



Data Mining

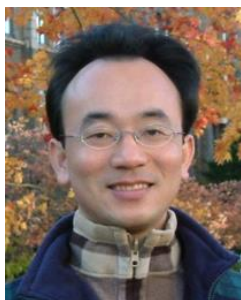
資料探勘

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



課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2015/02/25	資料探勘導論 (Introduction to Data Mining)
2	2015/03/04	關連分析 (Association Analysis)
3	2015/03/11	分類與預測 (Classification and Prediction)
4	2015/03/18	分群分析 (Cluster Analysis)
5	2015/03/25	個案分析與實作一 (SAS EM 分群分析) : Case Study 1 (Cluster Analysis – K-Means using SAS EM)
6	2015/04/01	教學行政觀摩日 (Off-campus study)
7	2015/04/08	個案分析與實作二 (SAS EM 關連分析) : Case Study 2 (Association Analysis using SAS EM)
8	2015/04/15	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)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
9	2015/04/22	期中報告 (Midterm Project Presentation)
10	2015/04/29	期中考試週 (Midterm Exam)
11	2015/05/06	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)
12	2015/05/13	海量資料分析 (Big Data Analytics)
13	2015/05/20	文字探勘與網頁探勘 (Text and Web Mining)
14	2015/05/27	期末報告 (Final Project Presentation)
15	2015/06/03	畢業考試週 (Final Exam)

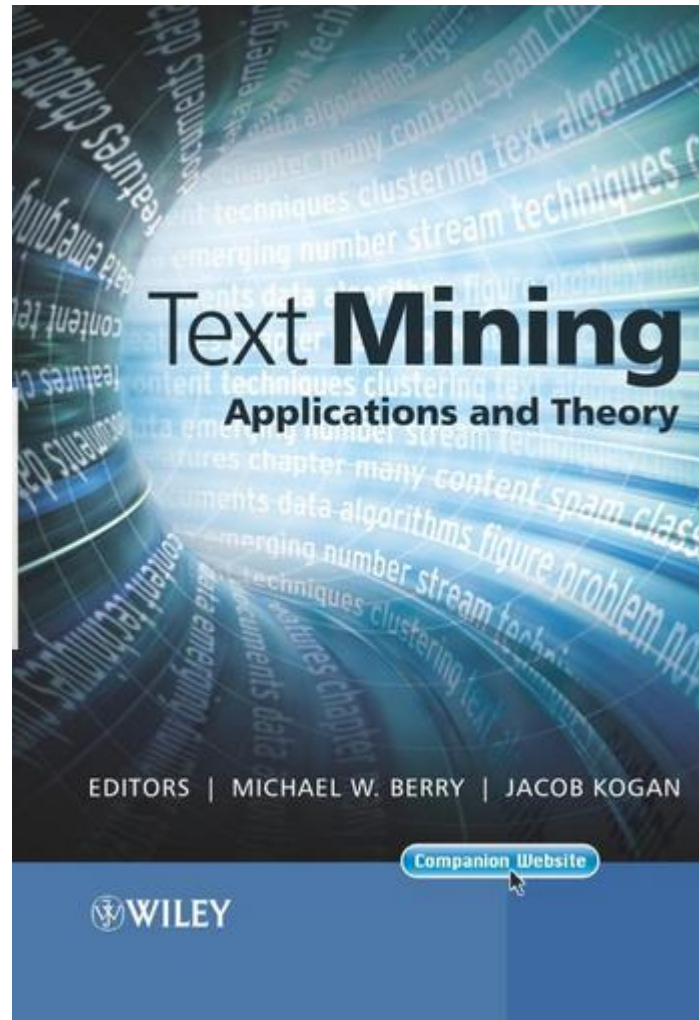
Learning Objectives

- 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)

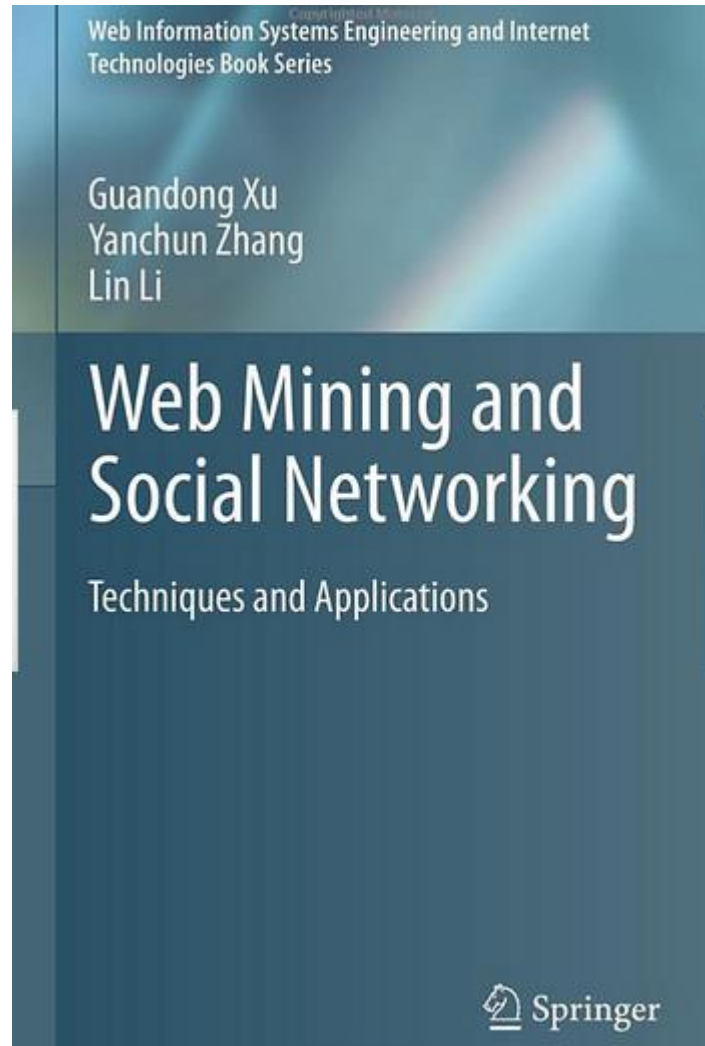
Text and Web Mining

- 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



Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

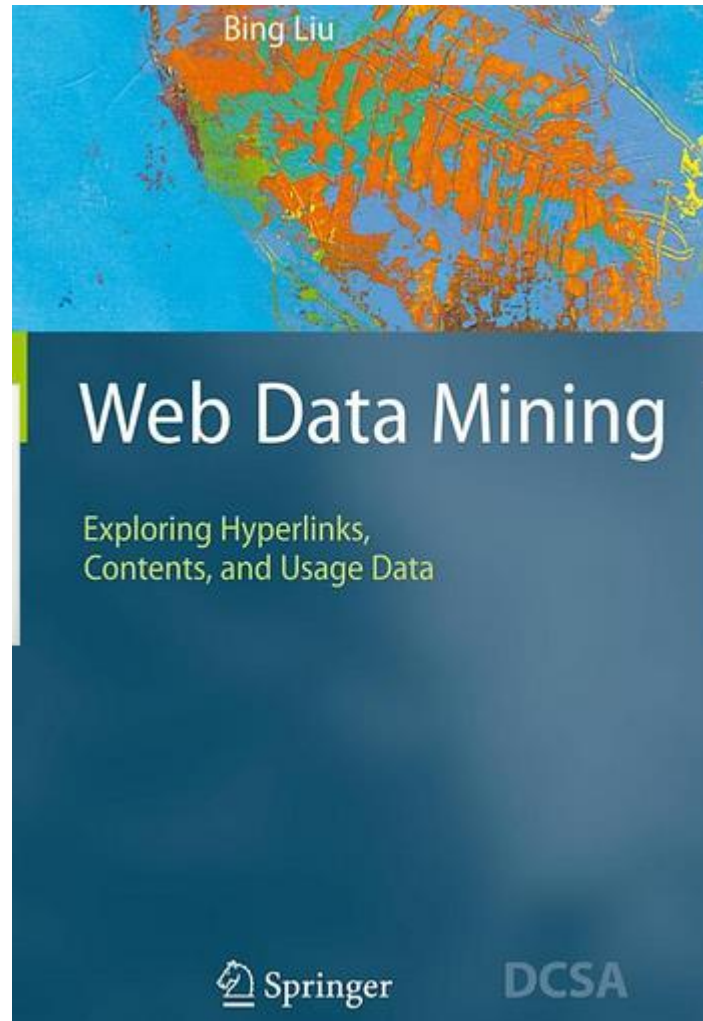
*Analyzing Data from Facebook, Twitter, LinkedIn,
and Other Social Media Sites*



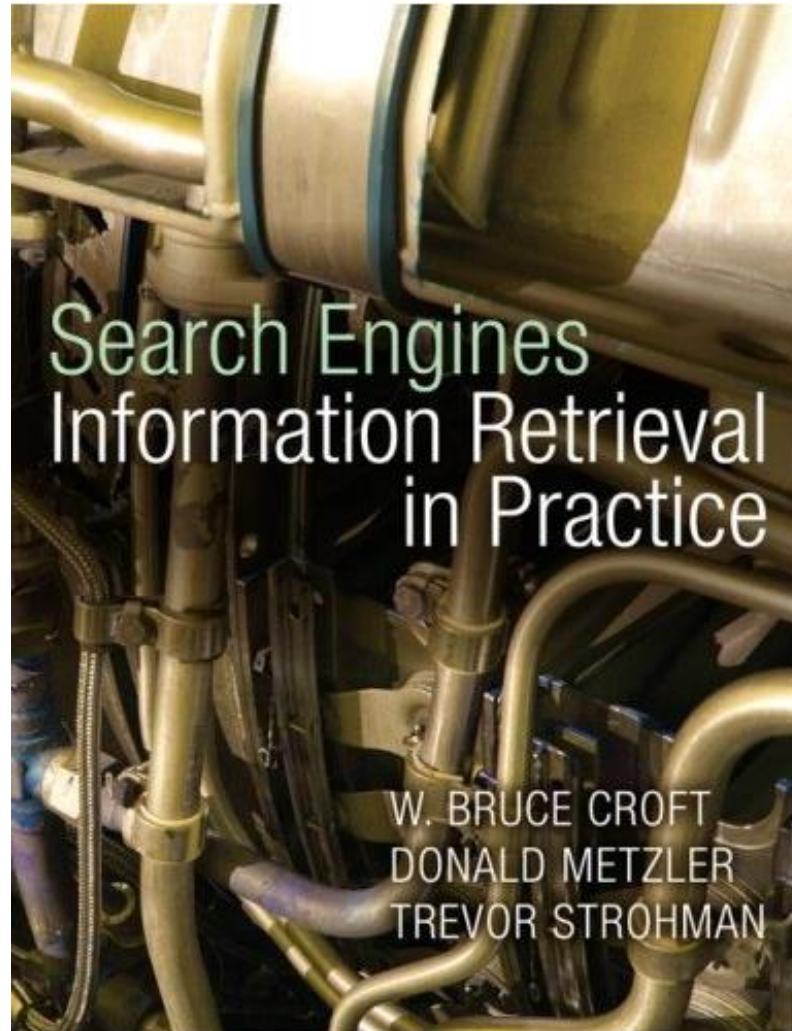
O'REILLY®

Matthew A. Russell

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



Search Engines: Information Retrieval in Practice



Text Mining

(text data mining)

**the process of
deriving
high-quality information
from text**

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.

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

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

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

Text Mining Application Area

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

Text Mining Terminology

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

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)

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

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?

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

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

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)

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

Text Mining Tools

- Commercial Software Tools
 - SPSS PASW Text Miner
 - SAS Enterprise Miner
 - Statistica Data Miner
 - ClearForest, ...
- Free Software Tools
 - RapidMiner
 - GATE
 - Spy-EM, ...

SAS Text Analytics



File:

I called up to **cancel** my card, because I have been **charged** an **annual fee** in every year and **complaint** about there being an **annual fee** is the card that my mother signed me up for when I was still under 18. I inherited it, so I was **charged** this **late fee** again and, because I was that billing cycle I was **charged** a **late fee** on top of that, so I owe whatever my balance is. I was kind of expecting... I have called, because and previous years that this has happened I've called them and **complaint** was just really busy I wasn't able to handle this in a **timely manner**. **annual fee charge**, which is basically the only **charge** on the card and asked to **cancel** and said that I **would pay** whatever that balance **cancel** it. I was kind of expecting that they might offer again to walk me through ahead with the procedure of **cancellation** this card.

Text Result Graphical Result

Test in rule-based model result is Negative
Probability to be positive is 0.239% with confidence 99.541%
43 matches:
0 matches for product definitions:
9 matches for feature definitions:
No 0 (62-71): orion-Fees : annual fee

SAS
THE POWER TO KNOW

0:16 / 4:50

SAS® Text Analytics Software Demo

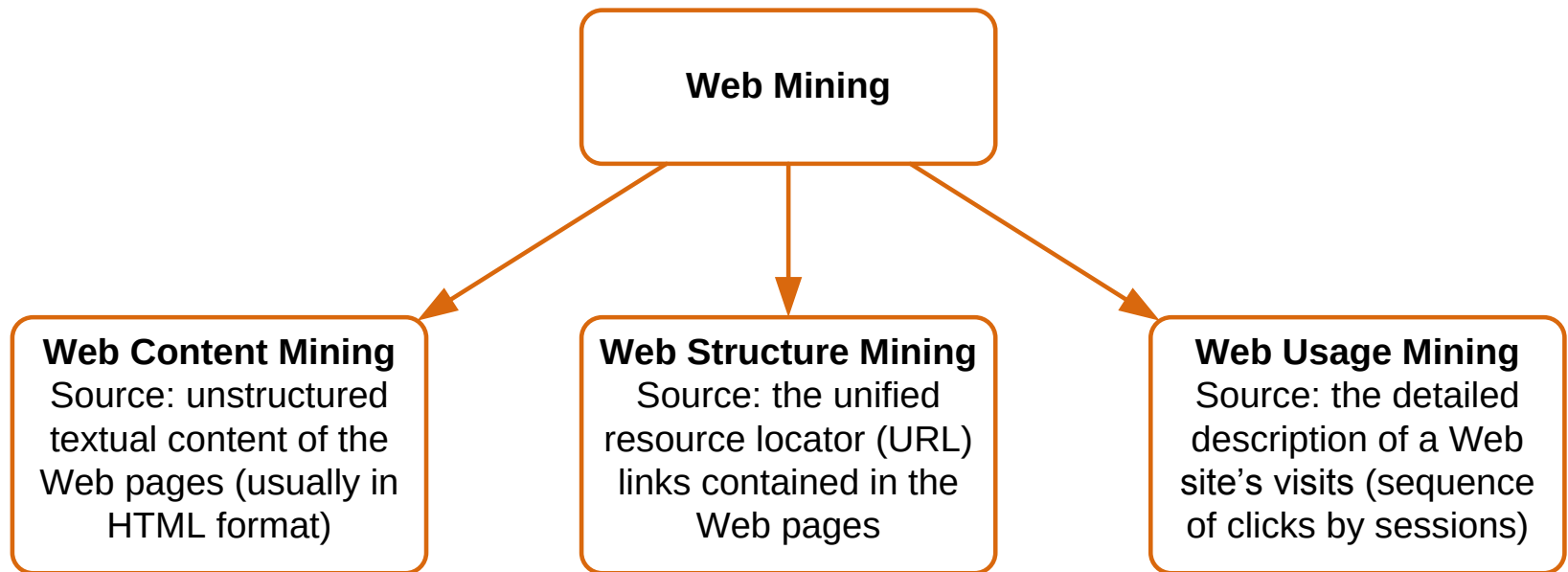


Web Mining Overview

- 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!

Web Mining

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



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

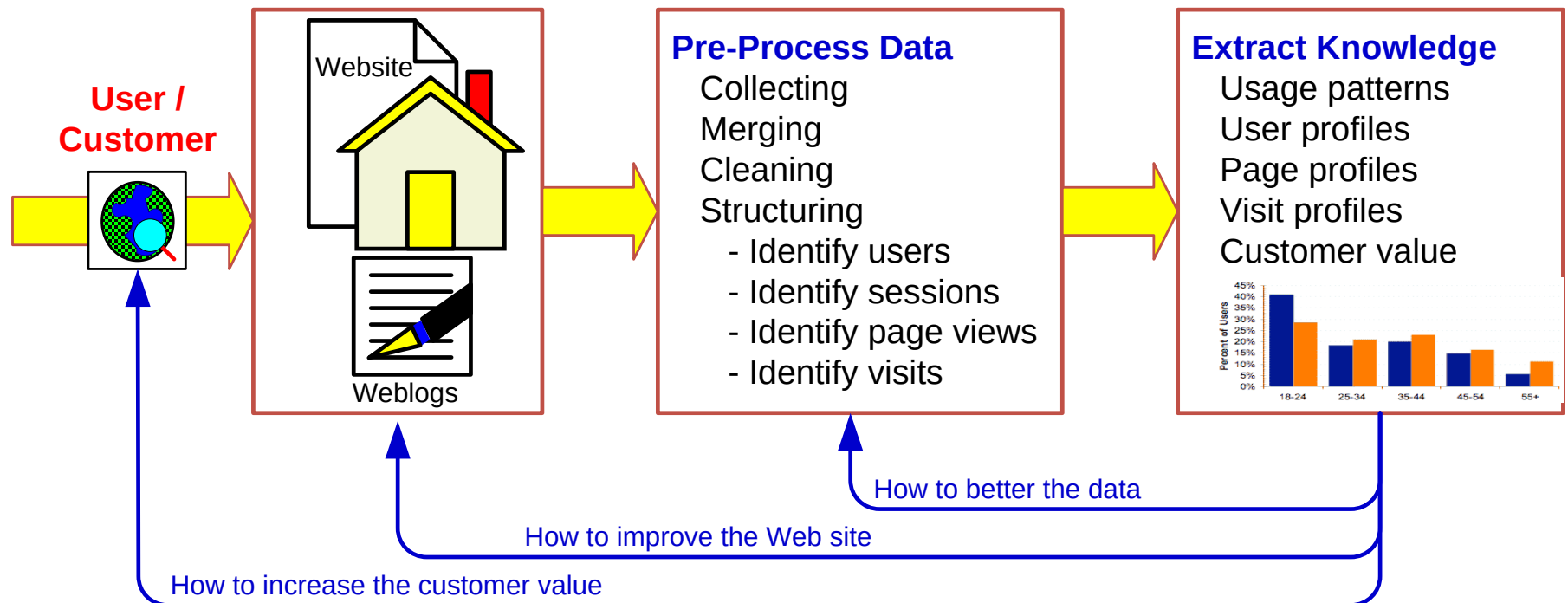
Web Usage Mining

- 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

Web Usage Mining

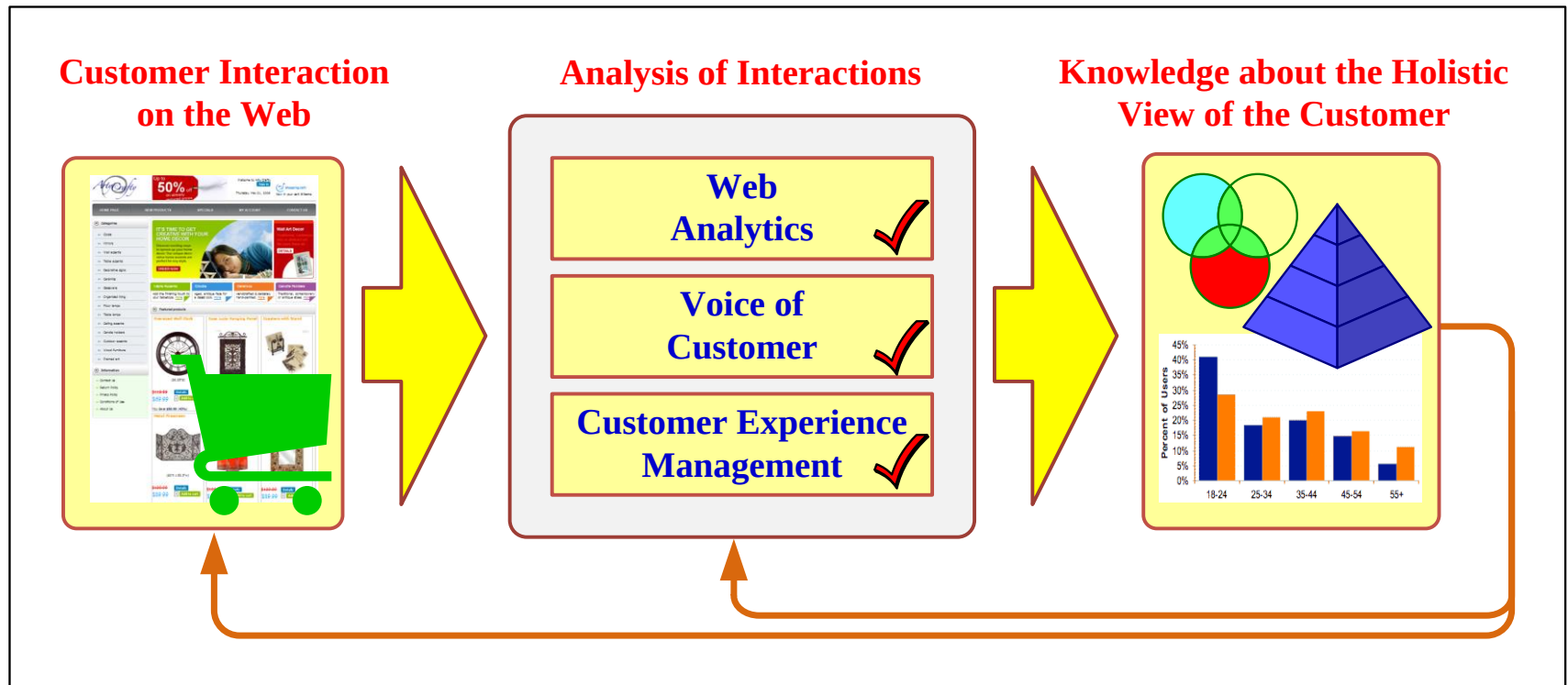
- Web usage mining applications
 - Determine the lifetime value of clients
 - Design cross-marketing strategies across products.
 - Evaluate promotional campaigns
 - Target electronic ads and coupons at user groups based on user access patterns
 - Predict user behavior based on previously learned rules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

Web Usage Mining (clickstream analysis)



Web Mining Success Stories

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



CKIP 中研院中文斷詞系統

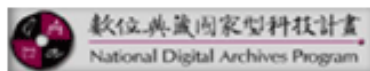
<http://ckipsvr.iis.sinica.edu.tw/>

中文斷詞系統

相關系統：[斷詞系統](#) | [剖析系統](#) | [詞首詞尾](#) | [平衡語料庫](#) | [廣義知網](#) | [句結構樹庫](#) | [錯字偵測](#)

- ➔ [簡介](#)
- ➔ [未知詞擷取做法](#)
- ➔ [詞類標記列表](#)
- ➔ [線上展示](#)
- ➔ [線上服務申請](#)
- ➔ [線上資源](#)
- ➔ [公告](#)
- ➔ [聯絡我們](#)

[隱私權聲明](#) | [版權聲明](#)



Copyright © National
Digital Archives Program,
Taiwan.
All Rights Reserved.

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 929135 篇文章

[送出](#)

[清除](#)

歐巴馬是美國的一位總統

歐巴馬是美國的一位總統

[文章的文字檔](#)

[擷取未知詞過程](#)

[包含未知詞的斷詞標記結果](#)

[未知詞列表](#)

歐巴馬(Nb) 是(SHI) 美國(Nc) 的(DE) 一(Neu) 位(Nf) 總統(Na)

中文文字處理：中文斷詞

抗氣候變遷 白宮籲採緊急行動



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，以及地層下陷等既有問題，將對經濟和生態帶來重大影響」。

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，以及地層下陷等既有問題，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

<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>

CKIP 中研院中文斷詞系統

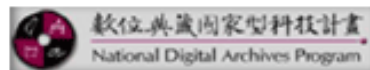
<http://ckipsvr.iis.sinica.edu.tw/>

中文斷詞系統

相關系統：[斷詞系統](#) | [剖析系統](#) | [詞首詞尾](#) | [平衡語料庫](#) | [廣義知網](#) | [句結構樹庫](#) | [錯字偵測](#)

- ➔ [簡介](#)
- ➔ [未知詞擷取做法](#)
- ➔ [詞類標記列表](#)
- ➔ [線上展示](#)
- ➔ [線上服務申請](#)
- ➔ [線上資源](#)
- ➔ [公告](#)
- ➔ [聯絡我們](#)

[隱私權聲明](#) | [版權聲明](#)



Copyright © National
Digital Archives Program,
Taiwan.
All Rights Reserved.

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

[送出](#)

[清除](#)

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，以及地層下陷等既有問題，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

CKIP 中研院中文斷詞系統

<http://ckipsvr.iis.sinica.edu.tw/>

中文斷詞系統

相關系統：斷詞系統 | 詞析系統 | 詞首詞尾 | 平衡語料庫 | 廣義知網 | 句結構樹庫 | 錯字偵測

- ➔ 簡介
- ➔ 未知詞擷取做法
- ➔ 詞類標記列表
- ➔ 線上展示
- ➔ 線上服務申請
- ➔ 線上資源
- ➔ 公告
- ➔ 聯絡我們

隱私權聲明 | 版權聲明



Copyright © National
Digital Archives Program,
Taiwan.
All Rights Reserved.

抗(VJ) 氣候(Na) 變遷(VH) 白宮(Nc) 籲(VE) 採(VC) 緊急(VH) 行動(Na) 中央社(Nc) 中央社(Nc) 2014年(Nd) 5月(Nd) 6日(Nd) 下午(Nd) 1
58(Neu) ((PARENTHESISCATEGORY) 中央社(Nc) 華盛頓(Nc) 6日(Nd) 綜合(A) 外電(Na) 報導(VE)) (PARENTHESISCATEGORY) 白宮(Nc) 今天(Nd)
呼籲(VE) 採取(VC) 緊急(VH) 行動(Na) 對抗(VC) 氣候(Na) 變遷(VH) 。(PERIODCATEGORY)
這(Nep) 份(Nf) 為期(VH) 4年(Nd) 的(DE) 調查(VE) 警告(VE) 。(COMMACATEGORY)
極端(VH) 氣候(Na) 事件(Na) 將(D) 對(P) 住家(Na) 、(PAUSECATEGORY) 基礎(VH) 設施(Na) 及(Caa) 產業(Na) 帶來(VC) 嚴重(VH) 威脅(Na) 。
美國(Nc) 總統(Na) 歐巴馬(Nb) 2008年(Nd) 當選(VG) 總統(Na) 時(Ng) 普(D) 在(P) 競選(VC) 造勢(VB) 時(Ng) 誓言(VE) 。(COMMACATEGORY)
要(D) 讓(VL) 美國(Nc) 成為(VG) 對抗(VC) 氣候(Na) 變遷(VH) 與(Caa) 相關(VH) 「(PARENTHESISCATEGORY) 安全(VH) 威脅(Na) 」(PARENTHESISCATEGORY)
但(Cbb) 歐巴馬(Nb) 在任(VH) 上(Ng) 一直(D) 未(D) 能(D) 說服(VF) 美國(Nc) 國會(Nc) 採取(VC) 重大(VH) 行動(Na) 。(PERIODCATEGORY)
在(P) 本(Nes) 週(Nf) 對(P) 這(Nep) 項(Nf) 議題(Na) 採取(VC) 的(DE) 新作(Na) 為(P) 中(Ncd) 。(COMMACATEGORY)
歐巴馬(Nb) 今天(Nd) 將(D) 與(P) 數(Neu) 名(Nf) 氣象學家(Na) 接受(VC) 電視(Na) 訪問(VC) 。(COMMACATEGORY)
討論(VE) 美國(Nc) 全國(Nc) 氣候(Na) 評估(VE) 第3(Neu) 版(Na) 調查(VE) 結果(Dk) 。(PERIODCATEGORY)
美國(Nc) 數百(Neu) 名(Nf) 來自(VJ) 政府(Na) 與(Caa) 民間(Nc) 的(DE) 頂尖(VH) 氣候(Na) 科學家(Na) 及(Caa) 技術(Na) 專家(Na) 。(COMMACATEGORY)
共同(A) 投入(VC) 這(Nep) 項(Nf) 研究(Na) 。(COMMACATEGORY)
檢視(VC) 氣候(Na) 變遷(VH) 對(P) 當今(Nd) 帶來(VC) 的(DE) 衝擊(Na) 並(D) 預測(VE) 將(D) 對(P) 下(Nes) 個(Nf) 世紀(Na) 帶來(VC) 何
研究(Na) 人員(Na) 警告(VE) 。(COMMACATEGORY)
加州(Nc) 可能(D) 發生(VJ) 旱災(Na) 、(PAUSECATEGORY) 奧克拉荷馬州(Nc) 發生(VJ) 草原(Na) 大火(Na) 。(COMMACATEGORY)
東岸(Nc) 則(D) 可能(D) 遭遇(VJ) 海平面(Na) 上升(VA) 。(COMMACATEGORY)
尤其(D) 佛羅里達(Nc) 。(COMMACATEGORY)
而(Cbb) 這些(Nega) 事件(Na) 多(D) 為(VG) 人類(Na) 造成(VK) 。(PERIODCATEGORY)
海平面(Na) 上升(VA) 也(D) 將(D) 吞噬(VC) 密西西比(Nb) 等(Cab) 低窪(VH) 地區(Nc) 。(PERIODCATEGORY)
至於(P) 超過(VJ) 8000萬(Neu) 人(Na) 居住(VA) 且(Cbb) 擁有(VJ) 全美(Nb) 部分(Nega) 成長(VH) 最(Dfa) 快(VH) 都會區(Nc) 的(DE) 東
「(PARENTHESISCATEGORY) 海平面(Na) 上升(VA) 加上(VC) 其他(Nega) 與(Caa) 氣候(Na) 變遷(VH) 有關(VJ) 的(DE) 衝擊(Na) 。(COMMACATEGORY)



The Stanford Natural Language Processing Group

[home](#) · [people](#) · [teaching](#) · [research](#) · [publications](#) · [software](#) · [events](#) · [local](#)

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 NLP Software

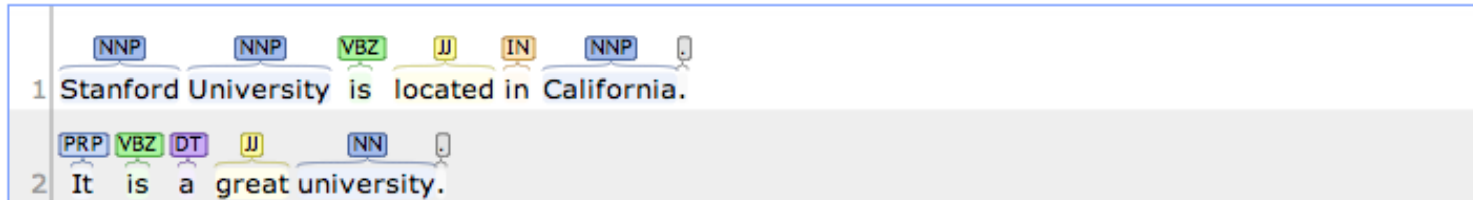
Stanford CoreNLP

Output format:

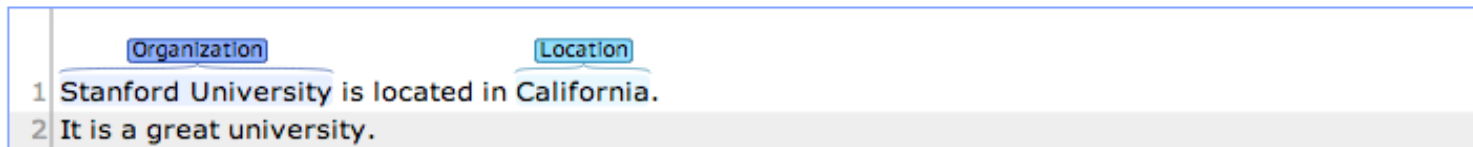
Please enter your text here:

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

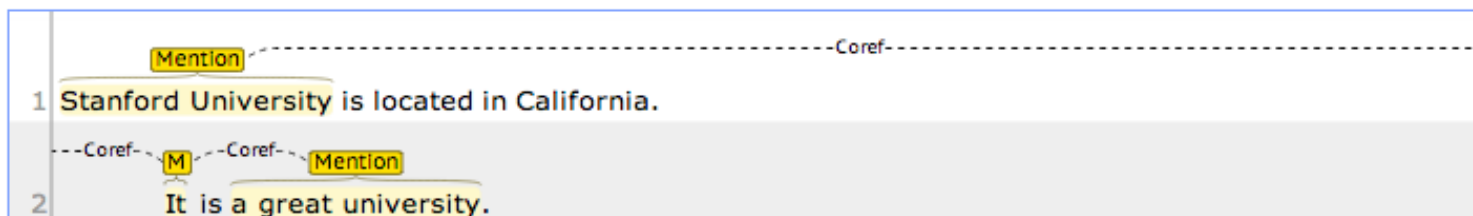
Part-of-Speech:



Named Entity Recognition:



Coreference:

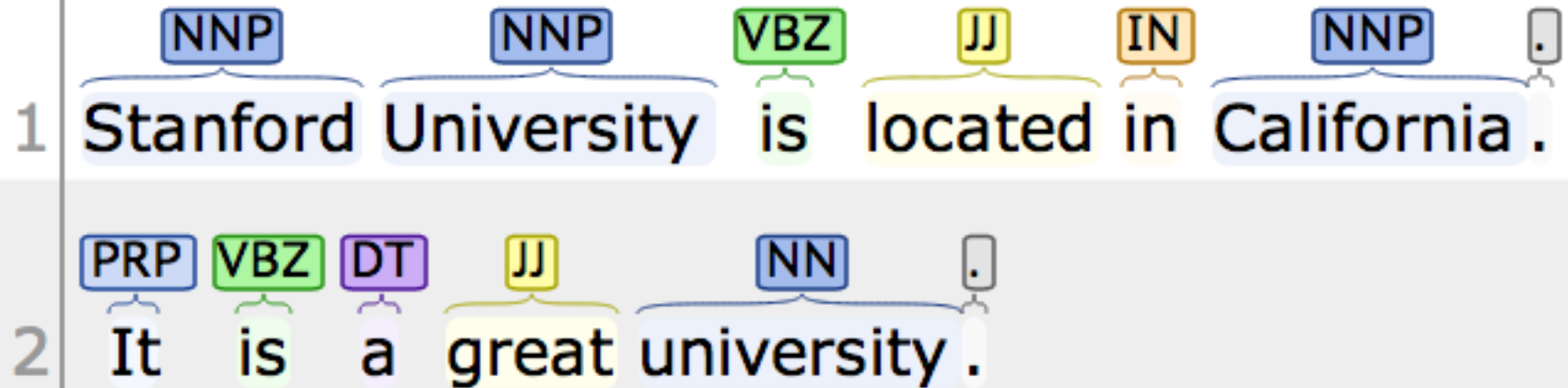


Stanford CoreNLP

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

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

Part-of-Speech:



Stanford CoreNLP

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

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

Named Entity Recognition:

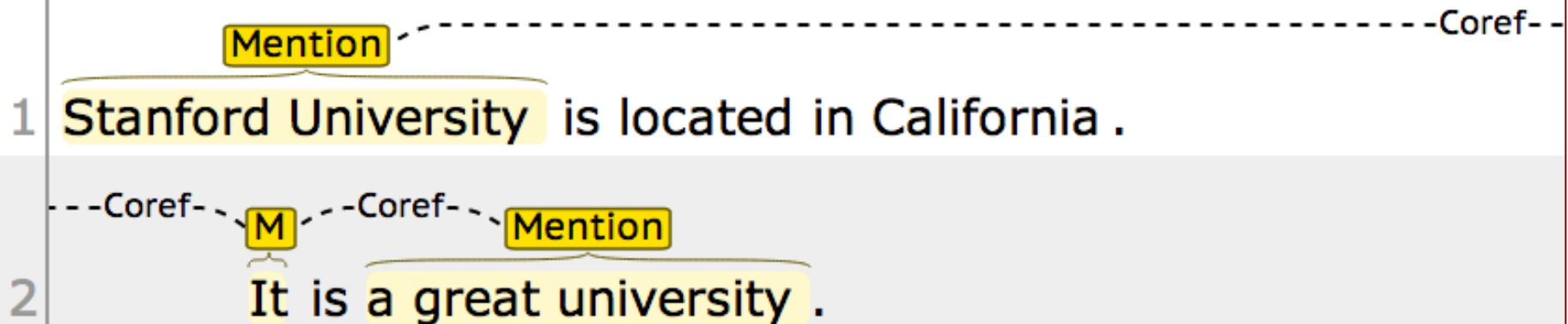
	Organization		Location
1	Stanford University	is located in	California .
2	It is a great university .		

Stanford CoreNLP

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

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

Coreference:

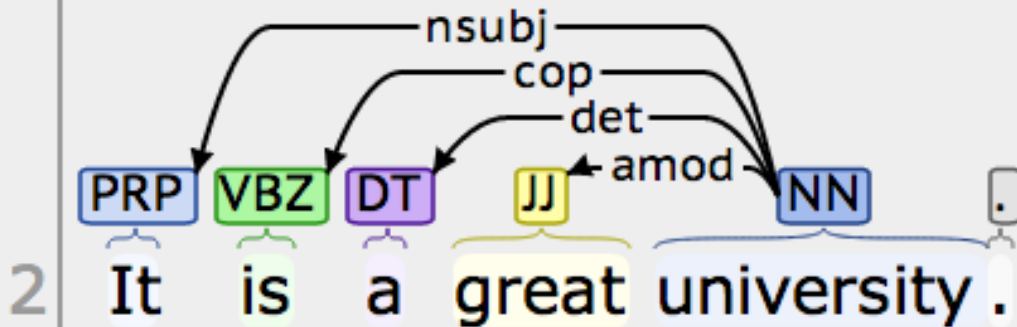
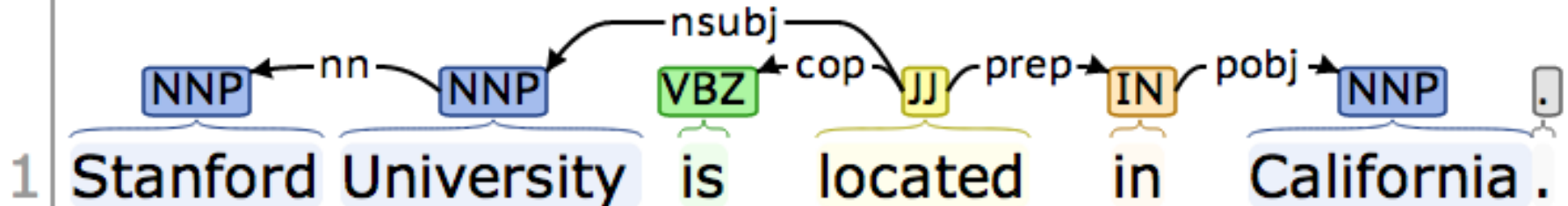


Stanford CoreNLP

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

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

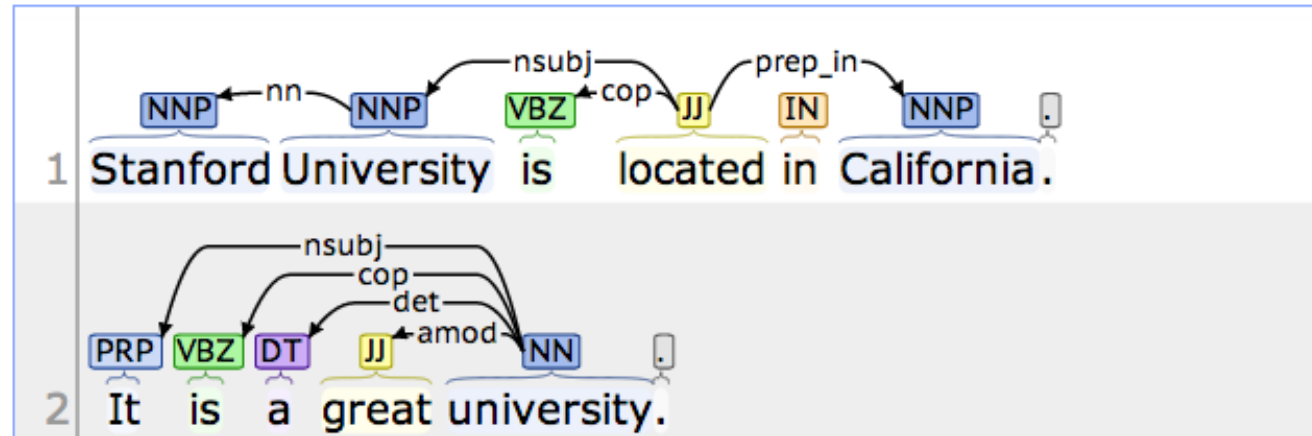
Basic dependencies:



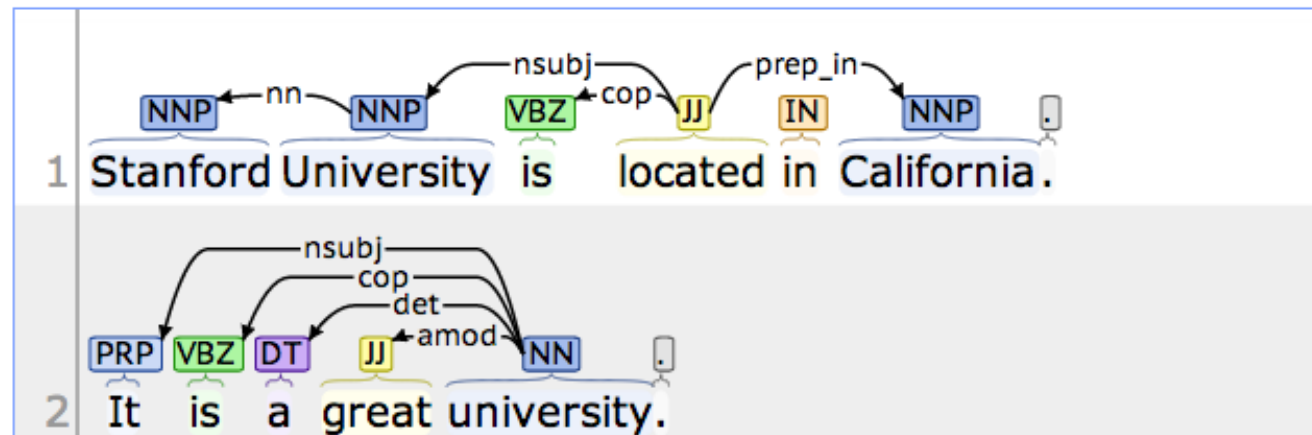
Stanford CoreNLP

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

Collapsed dependencies:



Collapsed CC-processed dependencies:



Visualisation provided using the [brat visualisation/annotation software](#).
Copyright © 2011, [Stanford University](#), All Rights Reserved.

Output format:

Please enter your text here:

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

Stanford CoreNLP XML Output

Document

Document Info

Sentences

Sentence #1

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PERO
2	University	University	9	19	NNP	ORGANIZATION		PERO
3	is	be	20	22	VBZ	O		PERO
4	located	located	23	30	JJ	O		PERO
5	in	in	31	33	IN	O		PERO
6	California	California	34	44	NNP	LOCATION		PERO
7	.	.	44	45	.	O		PERO

Parse tree

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

Stanford CoreNLP

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

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

Sentence #1

Tokens

<i>Id</i>	<i>Word</i>	<i>Lemma</i>	<i>Char begin</i>	<i>Char end</i>	<i>POS</i>	<i>NER</i>	<i>Normalized NER</i>	<i>Speaker</i>
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PERO
2	University	University	9	19	NNP	ORGANIZATION		PERO
3	is	be	20	22	VBZ	O		PERO
4	located	located	23	30	JJ	O		PERO
5	in	in	31	33	IN	O		PERO
6	California	California	34	44	NNP	LOCATION		PERO
7	.	.	44	45	.	O		PERO

Parse tree

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

Stanford CoreNLP

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

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

Sentence #2

Tokens

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

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.

Coreference resolution graph

1.

Sentence	Head	Text	Context
1	2 (gov)	Stanford University	
2	1	It	
2	5	a great university	

Tokens								
Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O	PER0	
4	located	located	23	30	JJ	O	PER0	
5	in	in	31	33	IN	O	PER0	
6	California	California	34	44	NNP	LOCATION	PER0	
7	.	.	44	45	.	O	PER0	

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 CoreNLP

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

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

Output format: XML

Please enter your text here:

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

Submit

Clear

```
<?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>PERO</Speaker>
          </token>
          <token id="2">
            <word>University</word>
            <lemma>University</lemma>
            <CharacterOffsetBegin>9</CharacterOffsetBegin>
            <CharacterOffsetEnd>19</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PERO</Speaker>
          </token>
```

NER for News Article

<http://money.cnn.com/2014/05/02/technology/gates-microsoft-stock-sale/index.html>

money.cnn.com/2014/05/02/technology/gates-microsoft-stock-sale/index.html

2K
TOTAL SHARES

461

1K

74

25

Bill Gates no longer Microsoft's biggest shareholder

By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Recommend



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

2K
TOTAL SHARES

461

1K

74

25

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

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.

Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

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)

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

Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

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)

Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, 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 <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>\$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>.

Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

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.

<wi num="0" entity="O">Bill</wi> <wi num="1" entity="O">Gates</wi> <wi num="2" entity="O">no</wi> <wi num="3" entity="O">longer</wi> <wi num="4" entity="ORGANIZATION">Microsoft</wi> <wi num="5" entity="O">'s</wi> <wi num="6" entity="O">biggest</wi> <wi num="7" entity="O">shareholder</wi> <wi num="8" entity="O">By</wi> <wi num="9" entity="PERSON">Patrick</wi> <wi num="10" entity="PERSON">M.</wi> <wi num="11" entity="PERSON">Sheridan</wi> <wi num="12" entity="O">@CNNTech</wi> <wi num="13" entity="DATE">May</wi> <wi num="14" entity="DATE">2</wi> <wi num="15" entity="DATE">,</wi> <wi num="16" entity="DATE">2014</wi> <wi num="17" entity="O">:</wi> <wi num="18" entity="O">5:46</wi> <wi num="19" entity="O">PM</wi> <wi num="20" entity="O">ET</wi> <wi num="21" entity="O">Bill</wi> <wi num="22" entity="O">Gates</wi> <wi num="23" entity="O">sold</wi> <wi num="24" entity="O">nearly</wi> <wi num="25" entity="O">8</wi> <wi num="26" entity="O">million</wi> <wi num="27" entity="O">shares</wi> <wi num="28" entity="O">of</wi> <wi num="29" entity="ORGANIZATION">Microsoft</wi> <wi num="30" entity="O">over</wi> <wi num="31" entity="O">the</wi> <wi num="32" entity="O">past</wi> <wi num="33" entity="O">two</wi> <wi num="34" entity="O">days</wi> <wi num="35" entity="O">.</wi> <wi num="0" entity="LOCATION">NEW</wi> <wi num="1" entity="LOCATION">YORK</wi> <wi num="2" entity="O">-LRB-</wi> <wi num="3" entity="O">CNNTech</wi> <wi num="4" entity="O">-RRB-</wi> <wi num="5" entity="O">For</wi> <wi num="6" entity="O">the</wi> <wi num="7" entity="O">first</wi> <wi num="8" entity="O">time</wi> <wi num="9" entity="O">in</wi> <wi num="10" entity="ORGANIZATION">Microsoft</wi> <wi num="11" entity="O">'s</wi> <wi num="12" entity="O">history</wi> <wi num="13" entity="O">,</wi> <wi num="14" entity="O">founder</wi> <wi num="15" entity="PERSON">Bill</wi> <wi num="16" entity="PERSON">Gates</wi> <wi num="17" entity="O">is</wi> <wi num="18" entity="O">no</wi> <wi num="19" entity="O">longer</wi> <wi num="20" entity="O">its</wi> <wi num="21" entity="O">largest</wi> <wi num="22" entity="O">individual</wi> <wi num="23" entity="O">shareholder</wi> <wi num="24" entity="O">.</wi> <wi num="0" entity="O">In</wi> <wi num="1" entity="O">the</wi> <wi num="2" entity="DATE">past</wi> <wi num="3" entity="DATE">two</wi> <wi num="4" entity="DATE">days</wi> <wi num="5" entity="O">Bill</wi> <wi num="6" entity="O">Gates</wi> <wi num="7" entity="O">has</wi> <wi num="8" entity="O">sold</wi>

Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

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 (CNNTech)

Bill/O Gates/O no/O longer/O Microsoft/ORGANIZATION's/O biggest/O shareholder/O By/O Patrick/PERSON M./PERSON Sheridan/PERSON @CNNTech/O May/DATE 2/DATE, /DATE 2014/DATE: /O 5:46/O PM/O ET/O Bill/O Gates/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION over/O the/O past/O two/O days/O. /O NEW/LOCATION YORK/LOCATION -LRB-/OCNNMoney/O-RRB-/O For/O the/O first/O time/O in/O Microsoft/ORGANIZATION's/O history/O, /O founder/O Bill/PERSON Gates/PERSON is/O no/O longer/O its/O largest/O individual/O shareholder/O. /O In/O the/O past/DATE two/DATE days/DATE, /O Gates/O has/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION -LRB-/OMSFT/ORGANIZATION, /O Fortune/O 500/O-RRB-/O, /O bringing/O down/O his/O total/O to/O roughly/O 330/O million/O. /O That/O puts/O him/O behind/O Microsoft/ORGANIZATION's/O former/O CEO/O Steve/PERSON Ballmer/PERSON who/O owns/O 333/O million/O shares/O. /O Related/O: /O Gates/O reclaims/O title/O of/O world/O's/O richest/O billionaire/O Ballmer/PERSON, /O who/O was/O Microsoft/ORGANIZATION's/O CEO/O until/O earlier/DATE this/DATE year/DATE, /O was/O one/O of/O Gates/O' /O first/O hires/O. /O It/O's/O a/O passing/O of/O the/O torch/O for/O Gates/O who/O has/O always/O been/O the/O largest/O single/O owner/O of/O his/O company/O's/O stock/O. /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. /O The/O foundation/O has/O spent/O \$/MONEY28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O back/O in/O 1997/DATE./O

Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

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)

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

MISC

Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

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)

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. **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

Stanford Named Entity Tagger (NER)

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

Stanford NER Output Format: inlineXML

Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, 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 <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>\$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>.

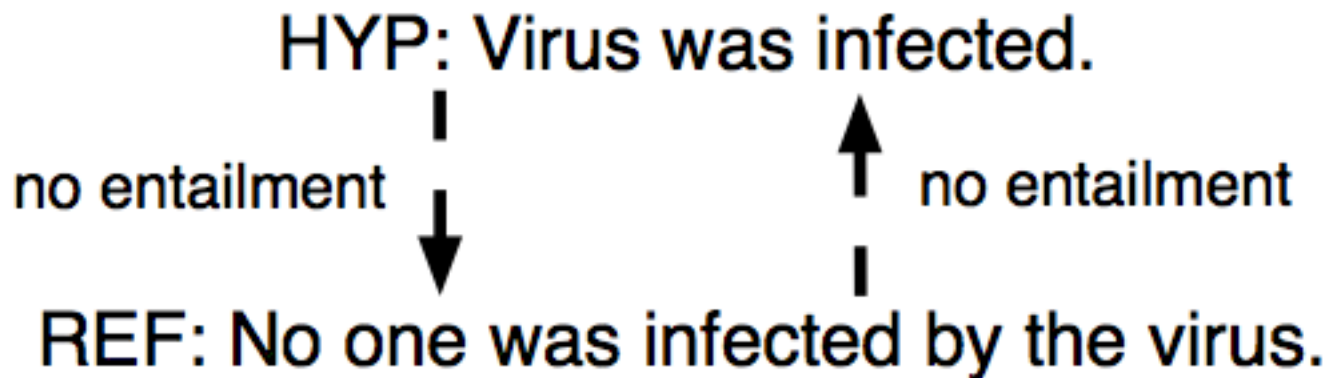
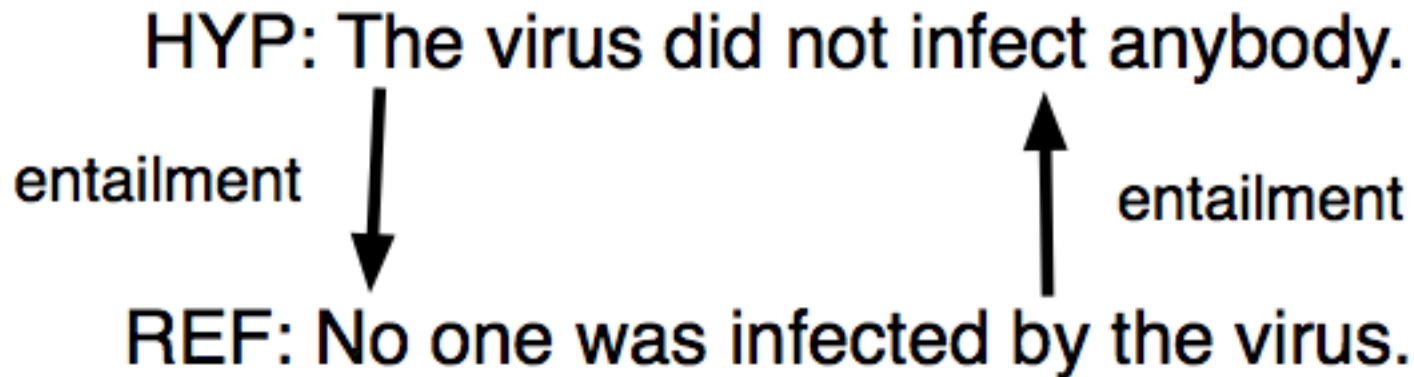
Stanford Named Entity Tagger (NER)

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

Stanford NER Output Format: slashTags

Bill/O Gates/O no/O longer/O Microsoft/ORGANIZATION's/O biggest/O shareholder/O By/O
Patrick/PERSON M./PERSON Sheridan/PERSON @CNNTech/O May/DATE 2/DATE,/DATE
2014/DATE:/O 5:46/O PM/O ET/O Bill/O Gates/O sold/O nearly/O 8/O million/O shares/O of/O
Microsoft/ORGANIZATION over/O the/O past/O two/O days/O./O NEW/LOCATION YORK/LOCATION
-LRB-/OCNNMoney/O-RRB-/O For/O the/O first/O time/O in/O Microsoft/ORGANIZATION's/O
history/O,/O founder/O Bill/PERSON Gates/PERSON is/O no/O longer/O its/O largest/O individual/O
shareholder/O./O In/O the/O past/DATE two/DATE days/DATE,/O Gates/O has/O sold/O nearly/O 8/O
million/O shares/O of/O Microsoft/ORGANIZATION -LRB-/OMSFT/ORGANIZATION,/O Fortune/O
500/O-RRB-/O,/O bringing/O down/O his/O total/O to/O roughly/O 330/O million/O./O That/O puts/O
him/O behind/O Microsoft/ORGANIZATION's/O former/O CEO/O Steve/PERSON Ballmer/PERSON
who/O owns/O 333/O million/O shares/O./O Related/O:/O Gates/O reclaims/O title/O of/O world/O's/O
richest/O billionaire/O Ballmer/PERSON,/O who/O was/O Microsoft/ORGANIZATION's/O CEO/O
until/O earlier/DATE this/DATE year/DATE,/O was/O one/O of/O Gates/O's/O first/O hires/O./O It/O's/O
a/O passing/O of/O the/O torch/O for/O Gates/O who/O has/O always/O been/O the/O largest/O
single/O owner/O of/O his/O company/O's/O stock/O./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./O The/O foundation/O has/O spent/O
\$/MONEY28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O
back/O in/O 1997/DATE./O

Textual Entailment Features for Machine Translation Evaluation



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/>

自然語言處理與資訊檢索研究資源

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

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: shallow 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)

References

- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.
- 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
- Bruce Croft, Donald Metzler, and Trevor Strohman, Search Engines: Information Retrieval in Practice, 2008, Addison Wesley,
<http://www.search-engines-book.com/>
- Text Mining, http://en.wikipedia.org/wiki/Text_mining