FutureLens

Gregory Shutt
November 20, 2008

Motivation

- Visualize tagged data
- Extract features from data
- Knowledge discovery

Data

- VAST 2007 Contest
- News stories
- SGML files tagged with different types of entities
  - Person, organization, money, date, location

<TIMEX TYPE="DATE">
Fri Aug 15 2003
</TIMEX>
<ENAMEX TYPE="PERSON">
Jon Zwickel
</ENAMEX>

wanted to create the ultimate B.C. hot dog. Hence the world has the

<ENAMEX TYPE="ORGANIZATION">
PNE Salmon Sausage
</ENAMEX>

, a new taste treat that will be unveiled when the Pacific National Exhibition opens

<TIMEX TYPE="DATE">**Saturday**</TIMEX>
<TIMEX TYPE="TIME">**morning**</TIMEX>.

"There's nothing more

<ENAMEX TYPE="LOCATION">
West Coast
</ENAMEX>

than salmon," said

<ENAMEX TYPE="PERSON">
Zwickel
</ENAMEX>

<TIMEX TYPE="DATE">
Wed Aug 6 2003
</TIMEX>

These [genetically engineered] products are absolutely safe. For the most part you wouldn't know [if you were eating them] but the point being that you wouldn't need to know.

<ENAMEX TYPE="PERSON">
Bryan Hurley
</ENAMEX>

,

<ENAMEX TYPE="ORGANIZATION">
Monsanto
</ENAMEX>

spokesperson

Sample Data

Data

- Nonnegative tensor factorization software used
- NTF software output 25 group files
- Each group was described by a number of interrelated entities and terms

Requirements

- Visualize VAST 2007 data
- Simple and easily modifiable
- Maintain UI responsiveness
- Allow viewing of individual group files

Group 9

Entities

0.3588235	$215 million
0.3588235	Cruz
0.3588235	Darla Banks
0.3588235	$25-30 million
0.3588235	Banks

Terms

0.3258677	banks
0.2219373	fishes
0.1687334	tropical
0.1465103	trafficking
0.1447117	brazil
0.1373243	illegal
0.1254398	poachers
0.1246595	fish

Relevant

Irrelevant

File: Week-of-Mon-20030630.xml.txt.p.NE

Something is rotten in the tropical **fish** import business and not just some dead **fish**. A southern environmentalist has succeeded in trapping **poachers** by conducting sting operations in **Brazil** - and **Darla Banks** loves doing this. She carries a concealed camera in her handbag and secretly films **illegal** freshwater fish collections, including the rare Black arwana and **Cruz**' Dwarf Pearlfish.

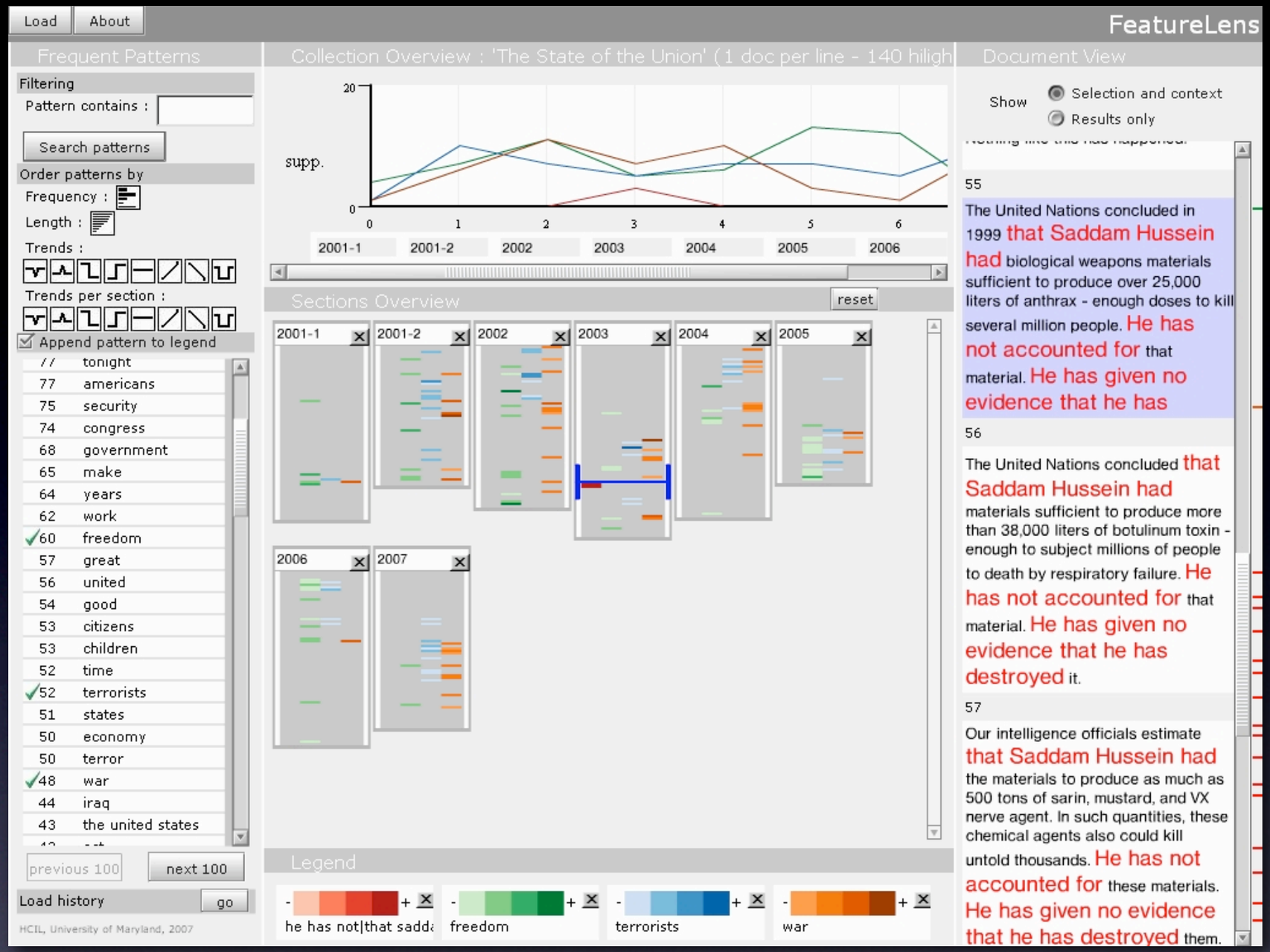
From 340-500 million **fishes** are kept in American homes (three times the total number of dogs and cats). Trade in **fish** grows every year. At least US **$ 215 million** in **tropical fishes** are handled every year in the US . The US imports 125 million of ornamental **fishes** per year - US **$ 25-30 million** /year.

File: Week-of-Mon-20030818.xml.txt.p.NE

Fall is time to see bighorns in Hells Canyon Out & About A herd of 18 bighorn sheep wanders along the **banks** of the Snake River in Hells Canyon, moseying from rock to rock, drinking from the river, and chewing on grass. Five rams, with massive horns curling over the sides of their heads, are in the herd. What a sight. A lamb bounds playfully but cautiously between the adult animals. Bighorn numbers started to decline as soon as the state began to be settled. They were easy to hunt and provided food for early miners and settlers. And they were -- and still are -- susceptible to diseases like scabies and pasturella, which are transmitted from domestic sheep. As homesteaders brought in more domestic sheep, the bighorns became sick and died. "The bottom line is there is no danger to domestic sheep from wild sheep," Cassirer says. "It's only one way."

Background

- Conceptually based on FeatureLens, a University of Maryland HCIL project
- Visualizes frequent terms and patterns in text over time



FeatureLens

FeatureLens

- Requires MySQL server, HTTP server, Adobe Flash enabled browser
- Written in Ruby and OpenLaszlo
- Difficult to modify
- Arbitrary data sets not loadable by end users
- Interface responsiveness is subpar

FutureLens

- Written in Java using SWT
- Cross platform with native look and feel
- Works with SGML and raw text
- Supports tagged entities
- Allows viewing of group files

Cross platform but it maintains the look and feel familiar to the user of the particular operating system

Most Java programs do not have this capability

Demonstration

- A demonstration of how a scenario can quickly be visualized using NTF output and the VAST 2007 data

Future Work

- Integrate data mining software
- Allow dynamic data sets
- Use machine learning to automate tasks

References

- Exploring and visualizing frequent patterns in text collections with FeatureLens. http://www.cs.umd.edu/hcil/textvis/featurelens. Visited November 2008.

- The MONK Project Wiki. https://apps.lis.uiuc.edu/wiki/display/MONK/The+MONK+Project+Wiki. Last edited August 2008.

- Brett W. Bader, Michael W. Berry, and Murray Brown. Discussion tracking in Enron email using PARAFAC. In M.W. Berry and M. Castellanos, editors, Survey of Text Mining II: Clustering, Classification, and Retrieval, pages 147–163. Springer-Verlag, London, 2008.

- Brett W. Bader, Michael W. Berry, and Amy N. Langville. Nonnegative matrix and tensor factorization for discussion tracking. In A. Srivastava and M. Sahami, editors, Text Mining: Theory, Applications, and Visualization. Chapman & Hall / CRC Press, 2008.

- Brett W. Bader, Andrey A. Puretskiy, and Michael W. Berry. Scenario discovery using nonnegative tensor factorization. In Jose Ruiz-Shulcloper and Walter G. Kropatsch, editors, Progress in Pattern Recognition, Image Analysis and Applications, Proceedings of the Thirteenth Iberoamerican Congress on Pattern Recognition, CIARP 2008, Havana, Cuba, Lecture Notes in Computer Science (LNCS) 5197, pages 791–805. Springer-Verlag, Berlin, 2008.

- A. Don, E. Zhelev, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. HCIL Technical report 2007-08, May 2007.