

FutureLens

A Visual Text Analysis Tool

G. Shutt, A. Puretskiy, M.W. Berry
Department of Electrical Engineering and Computer
Science
University of Tennessee, Knoxville

Background

- 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
```

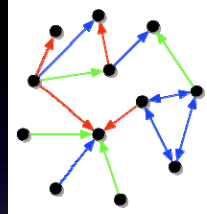
Sample Data

NTF

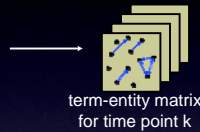
- Nonnegative tensor factorization software used to extract features through time
- Each feature corresponds to a particular subset of documents
 - NTF software outputs 25 group files
 - Each group is described by a number of interrelated entities and terms

NTF: Multidimensional Data Analysis

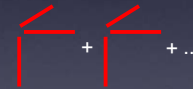
Fictitious News Articles



Build a 3-way array such that there is a term-entity matrix for each time point.



Multilinear algebra



Nonnegative PARAFAC

Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

NTF Outer Product Form

Outer (tensor) Product: $\mathbf{a} \circ \mathbf{b} = \mathbf{a} \mathbf{b}^T$

Third-order Tensor Representation: $\mathbf{X} \approx \sum_{l=1}^r \mathbf{a}_l \circ \mathbf{b}_l \circ \mathbf{c}_l$

Goal of NTF: $\min_{\mathbf{A}, \mathbf{B}, \mathbf{C}} \left\| \mathbf{X} - \sum_{l=1}^r \mathbf{a}_l \circ \mathbf{b}_l \circ \mathbf{c}_l \right\|$

Accomplished by solving a series of non-negative matrix factorization (NMF) subproblems:

$$\min_{\mathbf{A} \in \mathbb{R}_+^{m \times r}} \left\| \mathbf{X}^{(m \times np)} - \mathbf{A}(\mathbf{C} \circ \mathbf{B})^T \right\|_F$$

$$\min_{\mathbf{B} \in \mathbb{R}_+^{n \times r}} \left\| \mathbf{X}^{(n \times mp)} - \mathbf{B}(\mathbf{C} \circ \mathbf{A})^T \right\|_F$$

$$\min_{\mathbf{C} \in \mathbb{R}_+^{p \times r}} \left\| \mathbf{X}^{(p \times mn)} - \mathbf{C}(\mathbf{B} \circ \mathbf{A})^T \right\|_F$$

Sample NTF Output

```
##### Group 15 #####
Scores  Idx Name
0.2485621 7120 bruce longhorn 7120
0.2485621 7122 longhorn 7122
0.2485621 7128 chelmsworth 7128
0.2485621 7124 gil 7124
0.2485621 7121 virginia tech 7121
0.2485621 7125 mary ann ollesen 7125
...
Scores  Idx Term
0.2958673 6907 monkeypox
0.2054770 7468 outbreak
0.2008147 6358 longhorn
0.1594331 4644 gil
0.1552401 1856 chinchilla
0.1434742 11049 travel
0.1391984 9322 sars
0.1379675 1857 chinchillas
0.1342139 2372 continent
0.1294389 3888 expect
0.1215461 9711 sick
0.1161760 7469 outbreaks
0.1144558 3883 exotic
0.1122925 7824 pets
0.1026513 8088 pot-bellied
0.1026513 7229 novelty
0.1019125 1742 cesar
0.1004109 10280 strain
0.1000808 5878 jul
...
```

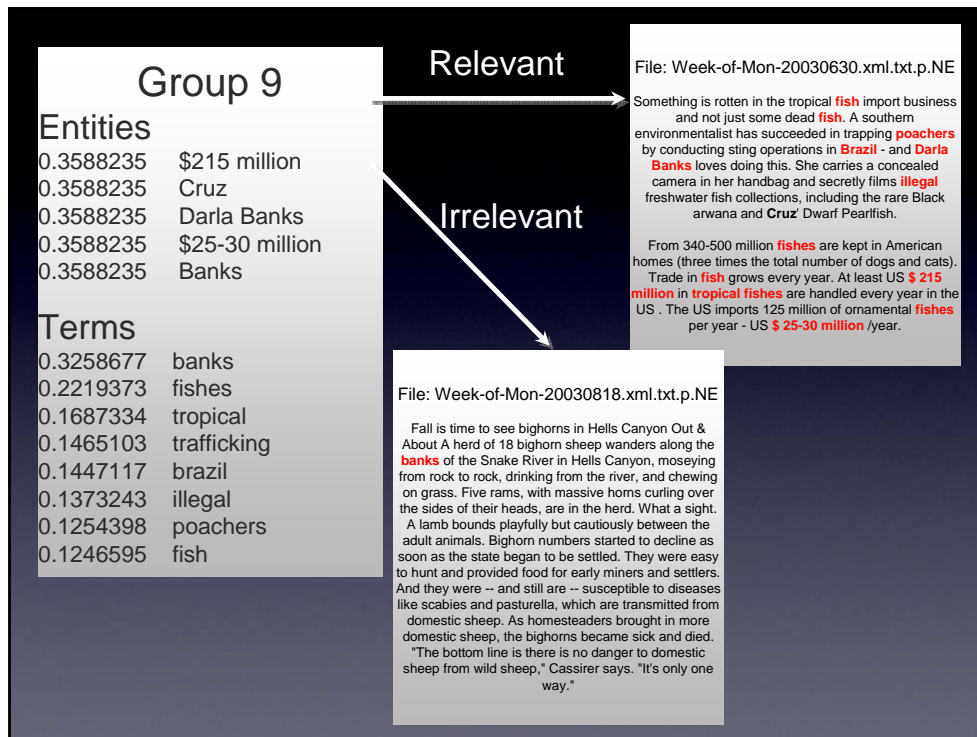
Sample NTF Output

```
##### Group 20 #####
Scores  Idx Name
0.2252609 4680 scott roberts 4680
0.2252609 4685 zhang 4685
0.2252609 4687 roberts 4687
0.2252609 4682 iron and steel statistics bureau 4682
0.2252609 4686 $13 billion 4686
0.2252609 4681 cambridge energy research associates
4681
0.2252609 4679 communist party 4679
0.2252609 4689 deutsche bank 4689
...
Scores  Idx Term
0.2140977 3644 energy
0.1915396 1855 china
0.1502502 8104 power
0.1321501 1011 beijing
0.1239235 7340 oil
0.1155490 9140 roberts
0.1130895 2146 communist
0.1129717 10203 steel
0.1057306 2023 coal
0.1018353 4842 growth
0.0983195 1529 cambridge
0.0977857 7920 plagued
0.0968663 5201 hoopla
0.0958164 11951 zhang
0.0948698 5247 hottest
0.0942696 1227 booming
0.0927391 3928 explosive
```

FutureLens

Motivation

- Inspired by FeatureLens, a University of Maryland Human-Computer Interaction Lab Project
- Visualization to Facilitate Analysis of Textual Data
- Visualization of NTF Output
 - Feature (event/activity) Tracking through Time
 - Knowledge discovery



FutureLens

Software Specifications

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

FutureLens

Features

- Automatically Loads All Terms Found in Input Dataset (except those on the list of exclusions)
- Ability to Search through Terms
- Ability to Sort Terms
- Ability to Create Collections of Terms
- Ability to Create Phrases

Demonstration

- A demonstration of how scenarios can be visualized using NTF-generated output based on the VAST 2007 dataset

Future Work

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

References

- 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. Pureskiy, 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.

References

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