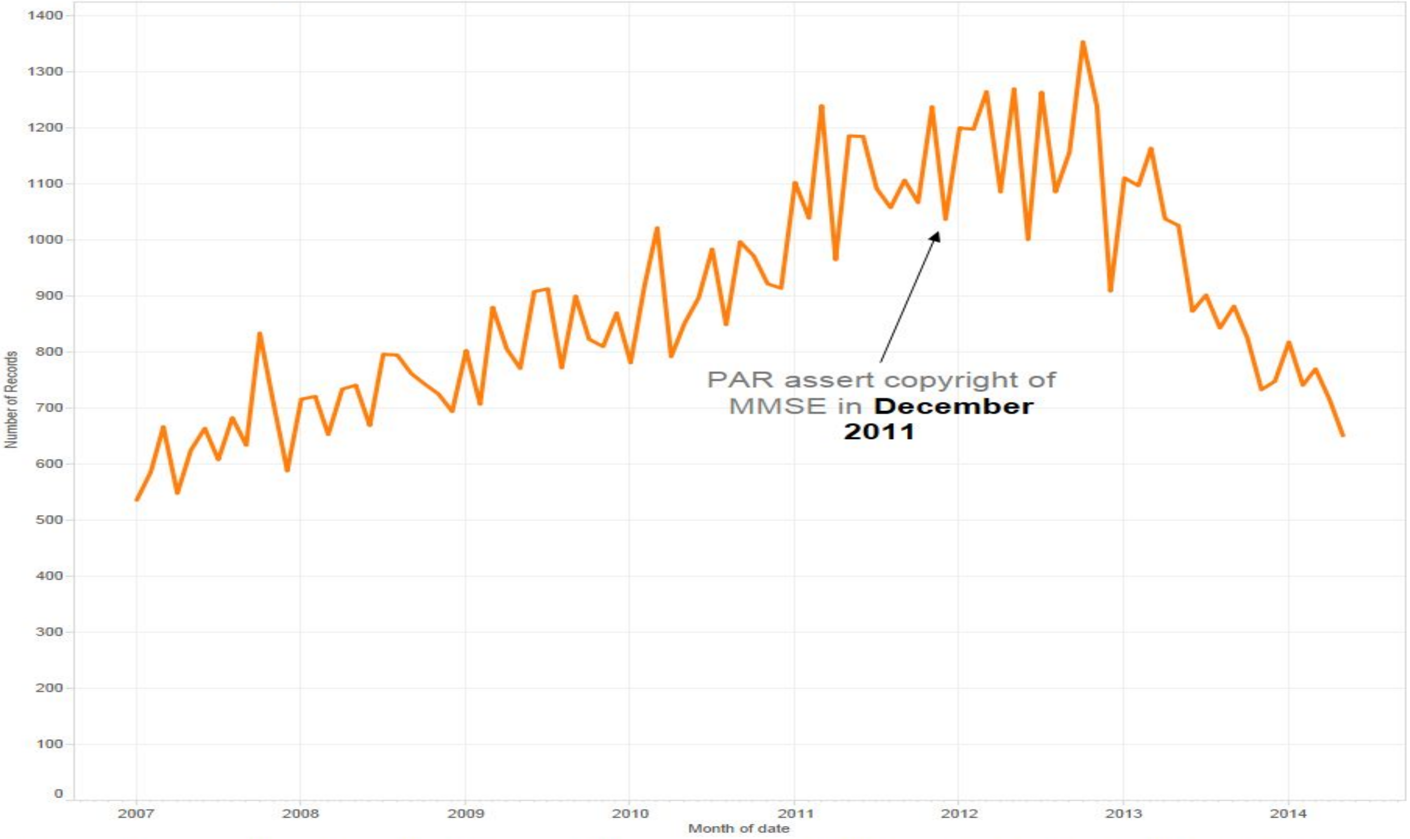


The trend of sum of Number of Records for date Month. Color shows details about word. The data is filtered on date, which includes the last 8 years. The filter associated with this field ranges from 01/01/2007 to 31/12/2014. The view is filtered on Exclusions (word,MONTH(date)) and word. The Exclusions (word,MONTH(date)) filter keeps 365 members. The word filter keeps mmse.

word

mmse

MMSE score information extraction rates 2007 - 2014

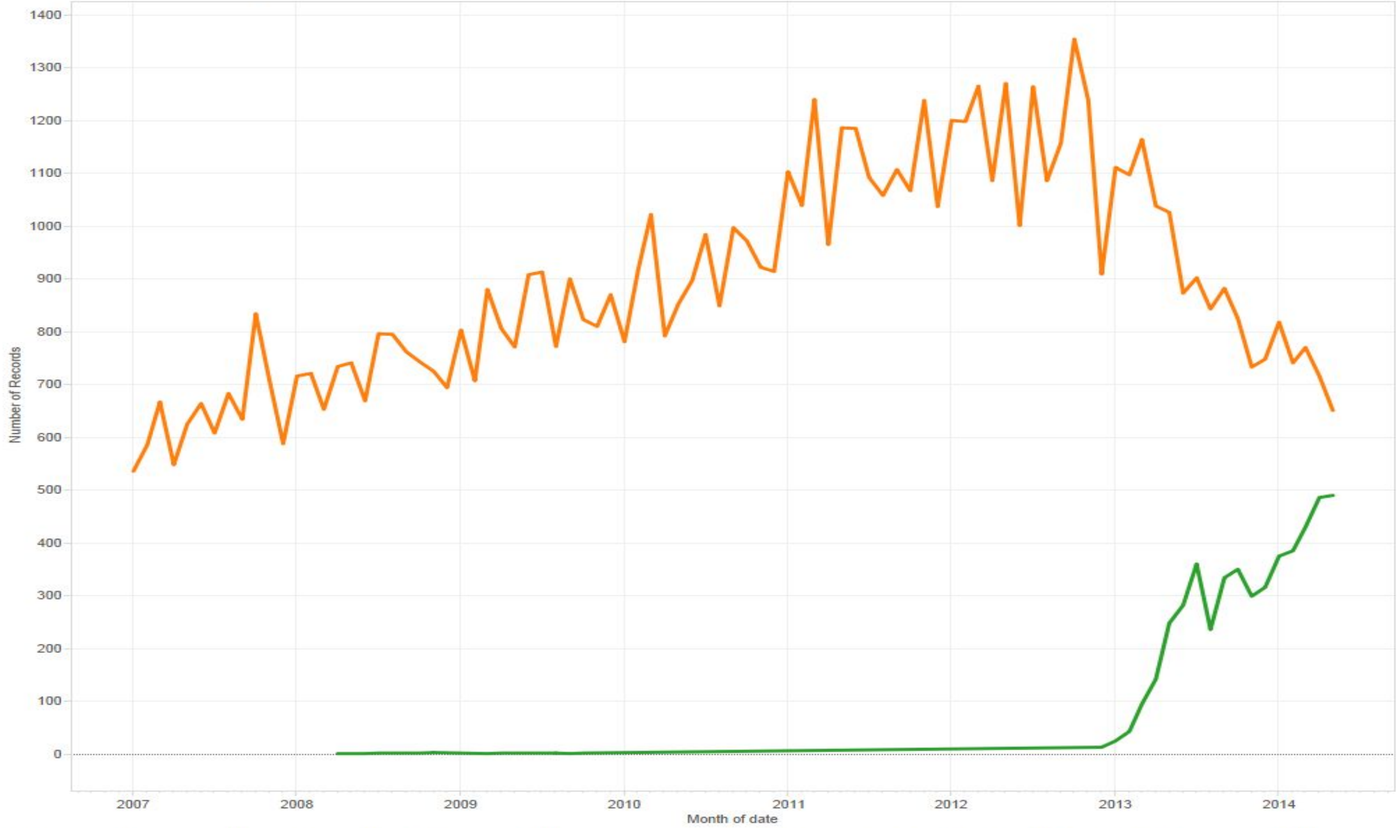


The trend of sum of Number of Records for date Month. Color shows details about word. The data is filtered on date, which includes the last 8 years. The filter associated with this field ranges from 01/01/2007 to 31/12/2014. The view is filtered on Exclusions (word,MONTH(date)) and word. The Exclusions (word,MONTH(date)) filter keeps 364 members. The word filter keeps mmse.

word

mmse

MMSE score information extraction rates 2007 - 2014

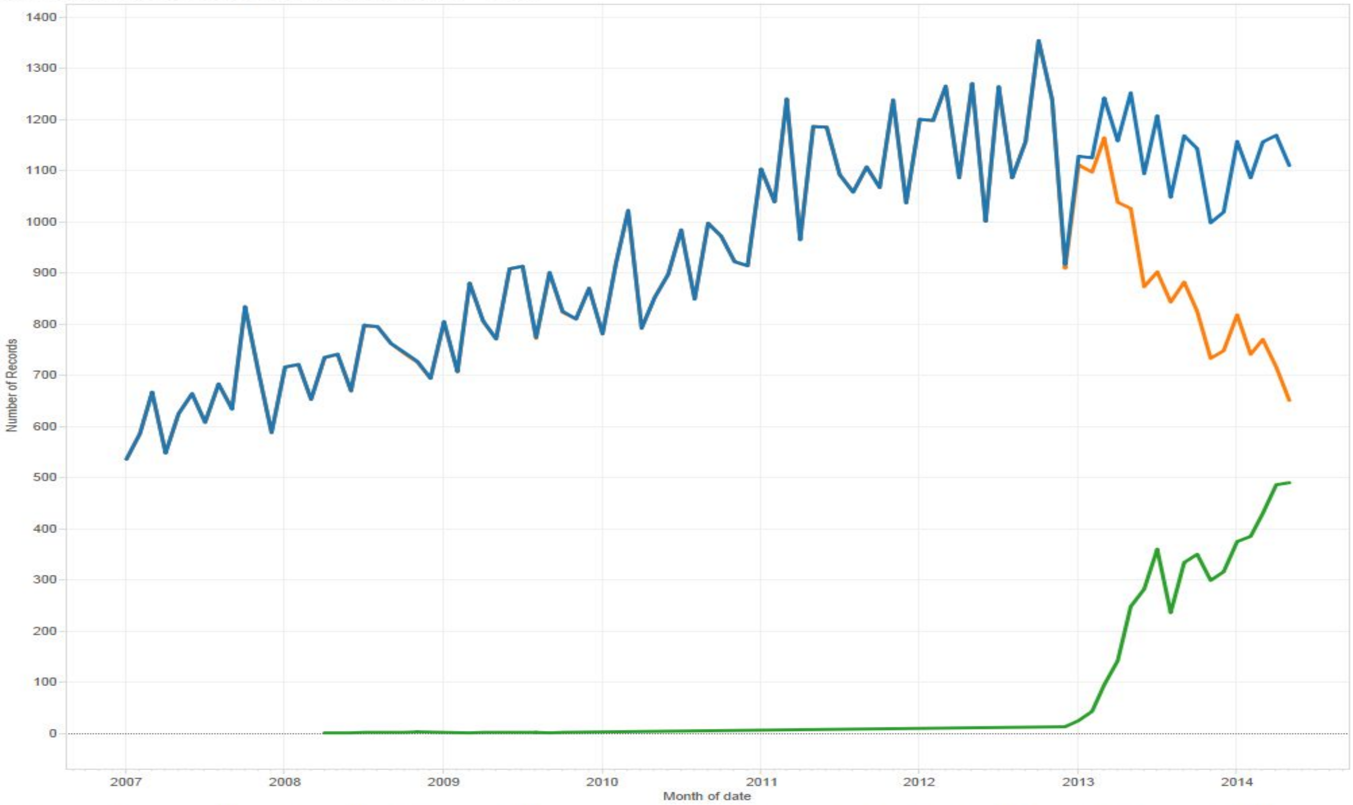


The trend of sum of Number of Records for date Month. Color shows details about word. The data is filtered on date, which includes the last 8 years. The filter associated with this field ranges from 01/01/2007 to 31/12/2014. The view is filtered on Exclusions (word,MONTH(date)) and word. The Exclusions (word,MONTH(date)) filter keeps 364 members. The word filter keeps mmse and smmse.

word

- mmse
- smmse

MMSE score information extraction rates 2007 - 2014



The trend of sum of Number of Records for date Month. Color shows details about word. The data is filtered on date, which includes the last 8 years. The filter associated with this field ranges from 01/01/2007 to 31/12/2014. The view is filtered on Exclusions (word,MONTH(date)) and word. The Exclusions (word,MONTH(date)) filter keeps 363 members. The word filter keeps both, mmse and smmse.

word
- both
- mmse
- smmse

Hard problems in clinical NLP

Temporality

context arises from

- document
- paragraph
- sentence
- unusual structures in sentences

Ontologies/nomenclatures

SNOMED/ICD10/UMLS etc

- Hard to adopt, even as structured sources
- need to map terminologies to ‘real world clinical language’
- language varies by region, even by hospital team!
- multiple levels of mapping possible. How much detail is needed?
- generally, practical solutions are adopted