Nonnegative Tensor Factorization for ~~Sentiment Analysis~~ Knowledge Discovery

TDA 2010 Workshop, Monopoli, Italy

Michael W. Berry and Andrey Puretskiy
Center for Intelligent Systems & Machine Learning
Department of Electrical Engineering & Computer Science
University of Tennessee, Knoxville

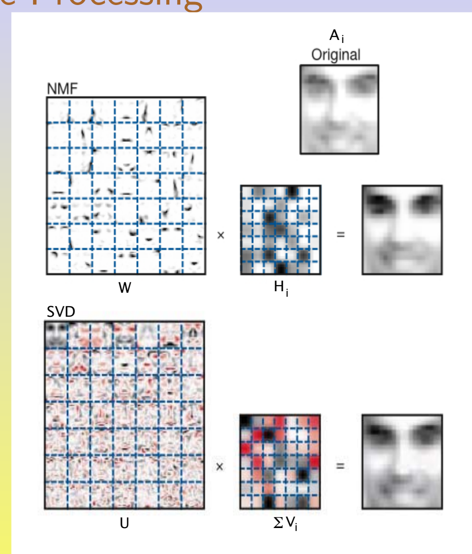
September 13, 2010

Outline of Presentation

NMF Origins

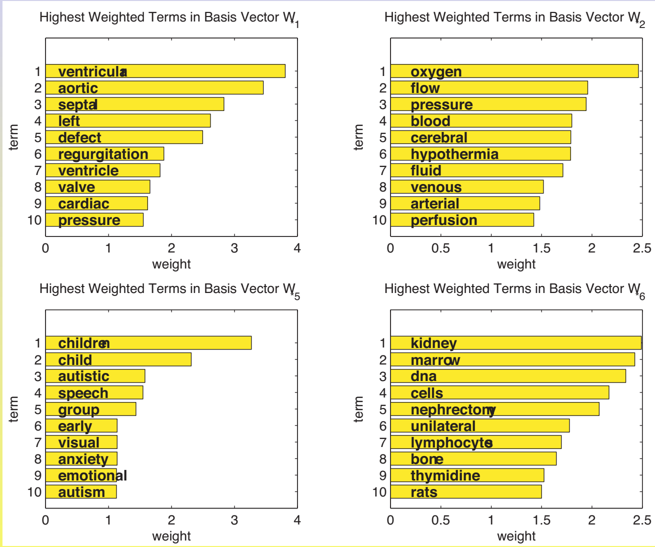
- NMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
- Lee and Seung (1999) demonstrated its use as a *sum-by-parts* representation of image data in order to both identify and classify image *features*.

NMF for Image Processing



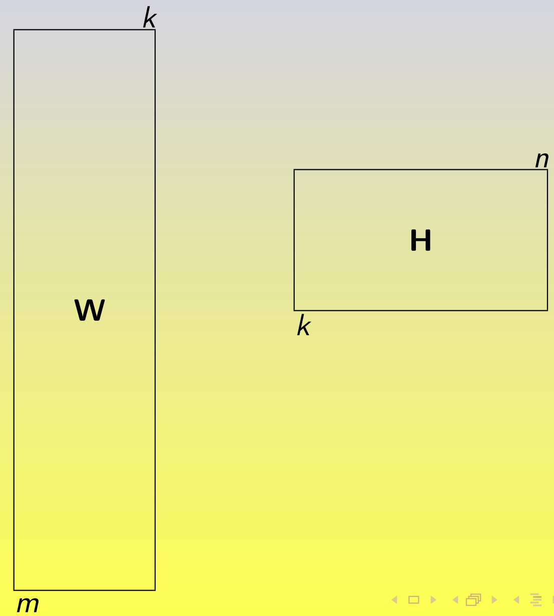
Sparse NMF versus Dense SVD Bases; Lee and Seung (1999)

NMF for Text Mining (Medlars)

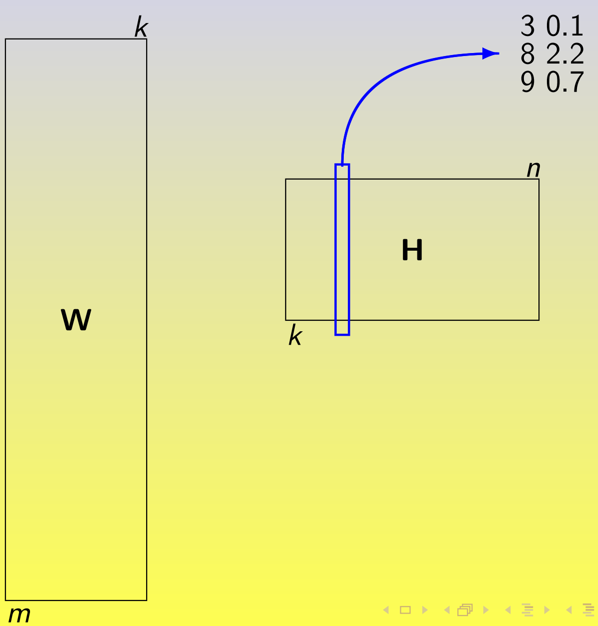


Interpretable NMF feature vectors; Langville et al. (2006)

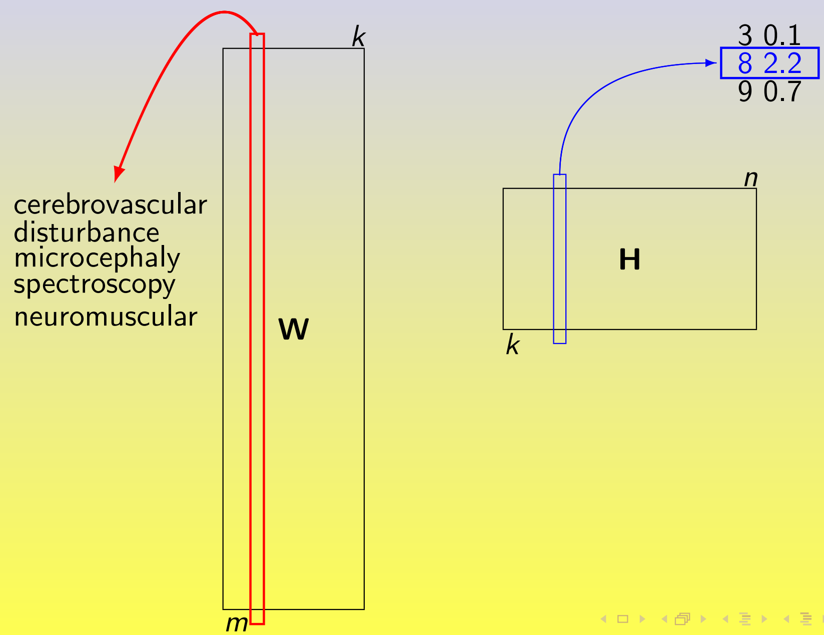
NNMF Schematic (Text Mining)



NNMF Schematic (Text Mining)



NNMF Schematic (Text Mining)



Derivation

- Given an $m \times n$ term-by-document (sparse) matrix X.
- Compute two reduced-dim. matrices W,H so that $X \simeq WH$; W is $m \times r$ and H is $r \times n$, with $r \ll n$.
- **Optimization problem**:

$$\min_{W,H} \|X - WH\|_F^2,$$

  subject to $W_{ij} \geq 0$ and $H_{ij} \geq 0$, $\forall i, j$.
- **General approach**: construct initial estimates for W and H and then improve them via alternating iterations.
- **Local Minima**: Non-convexity of functional $f(W,H) = \frac{1}{2}\|X - WH\|_F^2$ in both W and H.
- **Non-unique Solutions**: $WDD^{-1}H$ is nonnegative for any nonnegative (and invertible) D.

Multiplicative Method (MM)

- Multiplicative update rules for W and H (Lee and Seung, 1999):
  1. Initialize W and H with nonnegative values, and scale the columns of W to unit norm.
  2. Iterate for each c, j, and i until convergence or after k iterations:
     1. $H_{cj} \leftarrow H_{cj} \dfrac{(W^T X)_{cj}}{(W^T WH)_{cj} + \epsilon}$
     2. $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
     3. Scale the columns of W to unit norm.
- Setting $\epsilon = 10^{-9}$ will suffice to avoid division by zero.

Multiplicative Method (MM) contd.

Multiplicative Update MATLAB® Code for NMF

$\mathbf{W} = \text{rand}(m,k);$ % $\mathbf{W}$ initially random
$\mathbf{H} = \text{rand}(k,n);$ % $\mathbf{H}$ initially random
for i = 1 : maxiter
$\quad \mathbf{H} = \mathbf{H}$.* $(\mathbf{W^T A})$ ./ $(\mathbf{W^T W H} + \epsilon)$;
$\quad \mathbf{W} = \mathbf{W}$.* $(\mathbf{A H^T})$ ./ $(\mathbf{W H H^T} + \epsilon)$;
end

Improving Feature Interpretability

Gauging Parameters for Constrained Optimization

How sparse (or smooth) should factors (W, H) be to produce as many interpretable features as possible?

To what extent do different norms (L_1, L_2, L_∞) improve/degrade feature quality or span? At what cost?

Can a common nonnegative feature space be built from objects in both images and text? Are there opportunities for multimodal document similarity?

Enron Email Collection and Historical Events

- By-product of the FERC investigation of Enron (originally contained 15 million email messages).
- This study used the improved corpus known as the Enron Email set, which was edited by Dr. William Cohen at CMU.
- This set had over 500,000 email messages; most sent in the [1999,2001] time interval.

- Ongoing, problematic, development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra.
- Deregulation of the Calif. energy industry, which led to rolling electricity blackouts in summer of 2000.
- Revelation of Enron's deceptive business and accounting practices that led to collapse in Oct. 2001 and bankruptcy in Dec. 2001.

PRIVATE Collection

- Parsed all mail directories (of all 150 accounts) with the exception of `all_documents`, `calendar`, `contacts`, `deleted_items`, `discussion_threads`, `inbox`, `notes_inbox`, `sent`, `sent_items`, and `_sent_mail`; 495-term stoplist used and extracted terms must appear in more than 1 email and more than once globally; log-entropy term weighting used for elements of X.
- Distribution of messages sent in the year 2001:

Month	Msgs	Terms	Month	Msgs	Terms
Jan	3,621	17,888	Jul	3,077	17,617
Feb	2,804	16,958	Aug	2,828	16,417
Mar	3,525	20,305	Sep	2,330	15,405
Apr	4,273	24,010	Oct	2,821	20,995
May	4,261	24,335	Nov	2,204	18,693
Jun	4,324	18,599	Dec	1,489	8,097

Topics identified in PRIVATE Enron subcollection

- Identify rows of H from $X \simeq WH$ or H^k; $r = 50$ feature vectors (W_k):

Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
10	497	California	ca, **cpuc, gov, socalgas**, sempra, org, sce, gmssr, aelaw, ci
23	43	Louise Kitchen named top woman by Fortune	evp, **fortune,** britain, woman, **ceo**, avon, fiorina, cfo, hewlett, packard
26	231	Fantasy football	game, wr, qb, play, rb, season, injury, updated, fantasy, image

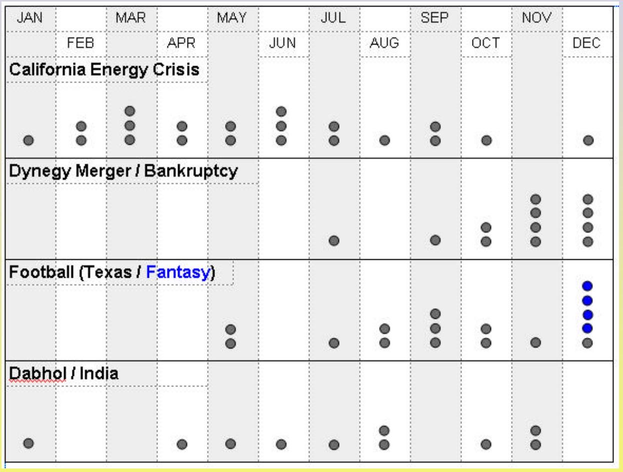
(Cluster size $\equiv$ no. of H^k elements $>$ $row_{max}/10$)

Topics identified in PRIVATE Enron subcollection, contd.

- Additional topic clusters of significant size:

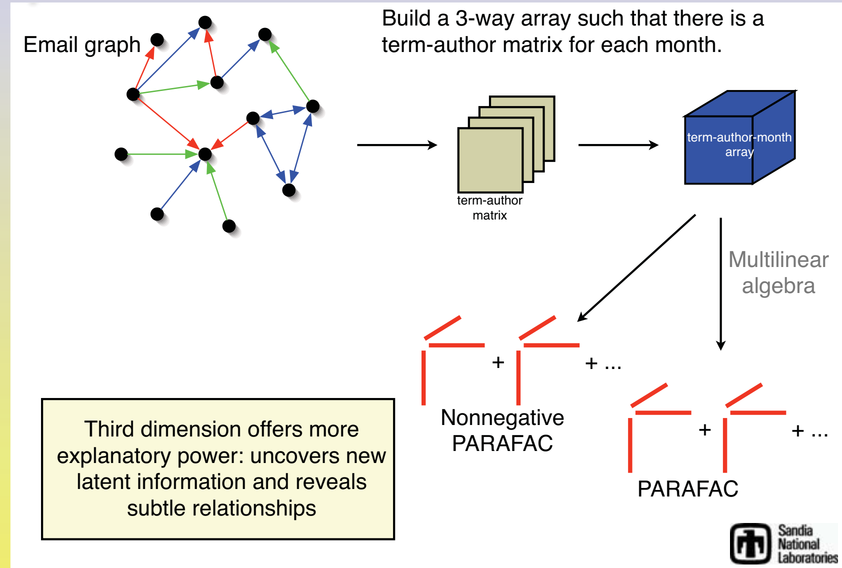
Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
33	233	Texas longhorn football newsletter	UT, orange, longhorn[s], texas, true, truorange, recruiting, oklahoma, defensive
34	65	Enron collapse	**partnership[s], fastow**, shares, **sec**, stock, shareholder, investors, equity, **lay**
39	235	Emails about India	**dabhol, dpc, india, mseb, maharashtra,** indian, lenders, delhi, foreign, minister

2001 Topics (NMF Features) Through Time



(*New York Times*, May 22, 2005)

Multidimensional Data Analysis via PARAFAC



Mathematical Notation

- Kronecker product

$$A \otimes B = \begin{pmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{pmatrix}$$

- Khatri-Rao product (columnwise Kronecker)

$$A \odot B = \begin{pmatrix} A_1 \otimes B_1 & \cdots & A_n \otimes B_n \end{pmatrix}$$

- Outer product

$$A_1 \circ B_1 = \begin{pmatrix} A_{11}B_{11} & \cdots & A_{11}B_{m1} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{11} & \cdots & A_{m1}B_{m1} \end{pmatrix}$$

PARAFAC Representations

- PARAllel FACtors (Harshman, 1970)
- Also known as CANDECOMP (Carroll & Chang, 1970)
- Typically solved by Alternating Least Squares (ALS)

Alternative PARAFAC formulations

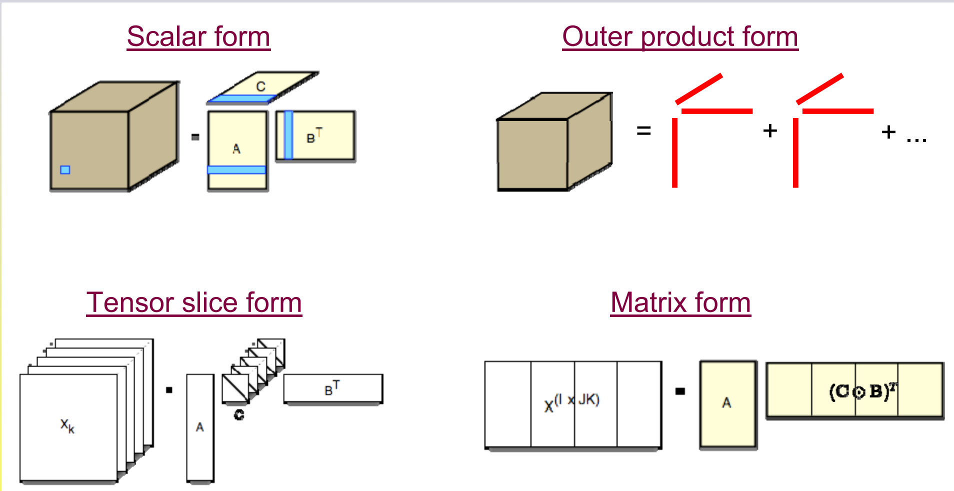
$$X_{ijk} \approx \sum_{i=1}^{r} A_{ir} B_{jr} C_{kr}$$

$$\mathcal{X} \approx \sum_{i=1}^{r} A_i \circ B_i \circ C_i, \text{ where } \mathcal{X} \text{ is a 3-way array (tensor).}$$

$$X_k \approx A \,\mathrm{diag}(C_{k:})\, B^T, \text{ where } X_k \text{ is a tensor slice.}$$

$$X^{I \times JK} \approx A(C \odot B)^T, \text{ where } X \text{ is matricized.}$$

PARAFAC (Visual) Representations



Nonnegative PARAFAC Algorithm

- Adapted from (Mørup, 2005) and based on NMF by (Lee and Seung, 2001)

$$\begin{aligned} \|X^{I\times JK} - A(C\odot B)^T\|_F &= \|X^{J\times IK} - B(C\odot A)^T\|_F \\ &= \|X^{K\times IJ} - C(B\odot A)^T\|_F \end{aligned}$$

- Minimize over A, B, C using multiplicative update rule:

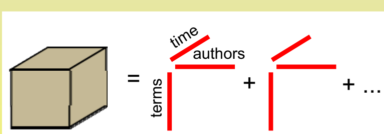
$$A_{i\rho} \leftarrow A_{i\rho}\frac{(X^{I\times JK}Z)_{i\rho}}{(AZ^TZ)_{i\rho}+\epsilon}, \quad Z=(C\odot B)$$

$$B_{j\rho} \leftarrow B_{j\rho}\frac{(X^{J\times IK}Z)_{j\rho}}{(BZ^TZ)_{j\rho}+\epsilon}, \quad Z=(C\odot A)$$

$$C_{k\rho} \leftarrow C_{k\rho}\frac{(X^{K\times IJ}Z)_{k\rho}}{(CZ^TZ)_{k\rho}+\epsilon}, \quad Z=(B\odot A)$$

Discussion Tracking Using Year 2001 Subset

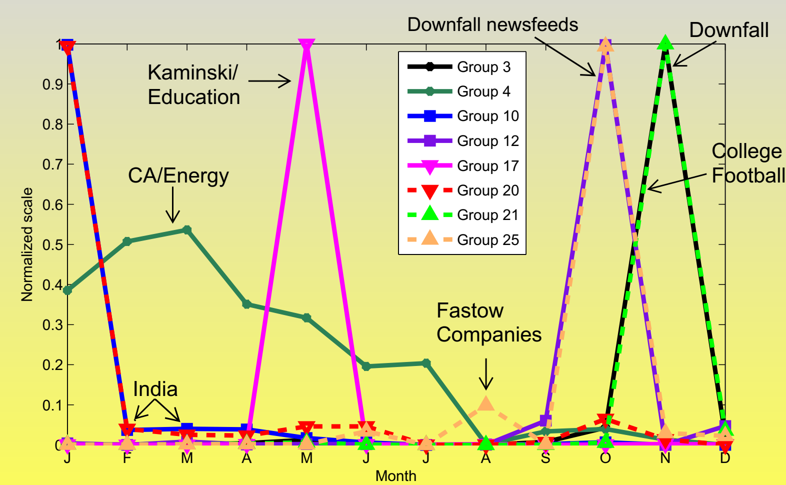
- 197 authors (`From:user_id@enron.com`) monitored over 12 months;
- Parsing 34,427 email subset with a base dictionary of 121,393 terms (derived from 517,431 emails) produced 69,157 unique terms; (term-author-month) array X has $\sim$ 1 million nonzeros.
- Rank-25 tensor: A $(69{,}157 \times 25)$, B (197×25), C (12×25)



Month	Emails	Month	Emails
Jan	7,050	Jul	2,166
Feb	6,387	Aug	2,074
Mar	6,871	Sep	2,192
Apr	7,382	Oct	5,719
May	5,989	Nov	4,011
Jun	2,510	Dec	1,382

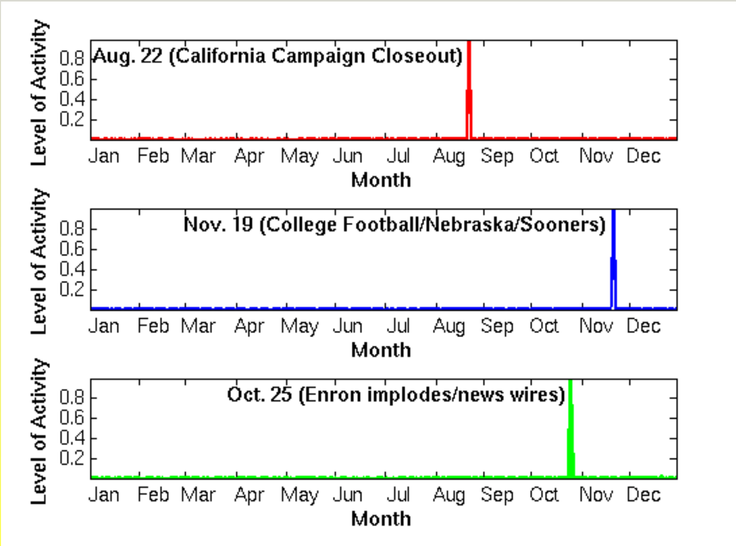
Tensor-Generated Group Discussions

- NTF Group Discussions in 2001
- 197 authors; 8 distinguishable discussions
- "Kaminski/Education" topic previously unseen



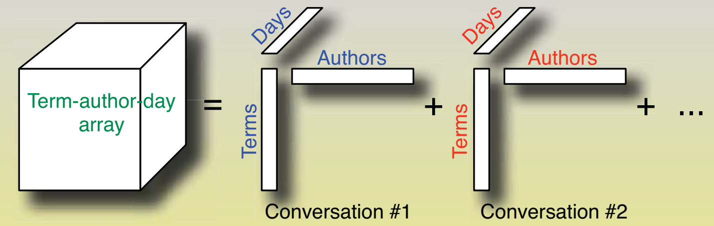
Day-level Analysis for NN-PARAFAC (Three Groups)

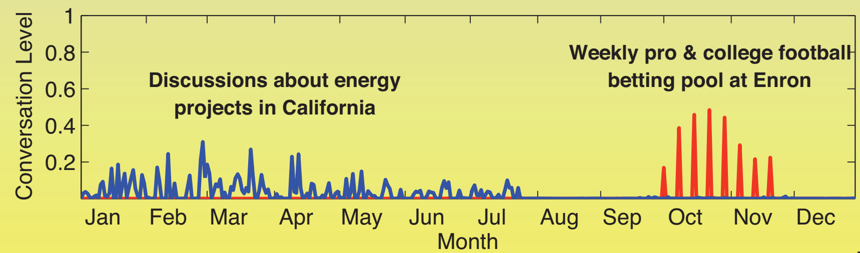
- Rank-25 tensor (best minimizer) for 357 out of 365 days of 2001: A $(69,157 \times 25)$, B (197×25), C (357×25)
- Groups 1,7,8 (out of 25 from C):



Day-level Analysis for NN-PARAFAC (Two Groups)

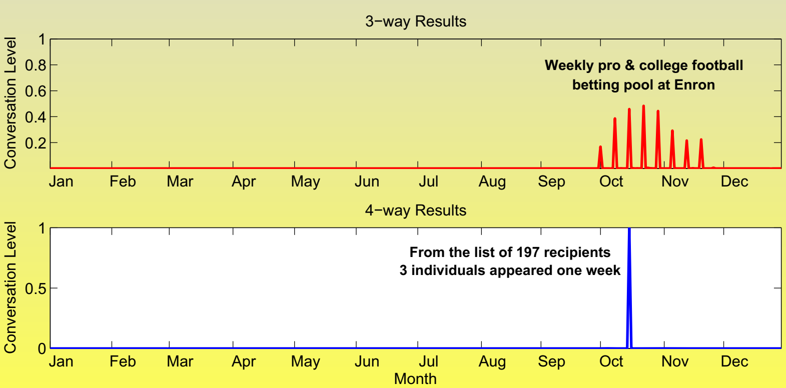
- Groups 20 (California Energy) and 9 (Football) (from C factor of best minimizer) in day-level analysis of 2001:





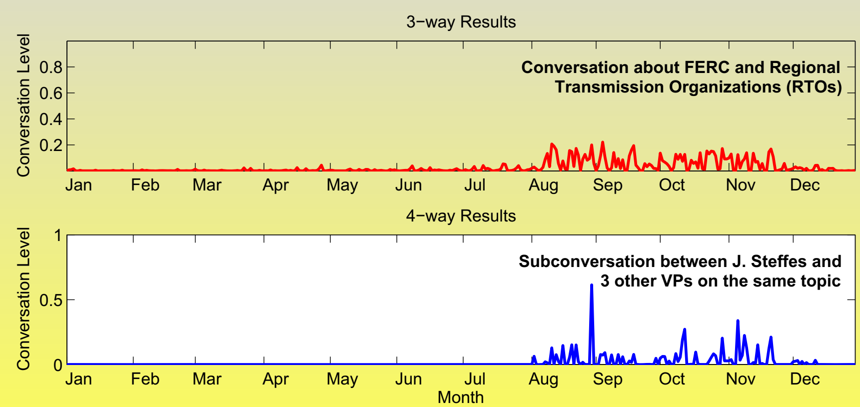
Four-way Tensor Results (Sept. 2007)

- Apply NN-PARAFAC to term-author-recipient-day array $(39,573 \times 197 \times 197 \times 357)$; construct a rank-25 tensor (best minimizer among 10 runs).
- Goal: track more focused discussions between individuals/ small groups; for example, betting pool (football).



Four-way Tensor Results (Sept. 2007)

- Four-way tensor may track subconversation already found by three-way tensor; for example, RTO (Regional Transmission Organization) discussions.



Improving Summarization and Steering

What versus why:

Extraction of textual concepts still requires human interpretation (in the absence of ontologies or domain-specific classifications).

How can previous knowledge or experience be captured for feature matching (or pruning)?

To what extent can feature vectors be annotated for future use or as the text collection is updated? What is the cost for updating NMF/NTF models?

Motivation and Software Design

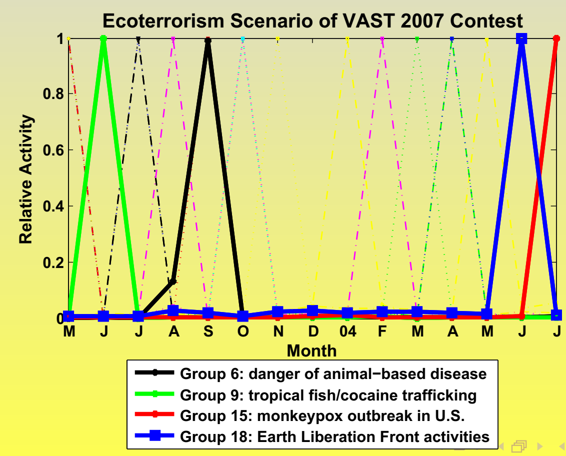
- Inspired by FeatureLens, a Univ. of Maryland HCI Lab Project
- Visualization to facilitate analysis of textual data (and NTF output)
- Feature (event/activity) tracking through time
- Written in Java using SWT.
- Cross platform with native look and feel.
- Works with tagged entities (SGML) and raw text.
- Allows viewing/interpretation of NTF (tensor) outputs.
- Can search/sort terms, create/find co-occurring terms and phrases [Shutt et al., 2009].

Available Datasets for FutureLens Testing

Name	No. of Files	Diskspace	Words/Doc.
Kenya/Factiva	900	3.6 MB	696
Bangladesh/Factiva	1,000	5.1 MB	848
ClimateGate	1,072	8.0 MB	214
VAST-2007	1,455	5.9 MB	391
VHM	3,257	12.7 MB	52
Somalia/Factiva	8,983	37.4 MB	1,005

NTF for Visual Analytics (VA)

- VAST 2007 Contest: $1,455$ news stories/emails/blog entries with underlying ecoterrorism activity to be uncovered.
- Who/What/When/Where questions using tagged entities (Person, Location, Organization, Money) and context (terms). (See http://www.cs.umd.edu/hcil/VASTcontest07)



VAST 2007 Contest Data

- Sample News Article

<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>

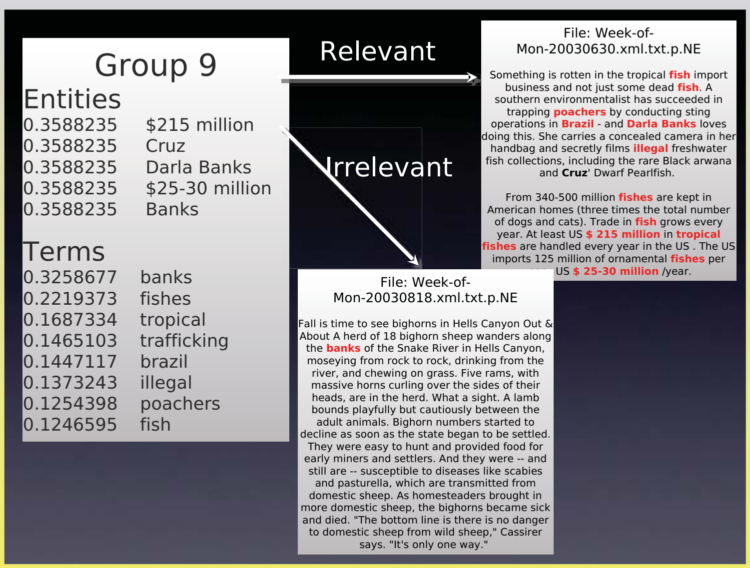
Sample NTF Group Output (No. 15)

Scores	Idx	Name
0.2485621	7120	bruce longhorn
0.2485621	7122	longhorn
0.2485621	7128	chelmsworth
0.2485621	7124	**gil**
0.2485621	7121	**virginia tech**
0.2485621	7125	mary ann ollesen

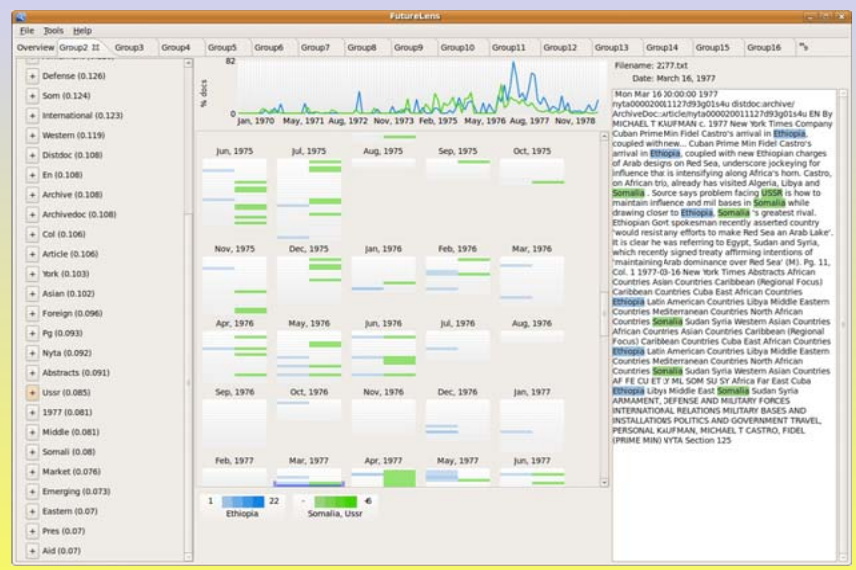
Scores	Idx	Name
0.2958673	6907	**monkeypox**
0.2054770	7468	**outbreak**
0.2008147	6358	longhorn
0.1594331	4644	**gil**
0.1552401	1856	**chinchilla**
0.1434742	11049	travel
0.1391984	9322	sars
0.1379675	1857	chinchillas
0.1342139	2372	continent
0.1294389	3888	expect
0.1215461	9711	sick

NTF Group to Document Matching

- Score documents against an interpretable NTF Group:



Demo using 2001-9 Factiva Articles on Kenya



- 900 docs, 11,652 terms, avg. doc length is 696 terms.

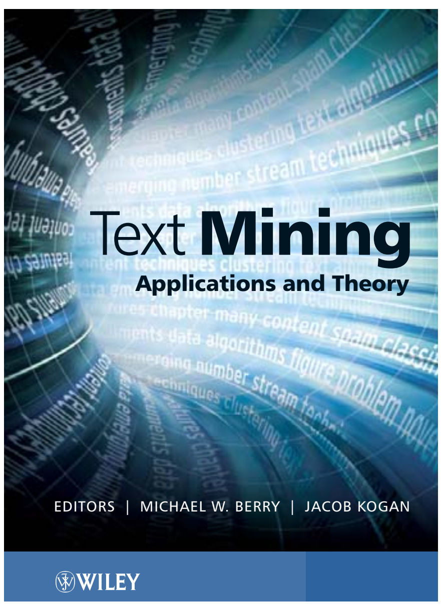
Acknowledgements: National Science Foundation

This research was in part by a Small Business Innovation Research (SBIR) Phase I award entitled *Weather/Climate Variability Impact on Energy, Water and Food Resources and Implications for Regional Stability*, Topic Number OSD09-HS1, Contract Number W913E5-10-C-0012.

For Further Reading (Recent to Past)

- A.A. Puretskiy, G.L. Shutt, and M.W. Berry.
  Survey of Text Visualization Techniques.
  in *Text Mining: Applications and Theory*, M.W. Berry and J. Kogan (Eds.), Wiley, Chichester, UK, 2010:107-127.
- G.L. Shutt, A.A. Puretskiy, and M.W. Berry.
  FutureLens: Software for Text Visualization and Tracking.
  Text Mining Workshop, Proc. Ninth SIAM Int'l Conf. on Data Mining, Sparks, NV, April 30-May 2, 2009.
- B.W. Bader, M.W. Berry, and M. Browne.
  Discussion Tracking in Enron Email Using PARAFAC.
  in *Survey of Text Mining II: Clus., Class., and Retr.*, M.W. Berry and M. Castellanos (Eds.), Springer-Verlag, 2008:147-163.
- M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons.
  Alg. and Applic. for Approx. Nonnegative Matrix Factorization.
  Comput. Stat. & Data Anal. 52(1):155-173, 2007.

John Wiley & Sons Ltd.

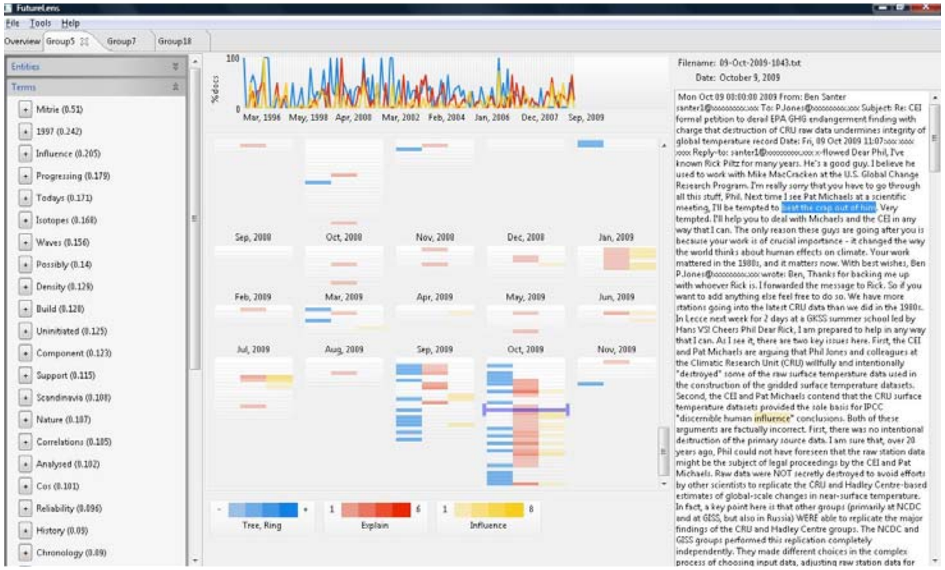


http://www.wiley.com/go/berry_mining

Extra FutureLens Examples

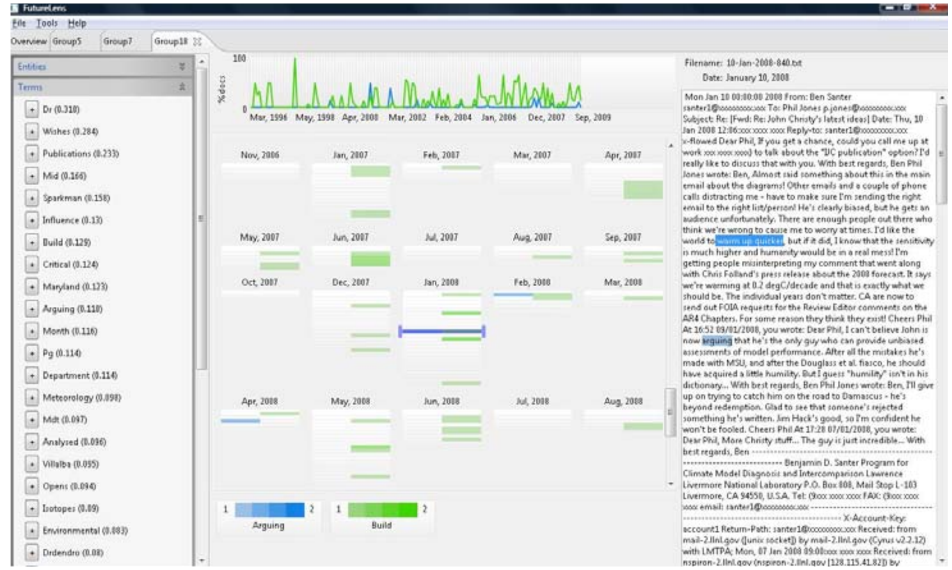
[Use if time for demo is limited]

FutureLens: ClimateGate emails dataset



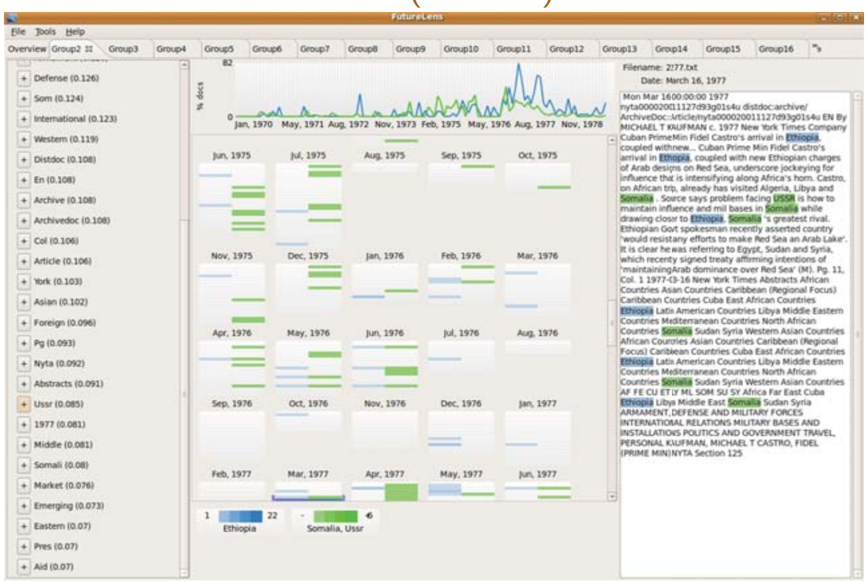
- Detecting *insults* to global warming skeptics.

FutureLens: ClimateGate emails dataset



- Detecting procedures for handling inconsistent data.

FutureLens: 1970s Somalia (Factiva) dataset



- Future event/activity prediction capability.