IBM Research - China

TIARA: A Visual Exploratory Text Analytic System

Furu Wei[+], Shixia Liu[+], Yangqiu Song[+], Shimei Pan[#], Michelle X. Zhou[*], Weihong Qian[+], Lei Shi[+], Li Tan[+] and Qiang Zhang[+]

\+ IBM Research – China, Beijing, China

\# IBM Research – T. J. Watson Center, Hawthorne, NY, USA

* IBM Research – Almaden Center, San Jose, CA, USA

KDD, Washington D.C, Jul 2010

Outline

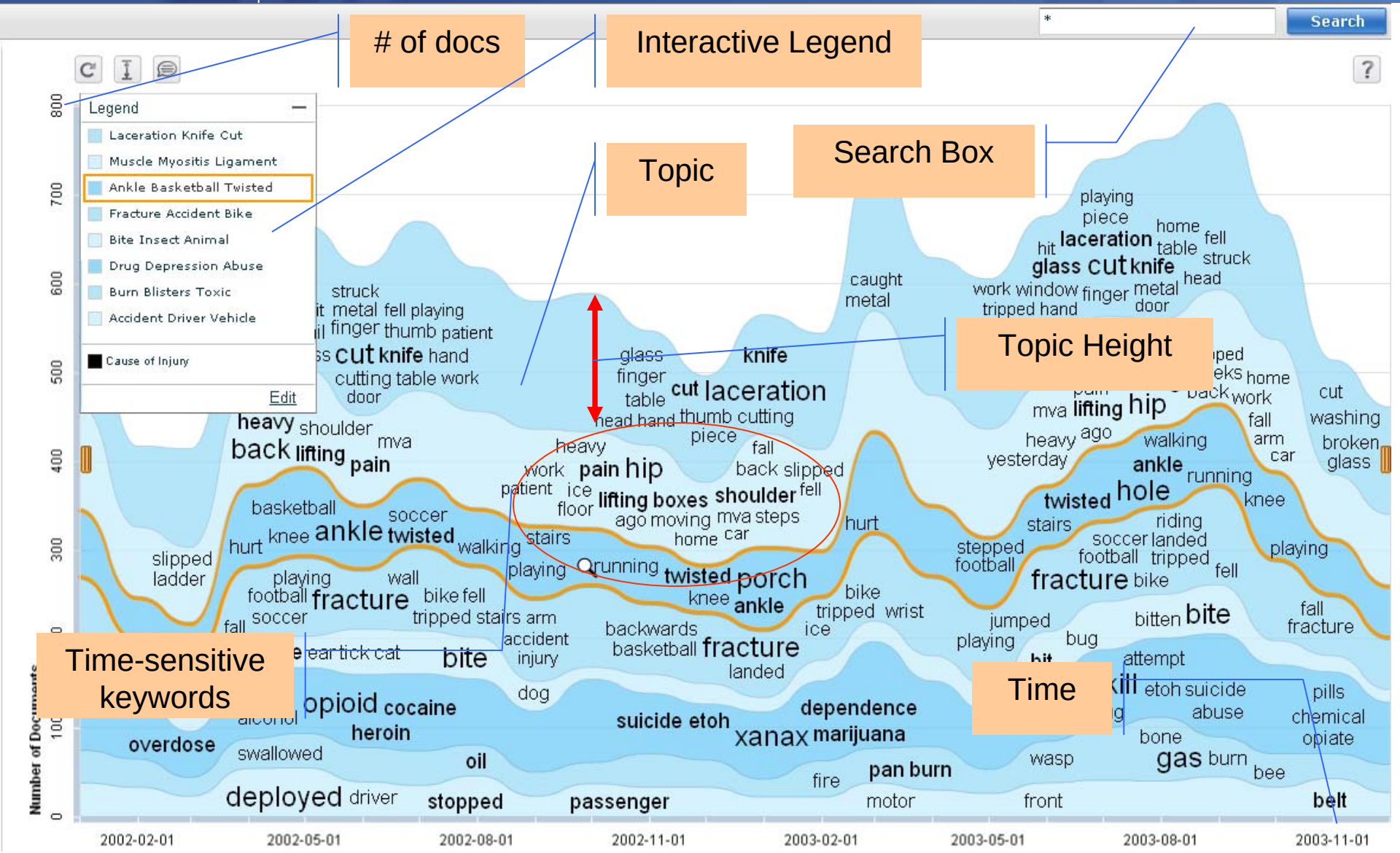
- Introduction
- The TIARA System
  - System Overview
  - Analytics
  - Experiments
- Demonstration
- Conclusion

What is TIARA

- **TIARA**: **T**ext **I**nsight via **A**utomated **R**esponsive **A**nalytics

- **Text analytics** + **interactive visualization** to visually analyze the topics in text collections and their content changes over time

- Visually revealing topic “strength” and “content” **evolution** over time



TIARA's visual summary of the "cause of injury" field of the 23,000+ emergency room records from 2002 to 2003

Outline

- Introduction
- The TIARA System
  - System Overview
  - Analytics
  - Experiments
- Demonstration
- Conclusion

Text Preprocessing

Content/Meta Extraction

Topic Analysis

Text/Topic Indexing

Offline

Lucene Index (Text, Metadata, Topic, etc.)

Online

Snippet Generation

Time-sensitive Keyword Re-ranking

Time-sensitive Keyword Extraction

Topic Ranking

Topic Summarization

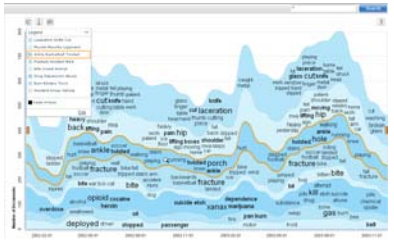
JSON Adapter

Index Parsing

Search

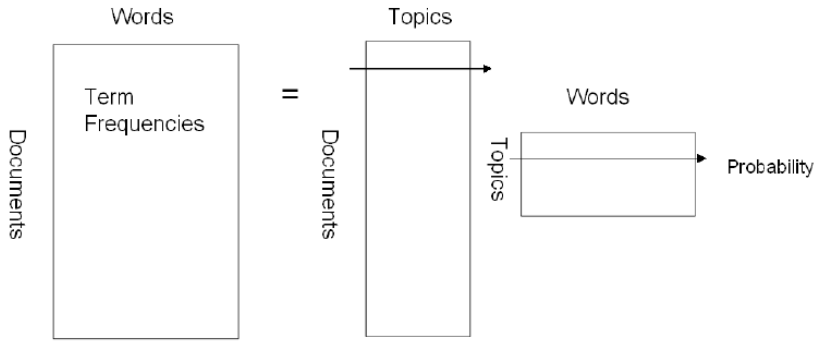
Query Parsing

Web Service



Analytics: Topic Analysis

- **Goal**: Extract **topics** from a set of documents
- **Solutions**
  - Topic Models (such as Latent Dirichlet Allocation, etc.)
  - Text Clustering (such as K-means etc.)



We utilize

- the document-topic distribution matrix $\Theta \in \mathbb{R}^{N \times K}$ and
- the topic-word distribution matrix $\Phi \in \mathbb{R}^{K \times V}$

to summarize the topic analysis results.

Analytics: Topic Ranking

- **Motivations**
  - Help user quickly locate the topics they are mostly interested in

$$\mu_i = \sum_{j=1}^{N} N_i \cdot \Theta_{j,i} / \sum_{j=1}^{N} N_i$$

- **Solutions**
  - Topic Content Coverage: Topics covering a significant portion of the corpus content a those covering less content

$$\sigma_i = \sqrt{\sum_{j=1}^{N} N_i \cdot (\Theta_{j,i} - \mu_i)^2 / \sum_{j=1}^{N} N_i}$$

  - Topic Variance: Topics appearing in all the documents are to be too generic to be interesting

$$r_i \triangleq (\mu_i)^{\lambda_1} \cdot (\sigma_i)^{\lambda_2}$$

Analytics: Keyword based Topic Summarization

- **Motivations**
  - Common/general words in topic analysis results (!= stop words!)
  - LDA on a financial news corpus: common words such as *Dow, Jones, Wall, Street* etc., are ranked high in many topics because they are relevan $weigt(w_m) = \Phi_{j,m} \cdot log \frac{\Phi_{j,m}}{(\prod_{k=1}^{K} \Phi_{k,m})^{\frac{1}{K}}}$
- **Solutions**
  - Topic Frequency: if a word occurs frequently in a topic, it is important
  - Inverse Topic Frequency: if the word also appears in many other topics, the word is not important because it is too common

Analytics: Time-sensitive Keyword Extraction

- **Motivations**
  - Allow user visually analyze content evolutions
- **Solutions**
  - Break the documents into several sub-collections, each of which is associated with a particular time interval
    - Most active time segments
    - Do not break a topic near the peaks of a topic layer
  - Extract time-sensitive keywords for each sub-collection

Analytics: Time-sensitive Keyword Extraction

- **Solutions (cont')**
  - Extract time-sensitive keywords for each sub-collection
    - Completeness: whether we can re-cover the original semantics of a topic by combining the semantics associated with each topic segment
    - Distinctiveness: whether we can distinguish one topic segment from another based on their associated keywords

$$\eta_1 \cdot \frac{TF_{j,i,m}}{\sum_i TF_{j,i,m}} + \eta_2 \cdot \Phi_{j,m} \cdot log \frac{\Phi_{j,m}}{(\prod_{k=1}^{K} \Phi_{k,m})^{\frac{1}{K}}}$$

Visualization

- **Visual Design**
  - Stacked graph based visual layout
  - Augment the stacked graph with layer ordering and tag clouds to generate a keyword based visual text summarization
  - See Liu et al. (CIKM 2009) for more details
- **Interactions**
  - See our (live) demo

Experiment Studies

- **Data Sets**
  - Email data set: Personal email collection with 8326 emails
  - Healthcare data set: Emergency room data set (ref. NHAMCS :National Hospital Ambulatory Medical Care Survey) containing 23,501 patient records from 2002 to 2003
- **Evaluation Plans**
  - Topic ranking
  - Keyword based topic summarization
  - **Topic segmentation and time-sensitive keyword selection**
  - **TIARA's online response time**

Song et al. (CIKM 2009)

Experiment Studies

- **Evaluation Criteria**
  - **Completeness**: F_1 measure between topic keywords and combination of time-sensitive keywords
  - **Distinctiveness**: KL-divergences of time-sensitive keyword distributions among different segments

$$D_{KL}(h_l^j||h_m^j) = \sum_{i=1}^{V} h_l^j(i) \log \frac{h_l^j(i)}{h_m^j(i)} \qquad (3)$$

Moreover, the symmetric Jensen-Shannon divergence is

$$D_{JS}(h_l^j||h_m^j) = \frac{1}{2} D_{KL}(h_l^j||\bar{h}^j) + \frac{1}{2} D_{KL}(h_m^j||\bar{h}^j) \qquad (4)$$

where $\bar{h}^j = \frac{1}{2}(h_l^j + h_m^j)$. Thus, we define the distinctiveness of topic j as

$$D(\{h_l^j\}_l^L) = \frac{1}{L(L-1)} \sum_l \sum_m D_{JS}(h_l^j||h_m^j) \qquad (5)$$

- **Baseline System**
  - Select time-sensitive keywords based on term frequencies

Experiment Results

	Completeness	Distinctiveness
Baseline	0.452 ± 0.043	0.182 ± 0.079
TIARA	$\mathbf{0.657 \pm 0.055}$	$\mathbf{0.315 \pm 0.082}$

Results on Email data (mean+std)

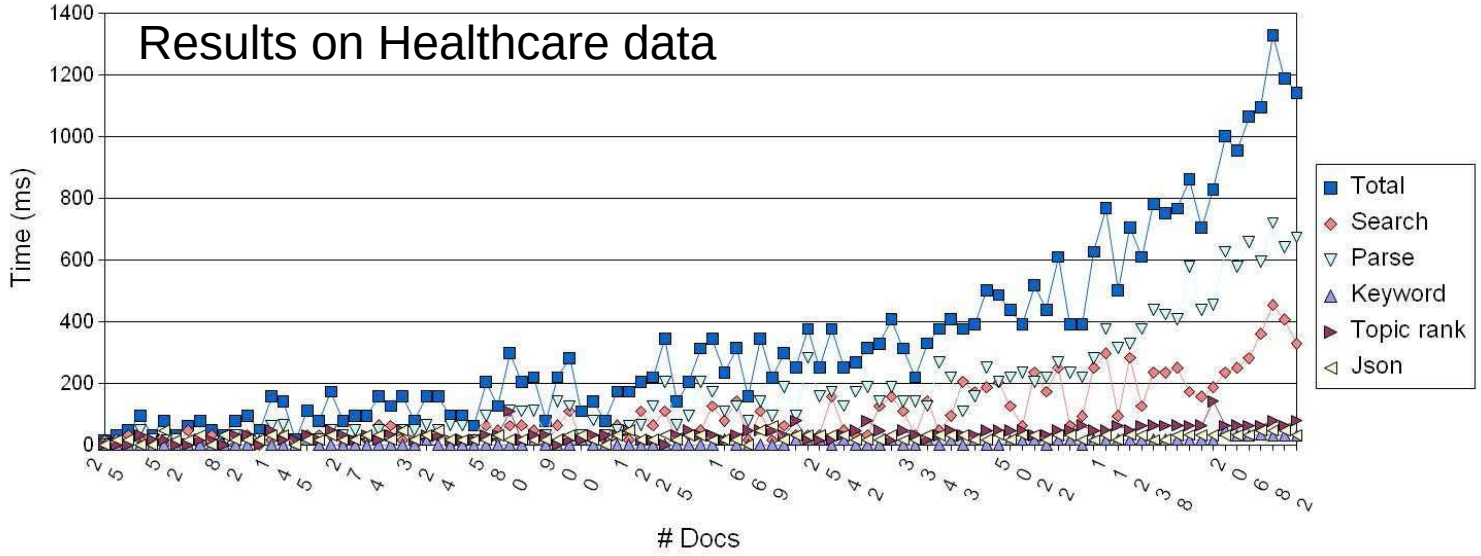
	Completeness	Distinctiveness
Baseline	0.578 ± 0.053	0.114 ± 0.087
TIARA	$\mathbf{0.740 \pm 0.073}$	$\mathbf{0.210 \pm 0.058}$

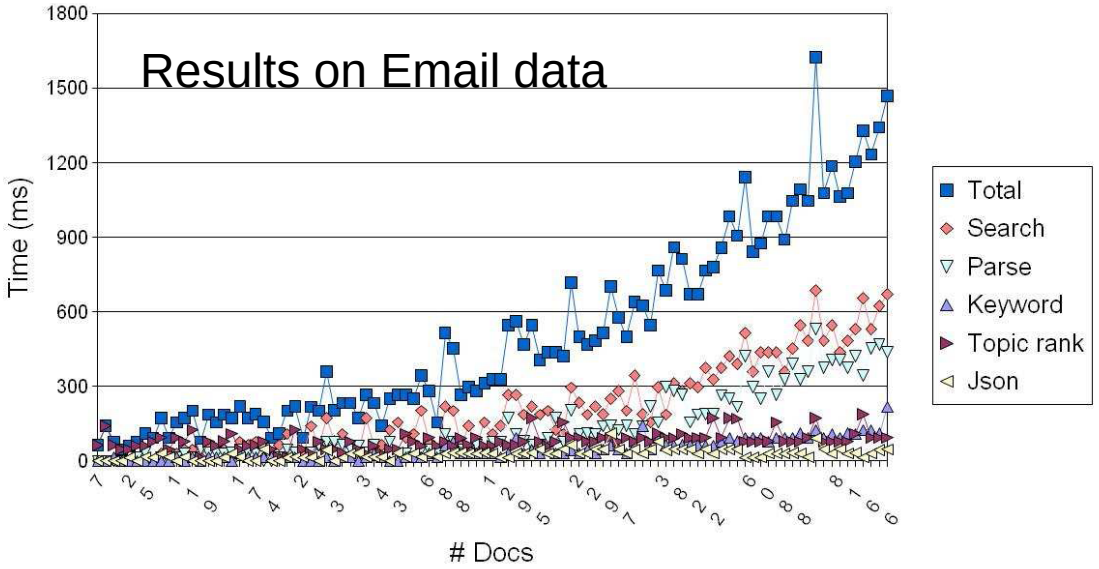
Results on Healthcare data (mean+std)

1. TIARA performs better than the baseline system

2. The topics derived from the emails **evolve more quickly** than those derived from the emergency room records

Experiment Results





1. The most time-consuming procedures are *Search* and *Parse*

2. Sampling, select top *N* documents, or indexing only the top *N* topic keywords

Outline

- Introduction
- The TIARA System
  - System Overview
  - Analytics
  - Experiments
- Demonstration
- Conclusion

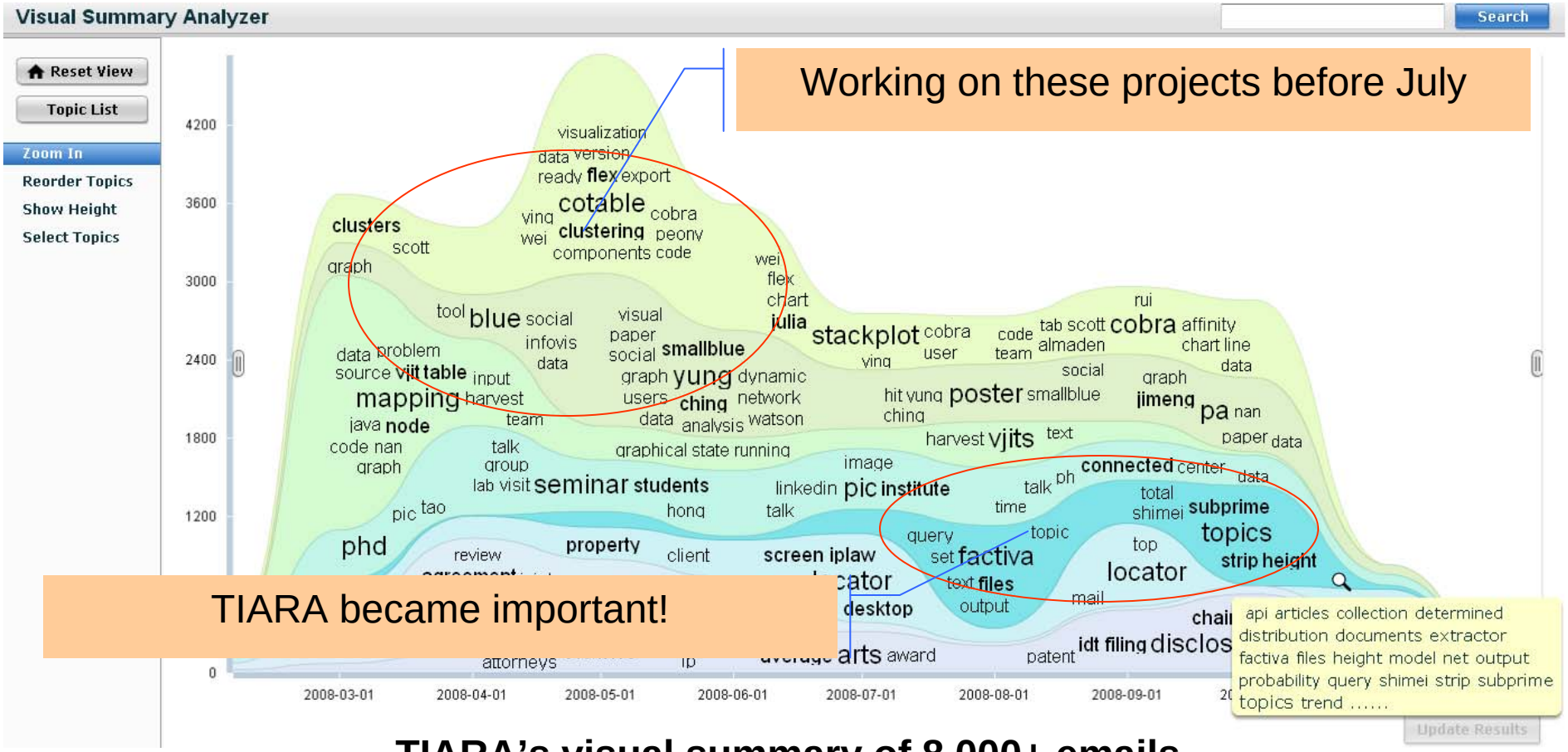
We have shown the live demo on Tuesday, 27 July

Application: Visual Email Summarization

- **Who**
  - Email owner: review his work (projects etc.) in 2008
- **Data**
  - Personal email collection with 8326 emails

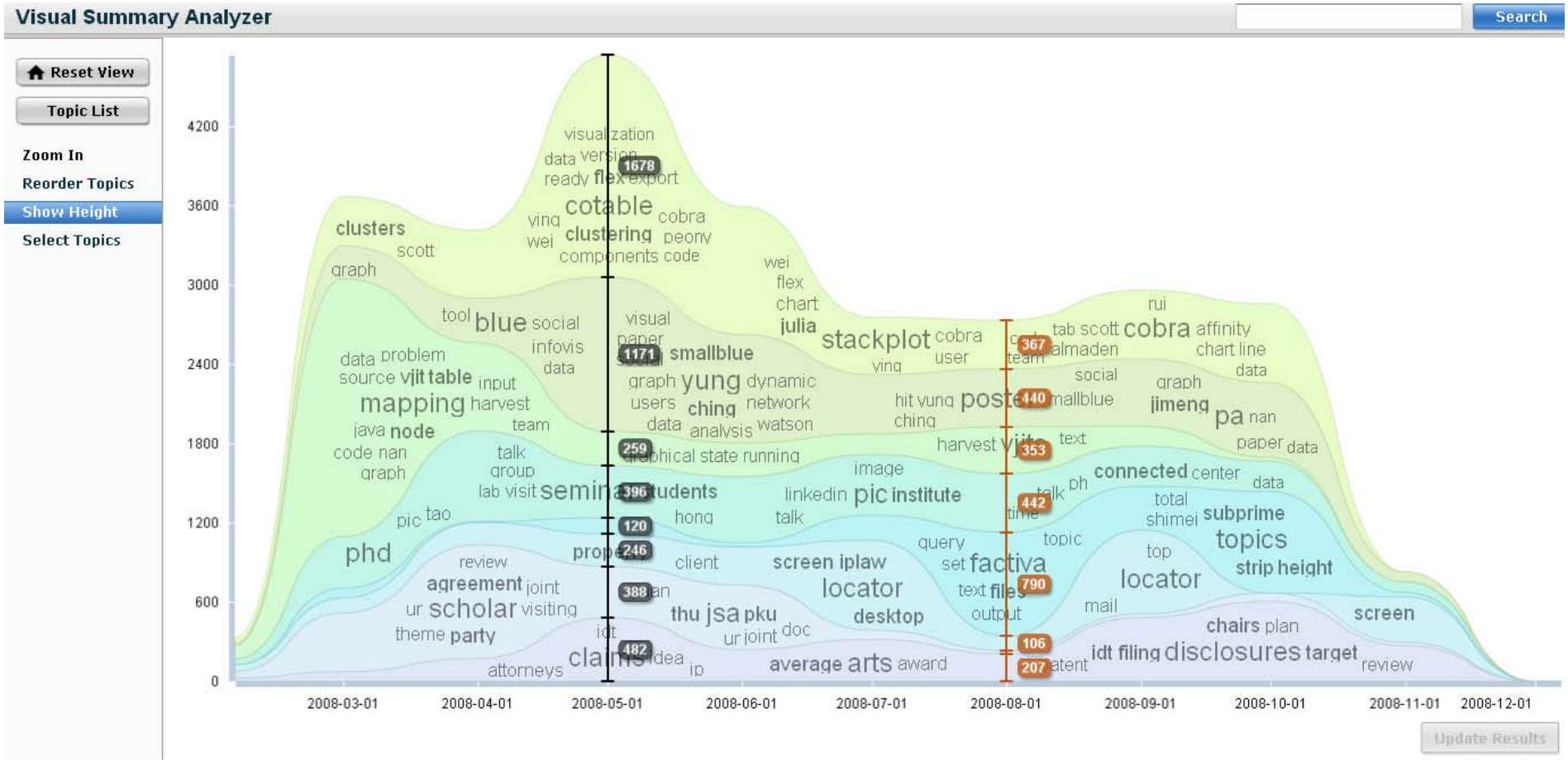
How could TIARA help him ?

Topic Evolution

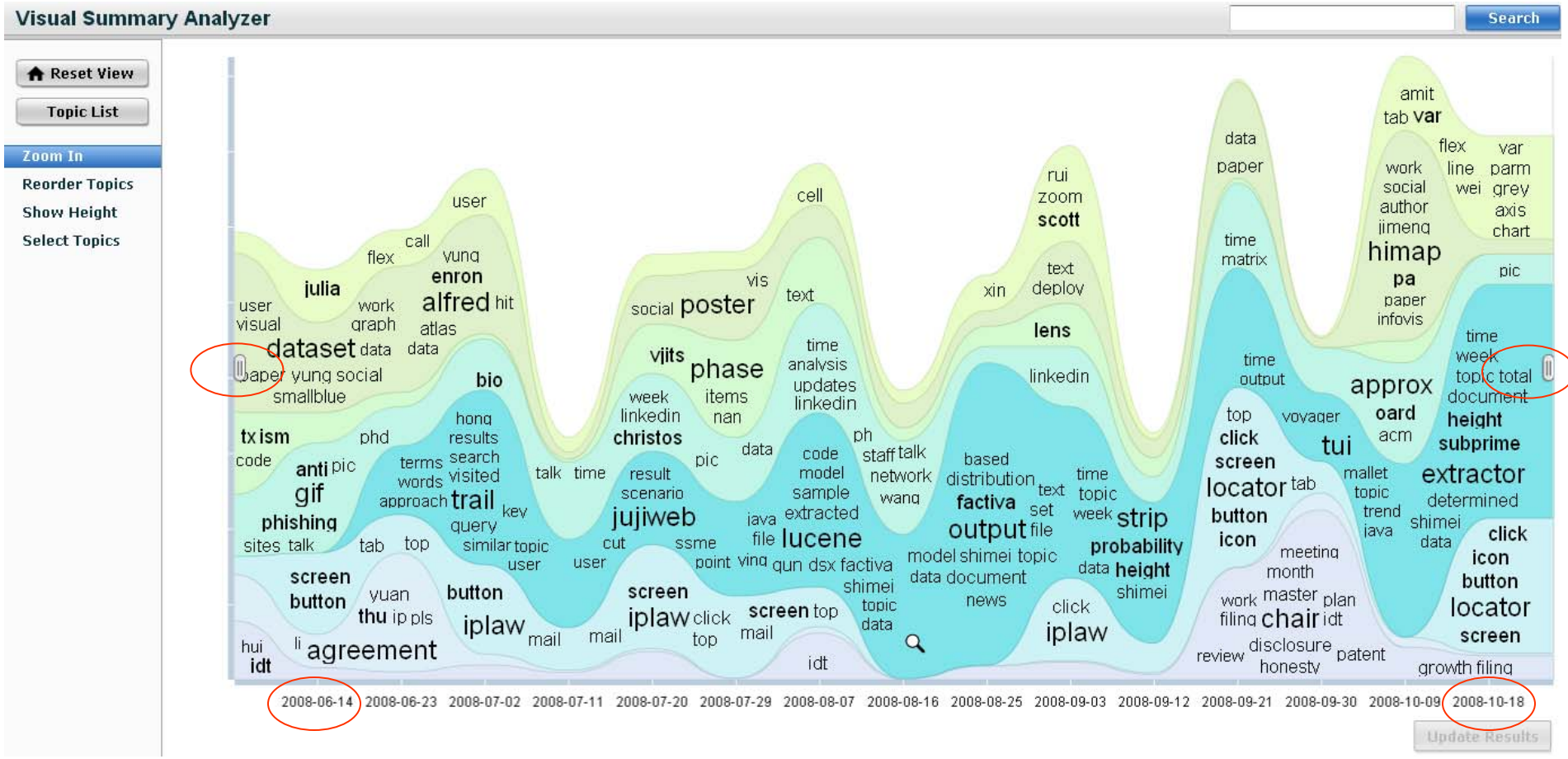


TIARA's visual summary of 8,000+ emails.

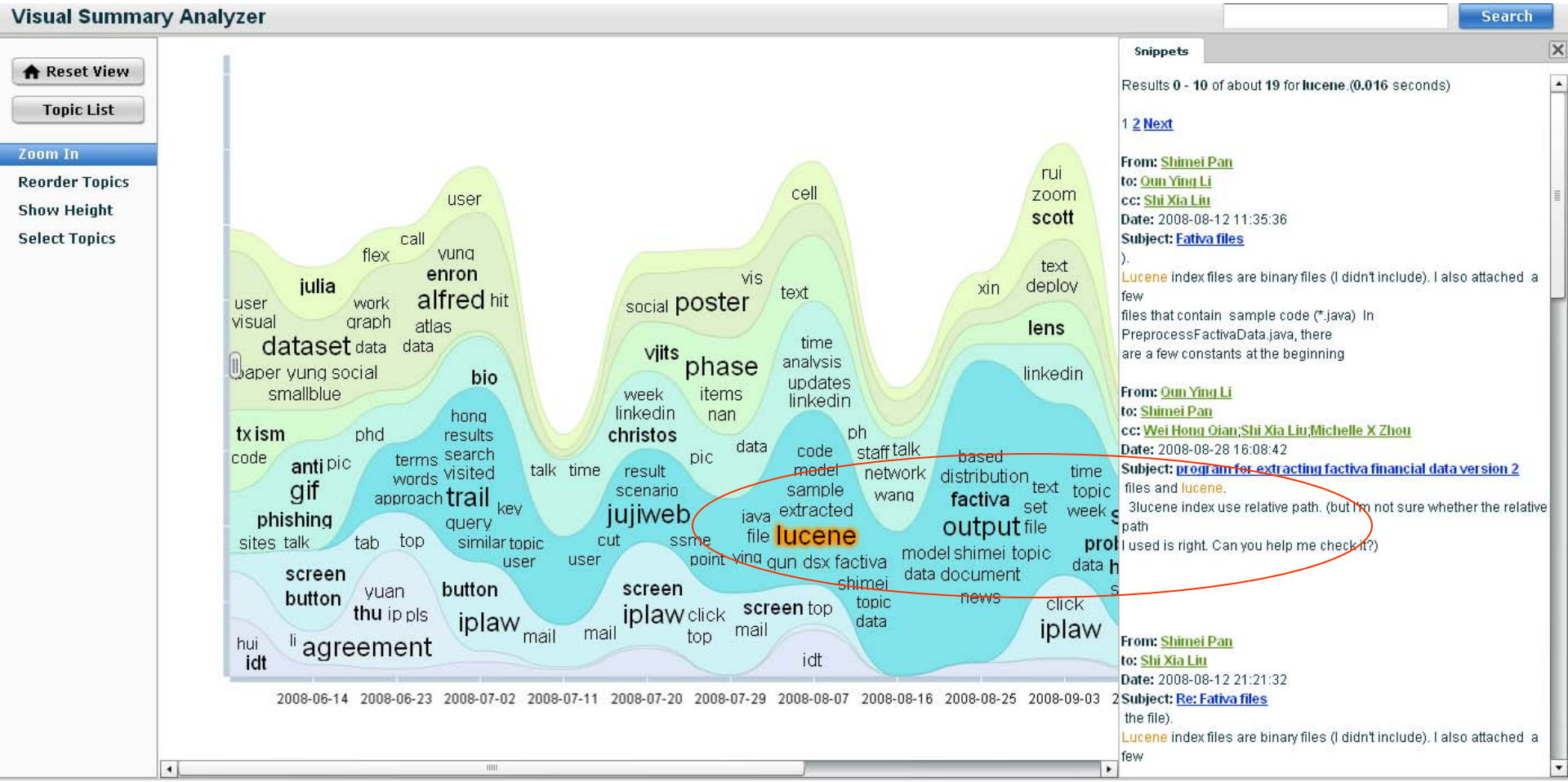
Show height



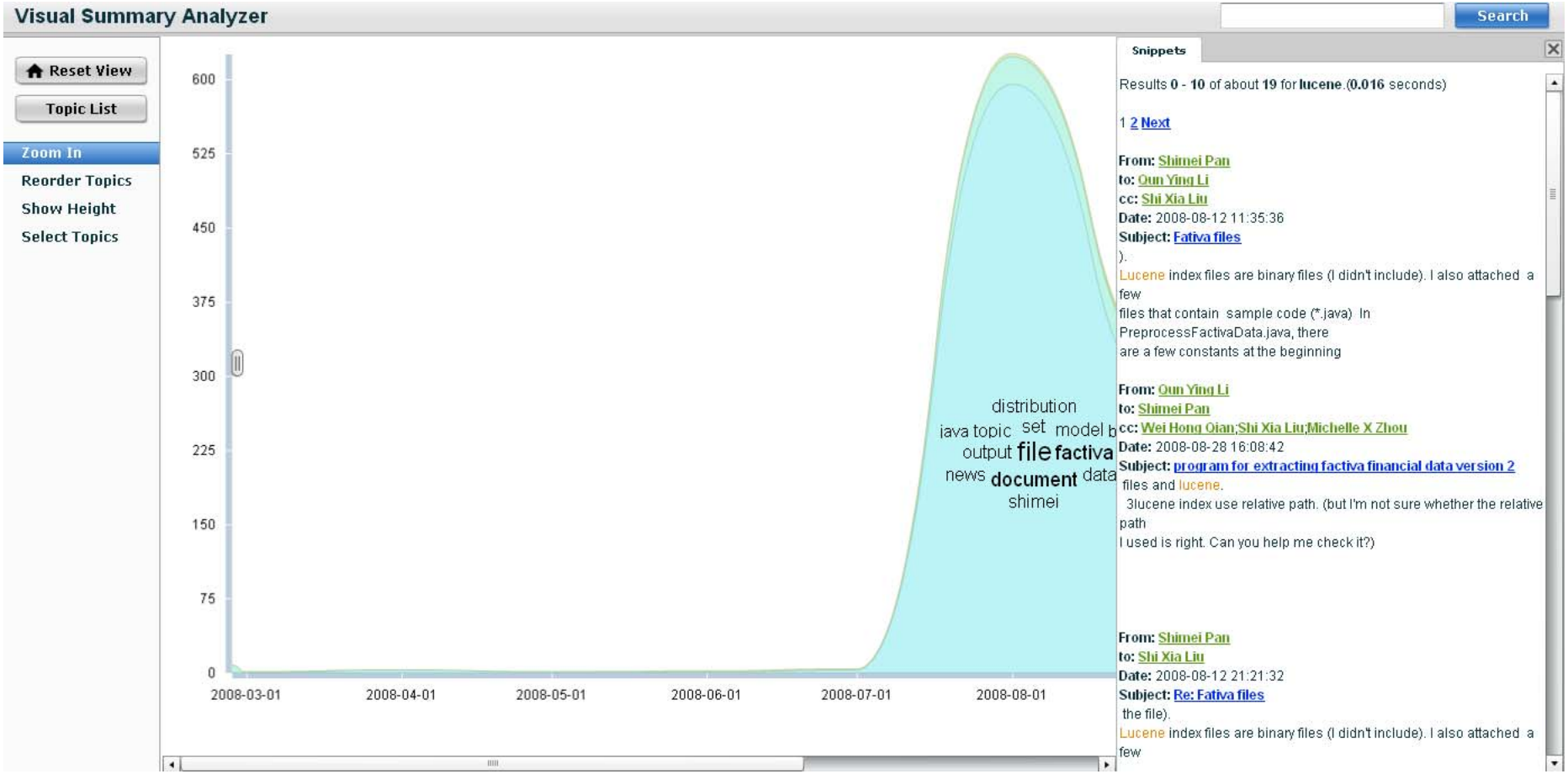
Zoom into details



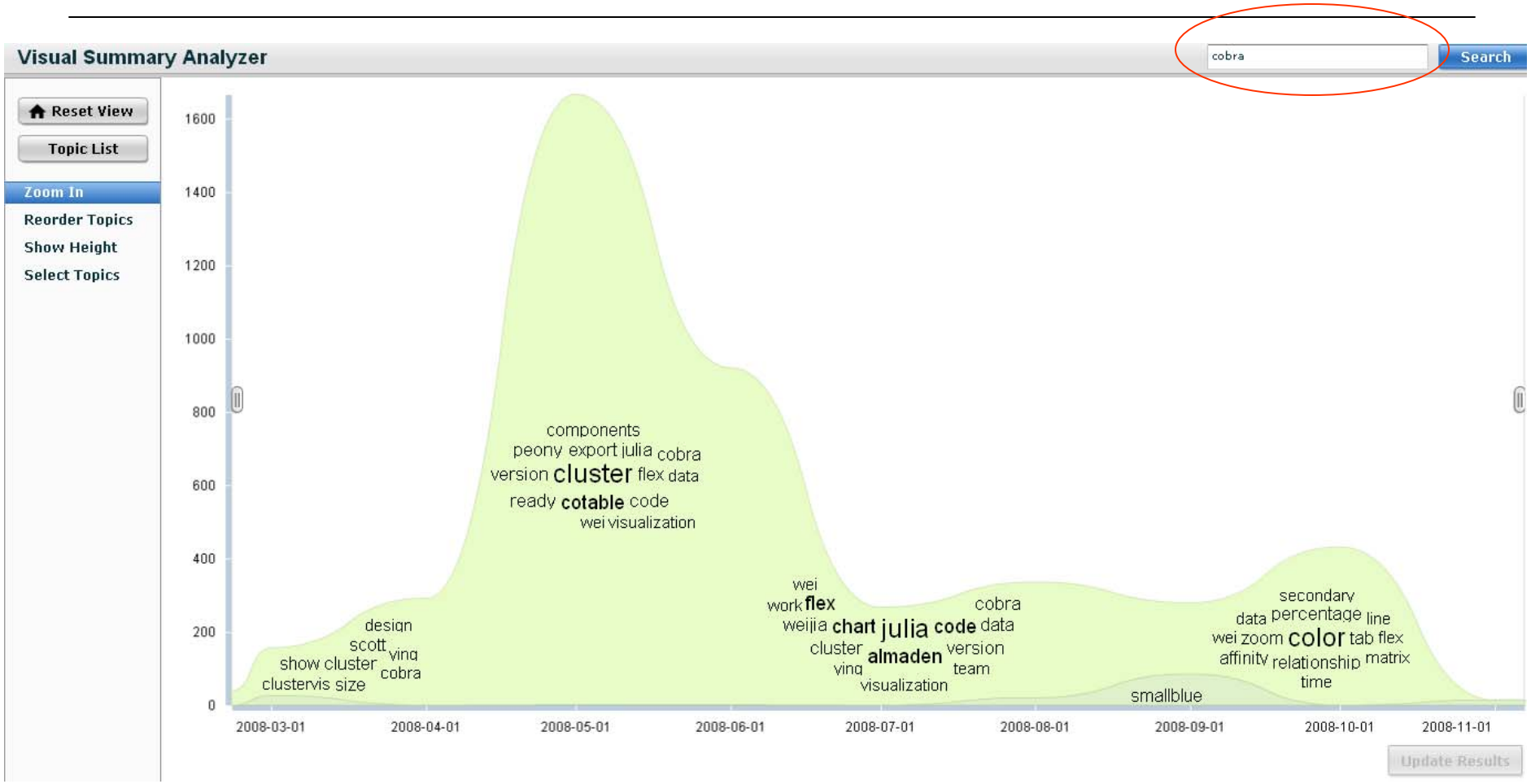
Drill down into the emails



Click “shimei pan”



Search “cobra”

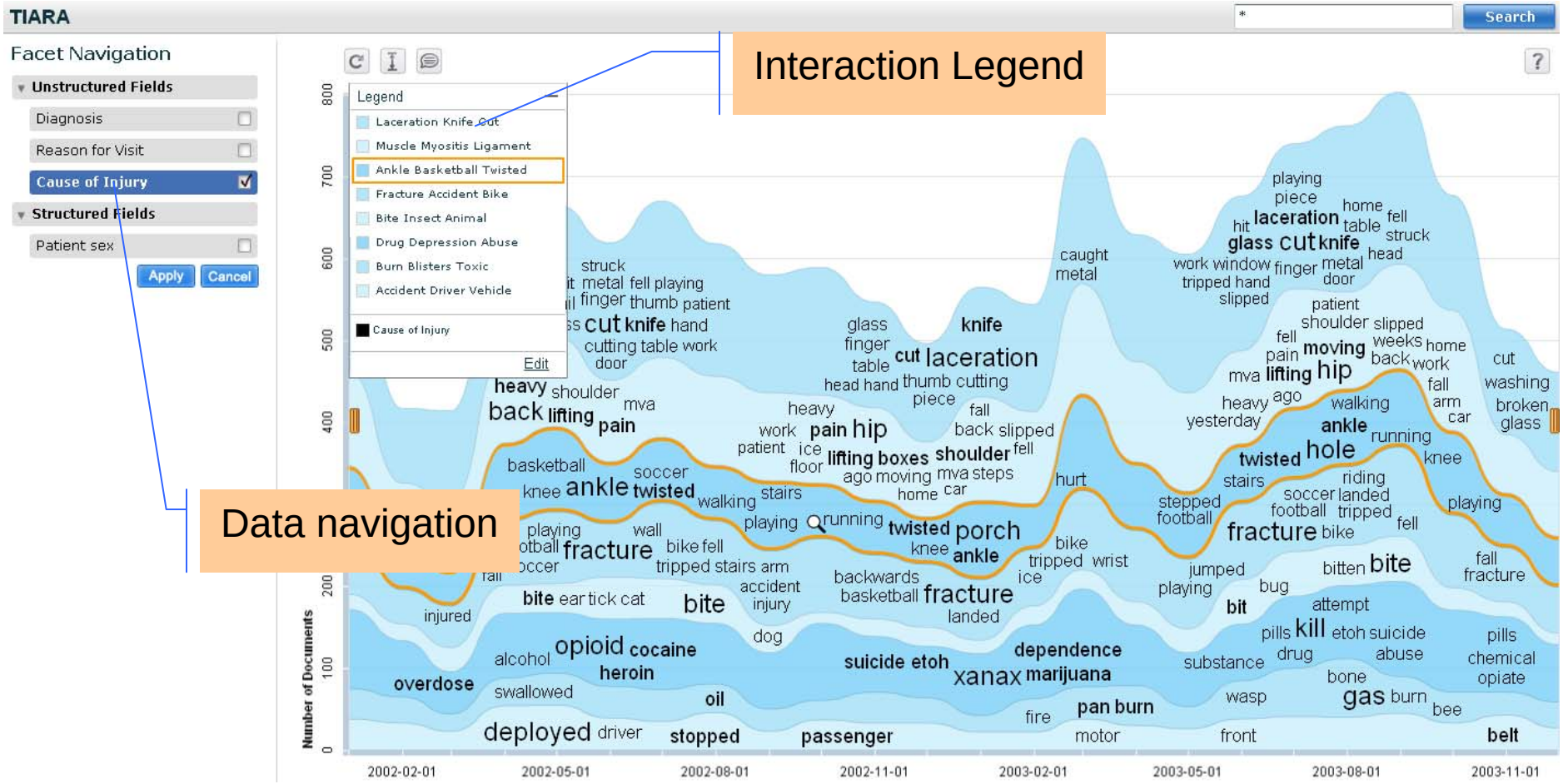


Application: Visual Patient Record Analysis

- **Who**
  - Alice: A government officer responsible for disease control and prevention, who is investigating the major causes and diagnosis for residential illnesses countrywide
- **Data**
  - Healthcare data: emergency room data set containing 23,501 patient records from 2002 to 2003
  - Free text fields: cause of injury, reason for visit, and diagnosis
  - Structured fields: patient sex, age, etc.

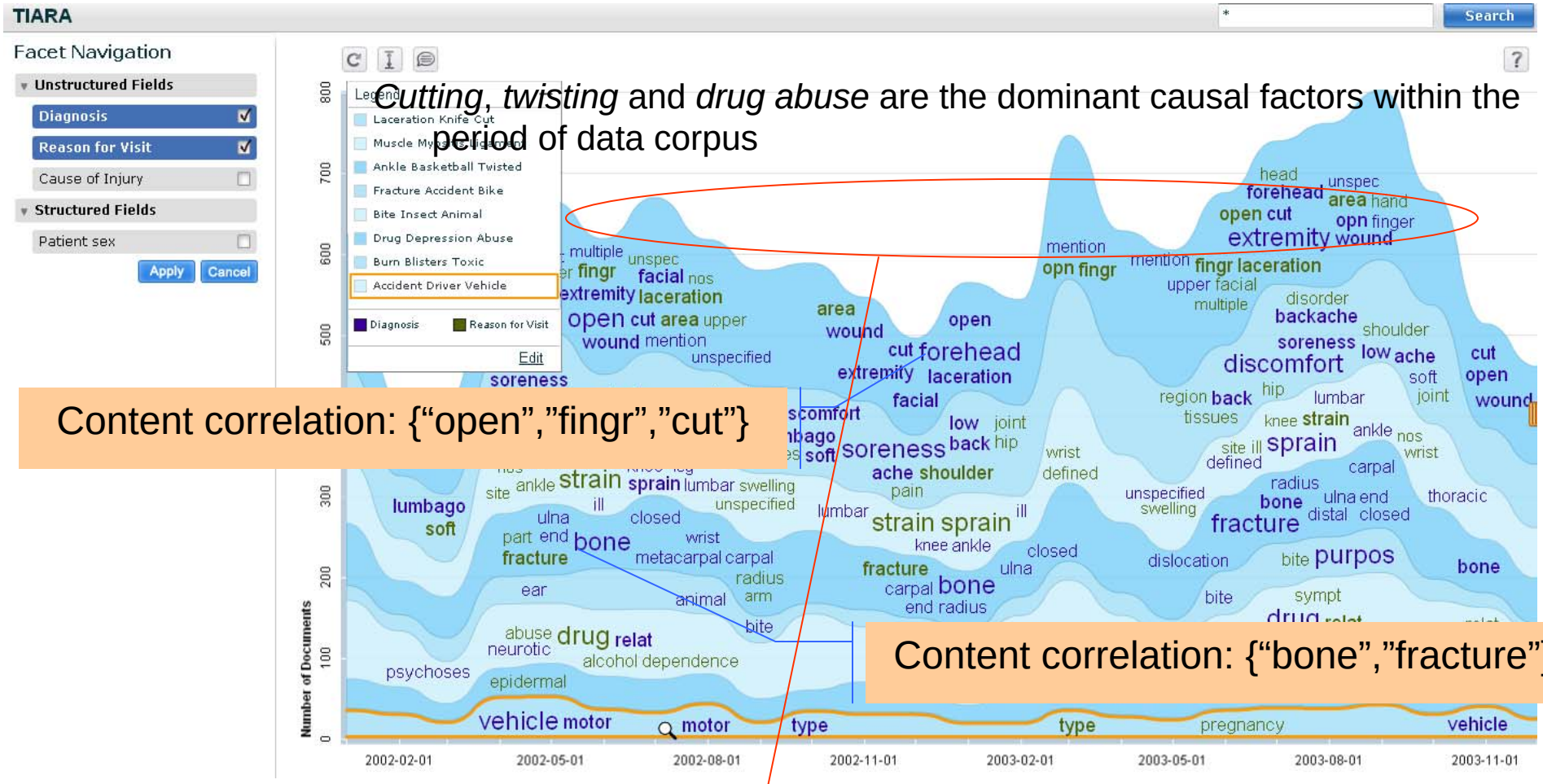
How could TIARA help her ?

Application: Visual Patient Record Analysis



TIARA's visual summary of the "cause of injury" field of the 23,000+ emergency room records from 2002 to 2003

Step 1: Select “reason for visit” and “diagnosis”



Cutting, *twisting* and *drug abuse* are the dominant causal factors within the period of data corpus

Content correlation: {“open”,”fingr”,”cut”}

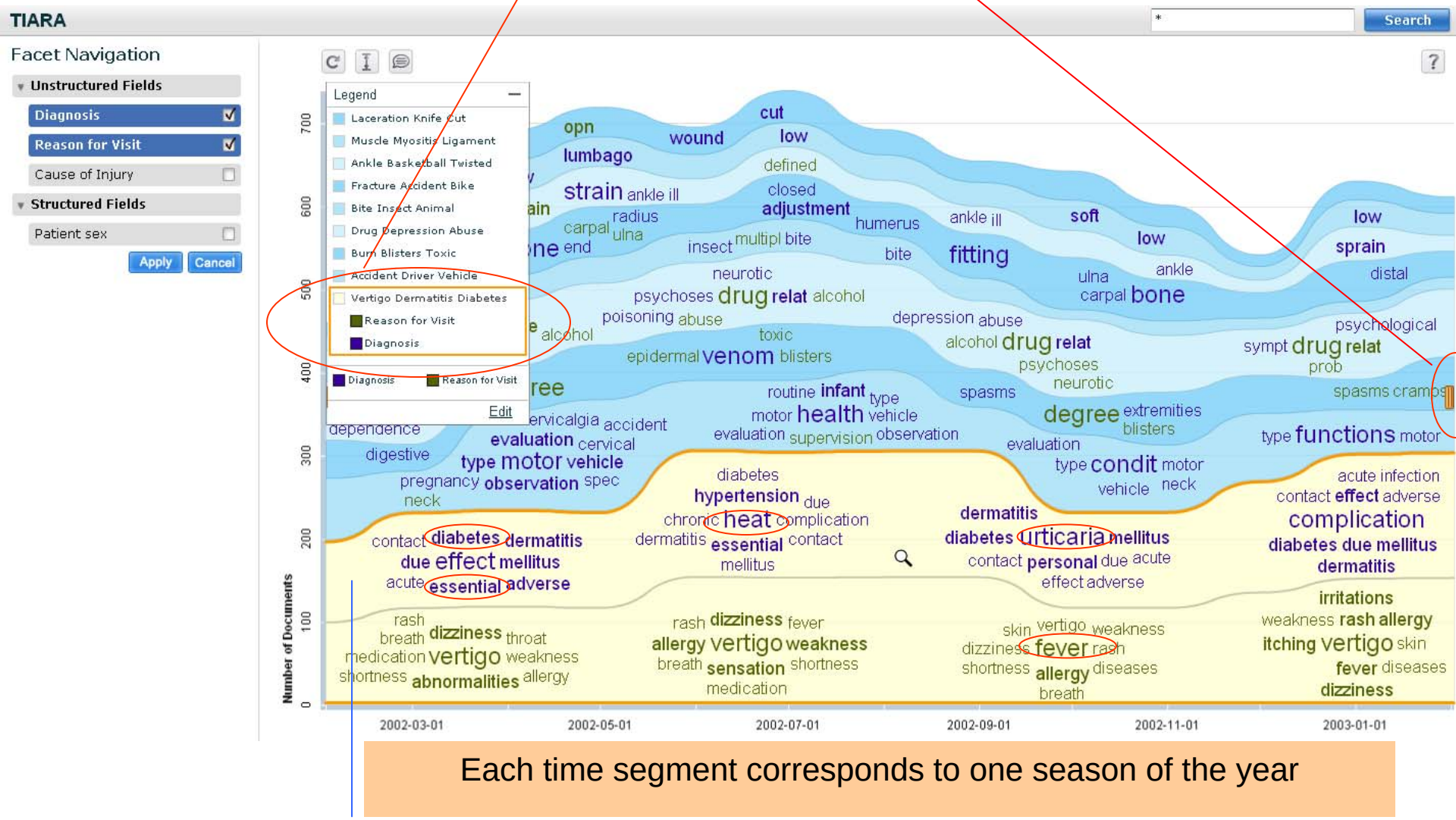
Content correlation: {“bone”,”fracture”}

the number of cases falls in each winter and rebounds from the spring

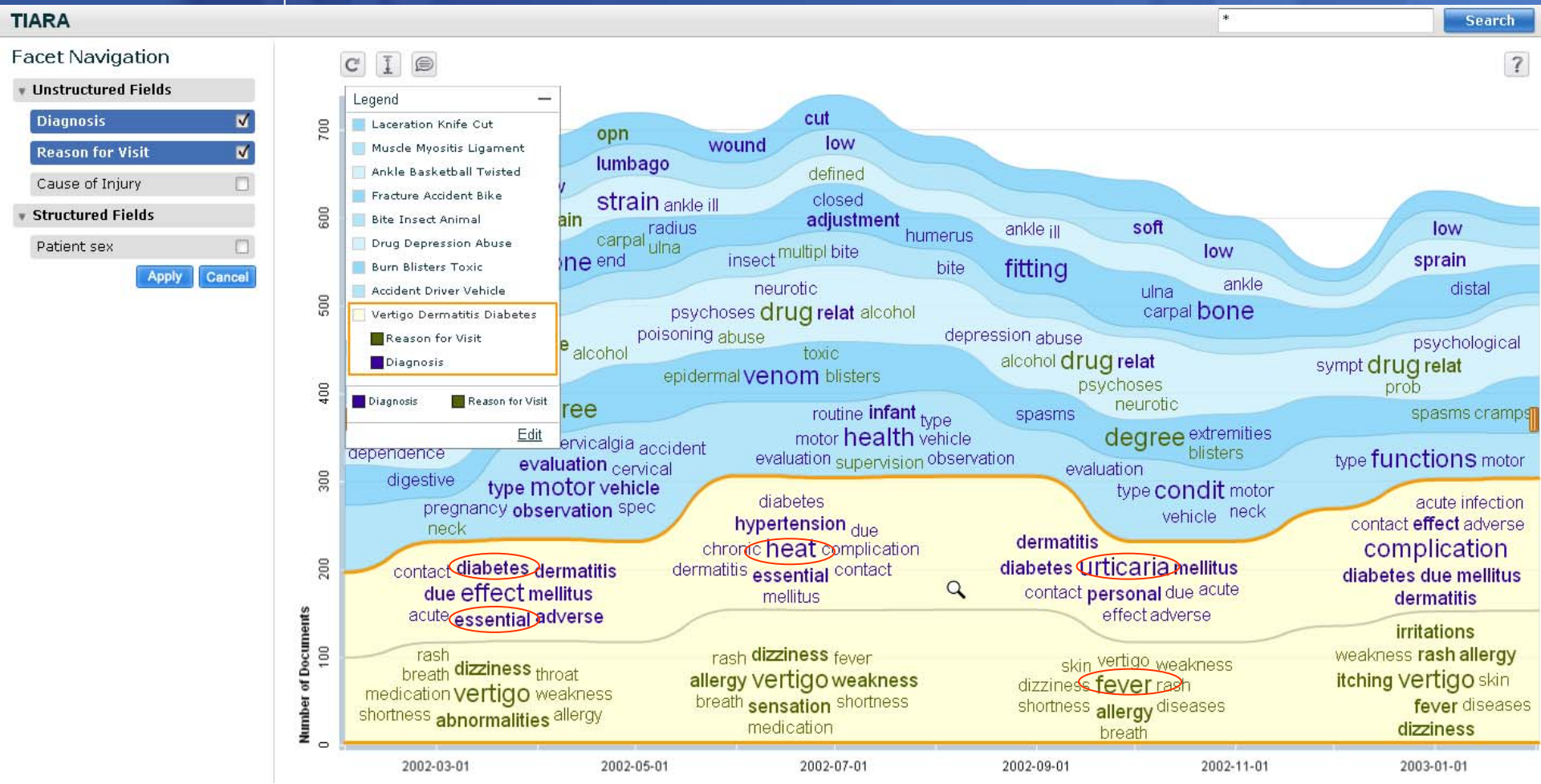
Step 2: Time zoom in (Feb 2002 to Jan 2003 within a whole year)

Step 3: Select the topic indicating "vertigo" illnesses from the full topic list (interactive legend)

Step 4: Click the "vertigo" topic trend to expand the view and show correlations

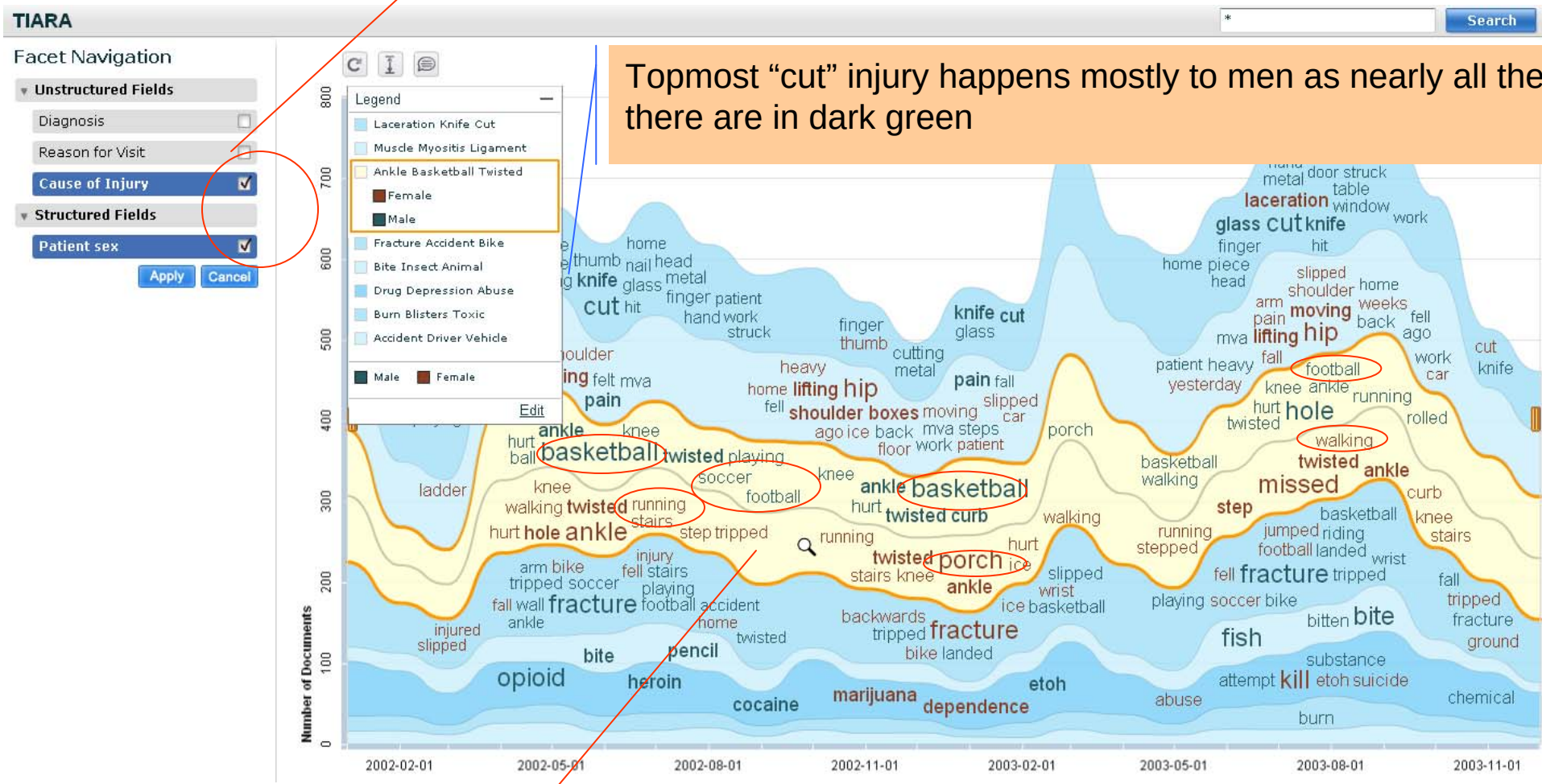


Each time segment corresponds to one season of the year



In spring, the patient suffers from adverse effect of drugs as well as some common diseases like diabetes and essential hypertension. While in summer, the high temperature which causes heat exhaustion turns out dominating. Further in winter, the same symptom may be ascribed to complications of common illnesses. The patterns in autumn are quite outstanding, where more urticaria and fever is found.

Step 5: Select the “cause of injury” field
Step 6: Select the “patient sex”

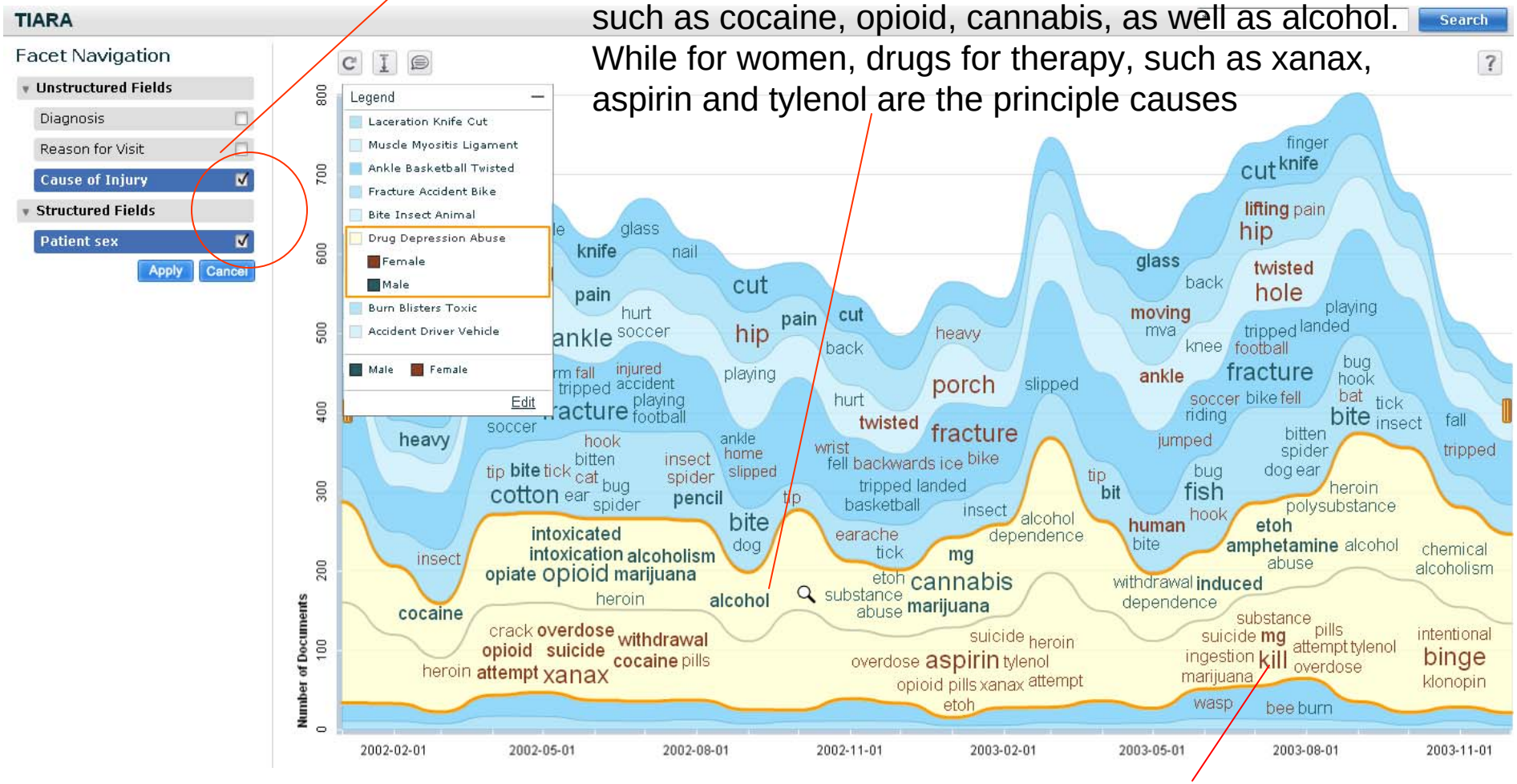


Topmost “cut” injury happens mostly to men as nearly all the there are in dark green

Men tend to twist their ankle during heavy sports including basketball, football and soccer. Women generally get their ankles hurt during walking, running in the porch or missing their steps downstairs

Step 5: Select the “cause of injury” field
Step 6: Select the “patient sex”

The cause of injury for men mainly refers to hard drugs, such as cocaine, opioid, cannabis, as well as alcohol. While for women, drugs for therapy, such as xanax, aspirin and tylenol are the principle causes



Suicide: “kill” and “attempt” almost exclusively for women

Outline

- Introduction (What)
- The TIARA System (How)
  - System Overview
  - Analytics
  - Experiments
- Demonstration
- Conclusion

Conclusions

- **TIARA to help users visually view, explore and analyze large collections of documents**

- **Lessons from our case studies and initial deployment**
  - TIARA is even more effective for business professionals
  - It is more effective for those who have some background knowledge (familiar with) on the data

What a topic is?

Future Work

- **Use any categorized facet as the topic layers**
  - Topics are difficult to be understood by common users
  - Classification labels, companies, hotels, ages, pos (part-of-speech), sentiment orientations, etc.
- **Visualize more meaningful text unit**
  - Besides keywords, we can show more meaningful text units: time-sensitive NEs (named entities), phrases, etc.
  - More NLP/IE and mining techniques are expected
- **Large scale data sets**
  - Sampling, select top *N* documents, or indexing only the top *N* topic keywords

If you want to see our live demo, please contact us!

THANK YOU