Data Mining
資料探勘

文字探勘與網頁探勘
(Text and Web Mining)

1032DM10
MI4
Wed, 7,8 (14:10-16:00) (B130)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2015-05-20
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2015/02/25</td>
<td>資料探勘導論 (Introduction to Data Mining)</td>
</tr>
<tr>
<td>2</td>
<td>2015/03/04</td>
<td>關連分析 (Association Analysis)</td>
</tr>
<tr>
<td>3</td>
<td>2015/03/11</td>
<td>分類與預測 (Classification and Prediction)</td>
</tr>
<tr>
<td>4</td>
<td>2015/03/18</td>
<td>分群分析 (Cluster Analysis)</td>
</tr>
<tr>
<td>5</td>
<td>2015/03/25</td>
<td>個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis – K-Means using SAS EM)</td>
</tr>
<tr>
<td>6</td>
<td>2015/04/01</td>
<td>教學行政觀摩日 (Off-campus study)</td>
</tr>
<tr>
<td>7</td>
<td>2015/04/08</td>
<td>個案分析與實作二 (SAS EM 關連分析)：Case Study 2 (Association Analysis using SAS EM)</td>
</tr>
<tr>
<td>8</td>
<td>2015/04/15</td>
<td>個案分析與實作三 (SAS EM 決策樹、模型評估)：Case Study 3 (Decision Tree, Model Evaluation using SAS EM)</td>
</tr>
<tr>
<td>週次 (Week)</td>
<td>日期 (Date)</td>
<td>內容 (Subject/Topics)</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>9</td>
<td>2015/04/22</td>
<td>期中報告 (Midterm Project Presentation)</td>
</tr>
<tr>
<td>10</td>
<td>2015/04/29</td>
<td>期中考試週 (Midterm Exam)</td>
</tr>
<tr>
<td>11</td>
<td>2015/05/06</td>
<td>個案分析與實作四 (SAS EM 迴歸分析、類神經網路)：Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)</td>
</tr>
<tr>
<td>12</td>
<td>2015/05/13</td>
<td>海量資料分析 (Big Data Analytics)</td>
</tr>
<tr>
<td>13</td>
<td>2015/05/20</td>
<td>文字探勘與網頁探勘 (Text and Web Mining)</td>
</tr>
<tr>
<td>14</td>
<td>2015/05/27</td>
<td>期末報告 (Final Project Presentation)</td>
</tr>
<tr>
<td>15</td>
<td>2015/06/03</td>
<td>畢業考試週 (Final Exam)</td>
</tr>
</tbody>
</table>
Learning Objectives

• Differentiate between text mining, Web mining and data mining
• Text mining
• Web mining
  – Web content mining
  – Web structure mining
  – Web usage mining
• Natural Language Processing (NLP)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text and Web Mining

- Text Mining: Applications and Theory
- Web Mining and Social Networking
- Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
- Search Engines – Information Retrieval in Practice
Text Mining

Web Mining and Social Networking
Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

http://www.amazon.com/Web-Data-Mining-Data-Centric-Applications/dp/3540378812
Search Engines: Information Retrieval in Practice
Text Mining
(text data mining)

the process of deriving high-quality information from text

http://en.wikipedia.org/wiki/Text_mining
Typical Text Mining Tasks

• Text categorization
• Text clustering
• Concept/entity extraction
• Production of granular taxonomies
• Sentiment analysis
• Document summarization
• Entity relation modeling
  – i.e., learning relations between named entities.

http://en.wikipedia.org/wiki/Text_mining
Web Mining

• Web mining
  – discover useful information or knowledge from the Web hyperlink structure, page content, and usage data.

• Three types of web mining tasks
  – Web structure mining
  – Web content mining
  – Web usage mining

Text Mining Concepts

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining versus Text Mining

• Both seek for novel and useful patterns
• Both are semi-automated processes
• Difference is the nature of the data:
  – Structured versus unstructured data
  – **Structured data:** in databases
  – **Unstructured data:** Word documents, PDF files, text excerpts, XML files, and so on
• Text mining – first, impose structure to the data, then mine the structured data

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Concepts

• Benefits of text mining are obvious especially in text-rich data environments
  – e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.

• Electronic communication records (e.g., Email)
  – Spam filtering
  – Email prioritization and categorization
  – Automatic response generation

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application Area

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Terminology

- Unstructured or semistructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Terminology

• Term dictionary
• Word frequency
• Part-of-speech tagging (POS)
• Morphology
• Term-by-document matrix (TDM)
  – Occurrence matrix
• Singular Value Decomposition (SVD)
  – Latent Semantic Indexing (LSI)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• Structuring a collection of text
  – Old approach: bag-of-words
  – New approach: natural language processing

• NLP is ...
  – a very important concept in text mining
  – a subfield of artificial intelligence and computational linguistics
  – the studies of "understanding" the natural human language

• Syntax versus semantics based text mining

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• What is “Understanding”?
  – Human understands, what about computers?
  – Natural language is vague, context driven
  – True understanding requires extensive knowledge of a topic

  – Can/will computers ever understand natural language the same/accurate way we do?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• Challenges in NLP
  – Part-of-speech tagging
  – Text segmentation
  – Word sense disambiguation
  – Syntax ambiguity
  – Imperfect or irregular input
  – Speech acts

• Dream of AI community
  – to have algorithms that are capable of automatically reading and obtaining knowledge from text

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Natural Language Processing (NLP)

• WordNet
  – A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  – A major resource for NLP
  – Need automation to be completed

• Sentiment Analysis
  – A technique used to detect favorable and unfavorable opinions toward specific products and services
  – CRM application

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
NLP Task Categories

- Information retrieval (IR)
- Information extraction (IE)
- Named-entity recognition (NER)
- Question answering (QA)
- Automatic summarization
- Natural language generation and understanding (NLU)
- Machine translation (ML)
- Foreign language reading and writing
- Speech recognition
- Text proofing
- Optical character recognition (OCR)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications

• Marketing applications
  – Enables better CRM

• Security applications
  – ECHELON, OASIS
  – Deception detection (…)

• Medicine and biology
  – Literature-based gene identification (…)

• Academic applications
  – Research stream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Tools

• Commercial Software Tools
  – SPSS PASW Text Miner
  – SAS Enterprise Miner
  – Statistica Data Miner
  – ClearForest, ...

• Free Software Tools
  – RapidMiner
  – GATE
  – Spy-EM, ...

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
SAS Text Analytics

https://www.youtube.com/watch?v=l1rYdrRCZJ4
Web Mining Overview

• Web is the largest repository of data
• Data is in HTML, XML, text format
• Challenges (of processing Web data)
  – The Web is too big for effective data mining
  – The Web is too complex
  – The Web is too dynamic
  – The Web is not specific to a domain
  – The Web has everything

• Opportunities and challenges are great!

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Mining

• Web mining (or Web data mining) is the **process** of discovering intrinsic relationships from Web data (textual, linkage, or usage)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Content/Structure Mining

• Mining of the textual content on the Web
• Data collection via Web crawlers

• Web pages include hyperlinks
  – Authoritative pages
  – Hubs
  – hyperlink-induced topic search (HITS) alg

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining

• Extraction of information from data generated through Web page visits and transactions...
  – data stored in server access logs, referrer logs, agent logs, and client-side cookies
  – user characteristics and usage profiles
  – metadata, such as page attributes, content attributes, and usage data

• Clickstream data

• Clickstream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining

• Web usage mining applications
  – Determine the lifetime value of clients
  – Design cross-marketing strategies across products.
  – Evaluate promotional campaigns
  – Target electronic ads and coupons at user groups based on user access patterns
  – Predict user behavior based on previously learned rules and users' profiles
  – Present dynamic information to users based on their interests and profiles...

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining
(clickstream analysis)

Pre-Process Data
- Collecting
- Merging
- Cleaning
- Structuring
  - Identify users
  - Identify sessions
  - Identify page views
  - Identify visits

Extract Knowledge
- Usage patterns
- User profiles
- Page profiles
- Visit profiles
- Customer value

How to better the data
How to improve the Web site
How to increase the customer value

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, ...
- Website Optimization Ecosystem

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
歐巴馬是美國的一位總統
抗氣候變遷 白宮籲採緊急行動

抗氣候變遷 白宮籲採緊急行動
中央社中央社 – 2014年5月6日 下午10:58
（中央社華盛頓6日綜合外電報導）白宮今天公布全球暖化對全美及美國經濟關鍵產業造成何種衝擊的新報告，呼籲採取緊急行動對抗氣候變遷。

這份為期4年的調查報告，極端氣候事件將對住家、基礎設施及產業帶來嚴重威脅。

美國總統歐巴馬2008年當選總統時曾在競選造勢時誓言，要讓美國成為對抗氣候變遷與相關「安全威脅」的領頭羊。

但歐巴馬在任上一直未能說服美國國會採取重大行動。

在本週對這項議題採取的新作為中，歐巴馬今天將與數名氣象學家接受電視訪問，討論美國全國氣候評估第3版調查結果。

美國數百名來自政府與民間的頂尖氣候科學家及技術專家，共同投入這項研究，檢視氣候變遷對當今帶來的衝擊並預測將對下個世紀帶來何種影響。

研究人員警告，加州可能發生旱災、奧克拉荷馬州發生草原大火，東岸則可能遭遇海平面上升，尤其佛羅里達，而這些事件多為人類造成。

海平面上升也將吞噬密西西比等低窪地區。

至於超過8000萬人居住且擁有全美部分成長最快都會區的東南部與加勒比海區，「海平面上升加上其他與氣候變遷有關的衝擊，以及地層下陷等既有問題，將對經濟和生態帶來重大影響」。

https://tw.news.yahoo.com/%E6%8A%97%E6%B0%A3%E5%80%99%E8%AE%8A%E9%81%B7-%E7%99%BD%E5%AE%AE%E7%B1%B2%E6%8E%A1%E7%B7%8A%E6%80%A5%E8%A1%8C%E5%8B%95-145804493.html
線上展示使用簡化詞類進行斷詞標記，僅供參考並且系統不再進行更新。線上服務斷詞和授權mirror site僅提供精簡詞類，結果也與舊版的展示系統不同。

自 2014/01/06 起，本斷詞系統已經處理過 929136 篇文章

抗氣候變遷 白宮籲採取緊急行動
中央社中央社 - 2014年5月6日 下午10:58
（中央社華盛頓6日綜合外電報導）白宮今天公布全球暖化對全美及美國經濟關鍵產業造成何種衝擊的新報告，呼籲採取緊急行動對抗氣候變遷。

這份為期4年的調查警告，極端氣候事件將對住家、基礎設施及產業帶來嚴重威脅。

美國總統歐巴馬2008年當選總統時曾在競選造勢時誓言，要讓美國成為對抗氣候變遷與相關「安全威脅」的領頭羊。

但歐巴馬在任上一直未能說服美國國會採取重大行動。

在本週對這項議題採取的新作為中，歐巴馬今天將與數名氣象學家接受電視訪問，討論美國全國氣候評估第3版調查結果。

美國數百名來自政府與民間的頂尖氣候科學家及技術專家，共同投入這項研究，檢視氣候變遷對當今帶來的衝擊並預測將對下個世紀帶來何種影響。

研究人員警告，加州可能發生旱災、奧克拉荷馬州發生草原大火，東岸則可能遭遇海平面上升，尤其佛羅里達，而這些事件多為人類造成。

海平面上升也將吞噬密西西比等低窪地區。

至於超過8000萬人居住且擁有全美部分成長最快都會區的東南部與加勒比海區，「海平面上升加上其他與氣候變遷有關的衝擊，以及地層下陷等具有問題，將對經濟和生態帶來重大影響」。

報告並說：「過去被認為是遙遠未來議題的氣候變遷，已著實成為當前議題。」（譯者：中央社蔡佳伶）1030506
中研院中文斷詞系統
http://ckipsvr.iis.sinica.edu.tw/
The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Jython), Ruby, Perl, Javascript, and F# or other .NET languages.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

All these software distributions are open source, licensed under the GNU General Public License (v2 or later). Note that this is the full GPL, which allows many free uses, but does not allow its incorporation into any type of distributed proprietary software, even in part or in translation. Commercial licensing is also available; please contact us if you are interested.

Stanford CoreNLP
An integrated suite of natural language processing tools for English and (mainland) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.

Stanford Parser
Implementations of probabilistic natural language parsers in Java: highly optimized PCFG and dependency parsers, a lexicalized PCFG parser, and a deep learning reranker. See also: Online parser demo, the Stanford Dependencies page, and Parser FAQ.

Stanford POS Tagger
A maximum-entropy (CMM) part-of-speech (POS) tagger for English,
Stanford University is located in California. It is a great university.

### Part-of-Speech:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stanford University</td>
<td>is</td>
<td>located in California.</td>
</tr>
<tr>
<td>2</td>
<td>It</td>
<td>is</td>
<td>a</td>
</tr>
</tbody>
</table>

### Named Entity Recognition:

1. Stanford University: Organization
2. California: Location

### Coreference:

1. Stanford University: Mention
   - Coreferent with: Stanford University
2. It: Mention
   - Coreferent with: Stanford University
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.

**Named Entity Recognition:**

1. Stanford University is located in California.
2. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Collapsed dependencies:

1. Stanford University is located in California.
2. It is a great university.

Collapsed CC-processed dependencies:

1. Stanford University is located in California.
2. It is a great university.

Visualisation provided using the brat visualisation/annotation software. Copyright © 2011, Stanford University, All Rights Reserved.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.

<table>
<thead>
<tr>
<th>Id</th>
<th>Word</th>
<th>Lemma</th>
<th>Char begin</th>
<th>Char end</th>
<th>POS</th>
<th>NER</th>
<th>Normalized NER</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>It</td>
<td>it</td>
<td>46</td>
<td>48</td>
<td>PRP</td>
<td>O</td>
<td></td>
<td>PERO</td>
</tr>
<tr>
<td>2</td>
<td>is</td>
<td>be</td>
<td>49</td>
<td>51</td>
<td>VBZ</td>
<td>O</td>
<td></td>
<td>PERO</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>52</td>
<td>53</td>
<td>DT</td>
<td>O</td>
<td></td>
<td>PERO</td>
</tr>
<tr>
<td>4</td>
<td>great</td>
<td>great</td>
<td>54</td>
<td>59</td>
<td>JJ</td>
<td>O</td>
<td></td>
<td>PERO</td>
</tr>
<tr>
<td>5</td>
<td>university</td>
<td>university</td>
<td>60</td>
<td>70</td>
<td>NN</td>
<td>O</td>
<td></td>
<td>PERO</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>70</td>
<td>71</td>
<td>.</td>
<td>O</td>
<td></td>
<td>PERO</td>
</tr>
</tbody>
</table>

Parse tree
(ROOT (S (NP (PRP It)) (VP (VBZ is) (NP (DT a) (JJ great) (NN university))) (.) .)))
Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California. It is a great university.
<table>
<thead>
<tr>
<th>Id</th>
<th>Word</th>
<th>Lemma</th>
<th>Char begin</th>
<th>Char end</th>
<th>POS</th>
<th>NER</th>
<th>Normalized NER</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stanford</td>
<td>Stanford</td>
<td>0</td>
<td>8</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td>PER0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>University</td>
<td>University</td>
<td>9</td>
<td>19</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td>PER0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>be</td>
<td>20</td>
<td>22</td>
<td>VBZ</td>
<td>O</td>
<td>PER0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>located</td>
<td>located</td>
<td>23</td>
<td>30</td>
<td>JJ</td>
<td>O</td>
<td>PER0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>in</td>
<td>in</td>
<td>31</td>
<td>33</td>
<td>IN</td>
<td>O</td>
<td>PER0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>California</td>
<td>California</td>
<td>34</td>
<td>44</td>
<td>NNP</td>
<td>LOCATION</td>
<td>PER0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>44</td>
<td>45</td>
<td>.</td>
<td>O</td>
<td>PER0</td>
<td></td>
</tr>
</tbody>
</table>

Parse tree
(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California)))))) (. .)))

Uncollapsed dependencies

root ( ROOT-0 , located-4 )
nn ( University-2 , Stanford-1 )
nsubj ( located-4 , University-2 )
cop ( located-4 , is-3 )
prep ( located-4 , in-5 )
pobj ( in-5 , California-6 )

Collapsed dependencies

root ( ROOT-0 , located-4 )
nn ( University-2 , Stanford-1 )
nsubj ( located-4 , University-2 )
cop ( located-4 , is-3 )
prep_in ( located-4 , California-6 )

Collapsed dependencies with CC processed

root ( ROOT-0 , located-4 )
nn ( University-2 , Stanford-1 )
nsubj ( located-4 , University-2 )
cop ( located-4 , is-3 )
prep_in ( located-4 , California-6 )

Stanford University is located in California. It is a great university.

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process
Stanford University is located in California. It is a great university.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xml-stylesheet href="CoreNLP-to-HTML.xsl" type="text/xsl"/>
<root>
<document>
<sentences>
<sentence id="1">
<tokens>
<token id="1">
<word>Stanford</word>
<lemma>Stanford</lemma>
<CharacterOffsetBegin>0</CharacterOffsetBegin>
<CharacterOffsetEnd>8</CharacterOffsetEnd>
<POS>NNP</POS>
<NER>ORGANIZATION</NER>
<Speaker>PER0</Speaker>
</token>
<token id="2">
<word>University</word>
<lemma>University</lemma>
<CharacterOffsetBegin>9</CharacterOffsetBegin>
<CharacterOffsetEnd>19</CharacterOffsetEnd>
<POS>NNP</POS>
<NER>ORGANIZATION</NER>
<Speaker>PER0</Speaker>
</token>
```

http://nlp.stanford.edu:8080/corenlp/process
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan  @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder.
In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million.

That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares.
Related: Gates reclaims title of world's richest billionaire
Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.
It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation.
The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET
Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE

Copyright © 2011, Stanford University, All Rights Reserved.
Bill Gates no longer Microsoft’s biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft’s history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft, bringing down his total to roughly 330 million. That puts him behind Microsoft’s former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world’s richest billionaire Ballmer, who was Microsoft’s CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates Foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON
- MISC
Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates Foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON
Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE
Bill Gates no longer Microsoft's biggest shareholder. By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder By Patrick Sheridan @CNNTech May 2, 2014 5:46 PM ET NEW YORK - For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft, bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft/ORGANIZATION's CEO until earlier this year, was one of Gates/O' first hires. It's a passing of the torch for Gates who has always been/O the/O largest/O single/O owner/O of/O his/O company/O's/O stock/O. Gates/O now/O spends/O his/O time/O and/O personal/O fortune/O helping/O run/O the/O Bill/ORGANIZATION &/ORGANIZATION Melinda/ORGANIZATION Gates/ORGANIZATION foundation/O. The/O foundation/O has/O spent/O $28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O back/O in/O 1997/O.
Textual Entailment Features for Machine Translation Evaluation

HYP: The virus did not infect anybody.
REF: No one was infected by the virus.

HYP: Virus was infected.
REF: No one was infected by the virus.

自然語言處理與資訊檢索研究資源

http://mail.tku.edu.tw/myday/resources/

淡江大學資訊管理學系
(Department of Information Management, Tamkang University)
自然語言處理與資訊檢索研究資源
(Resources of Natural Language Processing and Information Retrieval)

1. 中央研究院CKIP中文斷詞系統
   授權單位：中央研究院詞庫小組
   授權金額：免費授權學術使用。
   授權日期：2011.03.31。
   CKIP: http://ckipsvr.iis.sinica.edu.tw/

2. 「中央研究院中英雙語詞網」(The Academia Sinica Bilingual Wordnet)
   「中央研究院中英雙語詞網」(The Academia Sinica Bilingual Wordnet)，
   授權「淡江大學資訊管理學系」(Department of Information Management, Tamkang University)學術使用。
   授權單位：中央研究院，中華民國計算語言學學會
   授權金額：「中央研究院中英雙語詞網」(The Academia Sinica Bilingual Wordnet)國內非營利機構(1-10人使用) 非會員：NT$61,000元，
   授權日期：2011.05.16。
   Sinica BOW: http://bow.ling.sinica.edu.tw/
3. 開放式中研院專名問答系統 (OpenASQA)
授權單位：中央研究院資訊科學研究所智慧型代理人系統實驗室
授權金額：免費授權學術使用。
授權日期：2011.05.05。
ASQA: http://asqa.iis.sinica.edu.tw/
自然語言處理與資訊檢索研究資源

http://mail.tku.edu.tw/myday/resources/

4. 哈工大資訊檢索研究中心(HIT-CIR)語言技術平臺
語料資源
哈工大資訊檢索研究中心漢語依存樹庫〔HIT-CIR Chinese Dependency Treebank〕
哈工大資訊檢索研究中心同義詞詞林擴展版〔HIT-CIR Tongyici Cilin (Extended)〕

語言處理模組
斷句 (SplitSentence: Sentence Splitting)
詞法分析 (IRLAS: Lexical Analysis System)
基於SVMTool的詞性標注 (PosTag: Part-of-speech Tagging)
命名實體識別 (NER: Named Entity Recognition)
基於動態局部優化的依存句法分析 (Parser: Dependency Parsing)
基於圖的依存句法分析 (GParser: Graph-based DP)
全文詞義消歧 (WSD: Word Sense Disambiguation)
淺層語義標注模組 (SRL: hallow Semantics Labeling)

資料表示
語言技術置標語言 (LTML: Language Technology Markup Language)

視覺化工具
LTML視覺化XSL

授權單位：哈工大資訊檢索研究中心(HIT-CIR)
授權金額：免費授權學術使用。
授權日期：2011.05.03。
HIT IR: http://ir.hit.edu.cn/
Summary

- Differentiate between text mining, Web mining, and data mining
- Text mining
- Web mining
  - Web content mining
  - Web structure mining
  - Web usage mining
- Natural Language Processing (NLP)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
References

• Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
• Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, 2010, Wiley
• Guandong Xu, Yanchun Zhang, Lin Li, Web Mining and Social Networking: Techniques and Applications, 2011, Springer
• Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, O'Reilly Media
• Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2009, Springer