Deep Email Miner Application

Johannes Mager

University of Technology, Sydney

June 5, 2006

Outline

Outline

Introduction

- Email traffic is an important factor in organisations and companies
- Currently there is no specialised software for its analysis
- Idea: Develop a new software program for detailed analysis of an email corpus
- Questions: What is possible? What techniques are suitable?

Introduction

- Email traffic is an important factor in organisations and companies
- Currently there is no specialised software for its analysis
- Idea: Develop a new software program for detailed analysis of an email corpus
- Questions: What is possible? What techniques are suitable?

Introduction

- Email traffic is an important factor in organisations and companies
- Currently there is no specialised software for its analysis
- Idea: Develop a new software program for detailed analysis of an email corpus
- Questions: What is possible? What techniques are suitable?

Introduction

- Email traffic is an important factor in organisations and companies
- Currently there is no specialised software for its analysis
- Idea: Develop a new software program for detailed analysis of an email corpus
- Questions: What is possible? What techniques are suitable?

Outline

Enron Corporation

- Energy Company based in Texas, with around 21,000 employees and claimed revenues of US$ 101 billion in 2000.
- Fortune Business Magazine named Enron “America’s Most Innovative Company" for six consecutive years.
- In 2001, it was revealed that Enron was sustained mostly by institutionalized and systematic accounting fraud.
- May 25, 2006: Former CEOs Kenneth Lay and Jeff Skilling were convicted for conspiracy, fraud, false statements and insider trading.

Enron Corporation

- Energy Company based in Texas, with around 21,000 employees and claimed revenues of US$ 101 billion in 2000.
- Fortune Business Magazine named Enron “America’s Most Innovative Company" for six consecutive years.
- In 2001, it was revealed that Enron was sustained mostly by institutionalized and systematic accounting fraud.
- May 25, 2006: Former CEOs Kenneth Lay and Jeff Skilling were convicted for conspiracy, fraud, false statements and insider trading.

Enron Corporation

- Energy Company based in Texas, with around 21,000 employees and claimed revenues of US$ 101 billion in 2000.
- Fortune Business Magazine named Enron “America’s Most Innovative Company" for six consecutive years.
- In 2001, it was revealed that Enron was sustained mostly by institutionalized and systematic accounting fraud.
- May 25, 2006: Former CEOs Kenneth Lay and Jeff Skilling were convicted for conspiracy, fraud, false statements and insider trading.

Enron Corporation

- Energy Company based in Texas, with around 21,000 employees and claimed revenues of US$ 101 billion in 2000.
- Fortune Business Magazine named Enron “America’s Most Innovative Company" for six consecutive years.
- In 2001, it was revealed that Enron was sustained mostly by institutionalized and systematic accounting fraud.
- May 25, 2006: Former CEOs Kenneth Lay and Jeff Skilling were convicted for conspiracy, fraud, false statements and insider trading.

Enron's Emails

- March 2003: US Federal Energy Regulatory Commission made public more than 1.5 million e-mails from 176 Enron employees.
- "The public has a right to know the facts upon which the Enron investigation was based."
- Result: A unique and open real-life email corpus, suitable for all kinds of analysis.

Enron's Emails

- March 2003: US Federal Energy Regulatory Commission made public more than 1.5 million e-mails from 176 Enron employees.
- "The public has a right to know the facts upon which the Enron investigation was based."
- Result: A unique and open real-life email corpus, suitable for all kinds of analysis.

Enron's Emails

- March 2003: US Federal Energy Regulatory Commission made public more than 1.5 million e-mails from 176 Enron employees.
- "The public has a right to know the facts upon which the Enron investigation was based."
- Result: A unique and open real-life email corpus, suitable for all kinds of analysis.

Email Preprocessing

Preprocessing of the Enron corpus has been done by several research groups:

- Cleaned data and resolved integrity problems
- Deleted system emails and duplicates
- Included employee data
- Created database representation

Used data set (University of Southern California):

- 252,759 messages from 151 employees
- 2,064,442 recipients
- 500MB of text

Email Preprocessing

Preprocessing of the Enron corpus has been done by several research groups:

- Cleaned data and resolved integrity problems
- Deleted system emails and duplicates
- Included employee data
- Created database representation

Used data set (University of Southern California):

- 252,759 messages from 151 employees
- 2,064,442 recipients
- 500MB of text

Email Preprocessing

Preprocessing of the Enron corpus has been done by several research groups:

- Cleaned data and resolved integrity problems
- Deleted system emails and duplicates
- Included employee data
- Created database representation

Used data set (University of Southern California):

- 252,759 messages from 151 employees
- 2,064,442 recipients
- 500MB of text

Email Preprocessing

Preprocessing of the Enron corpus has been done by several research groups:

- Cleaned data and resolved integrity problems
- Deleted system emails and duplicates
- Included employee data
- Created database representation

Used data set (University of Southern California):

- 252,759 messages from 151 employees
- 2,064,442 recipients
- 500MB of text

Email Preprocessing

Preprocessing of the Enron corpus has been done by several research groups:

- Cleaned data and resolved integrity problems
- Deleted system emails and duplicates
- Included employee data
- Created database representation

Used data set (University of Southern California):

- 252,759 messages from 151 employees
- 2,064,442 recipients
- 500MB of text

Database Representation

```
TABLE employeelist (
      firstName varchar(31),
      lastName varchar(31),
      Email_id varchar(31),
      UNIQUE KEY (Email_id));

TABLE message (
      mid int(10),
      sender varchar(127),
      date datetime,
      subject text,
      body text,
      PRIMARY KEY (mid));

CREATE TABLE recipientinfo (
      rid int(10),
      mid int(10),
      rtype enum('TO','CC','BCC'),
      rvalue varchar(127),
      PRIMARY KEY (rid));
```

Database Representation

```
TABLE employeelist (
       firstName varchar(31),
       lastName varchar(31),
       Email_id varchar(31),
       UNIQUE KEY (Email_id));

TABLE message (
       mid int(10),
       sender varchar(127),
       date datetime,
       subject text,
       body text,
       PRIMARY KEY (mid));

CREATE TABLE recipientinfo (
       rid int(10),
       mid int(10),
       rtype enum('TO','CC','BCC'),
       rvalue varchar(127),
       PRIMARY KEY (rid));
```

Database Representation

```
TABLE employeelist (
       firstName varchar(31),
       lastName varchar(31),
       Email_id varchar(31),
       UNIQUE KEY (Email_id));

TABLE message (
       mid int(10),
       sender varchar(127),
       date datetime,
       subject text,
       body text,
       PRIMARY KEY (mid));

CREATE TABLE recipientinfo (
       rid int(10),
       mid int(10),
       rtype enum('TO','CC','BCC'),
       rvalue varchar(127),
       PRIMARY KEY (rid));
```

Outline

Program Features

- Social Network Analysis
- Email Analysis
- Text Mining
- Input/Output
- Software

Social Network Analysis

- Create social network with employees as nodes
- Create edge, if number of emails sent between two nodes is above threshold
- Find clusters of employees
- Rank nodes/edges on centrality measures
- Graph statistics

Social Network Analysis

- Create social network with employees as nodes
- Create edge, if number of emails sent between two nodes is above threshold
- Find clusters of employees
- Rank nodes/edges on centrality measures
- Graph statistics

Social Network Analysis

- Create social network with employees as nodes
- Create edge, if number of emails sent between two nodes is above threshold
- Find clusters of employees
- Rank nodes/edges on centrality measures
- Graph statistics

Social Network Analysis

- Create social network with employees as nodes
- Create edge, if number of emails sent between two nodes is above threshold
- Find clusters of employees
- Rank nodes/edges on centrality measures
- Graph statistics

Social Network Analysis

- Create social network with employees as nodes
- Create edge, if number of emails sent between two nodes is above threshold
- Find clusters of employees
- Rank nodes/edges on centrality measures
- Graph statistics

Email Analysis

- Find email threads
- Display emails
- Filter network based on email attributes and email number in the edges
- Email corpus statistics

Email Analysis

- Find email threads
- Display emails
- Filter network based on email attributes and email number in the edges
- Email corpus statistics

Email Analysis

- Find email threads
- Display emails
- Filter network based on email attributes and email number in the edges
- Email corpus statistics

Email Analysis

- Find email threads
- Display emails
- Filter network based on email attributes and email number in the edges
- Email corpus statistics

Text Mining

- Label emails as business/private
- Create word list
- Create word vectors

Text Mining

- Label emails as business/private
- Create word list
- Create word vectors

Text Mining

- Label emails as business/private
- Create word list
- Create word vectors

Input/Output

- Load network and single emails from database
- Cache email bodies to balance the amount of database-interaction and the program's memory usage
- Load/Save email labels to file
- Export word vectors to .csv
- Export network to Pajek .net

Input/Output

- Load network and single emails from database
- Cache email bodies to balance the amount of database-interaction and the program's memory usage
- Load/Save email labels to file
- Export word vectors to .csv
- Export network to Pajek .net

Input/Output

- Load network and single emails from database
- Cache email bodies to balance the amount of database-interaction and the program's memory usage
- Load/Save email labels to file
- Export word vectors to .csv
- Export network to Pajek .net

Input/Output

- Load network and single emails from database
- Cache email bodies to balance the amount of database-interaction and the program's memory usage
- Load/Save email labels to file
- Export word vectors to .csv
- Export network to Pajek .net

Input/Output

- Load network and single emails from database
- Cache email bodies to balance the amount of database-interaction and the program's memory usage
- Load/Save email labels to file
- Export word vectors to .csv
- Export network to Pajek .net

Software

- Written in Java, runnable everywhere
- Additional version as Win32 executable
- Configurable with properties-file

Software

- Written in Java, runnable everywhere
- Additional version as Win32 executable
- Configurable with properties-file

Software

- Written in Java, runnable everywhere
- Additional version as Win32 executable
- Configurable with properties-file

Used Software Libraries

- Java Universal Network/Graph Framework JUNG (University of California, Irvine)
- Statistical language modelling: WVTool 0.9.1 (University of Dortmund)
- JDBC driver for database connection: MySQL Connector/J 3.1
- launch4j - Cross-platform Java executable wrapper

Outline

Program Demo

Outline

Current Status

- Current Software Version: 1.0
- Available Documentation: Javadoc, Software Manual and Project Report
- Binaries, Documents and Source Code are available on sourceforge.net
- Project has been released under the GNU GPL
- Development is paused, but can be continued any time

Current Status

- Current Software Version: 1.0
- Available Documentation: Javadoc, Software Manual and Project Report
- Binaries, Documents and Source Code are available on sourceforge.net
- Project has been released under the GNU GPL
- Development is paused, but can be continued any time

Current Status

- Current Software Version: 1.0
- Available Documentation: Javadoc, Software Manual and Project Report
- Binaries, Documents and Source Code are available on sourceforge.net
- Project has been released under the GNU GPL
- Development is paused, but can be continued any time

Current Status

- Current Software Version: 1.0
- Available Documentation: Javadoc, Software Manual and Project Report
- Binaries, Documents and Source Code are available on sourceforge.net
- Project has been released under the GNU GPL
- Development is paused, but can be continued any time

Current Status

- Current Software Version: 1.0
- Available Documentation: Javadoc, Software Manual and Project Report
- Binaries, Documents and Source Code are available on sourceforge.net
- Project has been released under the GNU GPL
- Development is paused, but can be continued any time

(Possible) Future Developments

- Generalize labelling system
- Use machine learning to train classification
- Integrate text mining in clustering, filtering...
- Draw email tree based on a specific node
- Visualise network development over time
- ...WHAT ELSE?

(Possible) Future Developments

- Generalize labelling system
- Use machine learning to train classification
- Integrate text mining in clustering, filtering...
- Draw email tree based on a specific node
- Visualise network development over time
- ...WHAT ELSE?

(Possible) Future Developments

- Generalize labelling system
- Use machine learning to train classification
- Integrate text mining in clustering, filtering...
- Draw email tree based on a specific node
- Visualise network development over time
- ...WHAT ELSE?

(Possible) Future Developments

- Generalize labelling system
- Use machine learning to train classification
- Integrate text mining in clustering, filtering...
- Draw email tree based on a specific node
- Visualise network development over time
- ...WHAT ELSE?

(Possible) Future Developments

- Generalize labelling system
- Use machine learning to train classification
- Integrate text mining in clustering, filtering...
- Draw email tree based on a specific node
- Visualise network development over time
- ...WHAT ELSE?

(Possible) Future Developments

- Generalize labelling system
- Use machine learning to train classification
- Integrate text mining in clustering, filtering...
- Draw email tree based on a specific node
- Visualise network development over time
- ...WHAT ELSE?

Thanks to. . .

- Simeon for all the help, tips and advices
- Debbie for the help in the text mining part
- Paul & Les for the support with the database server

Thanks to. . .

- Simeon for all the help, tips and advices
- Debbie for the help in the text mining part
- Paul & Les for the support with the database server

Thanks to...

- Simeon for all the help, tips and advices
- Debbie for the help in the text mining part
- Paul & Les for the support with the database server

More Information

- Homepage: deepemailminer.sf.net
- Email: post@johannes-mager.de

- Questions?