Text Mining Approaches for Email Surveillance
Document Space Workshop, IPAM/UCLA

Michael W. Berry and Murray Browne
Department of Computer Science, UTK

January 23, 2006
Enron Background

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References
Text Mining Approaches for Email Surveillance

Enron Background

Email Collection

Historical Events

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References
Email Collection

- By-product of the FERC investigation of Enron (originally contained 15 million email messages).
Email Collection

- 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.
Email Collection

- 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. The majority were sent in the 1999 to 2001 timeframe.
Enron Background

Email Collection

Historical Events

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References
Enron Historical 1999-2001

- Ongoing, problematic, development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra.
Enron Historical 1999-2001

- 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 the summer of 2000 (and subsequent investigations).
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 the summer of 2000 (and subsequent investigations).

Revelation of Enron’s deceptive business and accounting practices that led to an abrupt collapse of the energy colossus in October, 2001; Enron filed for bankruptcy in December, 2001.
Enron Background

Non-Negative Matrix Factorization (NMF)
- Motivation
- Underlying Optimization Problem
- MM Method (Lee and Seung)
- Enforcing Statistical Sparsity
- Hybrid NMF Approach

Electronic Mail Surveillance

Conclusions and References
NMF Origins

NMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
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 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*.
- [Xu et al., 2003] demonstrated how NMF-based indexing could outperform SVD-based Latent Semantic Indexing (LSI) for some information retrieval tasks.
NMF for Image Processing

Sparse NMF verses Dense SVD Bases; Lee and Seung (1999)
Non-Negative Matrix Factorization (NMF)

Motivation

NMF for Text Mining (Medlars)

Highest Weighted Terms in Basis Vector $W_1$

1. ventricle
2. aortic
3. septal
4. left
5. defect
6. regurgitation
7. ventricle
8. valve
9. cardiac
10. pressure

weight

Highest Weighted Terms in Basis Vector $W_2$

1. oxygen
2. flow
3. pressure
4. blood
5. cerebral
6. hypothermia
7. fluid
8. venous
9. arterial
10. perfusion

weight

Highest Weighted Terms in Basis Vector $W_5$

1. children
2. child
3. autistic
4. speech
5. group
6. early
7. visual
8. anxiety
9. emotional
10. autism

weight

Highest Weighted Terms in Basis Vector $W_6$

1. kidney
2. marrow
3. dna
4. cells
5. nephrectomy
6. unilateral
7. lymphocyte
8. bone
9. thymidine
10. rats

weight

Interpretable NMF feature vectors; Langville et al. (2006)
Given an $m \times n$ term-by-message (sparse) matrix $X$. 
Given an $m \times n$ term-by-message (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 $r \ll n$. 
Derivation

- Given an $m \times n$ term-by-message (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 $H_{ij} \geq 0$, $\forall i,j$. 
Given an \( m \times n \) term-by-message (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.
Non-Negative Matrix Factorization (NMF)

MM Method (Lee and Seung)

Multiplicative Method (MM)

Multiplicative update rules for $W$ and $H$ (Lee and Seung, 1999):

1. Initialize $W$ and $H$ with non-negative values, and scale the columns of $W$ to unit norm.
2. Iterate for each $c$, $j$, and $i$ until convergence or after $k$ iterations:
   2.1 $H_{cj} \leftarrow H_{cj} \frac{(W^TX)_{cj}}{(W^TWH)_{cj} + \epsilon}$
   2.2 $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
   2.3 Scale the columns of $W$ to unit norm.

Setting $\epsilon = 10^{-9}$ will suffice [Shahnaz et al., 2006].
Multiplicative Method (MM)

- Multiplicative update rules for $W$ and $H$ (Lee and Seung, 1999):
  1. Initialize $W$ and $H$ with non-negative values, and scale the columns of $W$ to unit norm.
  2. Iterate for each $c$, $j$, and $i$ until convergence or after $k$ iterations:
    2.1 $H_{cj} \leftarrow H_{cj} \frac{(W^T X)_{cj}}{(W^T WH)_{cj} + \epsilon}$
    2.2 $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
    2.3 Scale the columns of $W$ to unit norm.

- Setting $\epsilon = 10^{-9}$ will suffice [Shahnaz et al., 2006].
Normalization, Complexity, and Convergence

- Important to normalize $X$ initially and the basis matrix $W$ at each iteration.
Normalization, Complexity, and Convergence

- Important to normalize $X$ initially and the basis matrix $W$ at each iteration.
- When optimizing on a unit hypersphere, the column (or feature) vectors of $W$, denoted by $W_k$, are effectively mapped to the surface of the hypersphere by repeated normalization.
Normalization, Complexity, and Convergence

- Important to normalize $X$ initially and the basis matrix $W$ at each iteration.
- When optimizing on a unit hypersphere, the column (or feature) vectors of $W$, denoted by $W_k$, are effectively mapped to the surface of the hypersphere by repeated normalization.
- MM implementation of NMF requires $O(rmn)$ operations per iteration; Lee and Seung (1999) proved that $\|X - WH\|_F^2$ is monotonically non-increasing with MM.
Important to normalize $X$ initially and the basis matrix $W$ at each iteration.

When optimizing on a unit hypersphere, the column (or feature) vectors of $W$, denoted by $W_k$, are effectively mapped to the surface of the hypersphere by repeated normalization.

MM implementation of NMF requires $O(rmn)$ operations per iteration; Lee and Seung (1999) proved that $\|X - WH\|_F^2$ is monotonically non-increasing with MM.

From a nonlinear optimization perspective, MM/NMF can be considered a diagonally-scaled gradient descent method.
From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix \( H \) in order to enhance the parts-based data representations in the matrix \( W \).
From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix $H$ in order to enhance the parts-based data representations in the matrix $W$. 

Mu et al. (2003) suggested a regularization approach to achieve statistical sparsity in the matrix $H$: **point count regularization**; penalize the *number* of nonzeros in $H$ rather than $\sum_{ij} H_{ij}$. 

Hoyer’s Method
From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix $H$ in order to enhance the parts-based data representations in the matrix $W$.

Mu et al. (2003) suggested a regularization approach to achieve statistical sparsity in the matrix $H$: **point count regularization**; penalize the *number* of nonzeros in $H$ rather than $\sum_{ij} H_{ij}$.

Goal of increased sparsity – better representation of *parts* or *features* spanned by the corpus ($X$) [Shahnaz et al., 2006].
First use MM to compute an approximation to $W$ for each iteration – a gradient descent (GD) optimization step.
First use MM to compute an approximation to $W$ for each iteration – a gradient descent (GD) optimization step.

Then, compute the weight matrix $H$ using a constrained least squares (CLS) model to penalize non-smoothness (i.e., non-sparsity) in $H$ – common Tikohonov regularization technique used in image processing (Prasad et al., 2003).
First use MM to compute an approximation to $W$ for each iteration – a gradient descent (GD) optimization step.

Then, compute the weight matrix $H$ using a constrained least squares (CLS) model to penalize non-smoothness (i.e., non-sparsity) in $H$ – common Tikohonov regularization technique used in image processing (Prasad et al., 2003).

Convergence to a non-stationary point evidenced (but no formal proof given to date).
GD-CLS Algorithm

1. Initialize $W$ and $H$ with non-negative values, and scale the columns of $W$ to unit norm.
GD-CLS Algorithm

1. Initialize $W$ and $H$ with non-negative values, and scale the columns of $W$ to unit norm.

2. Iterate until convergence or after $k$ iterations:
   2.1 $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for $c$ and $i$
   2.2 Rescale the columns of $W$ to unit norm.
   2.3 Solve the constrained least squares problem:

   $$\min_{H_j} \left\{ \|X_j - WH_j\|_2^2 + \lambda \|H_j\|_2^2 \right\},$$

   where the subscript $j$ denotes the $j^{th}$ column, for $j = 1, \ldots, m$. 
1. Initialize $W$ and $H$ with non-negative values, and scale the columns of $W$ to unit norm.

2. Iterate until convergence or after $k$ iterations:
   
   2.1 $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for $c$ and $i$
   
   2.2 Rescale the columns of $W$ to unit norm.
   
   2.3 Solve the constrained least squares problem:

   $$\min_{H_j} \{ \|X_j - WH_j\|_2^2 + \lambda \|H_j\|_2^2 \},$$

   where the subscript $j$ denotes the $j^{th}$ column, for $j = 1, \ldots, m$.

   ▶ Any negative values in $H_j$ are set to zero. The parameter $\lambda$ is a regularization value that is used to balance the reduction of the metric $\|X_j - WH_j\|_2^2$ with enforcement of smoothness and sparsity in $H$ [Shahnaz et al., 2006].
Enron Background

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance
  Message Parsing
  Term Weighting
  GD-CLS Benchmarks
  Clustering and Topic Extraction
  Topic Tracking (Through Time)
  Smoothing Effects Comparison

Conclusions and References
Parsed *inbox* folder of all 150 accounts (users) via GTP (General Text Parser); 495-term stoplist used and extracted terms must appear in more than 1 email and more than once globally.
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.
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.

- Distribution of messages sent in the year 2001:

<table>
<thead>
<tr>
<th>Month</th>
<th>Msgs</th>
<th>Terms</th>
<th>Month</th>
<th>Msgs</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>3,621</td>
<td>17,888</td>
<td>Jul</td>
<td>3,077</td>
<td>17,617</td>
</tr>
<tr>
<td>Feb</td>
<td>2,804</td>
<td>16,958</td>
<td>Aug</td>
<td>2,828</td>
<td>16,417</td>
</tr>
<tr>
<td>Mar</td>
<td>3,525</td>
<td>20,305</td>
<td>Sep</td>
<td>2,330</td>
<td>15,405</td>
</tr>
<tr>
<td>Apr</td>
<td>4,273</td>
<td>24,010</td>
<td>Oct</td>
<td>2,821</td>
<td>20,995</td>
</tr>
<tr>
<td>May</td>
<td>4,261</td>
<td>24,335</td>
<td>Nov</td>
<td>2,204</td>
<td>18,693</td>
</tr>
<tr>
<td>Jun</td>
<td>4,324</td>
<td>18,599</td>
<td>Dec</td>
<td>1,489</td>
<td>8,097</td>
</tr>
</tbody>
</table>
For $m \times n$ term-by-message matrix $X = [x_{ij}]$, define

$$x_{ij} = l_{ij} g_{i} d_{j},$$

where $l_{ij}$ is the local weight for term $i$ occurring in message $j$, $g_{i}$ is the global weight for term $i$ in the subcollection, and $d_{j}$ is a document normalization factor (set $d_{j} = 1$).
Term Weighting Schemes

- For $m \times n$ term-by-message matrix $X = [x_{ij}]$, define
  
  $$x_{ij} = l_{ij}g_id_j,$$

  where $l_{ij}$ is the local weight for term $i$ occurring in message $j$, $g_i$ is the global weight for term $i$ in the subcollection, and $d_j$ is a document normalization factor (set $d_j = 1$).

- Schemes used in parsing INBOX and PRIVATE subcollections:

<table>
<thead>
<tr>
<th>Name</th>
<th>Local</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>tx</td>
<td>Term Frequency</td>
<td>None</td>
</tr>
<tr>
<td>xx</td>
<td>$l_{ij} = f_{ij}$</td>
<td>$g_i = 1$</td>
</tr>
<tr>
<td>lex</td>
<td>Logarithmic</td>
<td>Entropy</td>
</tr>
<tr>
<td></td>
<td>$l_{ij} = \log(1 + f_{ij})$</td>
<td>$(\sum_j p_{ij} \log(p_{ij})) / \log n)$</td>
</tr>
</tbody>
</table>
Computational Complexity

- Rank-50 NMF ($X \simeq WH$) computed on a 450MHz (Dual) UltraSPARC-II processor using 100 iterations:

<table>
<thead>
<tr>
<th>Collection</th>
<th>Mail Messages</th>
<th>Dictionary Terms</th>
<th>$\lambda$</th>
<th>Time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INBOX</td>
<td>44,872</td>
<td>80,683</td>
<td>0.1</td>
<td>1,471</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>1,451</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td>1,521</td>
</tr>
<tr>
<td>PRIVATE</td>
<td>65,031</td>
<td>92,133</td>
<td>0.1</td>
<td>51,489</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>51,393</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td>51,562</td>
</tr>
</tbody>
</table>
PRIVATE with Log-Entropy Weighting

- Identify rows of $H$ from $X \simeq WH$ or $H^k$ with $\lambda = 0.1$; $r = 50$
- feature vectors ($W_k$) generated by GD-CLS:

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

(Cluster size $\equiv$ no. of $H^k$ elements $> row_{max}/10$)
## Additional topic clusters of significant size:

<table>
<thead>
<tr>
<th>Feature Index $(k)$</th>
<th>Cluster Size</th>
<th>Topic Description</th>
<th>Dominant Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>233</td>
<td>Texas longhorn football newsletter</td>
<td>UT, orange, longhorn[s], texas, true, truorange, recruiting, oklahoma, defensive</td>
</tr>
<tr>
<td>34</td>
<td>65</td>
<td>Enron collapse</td>
<td>partnership[s], fastow, shares, sec, stock, shareholder, investors, equity, lay</td>
</tr>
<tr>
<td>39</td>
<td>235</td>
<td>Emails about India</td>
<td>dabhol, dpc, india, mseb, maharashtra, indian, lenders, delhi, foreign, minister</td>
</tr>
</tbody>
</table>
2001 Topics Tracked by GD-CLS

<table>
<thead>
<tr>
<th>JAN</th>
<th>MAR</th>
<th>APR</th>
<th>MAY</th>
<th>JUN</th>
<th>JUL</th>
<th>AUG</th>
<th>OCT</th>
<th>NOV</th>
<th>DEC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$r = 50$ features, $\text{lex}$ term weighting, $\lambda = 0.1$
Introduce smoothing on $W_k$ (feature vectors) in addition to $H^k$:

$$\min_{W,H}\left\{ \| X - WH \|^2_F + \alpha \| W \|^2_F + \beta \| H \|^2_F \right\},$$

where $\| \cdot \|_F$ is the Frobenius norm.
Two Penalty Term Formulation

- Introduce smoothing on $W_k$ (feature vectors) in addition to $H^k$:

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

where $\| \cdot \|_F$ is the Frobenius norm.

- Constrained NMF (CNMF) iteration [Piper et al., 2004]:

$$
H_{cj} \leftarrow H_{cj} \frac{(W^T X)_{cj} - \beta H_{cj}}{(W^T WH)_{cj} + \epsilon}
$$

$$
W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic} - \alpha W_{ic}}{(WHH^T)_{ic} + \epsilon}
$$
## Term Distribution in Feature Vectors

<table>
<thead>
<tr>
<th>Terms</th>
<th>Wt</th>
<th>Lambda 0.1</th>
<th>Lambda 0.01</th>
<th>Lambda 0.001</th>
<th>Alpha 0.1</th>
<th>Alpha 0.01</th>
<th>Alpha 0.001</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blackouts</td>
<td>0.508</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cal</td>
</tr>
<tr>
<td>Stocks</td>
<td>0.511</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collapse</td>
</tr>
<tr>
<td>UT</td>
<td>0.517</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Texasfoot</td>
</tr>
<tr>
<td>Chronicle</td>
<td>0.523</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>India</td>
</tr>
<tr>
<td>Indian</td>
<td>0.527</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>India</td>
</tr>
<tr>
<td>Fastow</td>
<td>0.531</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collapse</td>
</tr>
<tr>
<td>Gas</td>
<td>0.531</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collapse</td>
</tr>
<tr>
<td>CFO</td>
<td>0.556</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kitchen</td>
</tr>
<tr>
<td>Californians</td>
<td>0.557</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cal</td>
</tr>
<tr>
<td>Solar</td>
<td>0.570</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partnerships</td>
<td>0.576</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collapse</td>
</tr>
<tr>
<td>Workers</td>
<td>0.577</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maharashtra</td>
<td>0.591</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>India</td>
</tr>
<tr>
<td>Mseb</td>
<td>0.605</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>India</td>
</tr>
<tr>
<td>Beach</td>
<td>0.611</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ljm</td>
<td>0.621</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collapse</td>
</tr>
<tr>
<td>Tues</td>
<td>0.626</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPPS</td>
<td>0.644</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cal</td>
</tr>
<tr>
<td>Rebates</td>
<td>0.647</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ljm2</td>
<td>0.688</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collapse</td>
</tr>
</tbody>
</table>

*Note: The table includes terms along with their weights and distribution across different topics.*
British National Corpus (BNC) Noun Conservation

- In collaboration with P. Keila and D. Skillicorn (Queens Univ.)
- 289,695 email subset (all mail folders - not just private)
- Smoothing solely applied to NMF $W$ matrix
  \( \alpha = 0.001, 0.01, 0.1, 0.25, 0.50, 0.75, 1.00 \) with \( \beta = 0 \)
- Log-entropy term weighting applied to the term-by-message matrix $X$
- Monitor top ten nouns for each feature vector (ranked by descending component values) and extract those appearing in two or more features; topics assigned manually.
## BNC Noun Distribution in Feature Vectors

<table>
<thead>
<tr>
<th>Noun</th>
<th>GF</th>
<th>Entropy</th>
<th>0.001</th>
<th>0.01</th>
<th>0.1</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waxman</td>
<td>680</td>
<td>0.424</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Downfall</td>
</tr>
<tr>
<td>Lieberman</td>
<td>915</td>
<td>0.426</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Downfall</td>
</tr>
<tr>
<td>Scandal</td>
<td>679</td>
<td>0.428</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Downfall</td>
</tr>
<tr>
<td>Nominee(s)</td>
<td>544</td>
<td>0.436</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Downfall</td>
</tr>
<tr>
<td>Barone</td>
<td>470</td>
<td>0.437</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Downfall</td>
</tr>
<tr>
<td>MEADE</td>
<td>456</td>
<td>0.437</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Downfall</td>
</tr>
<tr>
<td>Fichera</td>
<td>558</td>
<td>0.438</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>California blackout</td>
</tr>
<tr>
<td>Prabhu</td>
<td>824</td>
<td>0.445</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India-strong</td>
</tr>
<tr>
<td>Tata</td>
<td>778</td>
<td>0.448</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India-weak</td>
</tr>
<tr>
<td>Rupee(s)</td>
<td>323</td>
<td>0.452</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>India-strong</td>
</tr>
<tr>
<td>Soybean(s)</td>
<td>499</td>
<td>0.455</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India-weak</td>
</tr>
<tr>
<td>Rushing</td>
<td>891</td>
<td>0.486</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Football - college</td>
</tr>
<tr>
<td>Dlrs</td>
<td>596</td>
<td>0.487</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Janus</td>
<td>580</td>
<td>0.488</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>India-weak</td>
</tr>
<tr>
<td>BSES</td>
<td>451</td>
<td>0.498</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India-weak</td>
</tr>
<tr>
<td>Caracas</td>
<td>698</td>
<td>0.498</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>California/Blackout</td>
</tr>
<tr>
<td>Escondido</td>
<td>326</td>
<td>0.504</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Promoters</td>
<td>180</td>
<td>0.509</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>Energy/Scottish</td>
</tr>
<tr>
<td>Aramco</td>
<td>188</td>
<td>0.550</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>India-weak</td>
</tr>
<tr>
<td>DOORMAN</td>
<td>231</td>
<td>0.598</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>Bawdy/Real Estate</td>
</tr>
</tbody>
</table>
Enron Background

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References
   Conclusions
   Future Work
   References
Conclusions

- GD-CLS Algorithm can effectively produce a parts-based approximation $X \simeq WH$ of a sparse term-by-message matrix $X$. 
Conclusions

- GD-CLS Algorithm can effectively produce a *parts-based* approximation $X \simeq WH$ of a sparse term-by-message matrix $X$.

- Smoothing on the features matrix ($W$) as opposed to the weight matrix $H$ forces more reuse of higher weighted terms.
Conclusions

- GD-CLS Algorithm can effectively produce a \textit{parts-based} approximation \( X \simeq WH \) of a sparse term-by-message matrix \( X \).
- Smoothing on the features matrix \( W \) as opposed to the weight matrix \( H \) forces more reuse of higher weighted terms.
- Surveillance systems based on GD-CLS could be used to monitor discussions without the need to isolate or perhaps incriminate individual employees.
Conclusions

- GD-CLS Algorithm can effectively produce a parts-based approximation $X \simeq WH$ of a sparse term-by-message matrix $X$.
- Smoothing on the features matrix ($W$) as opposed to the weight matrix $H$ forces more reuse of higher weighted terms.
- Surveillance systems based on GD-CLS could be used to monitor discussions without the need to isolate or perhaps incriminate individual employees.
- Potential applications include the monitoring/tracking of company morale, employee feedback to policy decisions, and extracurricular activities.
Future Work

- Further work needed in determining effects of alternative term weighting schemes (for $X$) and choices of control parameters ($\alpha, \beta, \lambda$) on quality of the basis vectors $W_k$. 
Future Work

- Further work needed in determining effects of alternative term weighting schemes (for $X$) and choices of control parameters ($\alpha, \beta, \lambda$) on quality of the basis vectors $W_k$.

- How does document (or message) clustering change with different column ranks ($r$) in the matrix $W$?
Future Work

- Further work needed in determining effects of alternative term weighting schemes (for $X$) and choices of control parameters ($\alpha, \beta, \lambda$) on quality of the basis vectors $W_k$.

- How does document (or message) clustering change with different column ranks ($r$) in the matrix $W$?

- Use MailMiner and similar text mining software to produce a topic annotated Enron email subset for the public domain.
Future Work

▶ Further work needed in determining effects of alternative term weighting schemes (for $X$) and choices of control parameters ($\alpha, \beta, \lambda$) on quality of the basis vectors $W_k$.

▶ How does document (or message) clustering change with different column ranks ($r$) in the matrix $W$?

▶ Use MailMiner and similar text mining software to produce a topic annotated Enron email subset for the public domain.

▶ Explore use of NMF for automated gene classification; **Semantic Gene Organizer** (K. Heinrich, PhD Thesis 2006)
For Further Reading


SMD06 Text Mining Workshop

Hyatt Regency Bethesda
Bethesda, Maryland
April 22, 2006

to be held in conjunction with

Sixth SIAM International Conference on Data Mining (SDM 2006)

and also in conjunction with SIAM's Link Analysis, Counterterrorism Security Workshop, which is also being held on April 22, 2006.

Website: http://www.cs.utk.edu/tmw06