

Using NLP in Clinical Data Analytics

Tomasz Oliwa, PhD | toliwa@bsd.uchicago.edu

Brian Furner | bfurner@bsd.uchicago.edu

Center for Research Informatics

Biological Sciences Division

University of Chicago



Agenda

1. Goals
2. Lifecycle of a clinical note
3. Natural language processing (NLP)
 - a. Named-entity recognition
 - b. Information extraction
 - c. Search
 - d. Classification
4. Contact
5. References

Goals

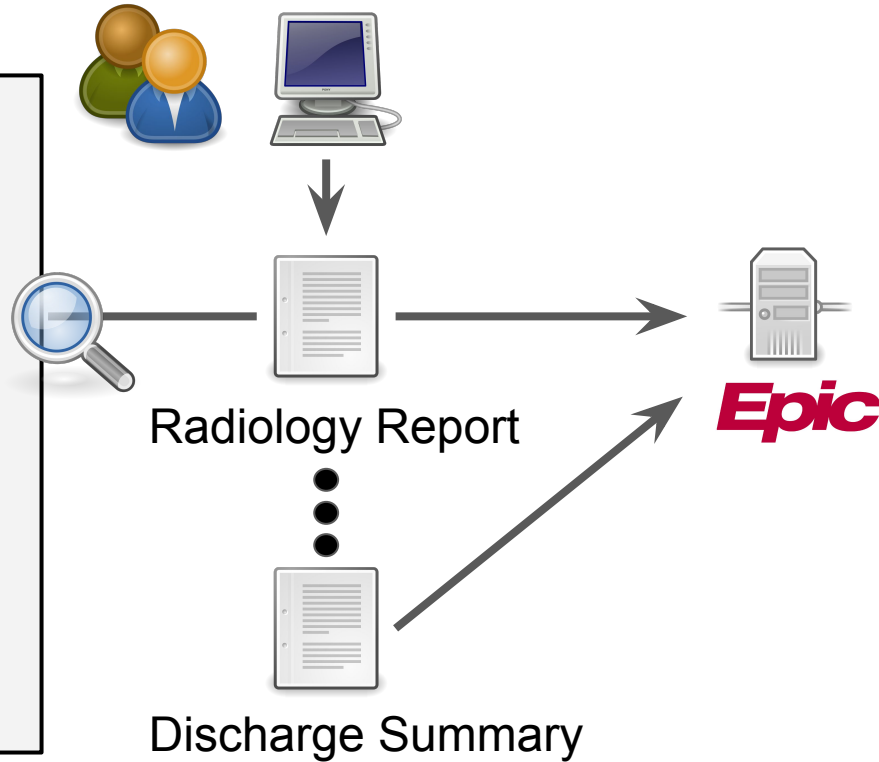
Explain NLP and what **types of problems** it can solve

Show **examples** of applied NLP for **clinical research / notes**

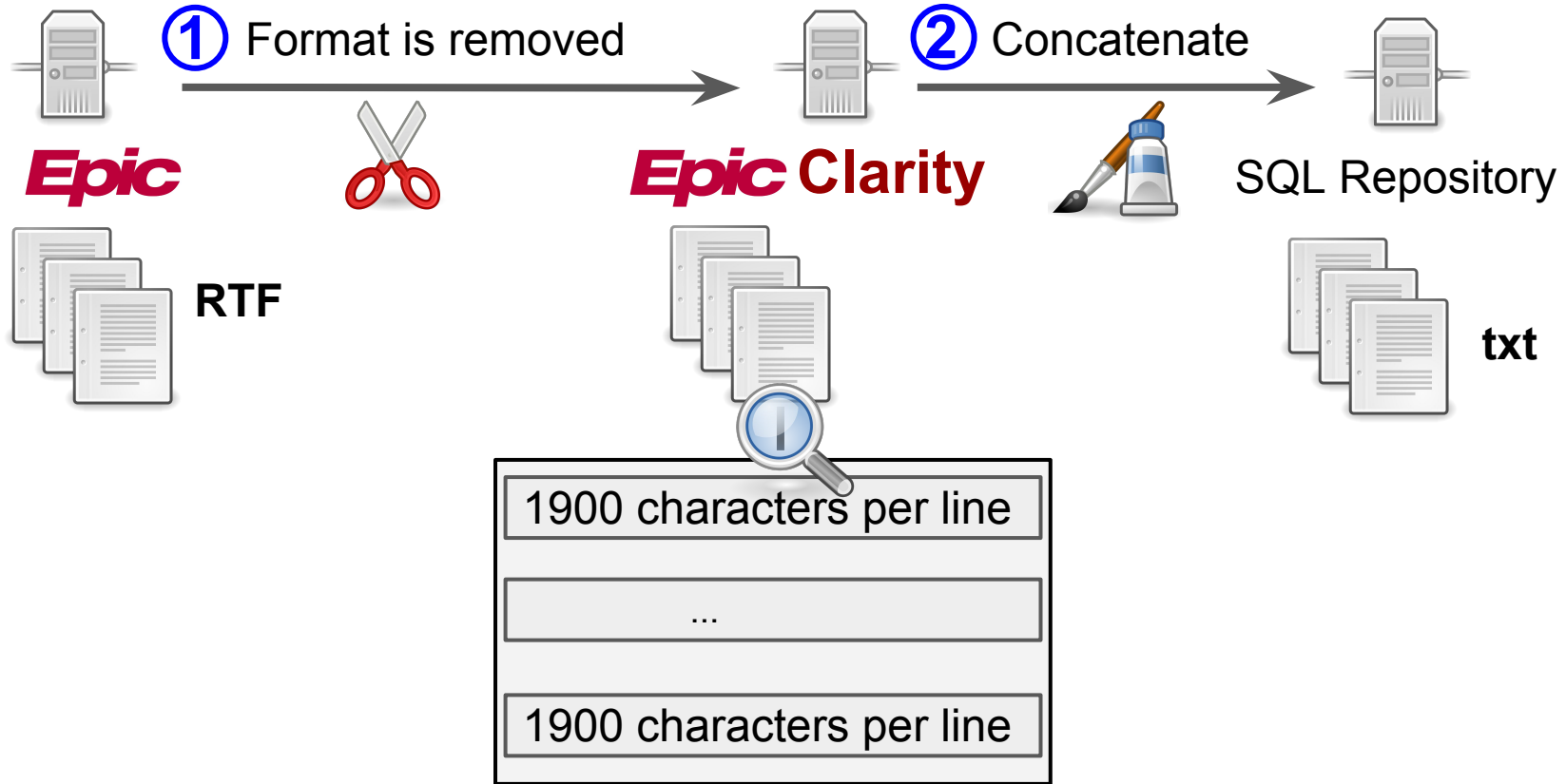
Provide an understanding of what **applications we can build** for you to **aid you in your research**

Lifecycle of a clinical note

Order Comment: ...
Exams: ...
CLINICAL DATA: ...
COMPARISON: ...
FINDINGS: ...
RESULT ID / ADDENDUM: ...
Ordering Physician: ...



Lifecycle of a clinical note



Natural language processing - NLP

Why NLP ?

Vast and **growing** number of **unstructured clinical notes**

NLP enables **computations** with natural (human) **languages**

- Retrieve **hidden information** and turn it into **knowledge**
- Harness **untapped** project-specific textual data sources

Named-entity recognition

Identify and **classify** words/phrases in unstructured text which might not all necessarily be known *a priori*

- **PHI** (**names, dates, locations, ...**)
- **Document sections**
- **Medical terms** (**diseases, symptoms, procedures, ...**)

Machine learning, rule-based, ontology-backed, hybrid

Named-entity recognition

Identify and classify

Synthetic note (fabricated patient info) named entities in text

History of Present Illness

Peter Miller is a 65 year old white male from New York with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

PAST MEDICAL HISTORY

Surgeries/procedures: Cardiac catheterization, post-MI, 10/11/2012 at Famous Hospital, RI [...]

History and Physical conducted by: Jeff York, MD

Named-entity recognition

Identify: PHI (names, locations ...)

Identify and classify
named entities in text

History of Present Illness

Peter Miller is a 65 year old white male from New York with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

PAST MEDICAL HISTORY

Surgeries/procedures: Cardiac catheterization, post-MI, 10/11/2012 at Famous Hospital, RI [...]

History and Physical conducted by: Jeff York, MD

Named-entity recognition

Identify and classify

Identify: PHI (names, locations ...) named entities in text

History of Present Illness

In practice, de-identification systems can mark all ages, not only > 89

Peter Miller is a **65** year old white male from **New York** with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

Not likely to be found in a dictionary

PAST MEDICAL HISTORY

Surgeries/procedures: Cardiac catheterization, post-MI, **10/11/2012** at

Famous Hospital RI [...]

History and Physical conducted by: **Jeff York** MD

Same token, different NE

Named-entity recognition

Redact named-entities

Identify and classify
named entities in text

History of Present Illness

****NAME<AAA>** is a ****AGE<in 60s>** year old white male from ****LOCATION** with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

PAST MEDICAL HISTORY

Surgeries/procedures: Cardiac catheterization, post-MI.

****DATE<[**2015-07-08**]>** at ****HOSPITAL** ****LOCATION** [...]

History and Physical conducted by: ****NAME<CCC>** MD

Named-entity recognition

Identify and classify
named entities in text

History of Present Illness

Peter Miller is a 65 year old white male from New York with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

PAST MEDICAL HISTORY

Surgeries/procedures: Cardiac catheterization, post-MI, 10/11/2012 at Famous Hospital, RI [...]

History and Physical conducted by: Jeff York, MD

Named-entity recognition

Identify: Document sections

Identify and **classify**
named entities in text

History of Present Illness

Peter Miller is a 65 year old white male from New York with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

PAST MEDICAL HISTORY

Surgeries/procedures: Cardiac catheterization, post-MI, 10/11/2012 at Famous Hospital, RI [...]

History and Physical conducted by: Jeff York, MD

Named-entity recognition

Identify: Document sections

Identify and classify named entities in text

History of Present Illness

Peter Miller is a 65 year old white male from New York with a past medical history significant for an MI and depression who presents today complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen. [...]

PAST MEDICAL HISTORY

Surgeries/procedures Cardiac catheterization, post-MI, 10/11/2012 at Famous Hospital, RI [...]

Ambiguity, NEs are project-specific

History and Physical conducted by Jeff York, MD

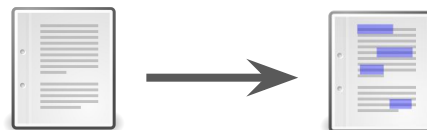
Named-entity recognition - Annotate

Browser based tool to accelerate the **project-specific** manual annotation process

The screenshot shows a web-based annotation tool interface. On the left, under the heading "Entity type", there is a list of entity categories: NAME, LOCATION, AGE, HOSPITAL, and DATE. Each category has a radio button next to it, and the "AGE" option is currently selected. A blue arrow points from the "AGE" option to the "65" in the text below. To the right of the list is a blue button labeled "New Annotation". Below this button is a text input field with the label "Text" and the value "65" entered, which is highlighted with a blue border.

3 Peter Miller is a 65 year old

The diagram illustrates the manual annotation process. The sentence "Peter Miller is a 65 year old" is shown with a vertical line on the left indicating the start of the text. "Peter Miller" is highlighted with a green box and labeled "NAME" in a yellow box above it. "65" is highlighted with a blue box, and a blue arrow points from this box to the "AGE" option in the "Entity type" list. A mouse cursor is shown pointing at the "65" box.



Named-entity recognition - Generate training data

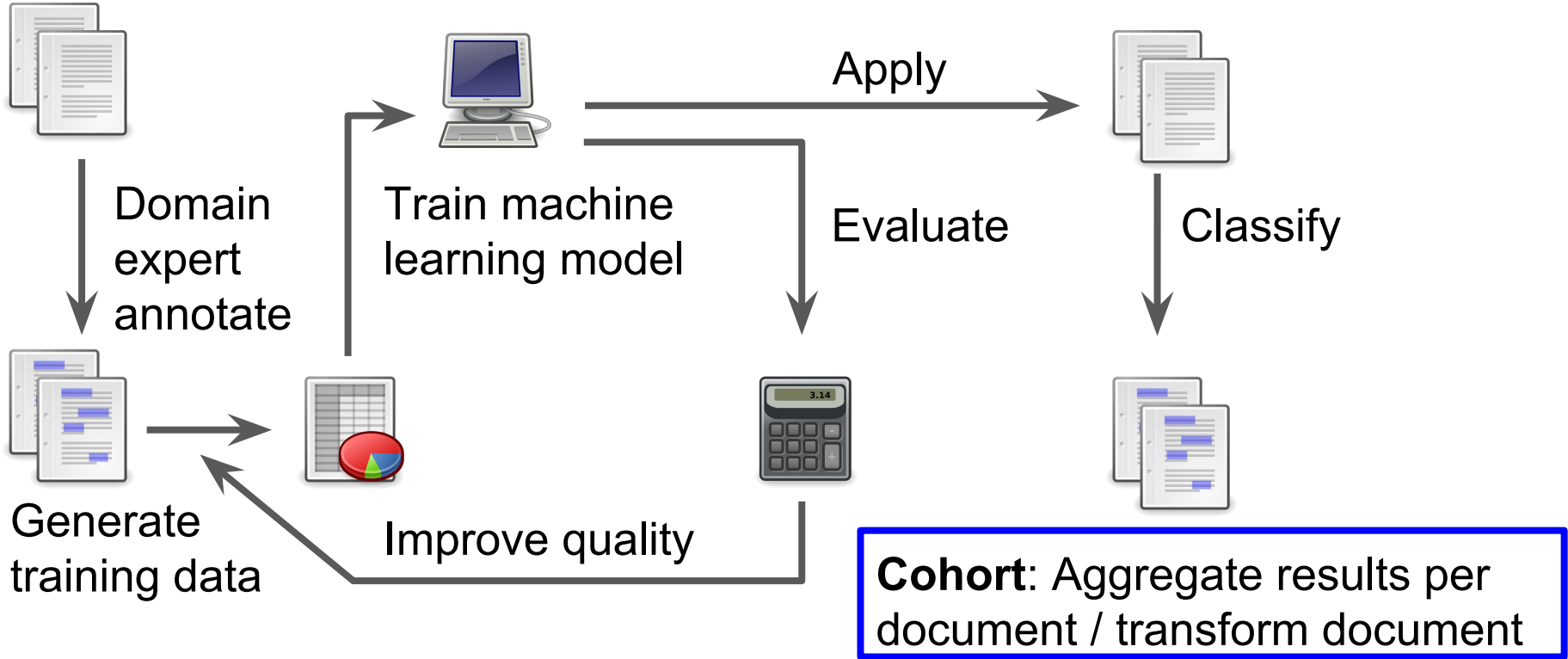
NER-TYPE	BEGIN END INDICES	NER-TEXT
NAME	28 40	Peter Miller
AGE	46 48	65
LOCATION	74 82	New York
DATE	410 420	10/11/2012
HOSPITAL	424 439	Famous Hospital
LOCATION	441 443	RI
NAME	487 496	Jeff York
...

Example features generated for each **token**

- The token itself
- The previous token
- The next token
- The last k-character token suffixes
- The first k-character token prefixes
- The token shape (upper or lower cases of each token character)
- The token gazetteer membership
- ...



Named-entity recognition - Train and apply



Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...

Synthetic note (fabricated patient info)

```
→ "begin":22,  
   "cui":["..."],  
→ "end":31,  
→ "highlight":"AL DIAGNOSIS: <em>Back pain</em>.  
   \n\nSECONDARY DI",  
   "isInsideLargerSameTypeMention":false,  
   "isPotentialAcronymFalsePositive":false,  
   "isSnomedCore":true,  
   "matchedText":"Back pain",  
   "matchedTextLength":9,  
   "mentionType":"SignSymptomMention",  
   "polarity":1,  
   "preferredText":"Back Pain",  
   "rxnorm":null,  
→ "section":"PRINCIPAL DIAGNOSIS",  
   "semGroup":"Sign or Symptom",  
   "snomedct_us":["..."],  
   "subject":"patient",  
   "synonyms":[  
   ],  
   "tui":["..."],
```

Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...



```
"begin":22,  
"cui":["..."],  
"end":31,  
"highlight":"AL DIAGNOSIS: <em>Back pain</em>.  
  \n\nSECONDARY DI",  
"isInsideLargerSameTypeMention":false,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":true,  
"matchedText":"Back pain",  
"matchedTextLength":9,  
"mentionType":"SignSymptomMention",  
"polarity":1,  
"preferredText":"Back Pain",  
"rxnorm":null,  
"section":"PRINCIPAL DIAGNOSIS",  
"semGroup":"Sign or Symptom",  
"snomedct_us":["..."],  
"subject":"patient",  
"synonyms":[  
],  
"tui":["..."],
```

unique ID for a metathesaurus
concept of UMLS

Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...

```
"begin":22,  
"cui":["..."],  
"end":31,  
"highlight":"AL DIAGNOSIS: <em>Back pain</em>.  
  \n\nSECONDARY DI",  
"isInsideLargerSameTypeMention":false,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":true,  
"matchedText":"Back pain",  
"matchedTextLength":9,  
"mentionType":"SignSymptomMention",  
"polarity":1,  
"preferredText":"Back Pain",  
"rxnorm":null,  
"section":"PRINCIPAL_DIAGNOSIS",  
"semGroup":"Sign or Symptom",  
"snomedct_us":["..."],  
"subject":"patient",  
"synonyms":[  
],  
"tui":["..."],
```

SNOMED Clinical Terms are
a collection of medical terms

Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...

```
"begin":22,  
"cui":["..."],  
"end":31,  
"highlight":"AL DIAGNOSIS: <em>Back pain</em>.  
  \n\nSECONDARY DI",  
"isInsideLargerSameTypeMention":false,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":true,  
"matchedText":"Back pain",  
"matchedTextLength":9,  
"mentionType":"SignSymptomMention",  
"polarity":1,  
"preferredText":"Back Pain",  
"rxnorm":null,  
"section":"PRINCIPAL_DIAGNOSIS",  
"semGroup":"Sign or Symptom",  
"snomedct_us":["..."],  
"subject":"patient",  
"synonyms":[  
],  
"tui":["..."],
```



CUIs can be assigned a
semantic group / tui code

Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...

```
"begin":27,  
"cui":["..."],  
"end":31,  
"highlight":"AGNOSIS: Back <em>pain</em>.  
  \n\nSECONDARY DI",  
"isInsideLargerSameTypeMention":true,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":true,  
"matchedText":"pain",  
"matchedTextLength":4,  
"mentionType":"SignSymptomMention",  
"polarity":1,  
"preferredText":"Pain",  
"rxnorm":null,  
"section":"PRINCIPAL DIAGNOSIS",  
"semGroup":"Sign or Symptom",  
"snomedct_us":["..."],  
"subject":"patient",  
"synonyms":[  
],  
"tui":["..."],
```

Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...

```
"begin":126,  
"cui":["..."],  
"end":129,  
"highlight":"est x-ray. 2. <em>MRI</em>  
of cervical ",  
"isInsideLargerSameTypeMention":true,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":false,  
"matchedText":"MRI",  
"matchedTextLength":3,  
"mentionType":"ProcedureMention",  
"polarity":1,  
"preferredText":"Magnetic Resonance Imaging",  
"rxnorm":null,  
"section":"PROCEDURES_PERFORMED",  
"semGroup":"Diagnostic Procedure",  
"snomedct_us":["..."],  
"subject":"patient",  
"synonyms":[  
],  
"tui":["..."],
```

Named-entity recognition - Ontologies and cTAKES

PRINCIPAL DIAGNOSIS: Back pain.

SECONDARY DIAGNOSIS: 1.
Gastroesophageal reflux disease.

PROCEDURES: 1. Chest x-ray. 2. MRI
of cervical spine.

FAMILY HISTORY: Father heart attack,
mom with hypertension and depression,
history of asthma on both sides of the
family.

PHYSICAL FINDINGS: Temperature
36.7, ...

```
"begin":199,  
"cui":["..."],  
"end":211,  
"highlight":"tack, mom with  
    <em>hypertension</em> and depression",  
"isInsideLargerSameTypeMention":false,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":true,  
"matchedText":"hypertension",  
"matchedTextLength":12,  
"mentionType":"DiseaseDisorderMention",  
"polarity":1,  
"preferredText":"Hypertensive disease",  
"rxnorm":null,  
"section":"FAMILY_HISTORY",  
"semGroup":"Disease or Syndrome",  
"snomedct_us":["..."],  
"subject":"family_member",  
"synonyms":[  
],  
"tui":["..."],
```


Named-entity recognition - Ontologies and cTAKES

IMPRESSION: No specific abnormality
... and no definite evidence of
pancreatitis

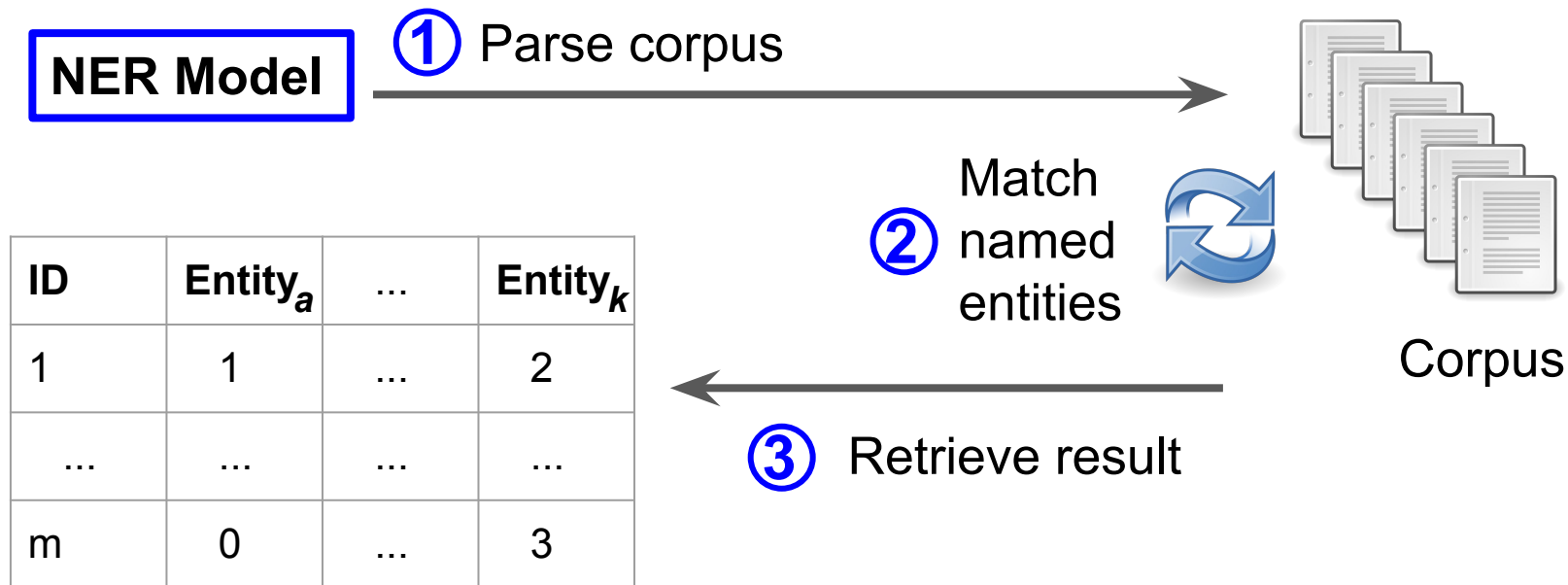
Synthetic note (fabricated patient info)

```
"begin":68,  
"cui":["..."],  
"end":80,  
"highlight":"te evidence of  
    <em>pancreatitis</em>.\n",  
"isInsideLargerSameTypeMention":false,  
"isPotentialAcronymFalsePositive":false,  
"isSnomedCore":true,  
"matchedText":"pancreatitis",  
"matchedTextLength":12,  
"mentionType":"DiseaseDisorderMention",  
"polarity":-1,  
"preferredText":"Pancreatitis",  
"rxnorm":null,  
"section":"IMPRESSION",  
"semGroup":"Disease or Syndrome",  
"snomedct_us":["..."],  
"subject":"patient",  
"synonyms":[  
],  
"tui":["..."],
```



Information Extraction

Generate document-entity count matrix from a corpus



Information Extraction - Rule-based example

Rule-based approach for **risk-factor** extraction

List of risk-factors



Clinical documents

Regex
NER

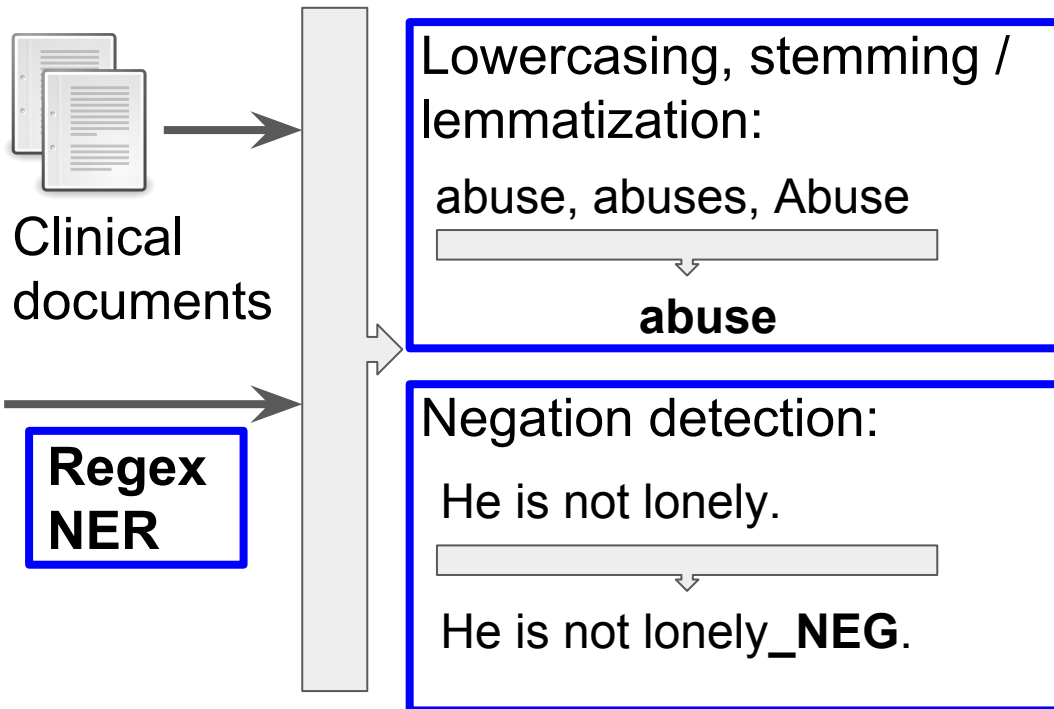
Social factors	Substance abuse	...
jail	alcohol abuse	...
lonely	(use abuse) ... substance	...
financial assistance	alcoholic	...
isolation

Information Extraction - Rule-based example

Rule-based approach for **risk-factor** extraction:

List of risk-factors

Social factors	Substance abuse	...
jail	alcohol abuse	...
lonely	(use abuse) ... substance	...
financial assistance	alcoholic	...
isolation



Information Extraction - Rule-based example



Token matching **pitfalls**:

List of risk-factors

Social factors	Medical factors	Others
aid	AIDS	hearing aids



Clinical documents



Lowercasing, stemming:
Looking to match financial aids but specifying only **aid**, falsely match AIDS.

Regex NER

Lowercasing, stemming / lemmatization:

AIDS, the acquired immune deficiency syndrome

... **aid** ...

Financial **aids** ...

Information Extraction - Rule-based example



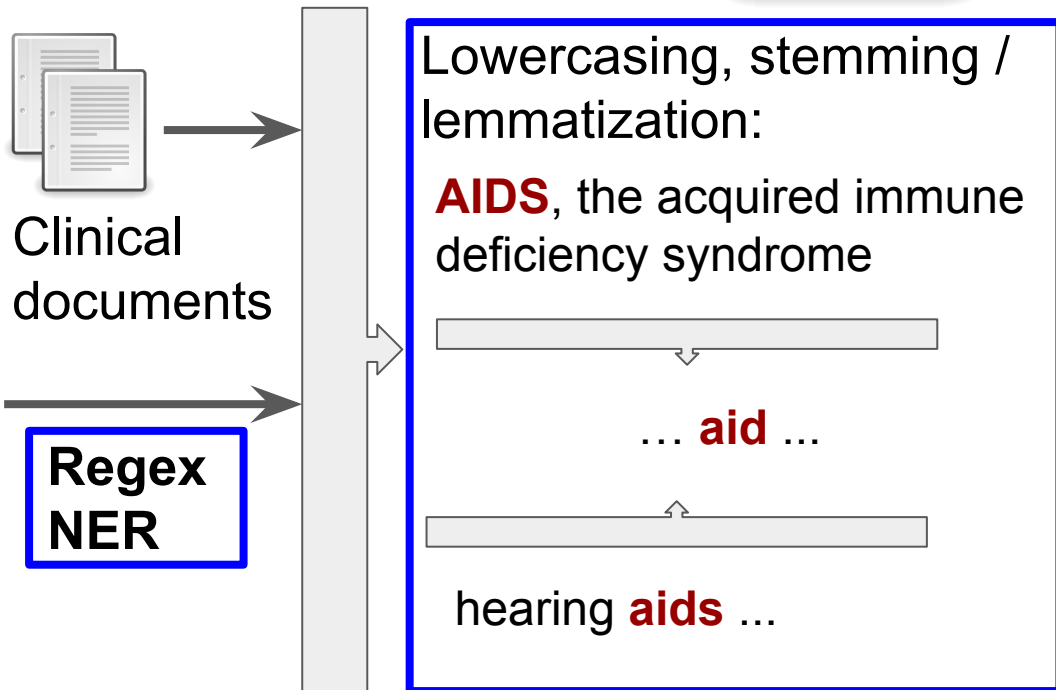
Token matching **pitfalls**:

List of risk-factors

Social factors	Medical factors	Others
aid	AIDS	hearing aids




Lowercasing: Looking to match **AIDS** and falsely match hearing aids.





Information Extraction - Rule-based example

Document-term/phrase **risk-factor** matrix:

Document-ID	Text	abuse alcohol	abuse_NEG alcohol_NEG	...
1		2	0	...
2		0	1	...
3		0	1	...
...

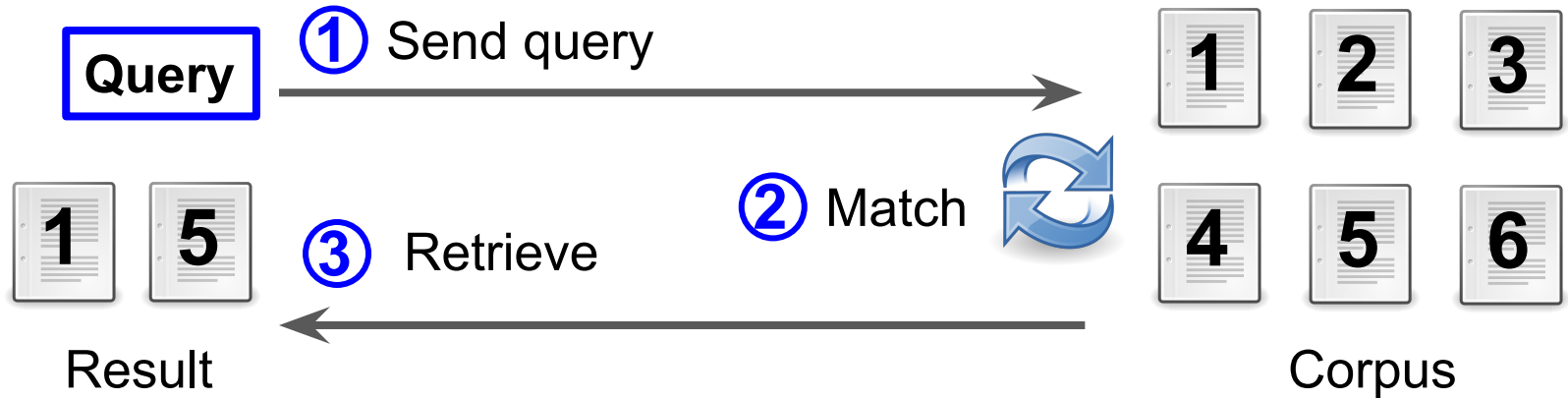
Information Extraction - Hybrid example with NER

Document-term/phrase **risk-factor** matrix:

Document-ID	Text	abuse alcohol	abuse_NEG alcohol_NEG	...	Named entities of class-1
1		2	0	...	0
2		0	1	...	2
3		0	1	...	0
...

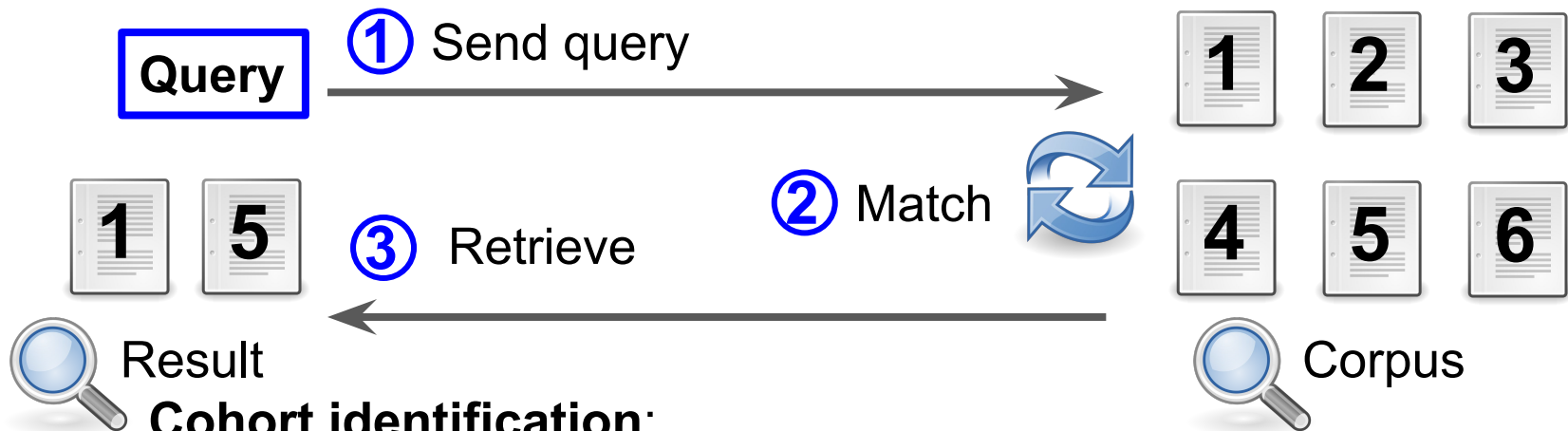
Search

Query arbitrary corpus to **retrieve matching** documents



Search

Query arbitrary corpus to **retrieve matching** documents



Cohort identification:

- Full document text
- Anonymized patient ID
- Associated discrete attributes

- Radiology reports
- Discharge summaries
- External project-specific corpus

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- Concept search

Search

Query arbitrary corpus to **retrieve matching** documents

Synthetic note (fabricated patient info)

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- Concept search

History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- **Token match**
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- Concept search



Query: **pain**

History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months duration. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- **Token match**
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- Concept search



History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Query: leg

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- **Boolean operators**
- Proximity phrase match
- Word stem match
- Section search
- Concept search



Query: pain **AND** epigastric

History of Present Illness
Pt is complaining of sharp, **epigastric** abdominal **pain** of 3-4 months duration. The **pain** is located in the **epigastric** region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- **Boolean operators**
- Proximity phrase match
- Word stem match
- Section search
- Concept search



History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months duration. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Query: pain **AND** leg

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- **Boolean operators**
- Proximity phrase match
- Word stem match
- Section search
- Concept search



Query: pain **OR** leg

History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months duration. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- **Boolean operators**
- Proximity phrase match
- Word stem match
- Section search
- Concept search



History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months duration. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Query: pain **AND** (**NOT** leg)



**Absence, not negation of
“leg”**

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- **Boolean operators**
- Proximity phrase match
- Word stem match
- Section search
- Concept search



History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months **duration**. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Query: pain **AND** (**NOT** duration)

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity **phrase match**
- Word stem match
- Section search
- Concept search



Query: “**abdominal pain**”

History of Present Illness
Pt is complaining of sharp, epigastric **abdominal pain** of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- Concept search



Query: “abdominal pain” AND region

History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity **phrase match**
- Word stem match
- Section search
- Concept search



Query: “epigastric pain”

History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- Concept search



Query: “epigastric pain”~1

History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- **Word stem match**
- Section search
- Concept search



Query: complaining

History of Present Illness
Pt is **complaining** of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- **Word stem match**
- Section search
- Concept search



Query: complain

History of Present Illness
Pt is **complaining** of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- **Word stem match**
- Section search
- Concept search



History of Present Illness
Pt is **complaining** of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Query: complains

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word **stem** match
- **Section search**
- Concept search



History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: **Cardiac** catheterization.

Query:

PAST_MEDICAL_HISTORY:
cardiac AND surgery

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- **Section search**
- Concept search



History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Query:

History_of_Present_Illness:
cardiac AND surgery

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- **Concept search**

Query: pain AND Age:[40 TO 60] AND Gender:Male



Patient
Data

History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months duration. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- **Concept search**

Query: pain AND Age:[20 TO 30] AND Gender:Male



Patient
Data

History of Present Illness
Pt is complaining of sharp, epigastric abdominal **pain** of 3-4 months duration. The **pain** is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: Cardiac catheterization.

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- **Concept search**



History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: **Cardiac** catheterization.

Query: cardiac OR heart OR coronary OR cor

UMLS
hand-craft

Search

Query arbitrary corpus to **retrieve matching** documents

- Token match
- Boolean operators
- Proximity phrase match
- Word stem match
- Section search
- **Concept search**



History of Present Illness
Pt is complaining of sharp, epigastric abdominal pain of 3-4 months duration. The pain is located in the epigastric region and left upper quadrant of the abdomen.

PAST MEDICAL HISTORY
Surgeries/procedures: **Cardiac** catheterization.

Query: concepts:"C0018787"

UMLS
cTAKES

Search

Query arbitrary corpus to **retrieve matching** documents

- Custom-build **browser-based search applications**
- Allow **interactive** data-driven exploration of project texts
- **Expose** NLP and machine learning results for **queries**

Search - Web Demo

ISEAR corpus - Survey to report situations of **emotions**:

- joy, fear, anger, sadness, disgust, shame, and guilt

Additionally reported:

- intensity, ergotropic arousal, coping, expected, fairness, ...

Id	Text	Emotion	Intensity	...
1	I am the secretary of an association, and during the last meeting I forgot to take the minutes.	guilt	2	...
2	Walking in the dark and thinking about ghost stories.	fear	3	...
3	Fighting with my father while drunk.	shame	4	...
...

Search - Web Demo

SEARCH QUERY

Q+

TOTAL HITS HITS

7,566

TEXT DOCVIEWER

90 of 100 available for paging

← Title →

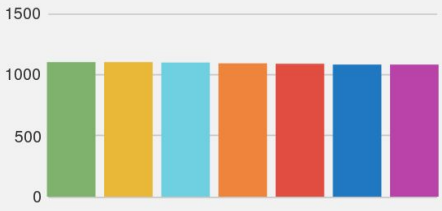
fear

Content

When biking and I felt very bad (problems with heart and respiration).

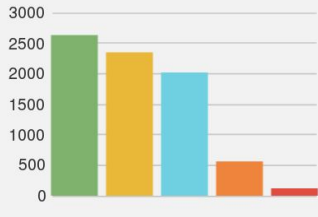
EMOTIONS TERMS

- joy (1,090)
- fear (1,090)
- anger (1,087)
- sadness (1,080)
- disgust (1,078)
- guilt (1,071)
- shame (1,070)



INTENSITY TERMS

- 3 (2,607)
- 4 (2,327)
- 2 (1,997)
- 1 (538)
- 0 (97)

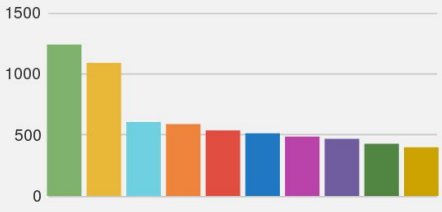


Filtering FILTERING

No filters available

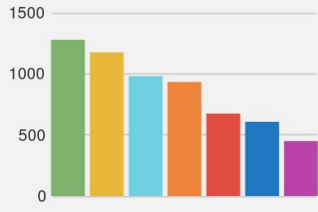
TOP FACETS TERMS

- friend (1,229)
- feel (1,079)
- time (594)
- go (578)
- get (526)
- tell (501)
- come (475)
- day (457)
- see (416)
- year (388)



TOP AGE TERMS

- 20 (1,268)
- 21 (1,164)
- 22 (970)
- 19 (923)
- 18 (663)
- 23 (596)
- 24 (438)



Search - Web Demo

The screenshot displays a search interface with several components:

- SEARCH:** A search bar with a green dot and a search icon.
- TEXT (DOCVIEWER):** A document viewer showing a document with a title and content. The title is "Title" and the content is "When biking and I felt very bad (problems with heart and respiration)". The viewer is highlighted with a blue border and has left and right navigation arrows.
- EMOTIONS:** A bar chart showing the frequency of various emotions. The legend includes: joy (1,090), fear (1,090), anger (1,087), sadness (1,080), disgust (1,078), guilt (1,071), and shame (1,070).
- INTENSITY:** A bar chart showing the frequency of different intensity levels. The legend includes: 3 (2,607), 4 (2,327), 2 (1,997), 1 (538), and 0 (97).
- TOP FACETS:** A bar chart showing the frequency of various terms. The legend includes: friend (1,229), feel (1,079), time (594), go (578), get (526), tell (501), come (475), day (457), see (416), and year (388).
- TOTAL HITS:** A summary bar showing 7,566 total hits.
- TOP AGE:** A bar chart showing the frequency of different age groups. The legend includes: 20 (1,268), 21 (1,164), 22 (970), 19 (923), 18 (663), 23 (596), and 24 (438).
- FILTERING:** A section indicating "No filters available".

Search - Web Demo

SEARCH QUERY

*** +

TOTAL HITS HITS

7,566

TEXT DOCVIEWER

90 of 100 available for paging

← Title →

fear

Content

When biking and I felt very bad (problems with heart and respiration).

EMOTIONS TERMS

- joy (1,090)
- fear (1,090)
- anger (1,087)
- sadness (1,080)
- disgust (1,078)
- guilt (1,071)
- shame (1,070)

Emotion	Count
joy	1,090
fear	1,090
anger	1,087
sadness	1,080
disgust	1,078
guilt	1,071
shame	1,070

INTENSITY TERMS

- 3 (2,607)
- 4 (2,327)
- 2 (1,997)
- 1 (538)
- 0 (97)

Intensity	Count
3	2,607
4	2,327
2	1,997
1	538
0	97

Filtering FILTERING

No filters available

TOP FACETS TERMS

- friend (1,229)
- feel (1,079)
- time (594)
- go (578)
- get (526)
- tell (501)
- come (475)
- day (457)
- see (416)
- year (388)

Facet	Count
friend	1,229
feel	1,079
time	594
go	578
get	526
tell	501
come	475
day	457
see	416
year	388

TOP AGE TERMS

- 20 (1,268)
- 21 (1,164)
- 22 (970)
- 19 (923)
- 18 (663)
- 23 (596)
- 24 (438)

Age	Count
20	1,268
21	1,164
22	970
19	923
18	663
23	596
24	438

Search - Web Demo

SEARCH QUERY

*** Q+

TOTAL HITS HITS

7,566

TEXT DOCVIEWER

90 of 100 available for paging

← Title →

fear

Content

When biking and I felt very bad (problems with heart and respiration).

EMOTIONS TERMS

- joy (1,090)
- fear (1,090)
- anger (1,087)
- sadness (1,080)
- disgust (1,078)
- guilt (1,071)
- shame (1,070)

Emotion	Count
joy	1,090
fear	1,090
anger	1,087
sadness	1,080
disgust	1,078
guilt	1,071
shame	1,070

INTENSITY TERMS

- 3 (2,607)
- 4 (2,327)
- 2 (1,997)
- 1 (538)
- 0 (97)

Intensity	Count
3	2,607
4	2,327
2	1,997
1	538
0	97

TOP FACETS TERMS

- friend (1,229)
- feel (1,079)
- time (594)
- go (578)
- get (526)
- tell (501)
- come (475)
- day (457)
- see (416)
- year (388)

Facet	Count
friend	1,229
feel	1,079
time	594
go	578
get	526
tell	501
come	475
day	457
see	416
year	388

NO FILTERS AVAILABLE FILTERING

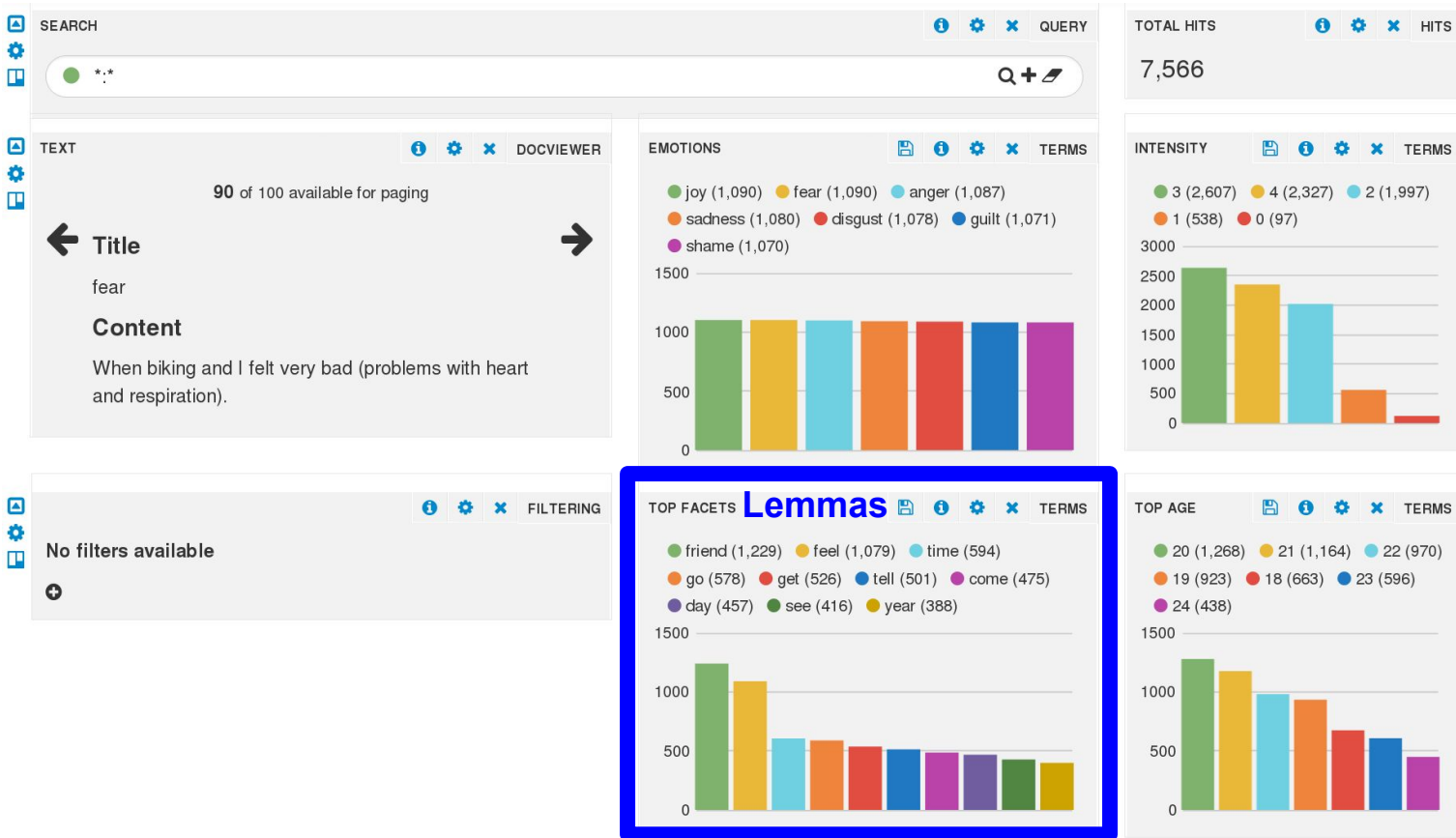
+

TOP AGE TERMS

- 20 (1,268)
- 21 (1,164)
- 22 (970)
- 19 (923)
- 18 (663)
- 23 (596)
- 24 (438)

Age	Count
20	1,268
21	1,164
22	970
19	923
18	663
23	596
24	438

Search - Web Demo



Search - Web Demo

SEARCH

SEARCH

7,566

TEXT

90 of 100 available for paging

Title

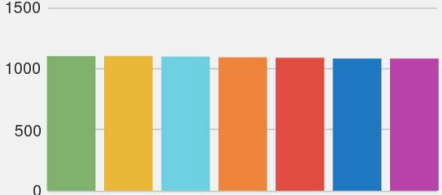
fear

Content

When biking and I felt very bad (problems with heart and respiration).

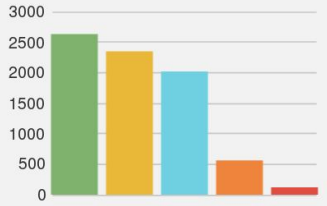
EMOTIONS

joy (1,090) fear (1,090) anger (1,087)
sadness (1,080) disgust (1,078) guilt (1,071)
shame (1,070)



INTENSITY

3 (2,607) 4 (2,327) 2 (1,997)
1 (538) 0 (97)

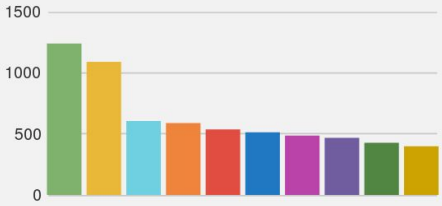


FILTERING

No filters available

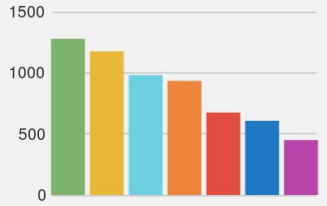
TOP FACETS

friend (1,229) feel (1,079) time (594)
go (578) get (526) tell (501) come (475)
day (457) see (416) year (388)

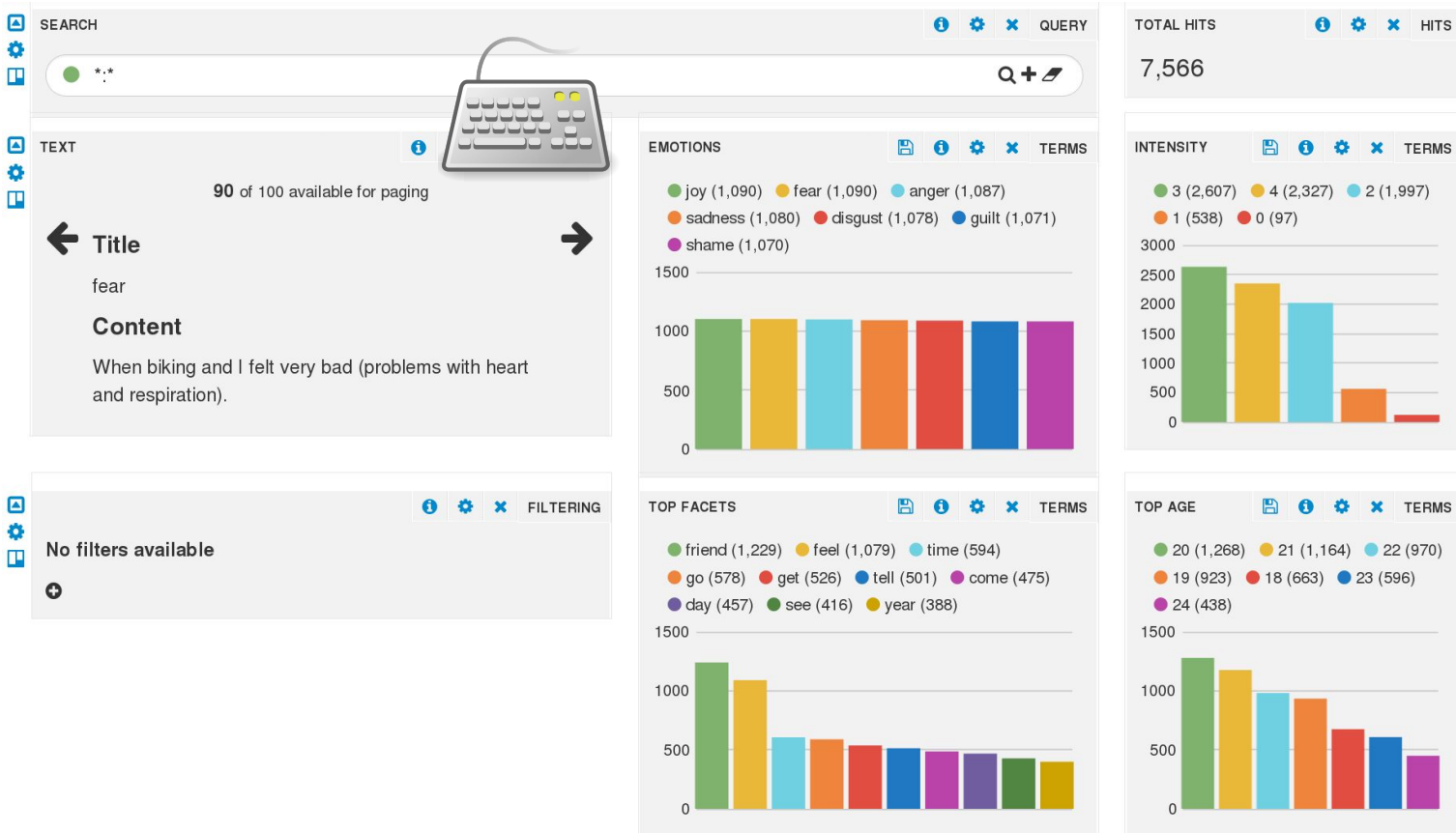


TOP AGE

20 (1,268) 21 (1,164) 22 (970)
19 (923) 18 (663) 23 (596)
24 (438)



Search - Web Demo



Search - Web Demo

SEARCH (afraid OR fear OR scared OR threatened)

TEXT 90 of 100 available for paging

Title

fear

Content

When biking and I felt very bad (problems with heart and respiration).

EMOTIONS

- joy (1,090)
- fear (1,090)
- anger (1,087)
- sadness (1,080)
- disgust (1,078)
- guilt (1,071)
- shame (1,070)

Emotion	Count
joy	1,090
fear	1,090
anger	1,087
sadness	1,080
disgust	1,078
guilt	1,071
shame	1,070

INTENSITY

- 3 (2,607)
- 4 (2,327)
- 2 (1,997)
- 1 (538)
- 0 (97)

Intensity	Count
3	2,607
4	2,327
2	1,997
1	538
0	97

TOP FACETS

- friend (1,229)
- feel (1,079)
- time (594)
- go (578)
- get (526)
- tell (501)
- come (475)
- day (457)
- see (416)
- year (388)

Facet	Count
friend	1,229
feel	1,079
time	594
go	578
get	526
tell	501
come	475
day	457
see	416
year	388

TOTAL HITS 7,566

INTENSITY

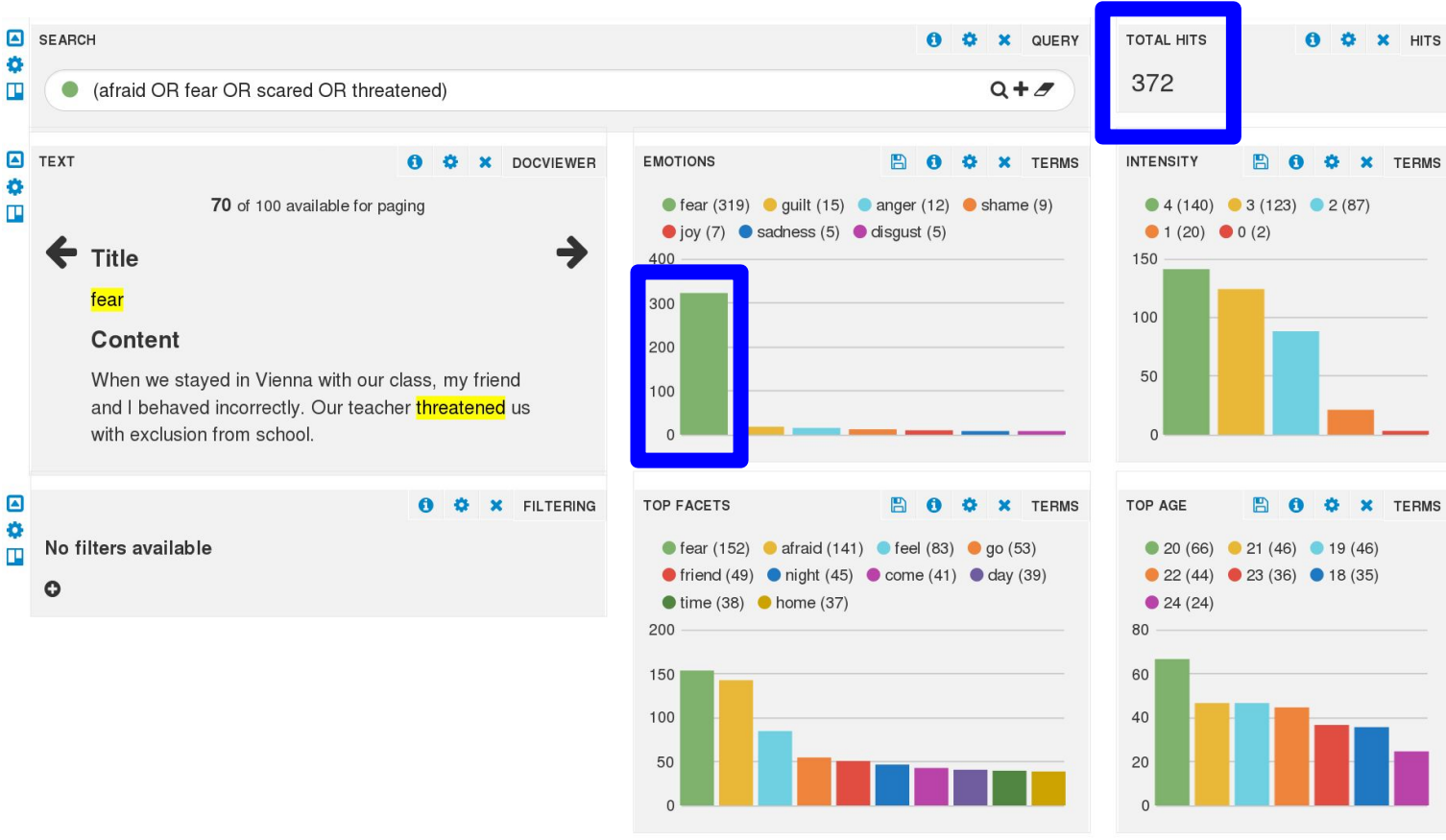
TOP AGE

- 20 (1,268)
- 21 (1,164)
- 22 (970)
- 19 (923)
- 18 (663)
- 23 (596)
- 24 (438)

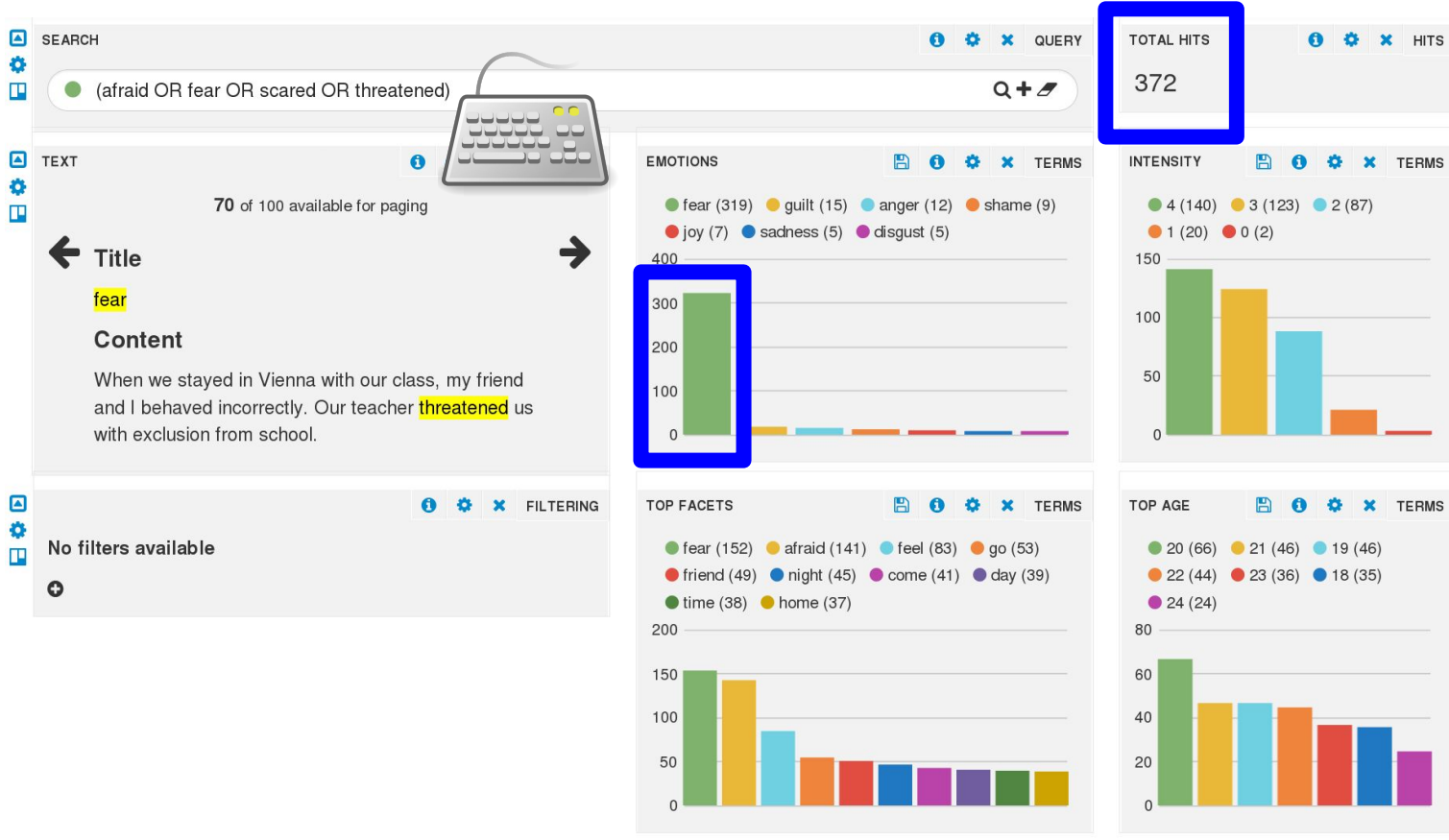
Age	Count
20	1,268
21	1,164
22	970
19	923
18	663
23	596
24	438

FILTERING No filters available

Search - Web Demo



Search - Web Demo



Search - Web Demo

SEARCH

SEARCH QUERY: (afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]

TOTAL HITS

372

TEXT

70 of 100 available for paging

Title

Content

When we stayed in Vienna with our class, my friend and I behaved incorrectly. Our teacher **threatened** us with exclusion from school.

EMOTIONS

Bar chart showing counts for various emotions:

Emotion	Count
fear	319
guilt	15
anger	12
shame	9
joy	7
sadness	5
disgust	5

INTENSITY

Bar chart showing counts for various intensity levels:

Intensity	Count
4	140
3	123
2	87
1	20
0	2

FILTERING

No filters available

TOP FACETS

Bar chart showing counts for various facets:

Facet	Count
fear	152
afraid	141
feel	83
go	53
friend	49
night	45
come	41
day	39
time	38
home	37

TOP AGE

Bar chart showing counts for various age groups:

Age	Count
20	66
21	46
19	46
22	44
23	36
18	35
24	24

Search - Web Demo

SEARCH QUERY **TOTAL HITS** HITS

(afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27] Q+

TEXT DOCVIEWER

20 of 100 available for paging

Title

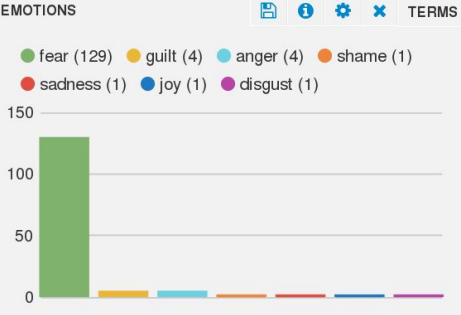
fear

Content

When my brother had an epileptic attack and I was scared as to what would happen to him.


EMOTIONS TERMS

- fear (129)
- guilt (4)
- anger (4)
- shame (1)
- sadness (1)
- joy (1)
- disgust (1)



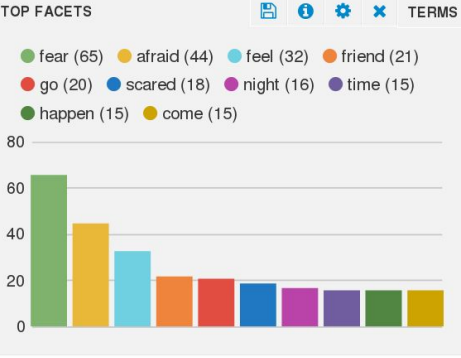
INTENSITY TERMS

- 4 (56)
- 3 (50)
- 2 (25)
- 1 (10)



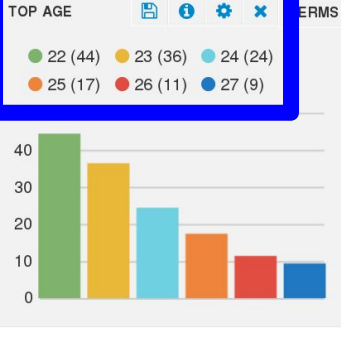
TOP FACETS TERMS

- fear (65)
- afraid (44)
- feel (32)
- friend (21)
- go (20)
- scared (18)
- night (16)
- time (15)
- happen (15)
- come (15)



TOP AGE TERMS

- 22 (44)
- 23 (36)
- 24 (24)
- 25 (17)
- 26 (11)
- 27 (9)



FILTERING

No filters available

Search - Web Demo

The screenshot displays a search interface with several panels:

- SEARCH:** Query: (afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]. Total Hits: 141.
- TEXT:** Document viewer showing a snippet with the word "fear" highlighted in yellow. The content reads: "When my brother had an epileptic attack and I was scared as to what would happen to him."
- EMOTIONS:** Bar chart showing the distribution of emotions. Legend: fear (129), guilt (4), anger (4), shame (1), sadness (1), joy (1), disgust (1).
- INTENSITY:** Bar chart showing the distribution of intensity levels. Legend: 4 (56), 3 (50), 2 (25), 1 (10).
- TOP FACETS:** Bar chart showing the distribution of top facets. Legend: fear (65), afraid (44), feel (32), friend (21), go (20), scared (18), night (16), time (15), happen (15), come (15).
- TOP AGE:** Bar chart showing the distribution of top age groups. Legend: 22 (44), 23 (36), 24 (24), 25 (17), 26 (11), 27 (9).

A mouse cursor is visible over the TOP FACETS chart.

Search - Web Demo

SEARCH

QUERY: (afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]

TEXT

20 of 100 available for paging

Title

fear

Content

When my brother had an epileptic attack and I was scared as to what would happen to him.

EMOTIONS

Emotion	Count
fear	129
guilt	4
anger	4
shame	1
sadness	1
joy	1
disgust	1

INTENSITY

Intensity	Count
4	56
3	50
2	25
1	10

TOP FACETS

Facet	Count
fear	65
afraid	44
feel	32
friend	21
go	20
scared	18
night	16
time	15
happen	15
come	15

TOTAL HITS

141

INTENSITY

Intensity	Count
4	56
3	50
2	25
1	10

TOP AGE

Age	Count
22	44
23	36
24	24
25	17
26	11
27	9

FILTERING

No filters available

EMOTIONS

feel (32)

Search - Web Demo

SEARCH (afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27] **TOTAL HITS** 32

TEXT 11 of 32 available for paging

Title

Content **Lemma**
Fear of the loss of a close friend, of feeling the ground slipping from under my feet.

EMOTIONS

Emotion	Count
fear	28
guilt	2
sadness	1
anger	1

INTENSITY

Intensity	Count
3	11
4	9
2	9
1	3

TOP FACETS

Facet	Count
feel	32
fear	20
afraid	9
happen	6
scared	4
night	4
friend	4
father	4
fail	4
child	3

TOP AGE

Age	Count
22	9
23	7
26	5
25	5
24	4
27	2

FILTERING

terms must
field : LEMMA_multis
value : feel

Search - Web Demo

SEARCH (afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]

TEXT 11 of 32 available for paging

Title

Content

EMOTIONS

Emotion	Count
fear	28
guilt	2
sadness	1
anger	1

INTENSITY

Intensity	Count
3	11
4	9
2	9
1	3

TOP FACETS

Facet	Count
feel	32
fear	20
afraid	9
happen	6
scared	4
night	4
friend	4
father	4
fail	4
child	3

TOP AGE

Age	Count
22	9
23	7
26	5
25	5
24	4
27	2

FILTERING

terms must

field : LEMMA_multis

value : feel

Search - Web Demo

SEARCH QUERY **TOTAL HITS** HITS

(afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]

20 of 100 available for paging

Title

Content

When my brother had an epileptic attack and I was **scared** as to what would happen to him.

No filters available

EMOTIONS

- fear (129)
- guilt (4)
- anger (4)
- shame (1)
- sadness (1)
- joy (1)
- disgust (1)

INTENSITY

- 4 (56)
- 3 (50)
- 2 (25)
- 1 (10)

TOP FACETS

- fear (65)
- afraid (44)
- feel (32)
- friend (21)
- go (20)
- scared (18)
- night (16)
- time (15)
- happen (15)
- come (15)

TOP AGE

- 22 (44)
- 23 (36)
- 24 (24)
- 25 (17)
- 26 (11)
- 27 (9)

The screenshot displays a search interface with a search bar containing the query "(afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]". The search results are shown in a text viewer, with the word "fear" highlighted in the title and "scared" highlighted in the content. The interface includes several analytics charts: "EMOTIONS" showing a bar chart for various emotions with "fear" being the most prominent; "INTENSITY" showing a bar chart for intensity levels with "4" being the most frequent; "TOP FACETS" showing a bar chart for various terms with "fear" being the most frequent; and "TOP AGE" showing a bar chart for age groups with "22" being the most frequent. A blue box highlights the "TOTAL HITS" section, which displays "141".

Search - Web Demo

The screenshot displays a search interface with the following components:

- SEARCH BAR:** Query: `(afraid OR fear OR scared OR threatened) AND AGE_i:[22 TO 27]`. Total Hits: 141 (highlighted with a blue box).
- TEXT PANEL:** Shows the first result with the word "fear" highlighted in the title and "scared" highlighted in the content. The content reads: "When my brother had an epileptic attack and I was scared as to what would happen to him."
- EMOTIONS FACET:** A bar chart showing the distribution of emotions. The legend includes: fear (129), guilt (4), anger (4), shame (1), sadness (1), joy (1), and disgust (1).
- INTENSITY FACET:** A bar chart showing the distribution of intensity levels. The legend includes: 4 (56), 3 (50), 2 (25), and 1 (10).

Remove previous panels



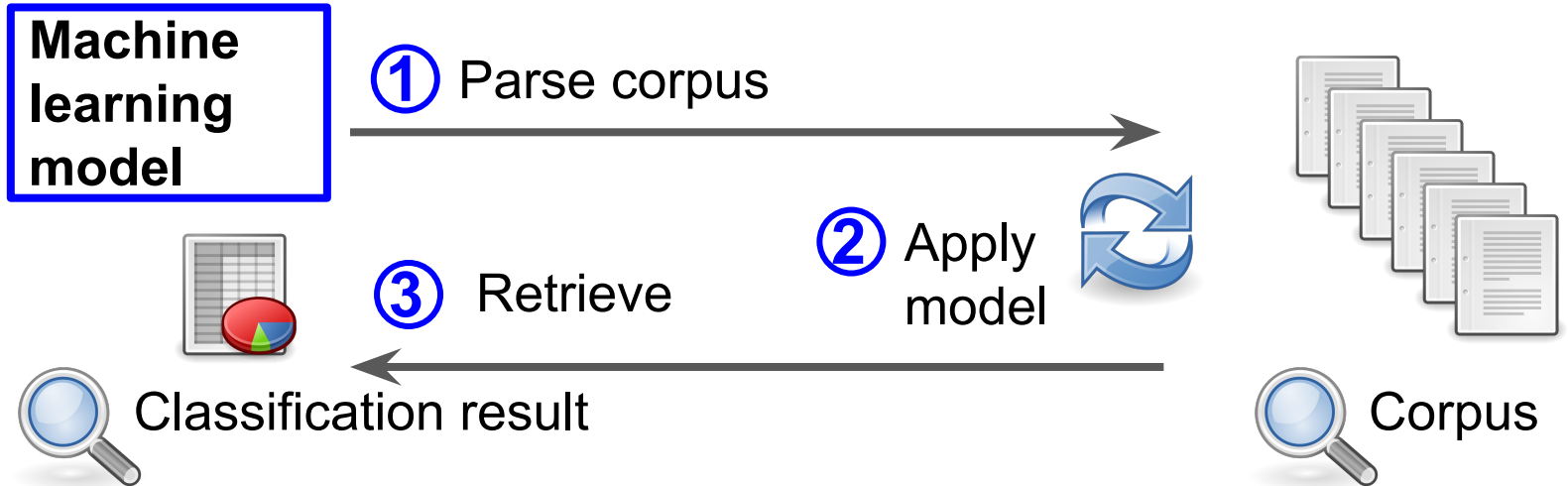
Add **part-of-speech** lemma facet panels

Search - Web Demo



Classification

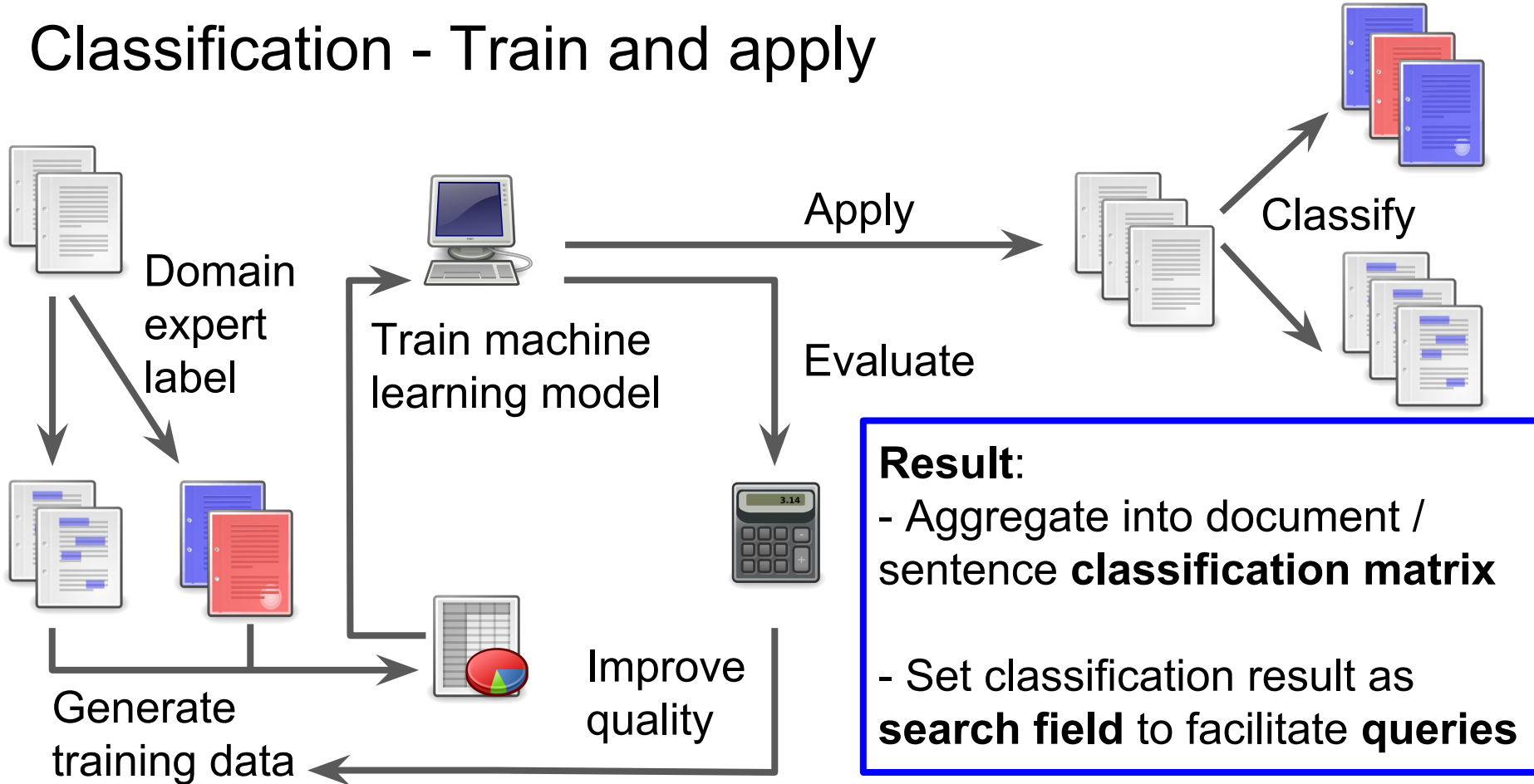
Assign labels to sentences/documents from a corpus



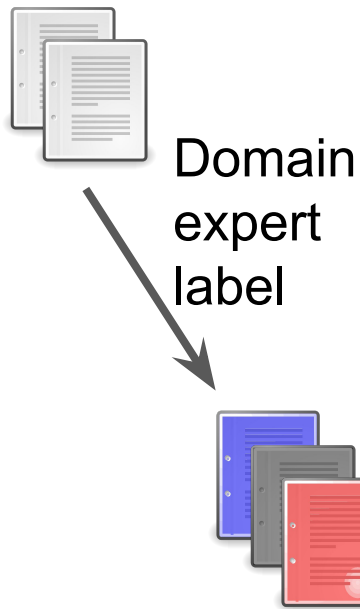
Class of **document** (Radiology/pathology report)
Sentiment of **sentences** (**negative**, **neutral**, **positive**)
Emotions of **sentences** (anger, fear, joy, disgust, ...)

- In-house medical
- Project-specific

Classification - Train and apply



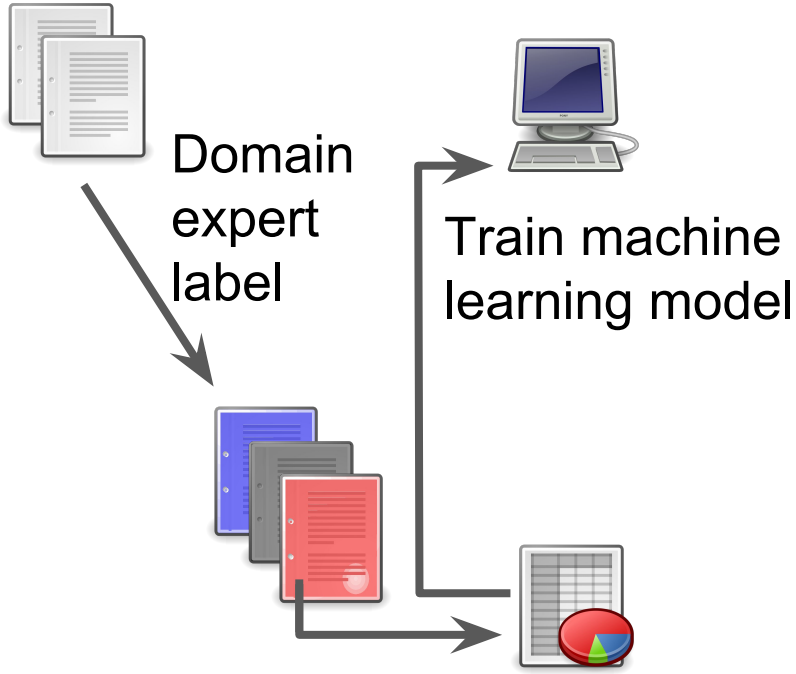
Classification - Example



Ohsumed (medical abstracts from *MeSH* categories) collection subset:

- Musculoskeletal Diseases
- Nutritional and Metabolic Diseases
- Eye Diseases

Classification - Example



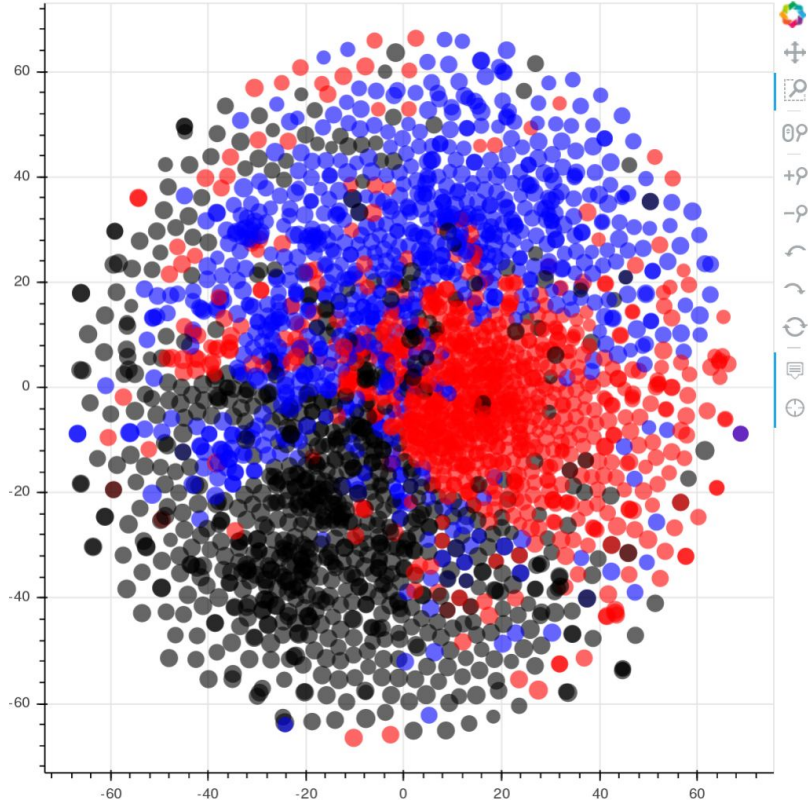
Ohsumed (medical abstracts from *MeSH* categories) collection subset:

- Musculoskeletal Diseases
- Nutritional and Metabolic Diseases
- Eye Diseases

Train machine learning model:

- Data: Abstract text
- Labels: Category

Classification - Example



Ohsumed (medical abstracts from *MeSH* categories) collection subset:

- Musculoskeletal Diseases
- Nutritional and Metabolic Diseases
- Eye Diseases

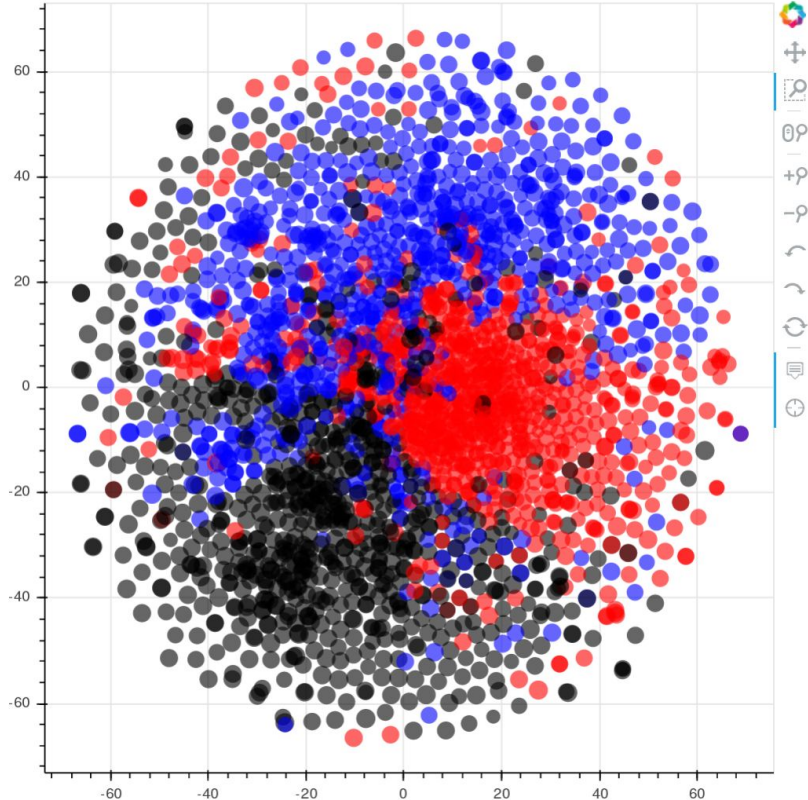
Train machine learning model:

- Data: Abstract text
- Labels: Category

Reduce the dimensionality:

- t-SNE

Classification - Example



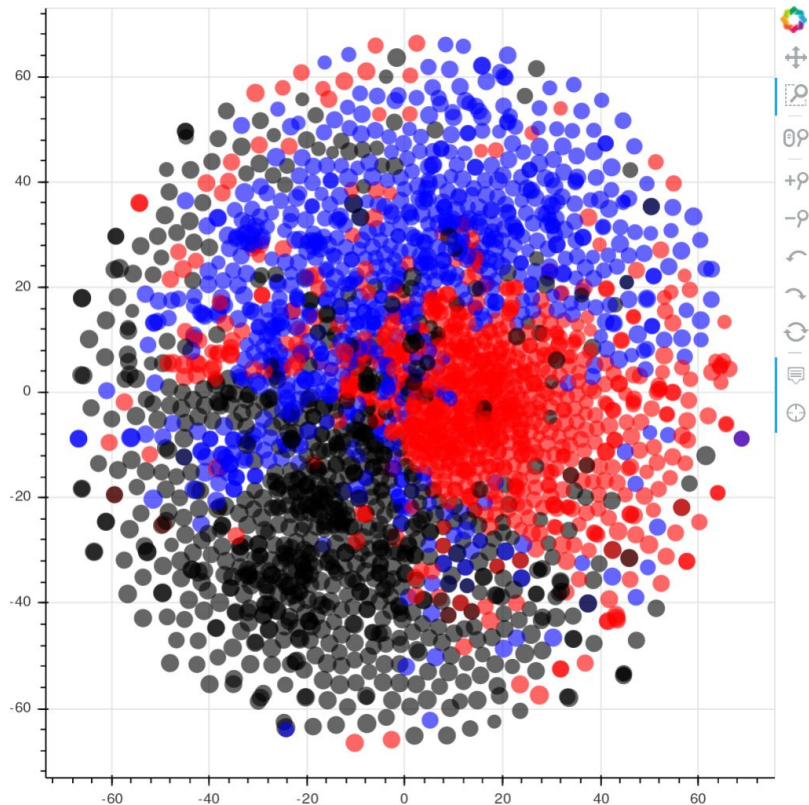
Top predictive features per category:

Musculoskeletal Diseases: bone arthritis
osteomyelitis synovial pain lumbar
myopathy scoliosis joint spine

Nutritional and Metabolic Diseases:
diabetic diabetes insulin obese glucose
malnutrition nutritional coronary renal
cholesterol

Eye Diseases: ocular retinal eye corneal
eyes uveitis graves glaucoma visual
cataract

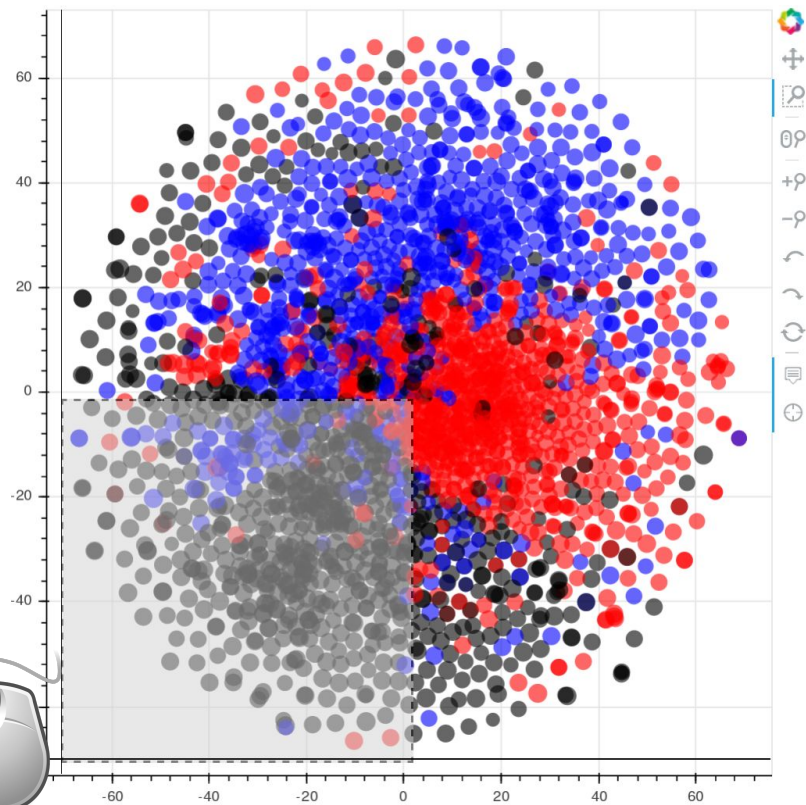
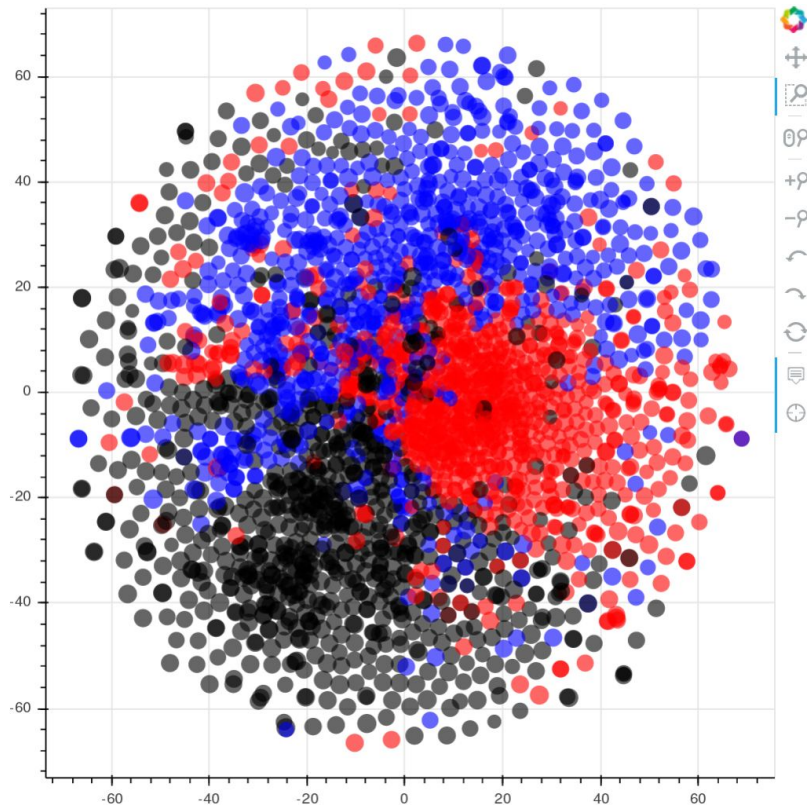
Post-classification - Interactive corpus exploration



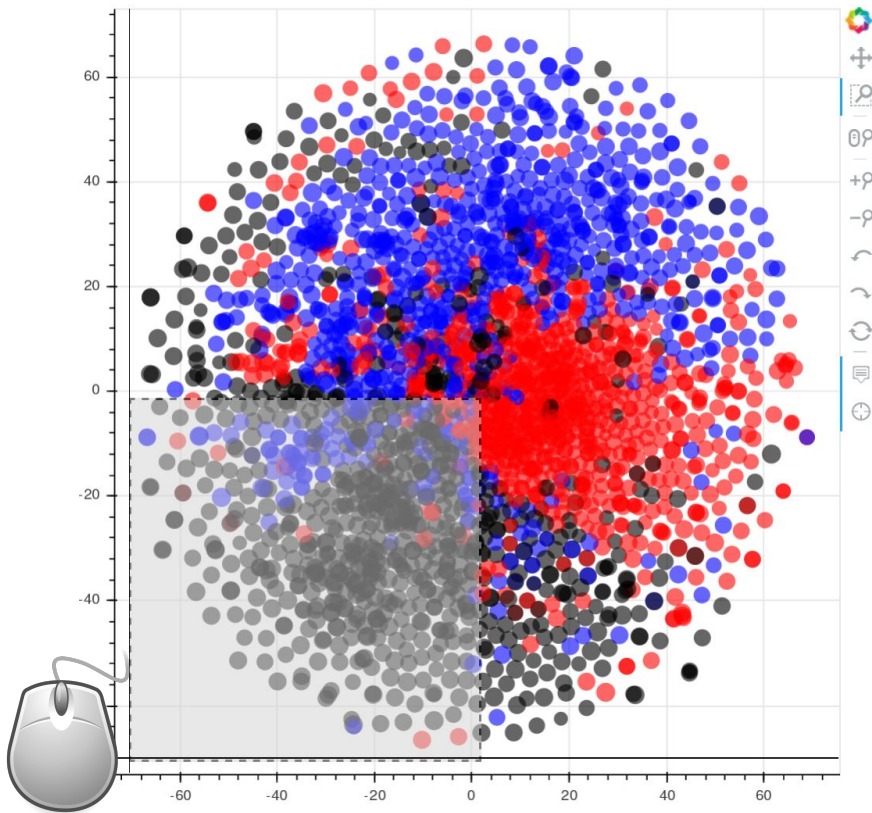
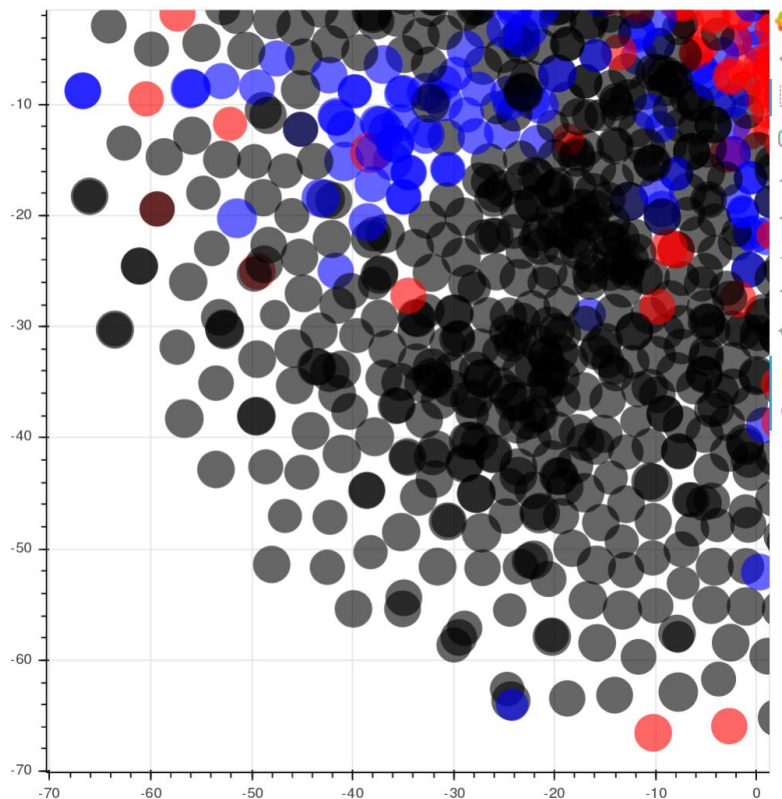
**Ohsumed (medical abstracts from
MeSH categories) collection subset:**

- Musculoskeletal Diseases
- Nutritional and Metabolic Diseases
- Eye Diseases

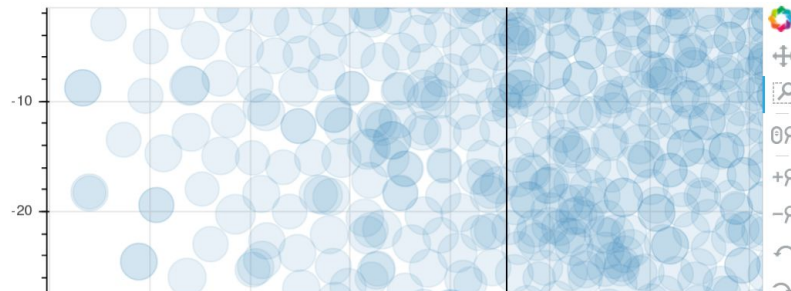
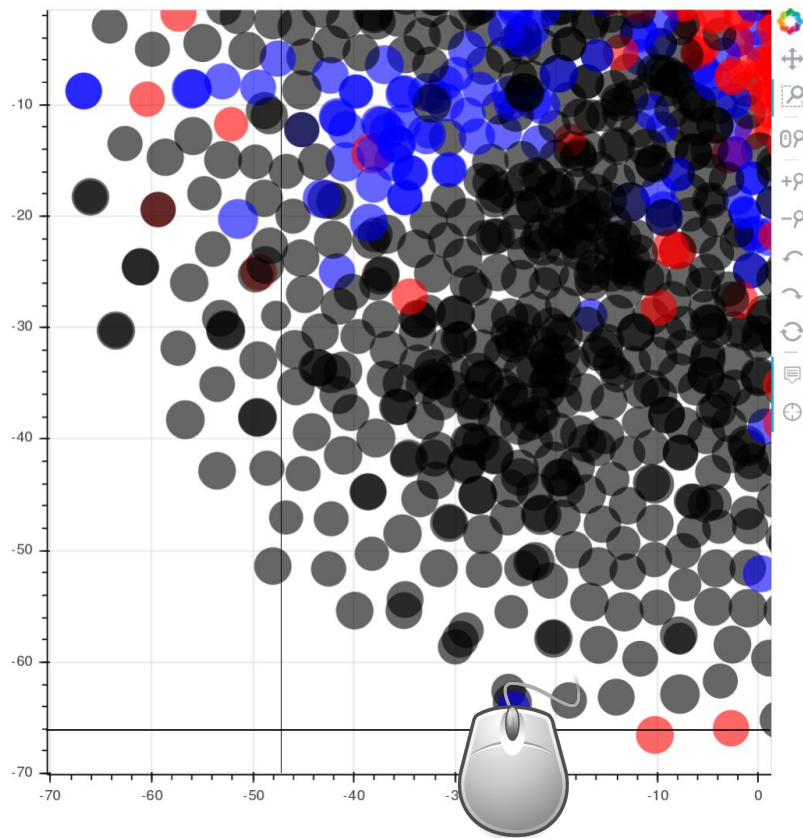
Post-classification - Interactive corpus exploration



Post-classification - Interactive corpus exploration



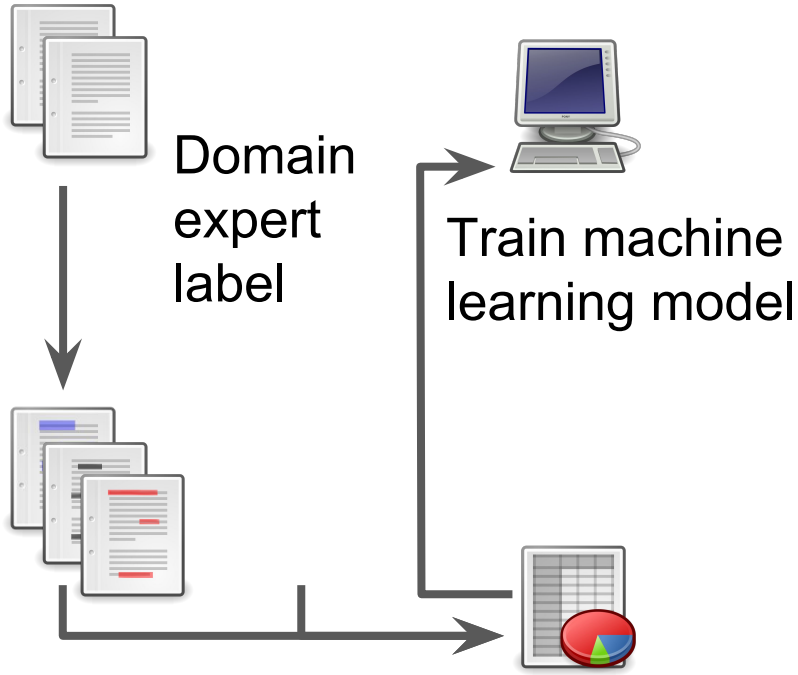
Post-classification - Interactive corpus exploration



index: 648
(x,y): (-24.266, -64.817)
color: blue
label: Musculoskeletal_Diseases
length: 605
text: Cerebrotendinous xanthomatosis: treatments with simvastatin, lovastatin, and chenodeoxycholic acid in 3 siblings. We report 3 sisters treated for cerebrotendinous xanthomatosis. We treated one, with a severe neurologic form of the illness, with chenodeoxycholic acid, then lovastatin and simvastatin. These drugs had different efficacy and tolerance, but induced no clinical improvement. Her sisters, without neurologic symptoms, received chenodeoxycholic acid, which normalized the cholestanol level. Optimal treatment of this illness must begin before there is significant clinical symptomatol

index: 547
(x,y): (-24.266, -64.817)
color: black
label: Nutritional_and_Metabolic_Diseases
length: 2322
text: Comparative evaluation of chenodeoxycholic and ursodeoxycholic acids in obese patients. Effects on biliary lipid metabolism during weight maintenance and weight reduction. Obesity is a condition associated with an increased frequency of gallstone disease. This study attempted to evaluate the comparative effects of two gallstone-dissolving agents, chenodeoxycholic acid and ursodeoxycholic acid, on bile acid metabolism and biliary lipid secretion in obese subjects in order to identify the bile acid of choice in preventing and treating gallstone disease in obesity. Twenty obese subjects (great

Classification - Sentence/short text example



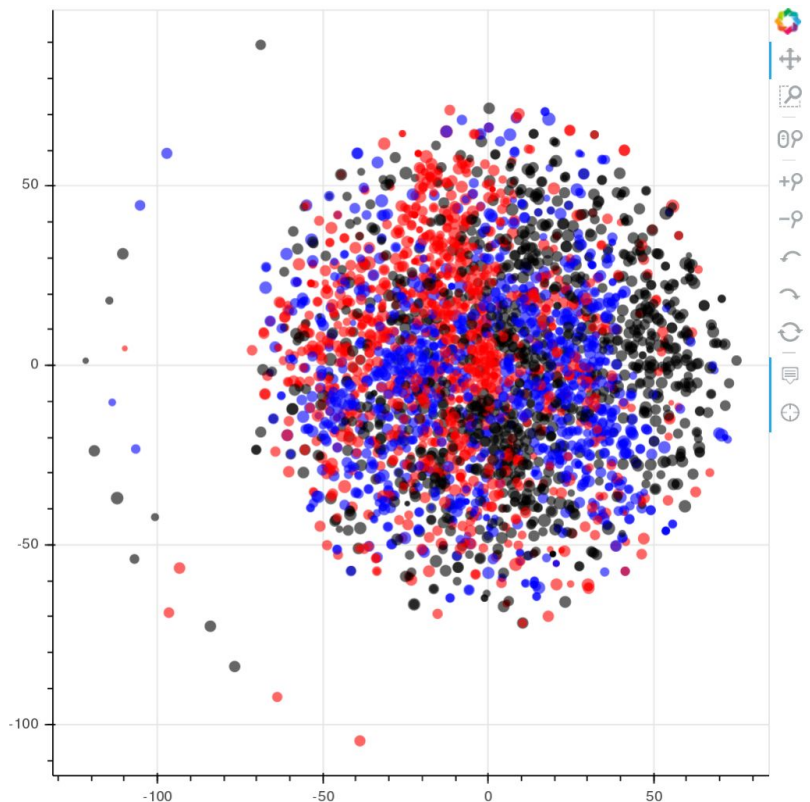
ISEAR corpus (sentences/short text) subset:

- Joy
- Fear
- Sadness

Train machine learning model:

- Data: Sentences/short text
- Labels: Emotion

Post-classification - Sentence/short text exploration



ISEAR corpus (sentences/short text) subset:

- Joy
- Fear
- Sadness

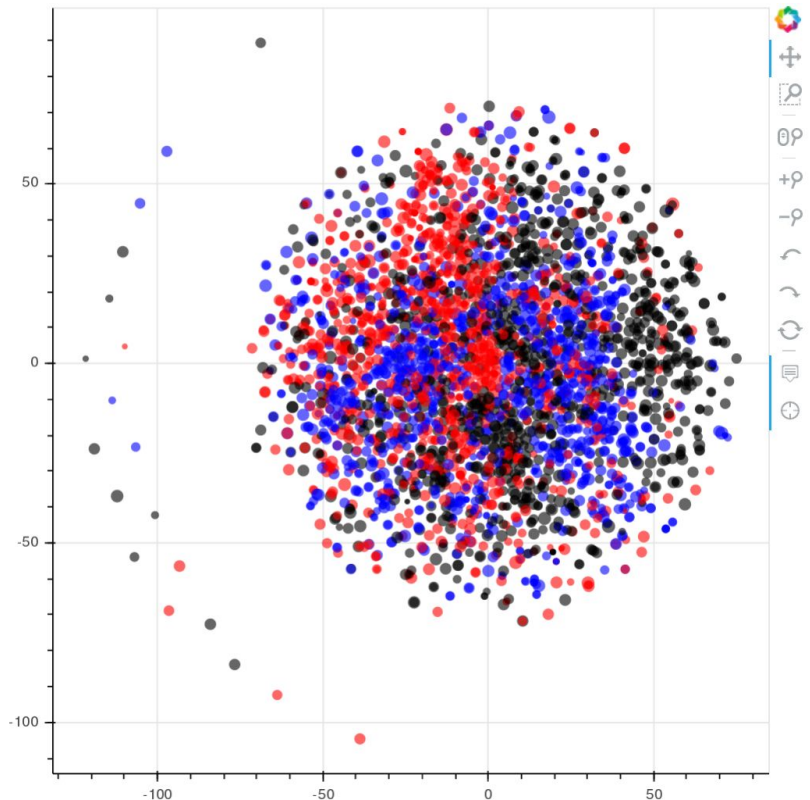
Train machine learning model:

- Data: Sentences/short text
- Labels: Emotion

Reduce the dimensionality:

- t-SNE

Post-classification - Sentence/short text exploration



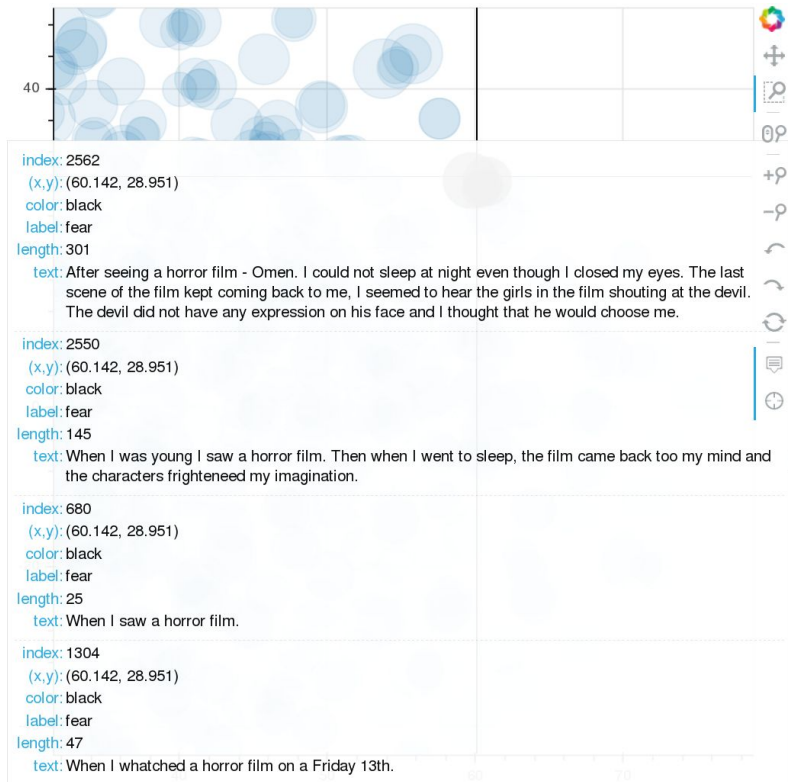
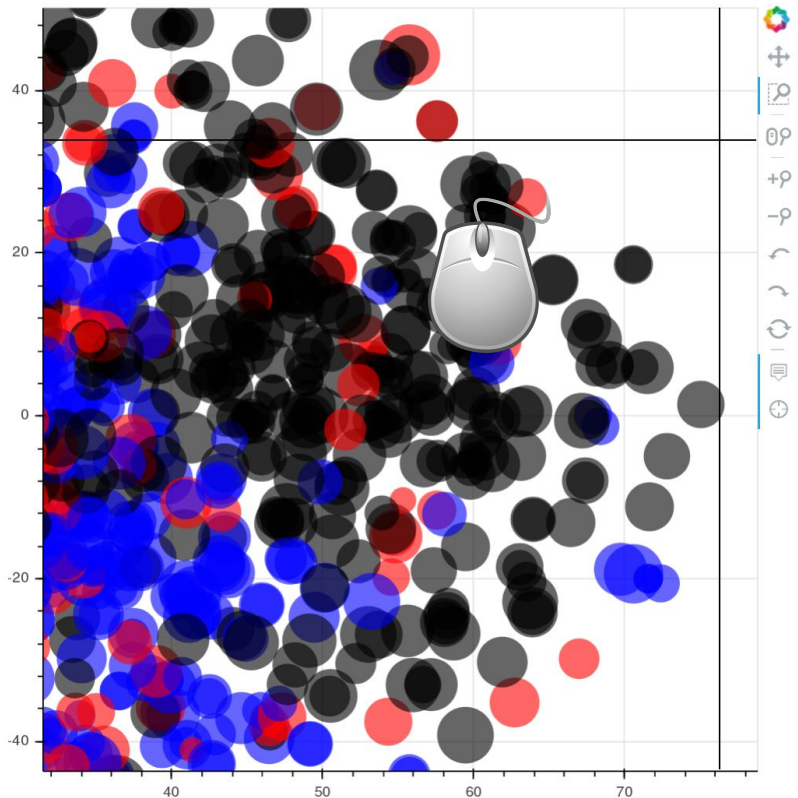
Top predictive features per category:

Joy joy happy passed glad won got
accepted birthday admitted wedding

Fear afraid fear scared night feared dark
threatened frightened friend_love house

Sadness sad died passed_away
sadness failed death separated left
leave relationship

Post-classification - Sentence/short text exploration



Search and classification - Augmented corpus

Apply classifiers on **project-specific** corpus + search



Apply trained classifiers on corpus

Augment document with classifier outputs

Index augmented corpus in search engine

Explore corpus and shape queries data-driven/interactively

Search and classification - Augmented corpus

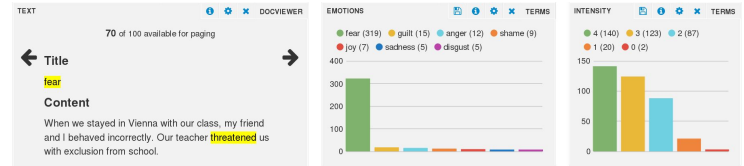
Apply classifiers on
project-specific
corpus + search



Application example query:

sadness_sentences_count:[2 TO *] **AND** document_type:"A"

Give me all documents with **at least two sentences**
containing **sadness AND** being of document type **A**



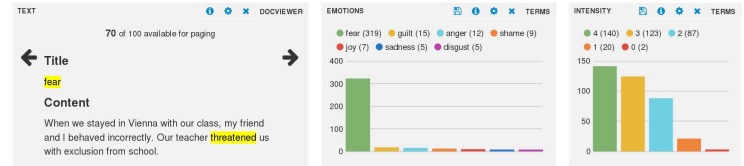
Search and classification - Augmented corpus

Apply classifiers on
project-specific
corpus + search



Application example query:

afraid **OR** fear **OR** scared **OR** night **OR** dark **OR** threatened



Give me all documents that contain at least one of my
domain-specific top predictive class features (Fear)

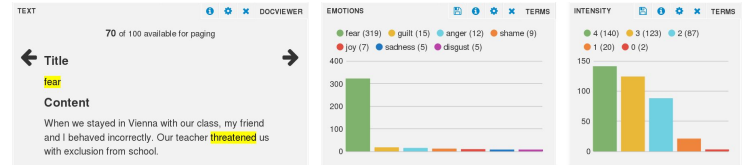
Search and classification - Augmented corpus

Apply classifiers on
project-specific
corpus + search



Application example query:

predicted_label:"class_A" **OR** predicted_label:"class_B"
AND document_date:[2012-01-01 TO 2014-12-31]
AND "some phrase"



Queries involving **date-ranges** are also possible

Summary

Research problems involving **clinical notes** cast as NLP tasks:

NER, information extraction, search, classification, (others ...)

We illustrated **what can be done**, now we would like you to bring us some interesting cases

We can help you with your research

Contact

Thank you for your attention!

We can help you with your research

You are encouraged to get in touch with us now or via email:

Tomasz Oliwa, PhD | toliwa@bsd.uchicago.edu

Brian Furner | bfurner@bsd.uchicago.edu

Center for Research Informatics

Biological Sciences Division

University of Chicago

References 1

- Epic, <http://www.epic.com/>
- Note text modified from example: <http://www.med.unc.edu/medselect/resources/sample-notes/sample-write-up-1> and at the bottom of this page:

Source

Rubin, R. and Strayer, D. Rubin's Pathology. 5th edition. Lippincott Williams and Wilkins, 2008.

- brat rapid annotation tool, <http://brat.nlplab.org/>
- UMLS, <https://www.nlm.nih.gov/research/umls/> , https://en.wikipedia.org/wiki/Unified_Medical_Language_System
- CUI, https://www.nlm.nih.gov/research/umls/new_users/glossary.html
- SNOMED Clinical Terms, <http://www.snomed.org/>, https://en.wikipedia.org/wiki/SNOMED_CT
- SNOMED CORE: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html
- Apache cTAKES, <http://ctakes.apache.org/>
- Apache Solr, <http://lucene.apache.org/solr/>
- Banana for Solr, <https://github.com/lucidworks/banana>
- Bokeh <https://bokeh.pydata.org/en/latest/>
- scikit-learn and t-SNE: <http://scikit-learn.org> and <https://lvdmaaten.github.io/tsne/>

References 2

Ohsumed dataset obtained from: <http://disi.unitn.it/moschitti/corpora.htm>

See also http://trec.nist.gov/data/t9_filtering.html and <http://trec.nist.gov/data/filtering/README.t9.filtering> for source:

(A) Description of the OHSUMED document collection (files: ohsumed.*)

The OHSUMED test collection is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The National Library of Medicine has agreed to make the MEDLINE references in the test database available for experimentation, restricted to the following conditions:

1. The data will not be used in any non-experimental clinical, library, or other setting.
2. Any human users of the data will explicitly be told that the data is incomplete and out-of-date.

The OHSUMED document collection was obtained by William Hersh (hersh@OHSU.EDU) and colleagues for the experiments described in the papers below:

Hersh WR, Buckley C, Leone TJ, Hickam DH, OHSUMED: An interactive retrieval evaluation and new large test collection for research, Proceedings of the 17th Annual ACM SIGIR Conference, 1994, 192-201.

Hersh WR, Hickam DH, Use of a multi-application computer workstation in a clinical setting, Bulletin of the Medical Library Association, 1994, 82: 382-389.

References 3

ISEAR attribution:

Copyright, disclaimer, license, author's website:

<http://www.affective-sciences.org/home/research/materials-and-online-research/research-material/>

License notice on website: All these materials are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.:

<https://creativecommons.org/licenses/by-nc-sa/3.0/>

Title: International Survey On Emotion Antecedents And Reactions (ISEAR)

Short description

Over a period of many years during the 1990s, a large group of psychologists all over the world collected data in the ISEAR project, directed by Klaus R. Scherer and Harald Wallbott. Student respondents, both psychologists and non-psychologists, were asked to report situations in which they had experienced all of 7 major emotions (joy, fear, anger, sadness, disgust, shame, and guilt). In each case, the questions covered the way they had appraised the situation and how they reacted. The final data set thus contained reports on seven emotions each by close to 3000 respondents in 37 countries on all 5 continents.

References

The following publications describe the procedures and report the major patterns of results:

Wallbott, H.G., & Scherer, K. R. (1986). How universal and specific is emotional experience? *Social Science Information*, 24, 763-795.

Matsumoto, D., Kudoh, T., Scherer, K. R., & Wallbott, H.G. (1988). Antecedents of and reactions to emotions in the US and Japan. *Journal of Cross-Cultural Psychology*, 19, 267-286.

Wallbott, H.G., & Scherer, K. R. (1988). Emotion and economic development - Data and speculations concerning the relationships between economic factors and emotional experience. *European Journal of Social Psychology*, 18, 267-273.

Scherer, K. R., & Wallbott, H.G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66, 310-328.

Scherer, K. R. (1997). Profiles of emotion-antecedent appraisal: testing theoretical predictions across cultures. *Cognition and Emotion*, 11, 113-150.

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, 73, 902-922.

Mikula, G., Scherer, K. R., & Athenstaedt, U. (1998). The role of injustice in the elicitation of differential emotional reactions. *Personality and Social Psychology Bulletin*, 24(7), 769-783.