

# Social Computing and Big Data Analytics

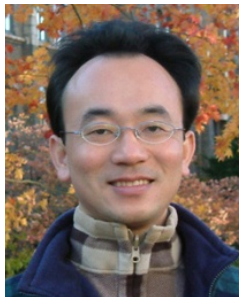
## 社群運算與大數據分析

# Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理)

1042SCBDA07

MIS MBA (M2226) (8628)

Wed, 8,9, (15:10-17:00) (B309)



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/>

2016-03-30



# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2016/02/17	Course Orientation for Social Computing and Big Data Analytics (社群運算與大數據分析課程介紹)
2	2016/02/24	Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data (資料科學與大數據分析： 探索、分析、視覺化與呈現資料)
3	2016/03/02	Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem (大數據基礎：MapReduce典範、 Hadoop與Spark生態系統)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
4	2016/03/09	Big Data Processing Platforms with SMACK: Spark, Mesos, Akka, Cassandra and Kafka (大數據處理平台SMACK： Spark, Mesos, Akka, Cassandra, Kafka)
5	2016/03/16	Big Data Analytics with Numpy in Python (Python Numpy 大數據分析)
6	2016/03/23	Finance Big Data Analytics with Pandas in Python (Python Pandas 財務大數據分析)
7	2016/03/30	Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理)
8	2016/04/06	Off-campus study (教學行政觀摩日)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
9	2016/04/13	Social Media Marketing Analytics (社群媒體行銷分析)
10	2016/04/20	期中報告 (Midterm Project Report)
11	2016/04/27	Deep Learning with Theano and Keras in Python (Python Theano 和 Keras 深度學習)
12	2016/05/04	Deep Learning with Google TensorFlow (Google TensorFlow 深度學習)
13	2016/05/11	Sentiment Analysis on Social Media with Deep Learning (深度學習社群媒體情感分析)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
14	2016/05/18	Social Network Analysis (社會網絡分析)
15	2016/05/25	Measurements of Social Network (社會網絡量測)
16	2016/06/01	Tools of Social Network Analysis (社會網絡分析工具)
17	2016/06/08	Final Project Presentation I (期末報告 I)
18	2016/06/15	Final Project Presentation II (期末報告 II)

# **Text Mining Techniques**

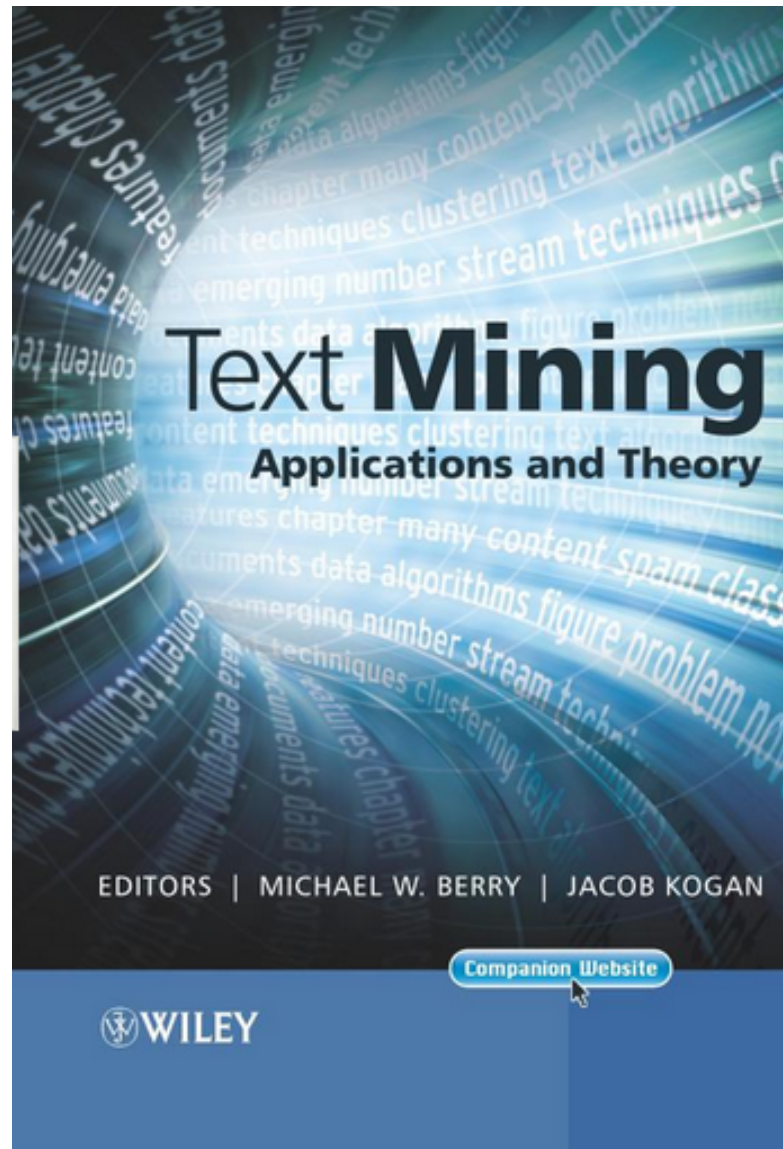
# Natural Language Processing (NLP)

# Outline

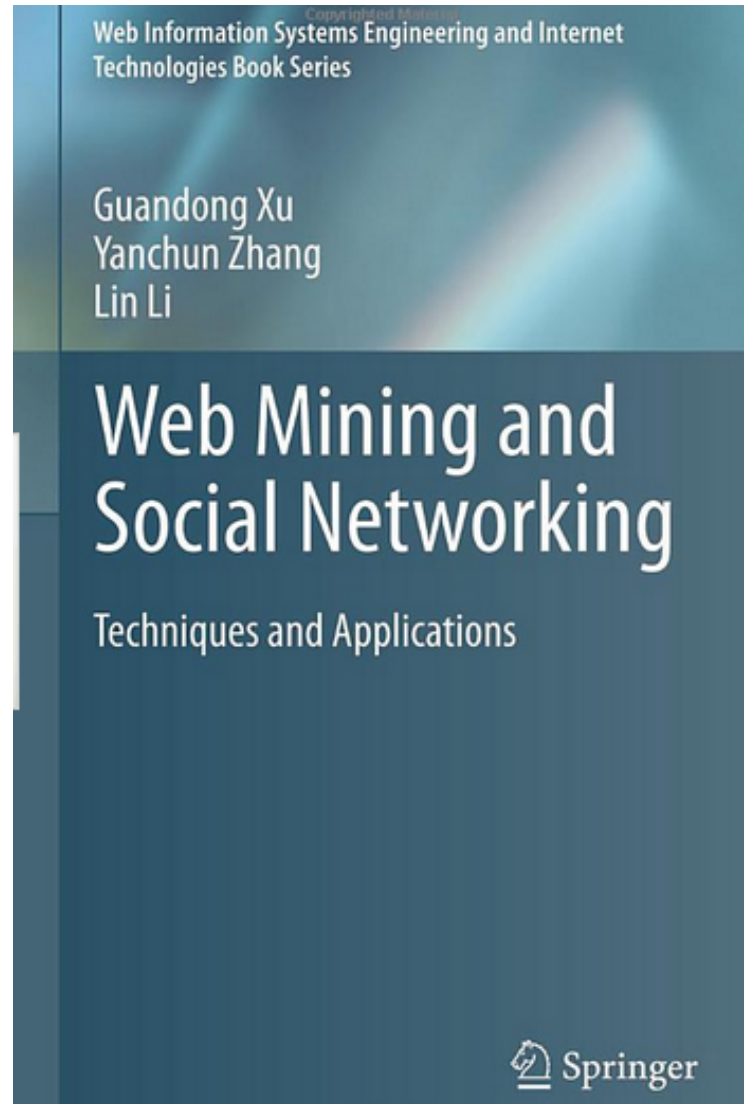
- Differentiate between text mining, Web mining and data mining
- Text mining
- Web mining
  - Web content mining
  - Web structure mining
  - Web usage mining
- Natural Language Processing (NLP)
- Natural Language Processing with NLTK in Python



# 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*

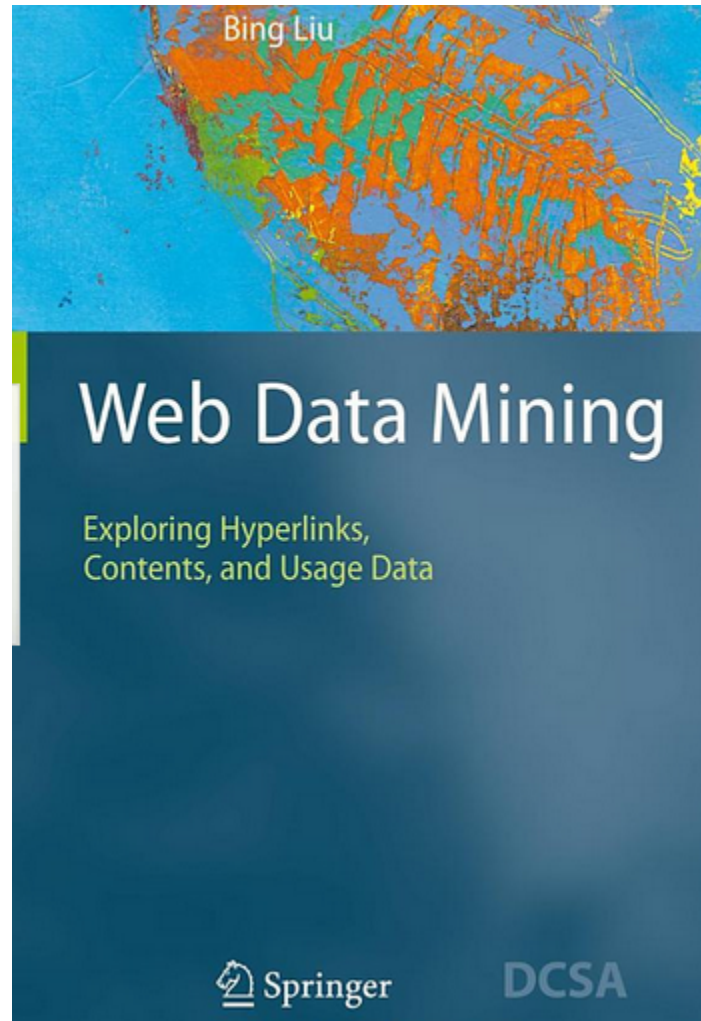


Mining the  
Social Web

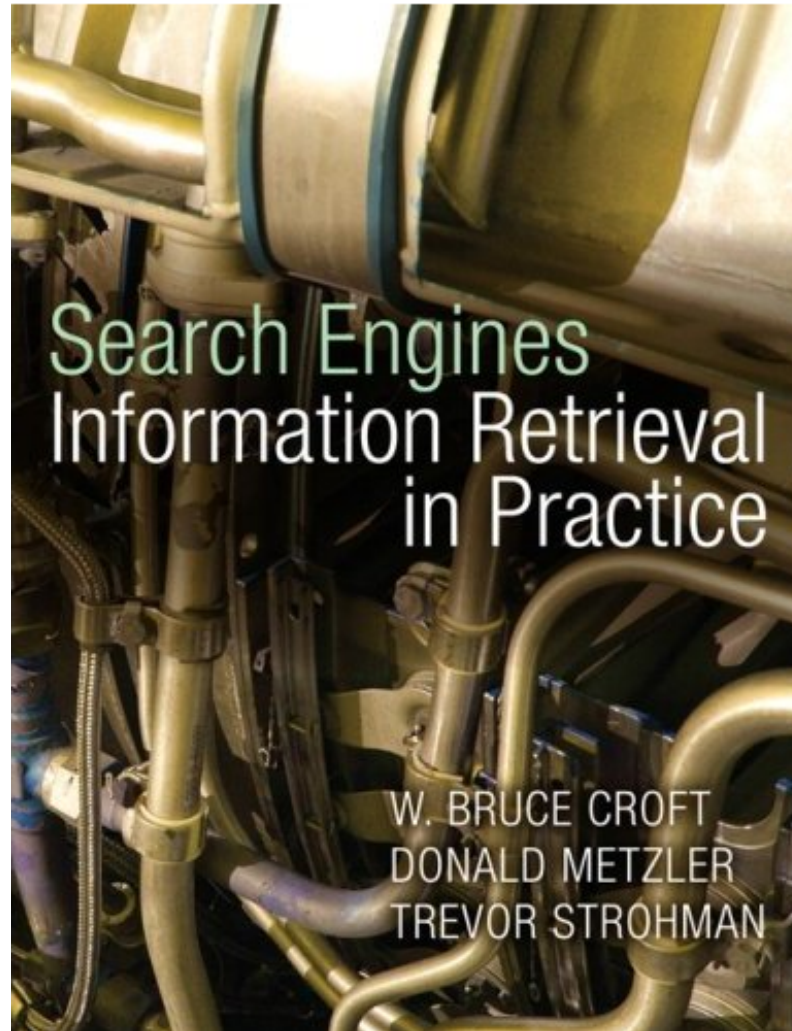
O'REILLY®

*Matthew A. Russell*

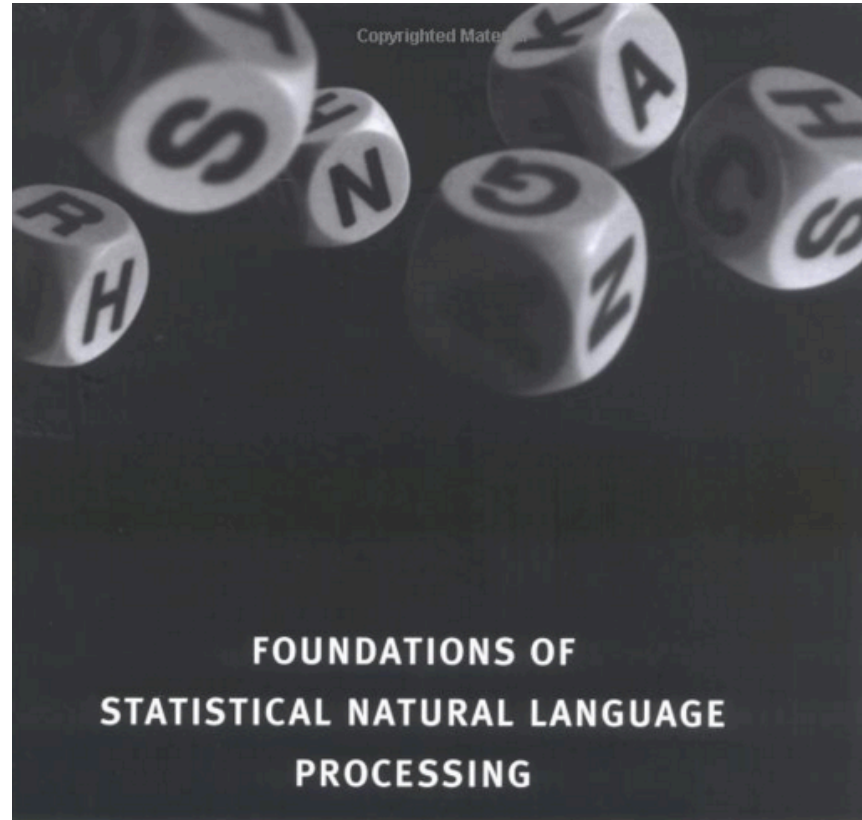
# Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data



# Search Engines: Information Retrieval in Practice

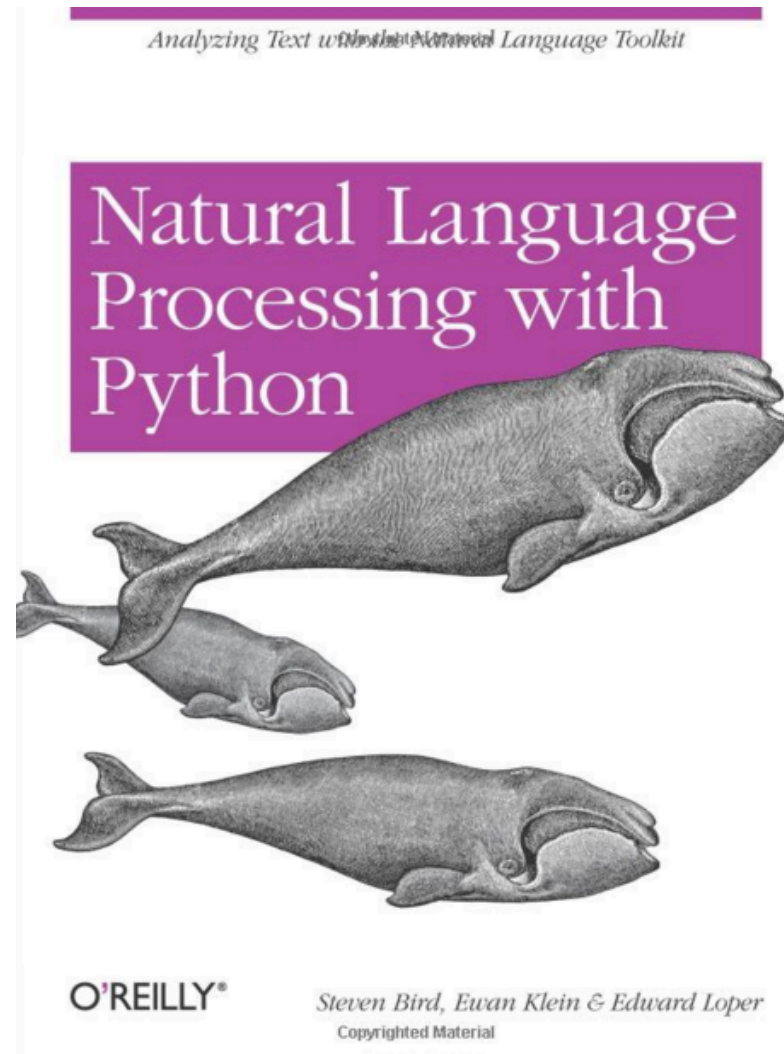


Christopher D. Manning and Hinrich Schütze (1999),  
**Foundations of  
Statistical Natural Language Processing,**  
The MIT Press



**CHRISTOPHER D. MANNING AND  
HINRICH SCHÜTZE**

Steven Bird, Ewan Klein and Edward Loper (2009),  
**Natural Language Processing with Python,**  
O'Reilly Media



# Natural Language Processing with Python

## – Analyzing Text with the Natural Language Toolkit

← → ↻ [www.nltk.org/book/](http://www.nltk.org/book/)



# Natural Language Processing with Python

## – Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

*The NLTK book is currently being updated for Python 3 and NLTK 3. This is work in progress; chapters that still need to be updated are indicated. The first edition of the book, published by O'Reilly, is available at [http://nltk.org/book\\_1ed/](http://nltk.org/book_1ed/). A second edition of the book is anticipated in early 2016.*

0. [Preface](#)
1. [Language Processing and Python](#)
2. [Accessing Text Corpora and Lexical Resources](#)
3. [Processing Raw Text](#)
4. [Writing Structured Programs](#)
5. [Categorizing and Tagging Words](#) (minor fixes still required)
6. [Learning to Classify Text](#)
7. [Extracting Information from Text](#)
8. [Analyzing Sentence Structure](#)
9. [Building Feature Based Grammars](#)
10. [Analyzing the Meaning of Sentences](#) (minor fixes still required)
11. [Managing Linguistic Data](#) (minor fixes still required)
12. [Afterword: Facing the Language Challenge](#)

[Bibliography](#)

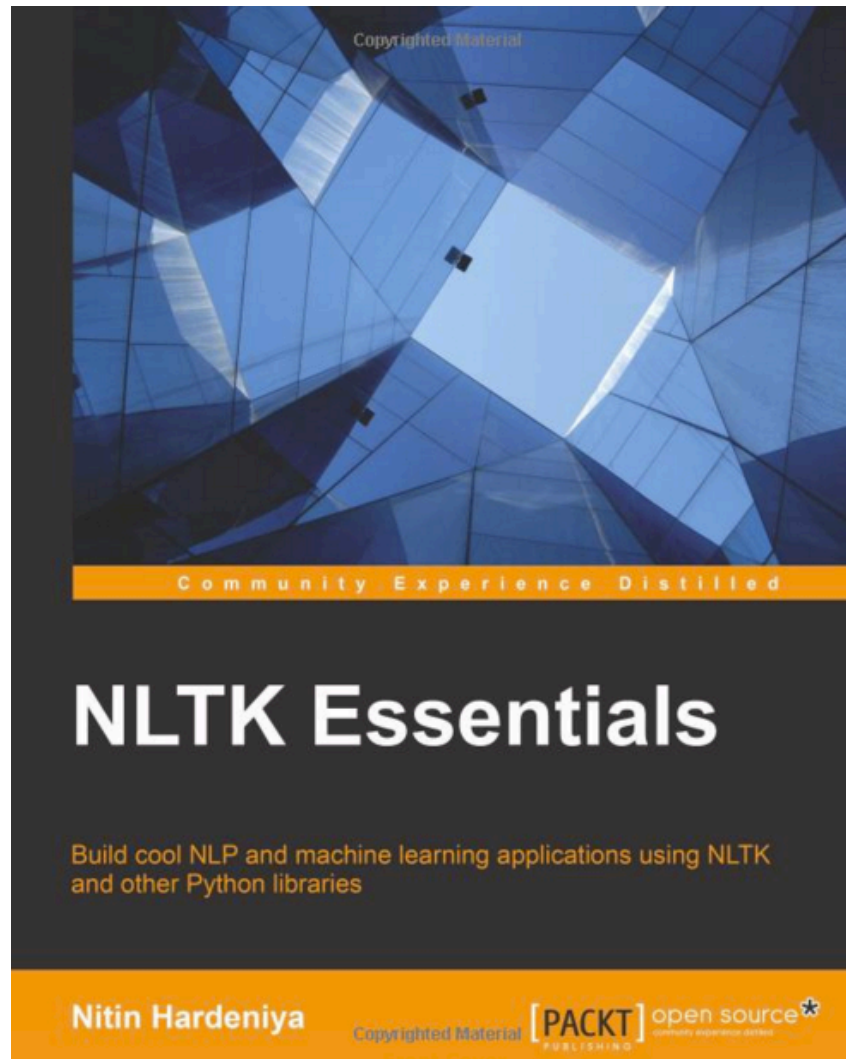
[Term Index](#)

*This book is made available under the terms of the [Creative Commons Attribution Noncommercial No-Derivative-Works 3.0 US License](#). Please post any questions about the materials to the [nltk-users](#) mailing list. Please report any errors on the [issue tracker](#).*

<http://www.nltk.org/book/>



# Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing



<http://www.amazon.com/NLTK-Essentials-Nitin-Hardeniya/dp/1784396907>

# **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

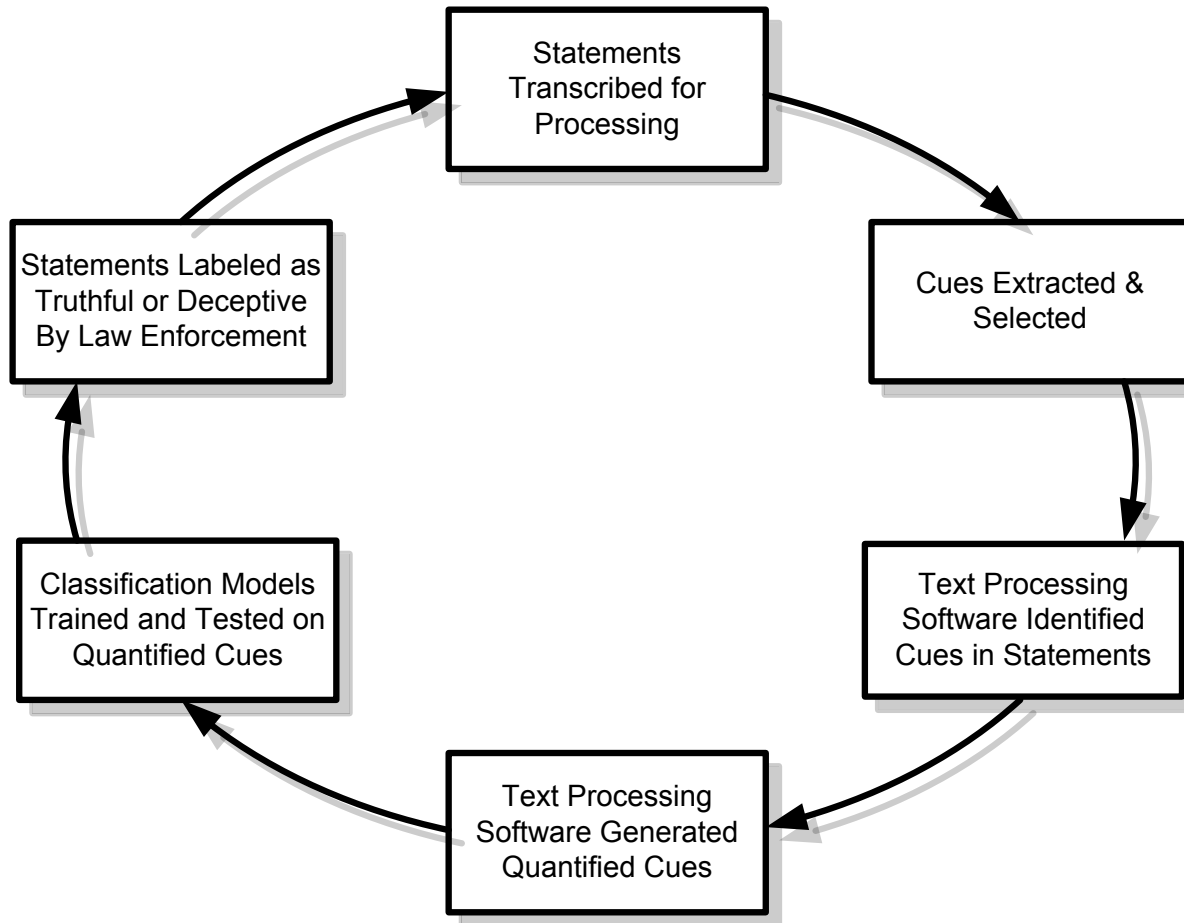


# Text Mining Applications

- Application Case: Mining for Lies
- Deception detection
  - A difficult problem
  - If detection is limited to only text, then the problem is even more difficult
- The study
  - analyzed text based testimonies of person of interests at military bases
  - used only text-based features (cues)

# Text Mining Applications

- Application Case: Mining for Lies



# Text Mining Applications

- Application Case: Mining for Lies

---

<b>Category</b>	<b>Example Cues</b>
Quantity	Verb count, noun-phrase count, ...
Complexity	Avg. no of clauses, sentence length, ...
Uncertainty	Modifiers, modal verbs, ...
Nonimmediacy	Passive voice, objectification, ...
Expressivity	Emotiveness
Diversity	Lexical diversity, redundancy, ...
Informality	Typographical error ratio
Specificity	Spatiotemporal, perceptual information ...
Affect	Positive affect, negative affect, etc.

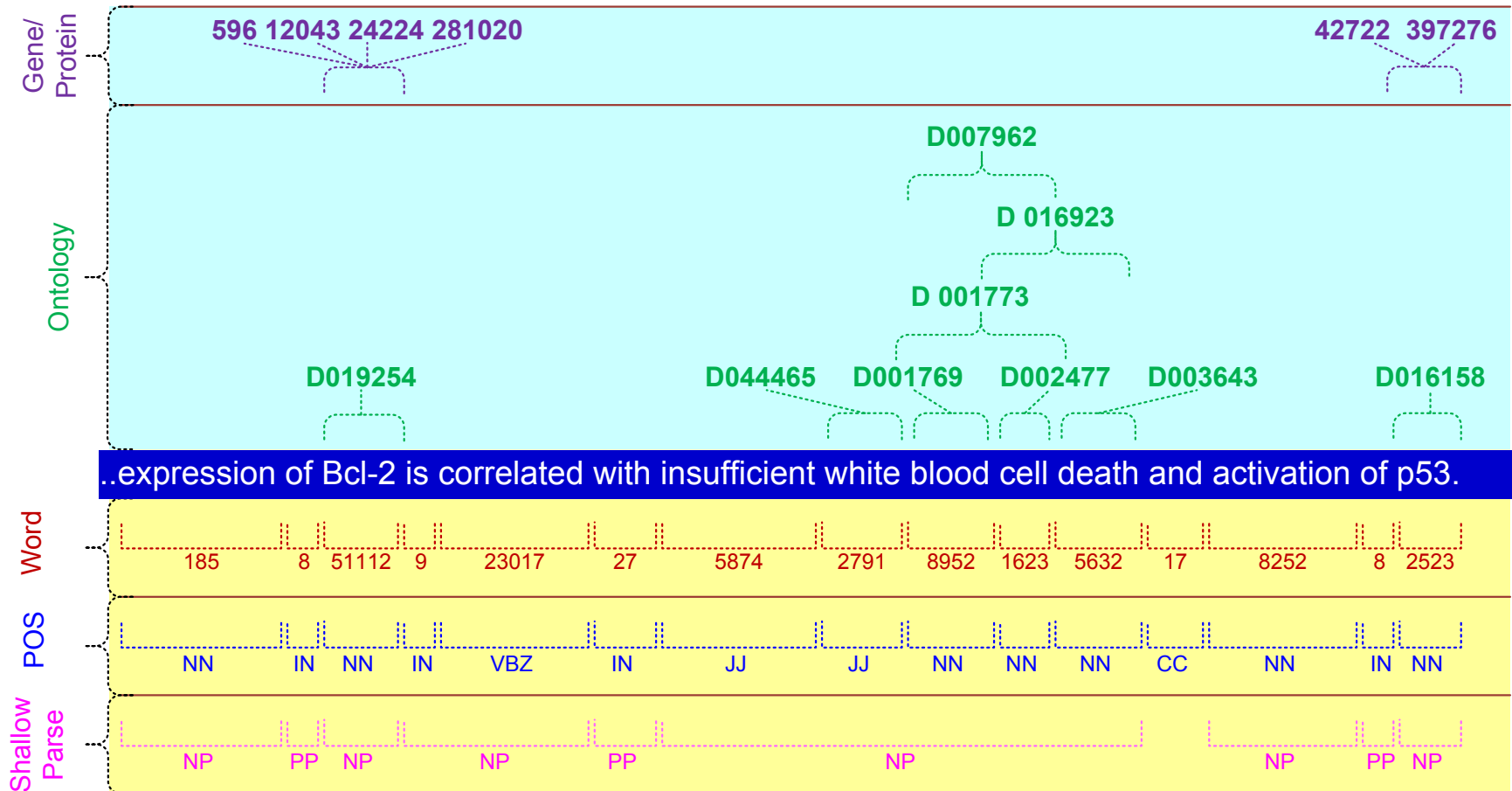
---

# Text Mining Applications

- Application Case: Mining for Lies
  - 371 usable statements are generated
  - 31 features are used
  - Different feature selection methods used
  - 10-fold cross validation is used
  - Results (overall % accuracy)
    - Logistic regression                      67.28
    - Decision trees                              71.60
    - Neural networks                            73.46

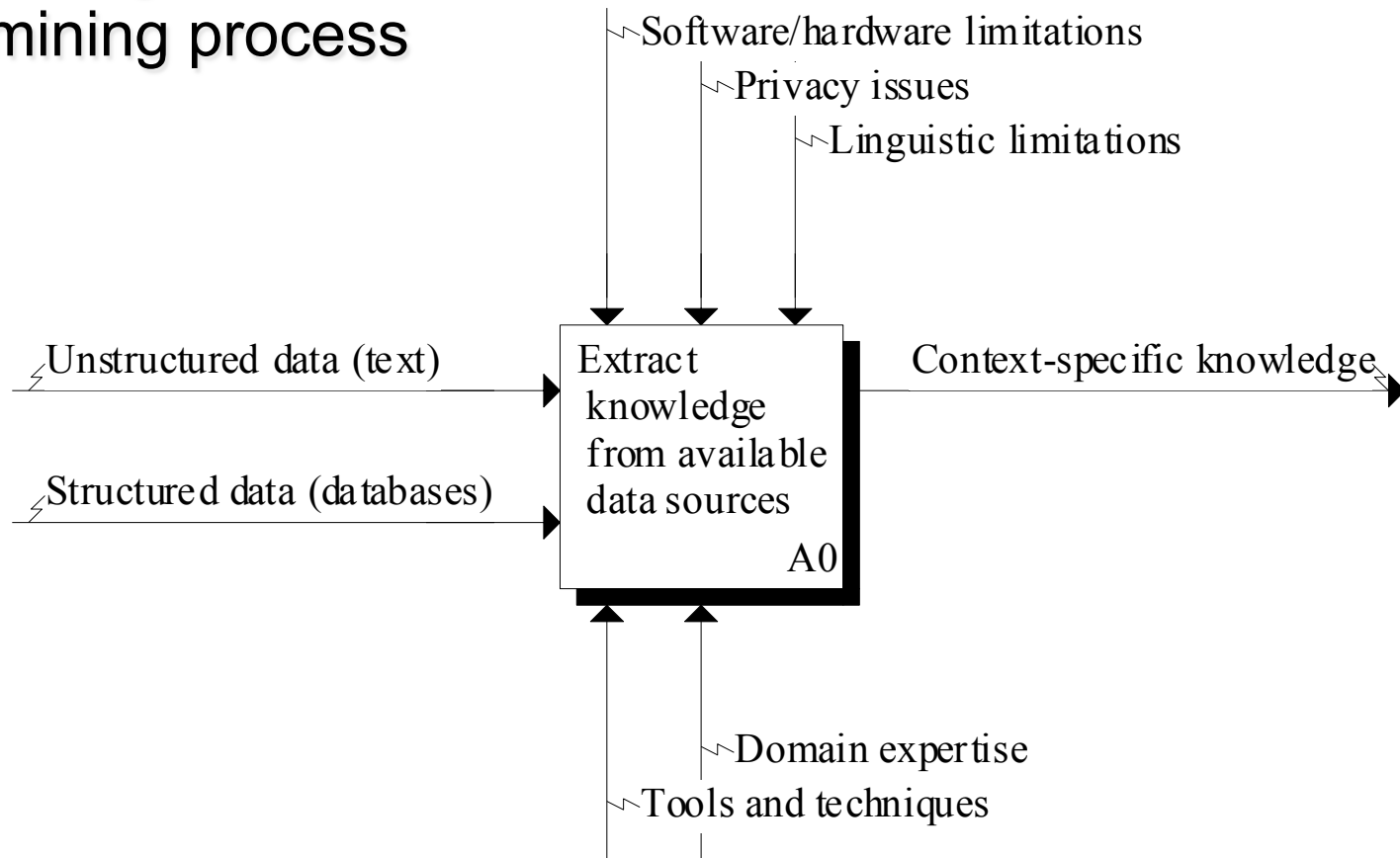
# Text Mining Applications

## (gene/protein interaction identification)

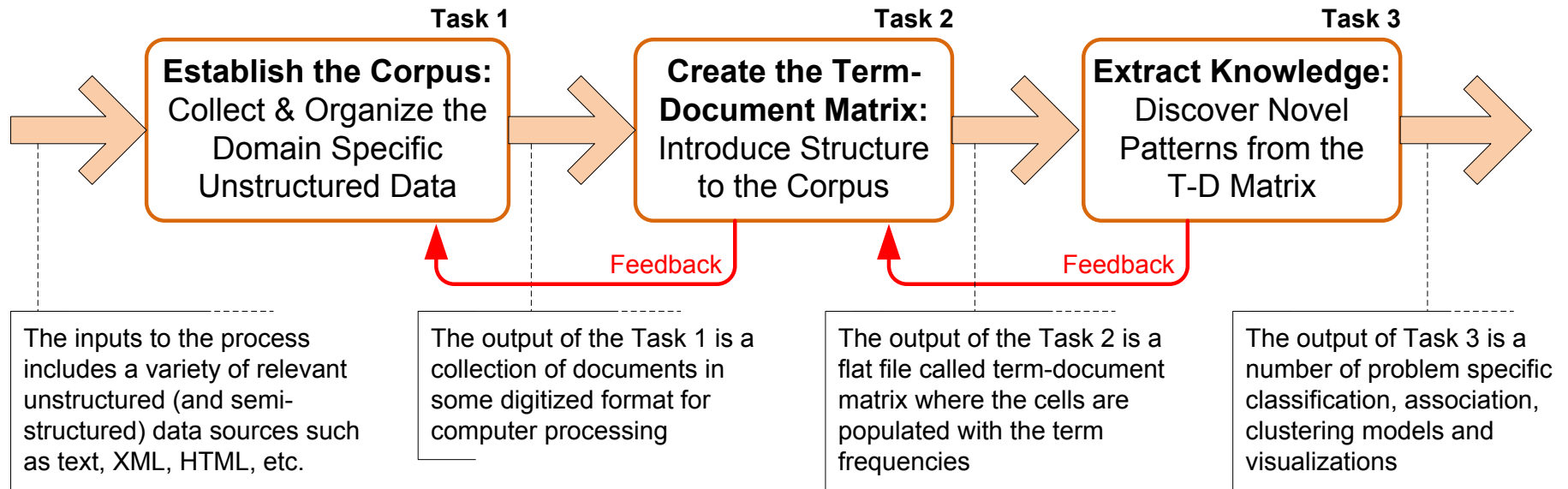


# Text Mining Process

Context diagram for the text mining process



# Text Mining Process



The three-step text mining process

# Text Mining Process

- **Step 1:** Establish the corpus
  - Collect all relevant unstructured data (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings...)
  - Digitize, standardize the collection (e.g., all in ASCII text files)
  - Place the collection in a common place (e.g., in a flat file, or in a directory as separate files)



# Text Mining Process

- **Step 2:** Create the Term-by-Document Matrix

Terms Documents	investment risk	project management	software engineering	development	SAP	...
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
...						

# Text Mining Process

- **Step 2:** Create the Term–by–Document Matrix (TDM), cont.
  - Should all terms be included?
    - Stop words, include words
    - Synonyms, homonyms
    - Stemming
  - What is the best representation of the indices (values in cells)?
    - Row counts; binary frequencies; log frequencies;
    - Inverse document frequency

# Text Mining Process

- **Step 2:** Create the Term-by-Document Matrix (TDM), cont.
  - TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    - Manual - a domain expert goes through it
    - Eliminate terms with very few occurrences in very few documents (?)
    - Transform the matrix using singular value decomposition (SVD)
    - SVD is similar to principle component analysis

# Text Mining Process

- **Step 3: Extract patterns/knowledge**
  - Classification (text categorization)
  - Clustering (natural groupings of text)
    - Improve search recall
    - Improve search precision
    - Scatter/gather
    - Query-specific clustering
  - Association
  - Trend Analysis (...)

# Text Mining Application

## (research trend identification in literature)

- Mining the published IS literature
  - MIS Quarterly (MISQ)
  - Journal of MIS (JMIS)
  - Information Systems Research (ISR)
  - Covers 12-year period (1994-2005)
  - 901 papers are included in the study
  - Only the paper abstracts are used
  - 9 clusters are generated for further analysis

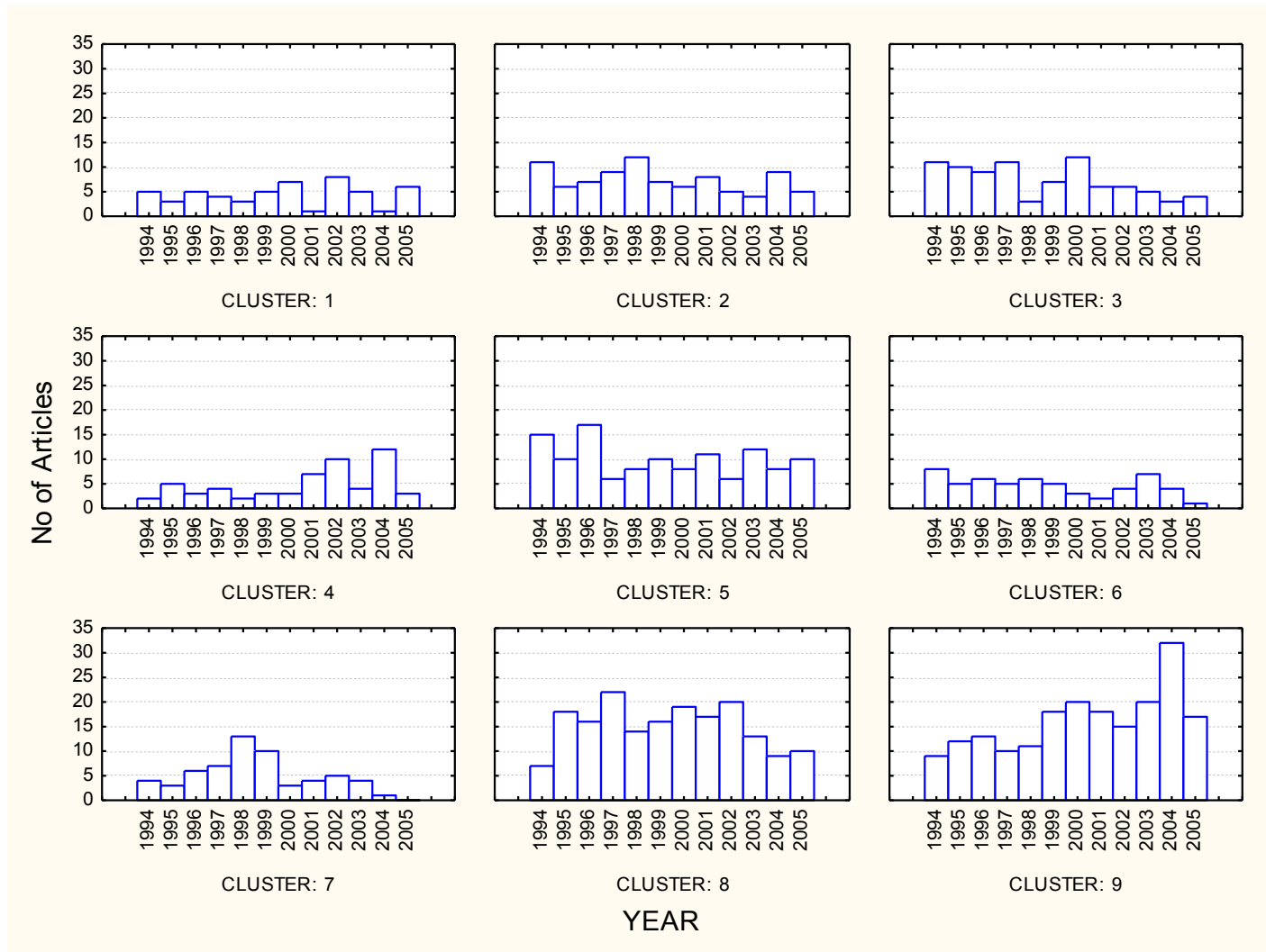
# Text Mining Application

## (research trend identification in literature)

Journal	Year	Author(s)	Title	Vol/No	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gosain and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145-187	knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing
ISR	1999	D. Robey and M. C. Boudreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications	2-Oct	167-185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment in quality and investment in self-promotion for information products	18/2	65-88	information products internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of
...	...	...	...	...	...	...	...

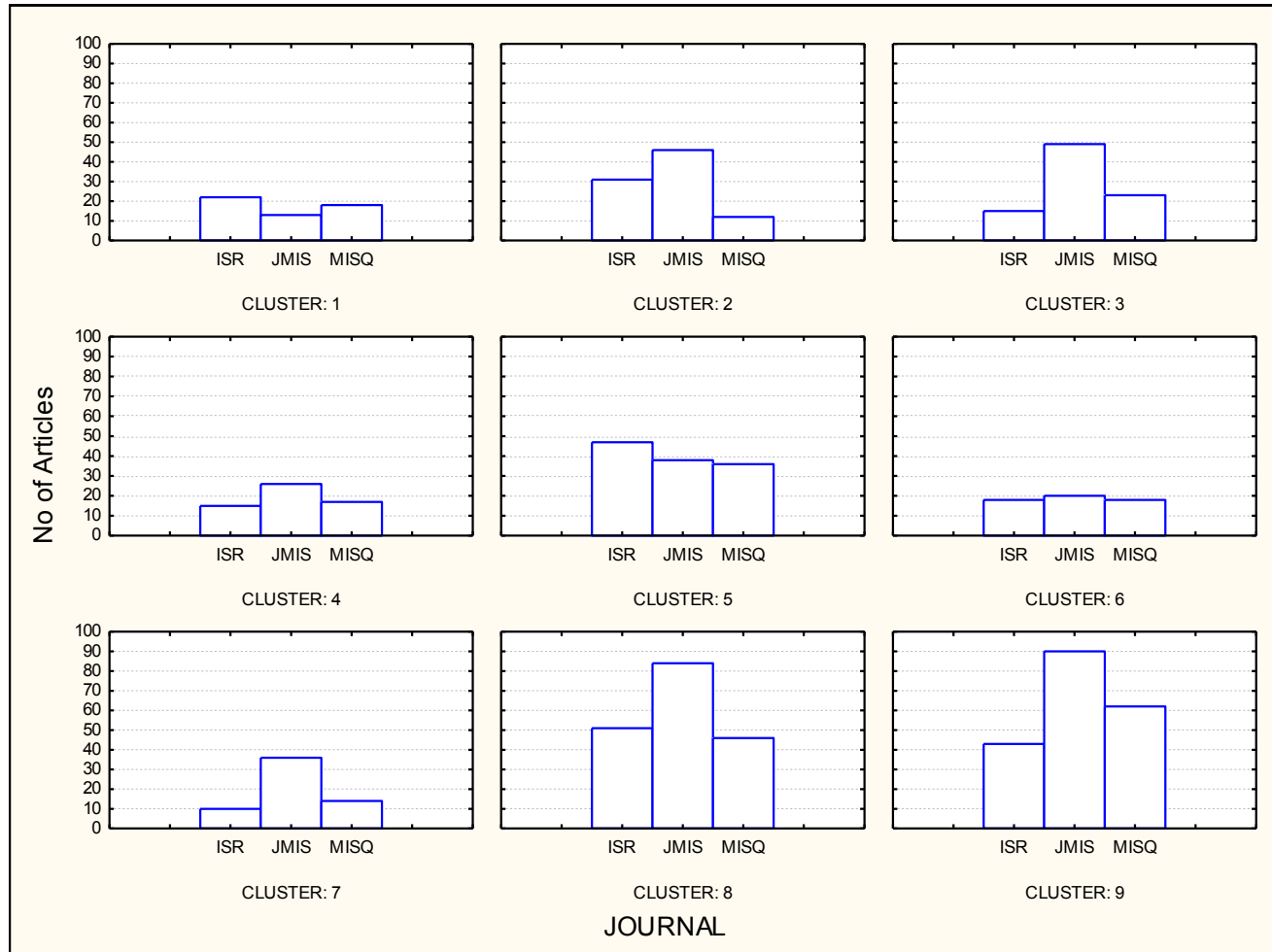
# Text Mining Application

## (research trend identification in literature)



# Text Mining Application

(research trend identification in literature)





# 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: [input field]

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

Text Result: [selected] Graphical Result

Test in rule-based model result is Negative  
Probability to be positive is 0.239% with confidence 99.54%  
43 matches:  
0 matches for product definitions:  
9 matches for feature definitions:  
No 0 (62-71): orion-Fees : annual fee

SAS® Text Analytics

SAS THE POWER TO KNOW.

0:16 / 4:50

SAS® Text Analytics Software Demo

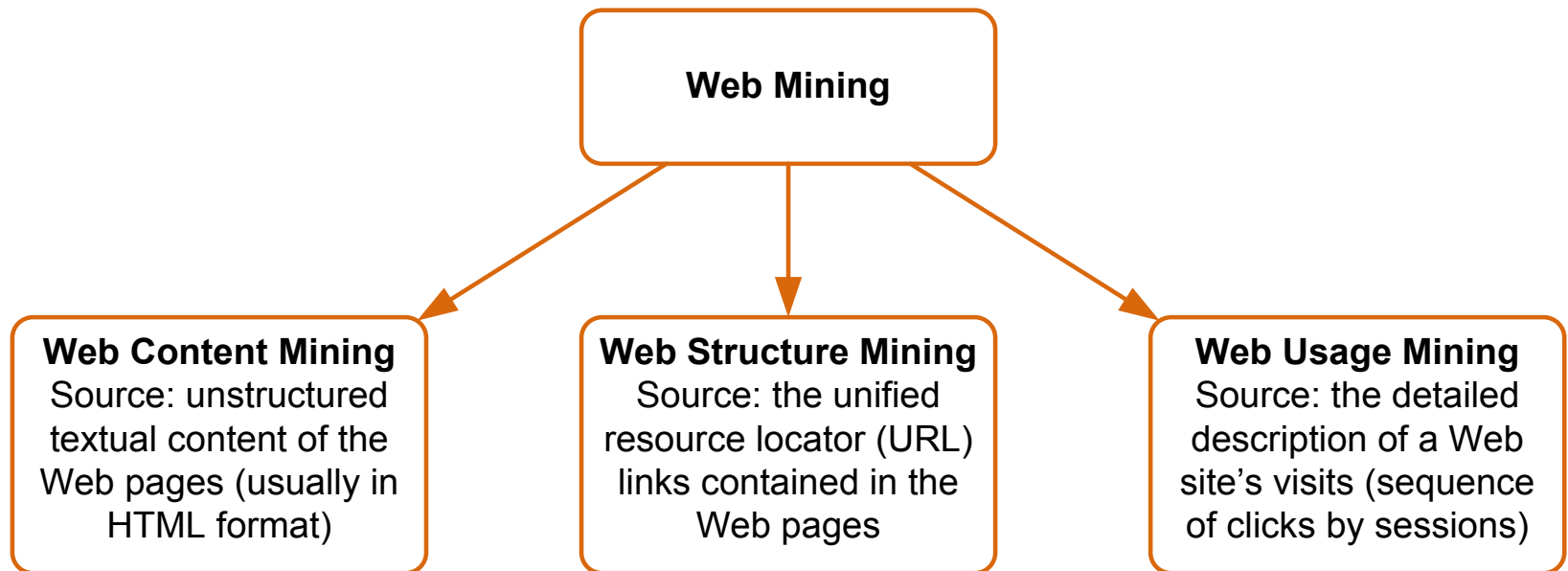


# Web Mining Overview

- 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

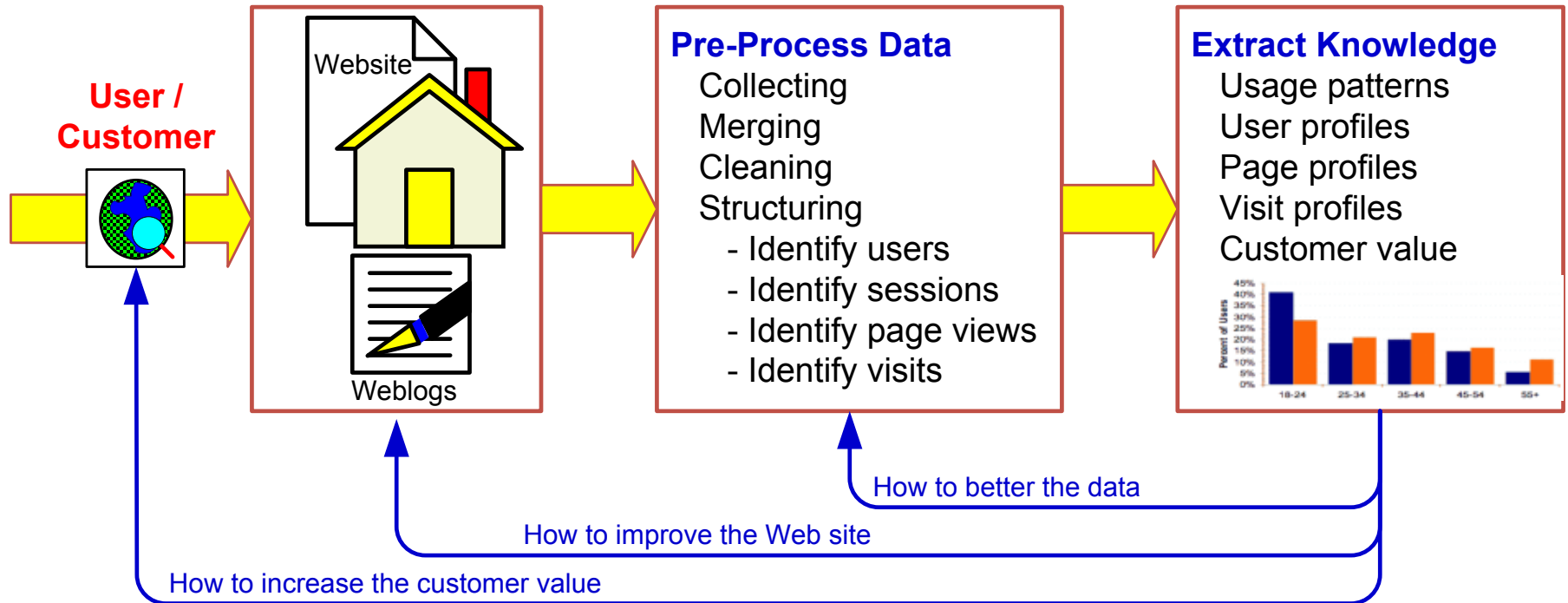
# Web Usage Mining

- 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...

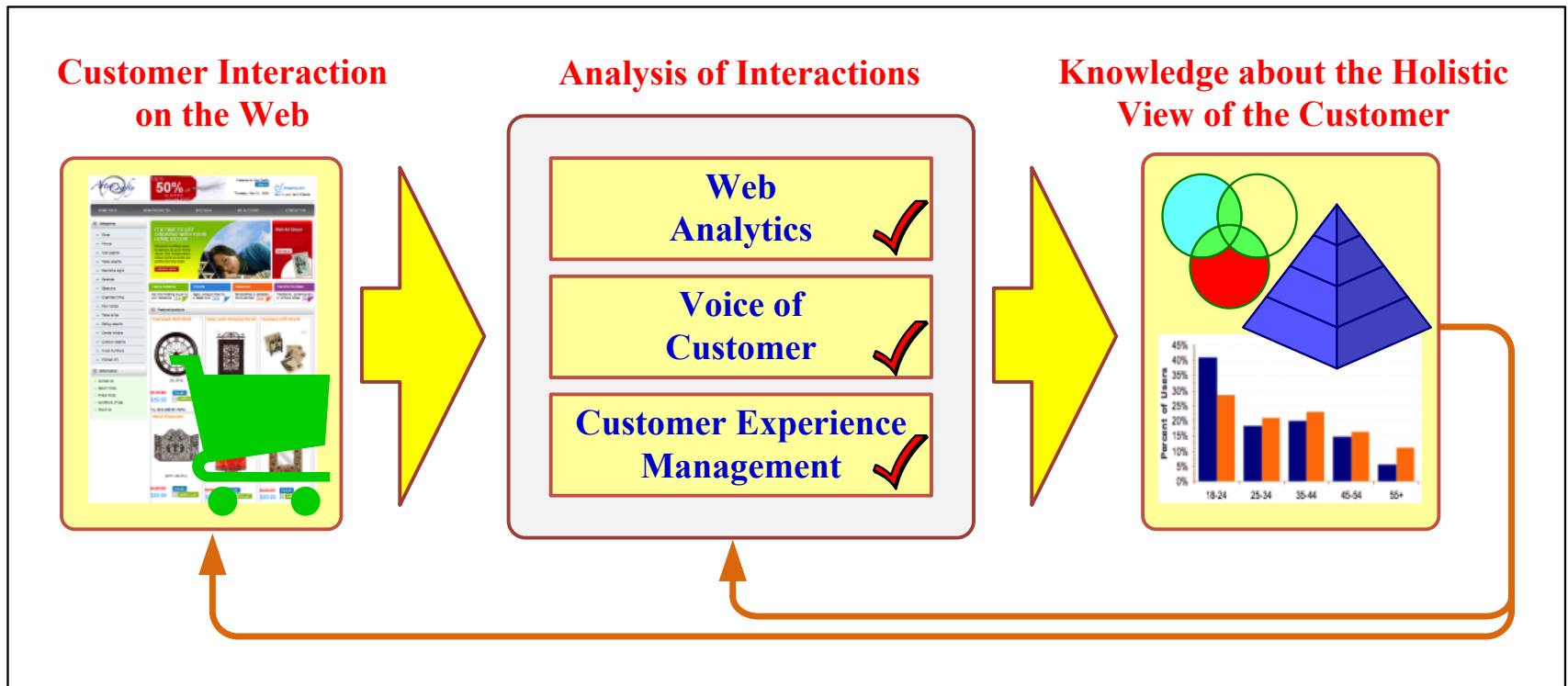
# Web Usage Mining (clickstream analysis)





# Web Mining Success Stories

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



# CKIP 中研院中文斷詞系統

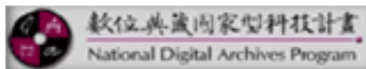
<http://ckipsvr.iis.sinica.edu.tw/>

## 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

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

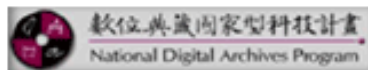
<http://ckipsvr.iis.sinica.edu.tw/>

## 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](#) | [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.

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(VE) 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: Visualise

Please enter your text here:

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

Submit

Clear

### Part-of-Speech:

	NP	NP	VBZ	JJ	IN	NP	.
1	Stanford	University	is	located	in	California	.
2	PRP	VBZ	DT	JJ	NN	.	
	It	is	a	great	university	.	

### Named Entity Recognition:

	Organization	Location
1	Stanford University	California
2		

### Coreference:

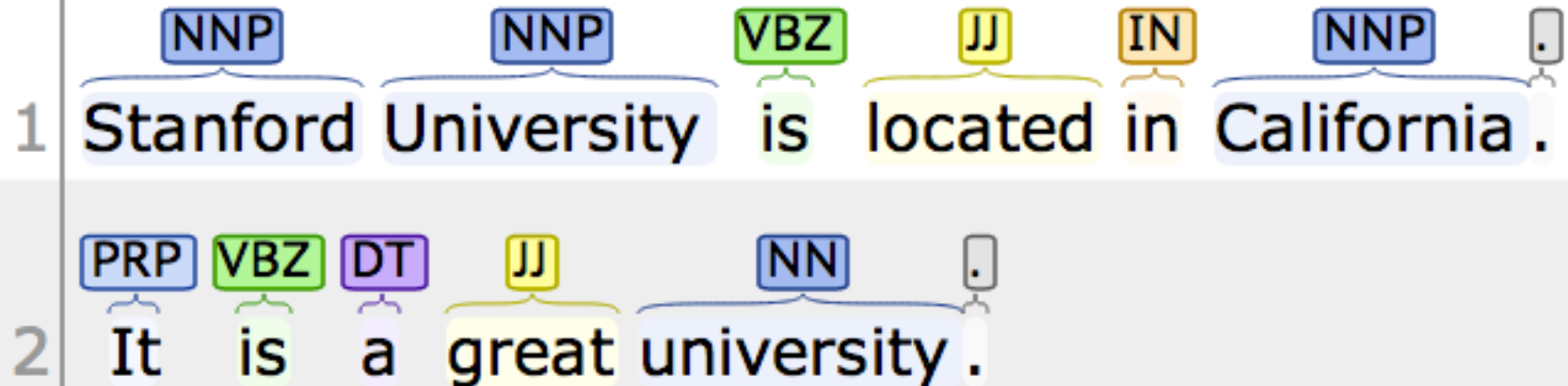
	Mention	Coref	
1	Stanford University		
2	M	Coref	Mention
	It		

# 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:

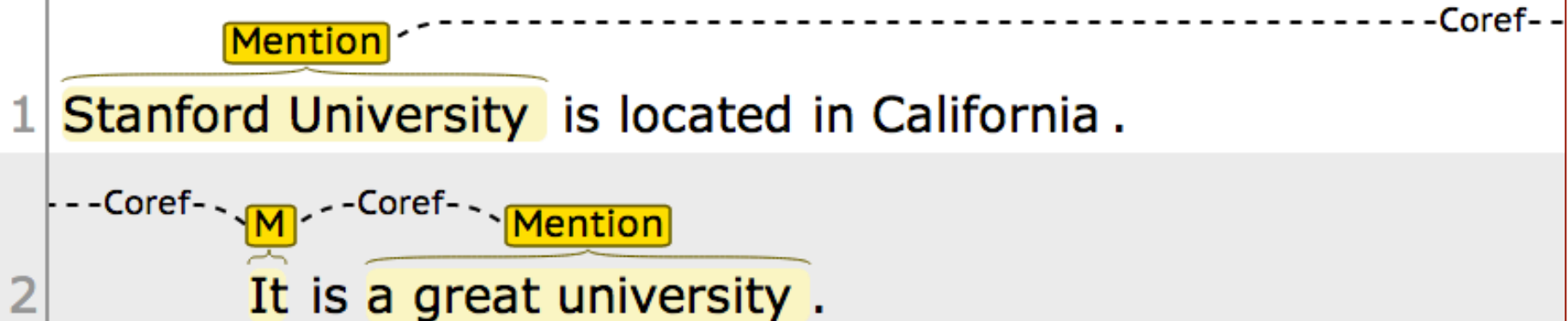
	<b>Organization</b>		<b>Location</b>
1	Stanford University	is located in	California .
2	It is a great university .		

# Stanford CoreNLP

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

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

## Coreference:

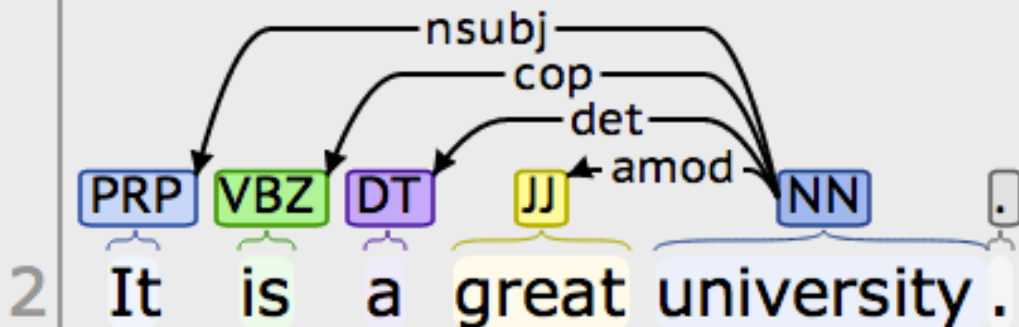
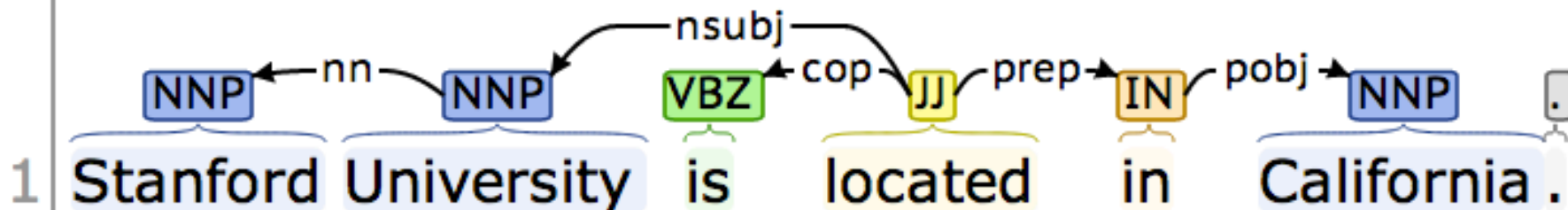


# Stanford CoreNLP

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

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

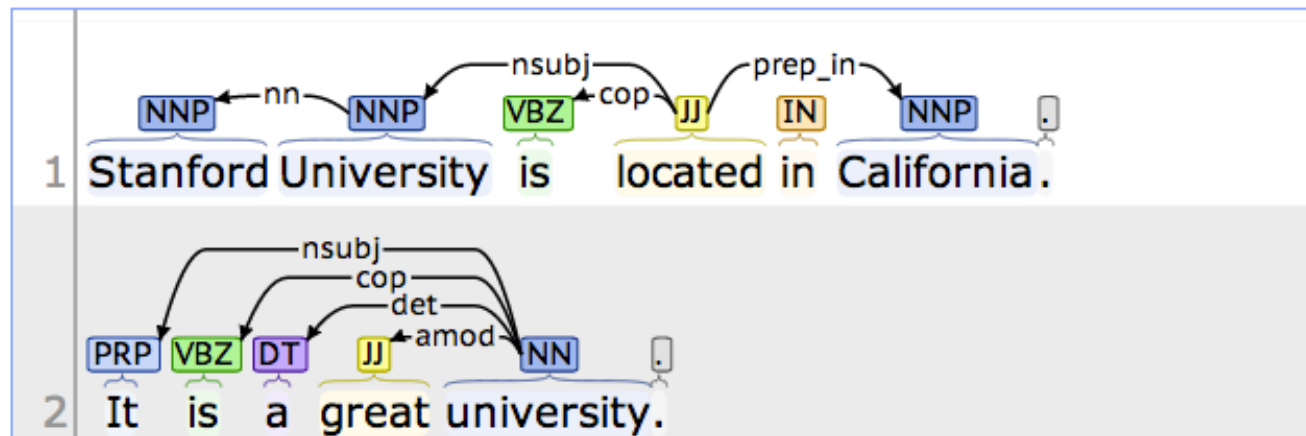
## Basic dependencies:



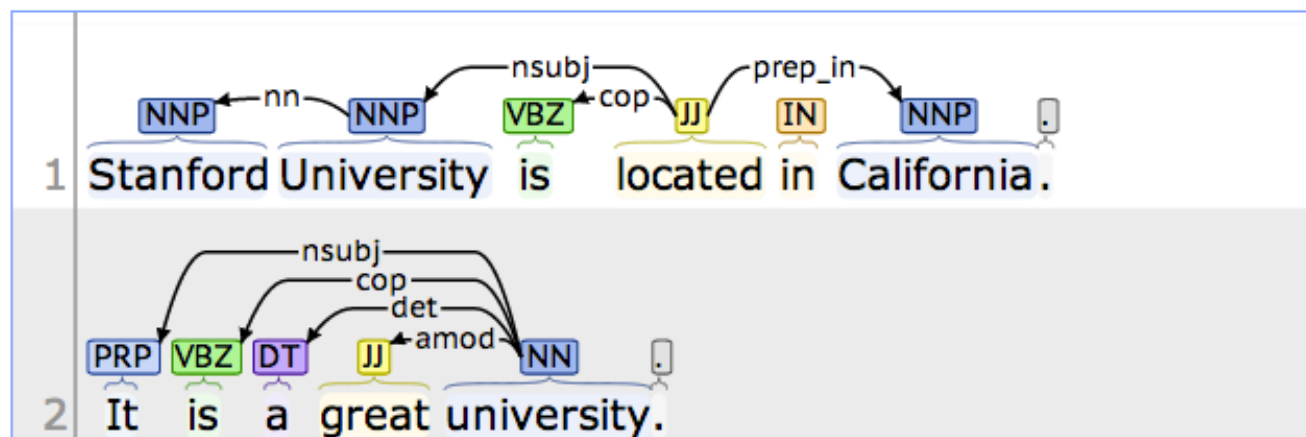
# Stanford CoreNLP

<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								
<b>Sentence #1</b>								
<i>Tokens</i>								
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
<i>Parse tree</i>								
(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*

<b>Id</b>	<b>Word</b>	<b>Lemma</b>	<b>Char begin</b>	<b>Char end</b>	<b>POS</b>	<b>NER</b>	<b>Normalized NER</b>	<b>Speaker</b>
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*

<b>Id</b>	<b>Word</b>	<b>Lemma</b>	<b>Char begin</b>	<b>Char end</b>	<b>POS</b>	<b>NER</b>	<b>Normalized NER</b>	<b>Speaker</b>
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: 

Please enter your text here:

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

```
<?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>
          ...
        </tokens>
      </sentence>
    </sentences>
  </document>
</root>
```

# 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 1.2k



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) — 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**

# 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">&apos;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">CNNMoney</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">&apos;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">,</wi> <wi num="6" entity="O">Gates</wi> <wi num="7" entity="O">has</wi> <wi num="8" entity="O">sold</wi> <wi num="9" entity="O">nearly</wi> <wi num="10" entity="O">8</wi> <wi num="11" entity="O">million</wi> <wi num="12" entity="O">shares</wi> <wi num="13" entity="O">of</wi> <wi num="14" entity="O">Microsoft</wi> <wi num="15" entity="O">over</wi> <wi num="16" entity="O">the</wi> <wi num="17" entity="O">past</wi> <wi num="18" entity="O">two</wi> <wi num="19" entity="O">days</wi> .</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 (CNNMoney) —

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

## Classifier: english.muc.7class.distsim.crf.ser.gz

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** (CNMoney) 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**

## Classifier: english.all.3class.distsim.crf.ser.gz

**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** (CNMoney) 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'/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

# Natural Language Processing with NLTK in Python

# NLTK (Natural Language Toolkit)

## NLTK 3.0 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

## Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

[Natural Language Processing with Python](#) provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at [http://nltk.org/book\\_1ed](http://nltk.org/book_1ed).)

## Some simple things you can do with NLTK

Tokenize and tag some text:

```
>>> import nltk
```

## TABLE OF CONTENTS

[NLTK News](#)

[Installing NLTK](#)

[Installing NLTK Data](#)

[Contribute to NLTK](#)

[FAQ](#)

[Wiki](#)

[API](#)

[HOWTO](#)

## SEARCH

Enter search terms or a module, class or function name.

# jupyter notebook



```
imyday — jupyter-notebook — 90x7
[iMydaytekiMacBook-Pro:~ imyday$ jupyter notebook
[I 05:00:21.870 NotebookApp] Serving notebooks from local directory: /Users/imyday
[I 05:00:21.870 NotebookApp] 0 active kernels
[I 05:00:21.870 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
[I 05:00:21.870 NotebookApp] Use Control-C to stop this server and shut down all kernels (
twice to skip confirmation).
█
```

# Jupyter New Terminal

The image shows a web browser window at localhost:8888/tree# displaying the JupyterLab interface. The 'Files' tab is active, showing a file tree with folders like Applications, Desktop, Development, Documents, and Downloads. In the top right of the file tree area, there are buttons for 'Upload', 'New', and a refresh icon. The 'New' button is open, showing a dropdown menu with options: 'Text File', 'Folder', 'Terminal', 'Notebooks', and 'Python 2'. The 'Terminal' option is highlighted with a red box. The 'New' button itself is also highlighted with a red box.

localhost:8888/tree#

jupyter

Files Running Clusters

Select items to perform actions on them.

Upload New

- Text File
- Folder
- Terminal
- Notebooks
- Python 2

Applications Desktop Development Documents Downloads



# conda list

localhost:8888/terminals/1

jupyter

```
bash-3.2$ conda list
# packages in environment at //anaconda:
#
abstract-rendering      0.5.1      np110py27_0
alabaster                0.7.7      py27_0
anaconda                 2.5.0      np110py27_0
anaconda-client         1.2.2      py27_0
appnope                 0.1.0      py27_0
appscript               1.0.1      py27_0
argcomplete             1.0.0      py27_1
astrophy                1.1.1      np110py27_0
babel                   2.2.0      py27_0
backports-abc           0.4        <pip>
backports.ssl-match-hostname 3.4.0.2    <pip>
backports_abc           0.4        py27_0
beautifulsoup4          4.4.1      py27_0
bitarray                0.8.1      py27_0
blaze                   0.9.0      <pip>
blaze-core              0.9.0      py27_0
bokeh                   0.11.0     py27_0
boto                    2.39.0     py27_0
bottleneck              1.0.0      np110py27_0
cdecimal                2.3        py27_0
cffi                    1.2.1      py27_0
clyent                  1.2.0      py27_0
colorama                0.3.6      py27_0
conda                   4.0.5      py27_0
conda-build             1.19.0     py27_0
conda-env               2.4.5      py27_0
```

# conda list

nlTK 3.1 py27\_0

localhost:8888/terminals/1

jupyter

```
nlTK 3.1 py27_0
node-webkit 0.10.1 0
nose 1.3.7 py27_0
notebook 4.1.0 py27_0
numba 0.23.1 np110py27_0
numexpr 2.4.6 np110py27_1
numpy 1.10.4 py27_0
odo 0.4.0 py27_0
openpyxl 2.3.2 py27_0
openssl 1.0.2g 0
pandas 0.18.0 np110py27_0
path.py 8.1.2 py27_1
patsy 0.4.0 np110py27_0
pep8 1.7.0 py27_0
pexpect 3.3 py27_0
pickleshare 0.5 py27_0
pillow 3.1.0 py27_0
pip 8.1.0 py27_0
ply 3.8 py27_0
psutil 3.4.2 py27_0
ptyprocess 0.5 py27_0
py 1.4.31 py27_0
pyasn1 0.1.9 py27_0
pyaudio 0.2.7 py27_0
pycosat 0.6.1 py27_0
pyparser 2.14 py27_0
pycrypto 2.6.1 py27_0
pycurl 7.19.5.3 py27_0
pyflakes 1.0.0 py27_0
```



# help( 'modules' )

In [2]: help('modules')



```
__builtin__
_Qt
_Res
_Scrap
_Snd
_TE
_Win
__builtin__
__future__
__abcoll
__ast
__bisect
__builtinSuites
__cffi_backend
__codecs
__codecs_cn
__codecs_hk
__codecs_iso2022
__codecs_jp
__codecs_kr
__codecs_tw
__cookiejar
copy
copy_reg
copyreg
crypt
cryptography
csv
ctypes
curl
curses
cyclcr
cython
cythonmagic
cytoolz
datashape
datetime
dateutil
dbhash
dbm
decimal
decorator
nltk
nntplib
nose
notebook
ntpath
nturl2path
numba
numbers
numexpr
numpy
odo
opcode
openpyxl
operator
optparse
os
os2emxpath
osax
pandas
parser
cabinetry
tarfile
telnetlib
tempfile
terminado
terminalcommand
termios
test_path
test_pycosat
tests
textwrap
this
thread
threading
time
timeit
tkColorChooser
tkCommonDialog
tkFileDialog
tkFont
tkMessageBox
```

# import nltk

localhost:8888/notebooks/TextMiningNLP.ipynb

 jupyter TextMiningNLP (autosaved) 

File Edit View Insert Cell Kernel Help

 Python 2 

          Code   CellToolbar

```
In [ ]: import n|
```

- nltk
- nntplib
- nose
- notebook
- ntpath
- nturl2pat
- numba
- numbers
- numexpr
- numpy

```
import nltk
nltk.download()
```

The screenshot shows a Jupyter Notebook interface with a code cell containing the following Python code:

```
In [*]: import nltk
nltk.download()
```

Below the code cell, the NLTK Downloader window is open, displaying a table of available collections. The table has four columns: Identifier, Name, Size, and Status. The data is as follows:

Identifier	Name	Size	Status
all	All packages	n/a	not installed
all-corpora	All the corpora	n/a	not installed
book	Everything used in the NLTK Book	n/a	not installed

At the bottom of the window, there are input fields for the Server Index and Download Directory, and buttons for Download and Refresh.

Server Index:

Download Directory:

Source: <http://www.nltk.org/>

```
import nltk
nltk.download()
```

NLTK Downloader

Collections Corpora Models All Packages


Identifier	Name	Size	Status
all	All packages	n/a	partial
all-corpora	All the corpora	n/a	partial
book	Everything used in the NLTK Book	n/a	partial

Cancel Refresh

Server Index:

Download Directory:

Downloading package u'cess\_esp'



```
import nltk
nltk.download()
```

```
In [*]: import nltk
nltk.download()
```

```
In [ ]:
```

NLTK Downloader

**Collections** Corpora Models All Packages


Identifier	Name	Size	Status
all	All packages	n/a	partial
all-corpora	All the corpora	n/a	partial
book	Everything used in the NLTK Book	n/a	installed

Cancel Refresh

Server Index:

Download Directory:

Downloading package u'panlex\_lite'



# nltk\_data



chunkers



corpora



grammars



help



models



stemmers



taggers



tokenizers



At eight o'clock on  
Thursday morning Arthur  
didn't feel very good.

```
[ ('At', 'IN'),  
  ('eight', 'CD'),  
  ("o'clock", 'NN'),  
  ('on', 'IN'),  
  ('Thursday', 'NNP'),  
  ('morning', 'NN'),  
  ('Arthur', 'NNP'),  
  ('did', 'VBD'),  
  ("n't", 'RB'),  
  ('feel', 'VB'),  
  ('very', 'RB'),  
  ('good', 'JJ'),  
  ('.', '.')] ]
```

```
import nltk
sentence = "At eight o'clock on Thursday morning Arthur didn't feel very good."
tokens = nltk.word_tokenize(sentence)
tokens
```

```
print(tokens)
```

```
In [1]: import nltk
sentence = "At eight o'clock on Thursday morning Arthur didn't feel very good."
tokens = nltk.word_tokenize(sentence)
tokens
```

```
Out[1]: ['At',
'eight',
'o'clock',
'on',
'Thursday',
'morning',
'Arthur',
'did',
'n't',
'feel',
'very',
'good',
'.']
```

```
In [2]: print(tokens)
```

```
['At', 'eight', 'o'clock', 'on', 'Thursday', 'morning', 'Arthur', 'did', 'n't', 'feel', 'ver
y', 'good', '.']
```

```
tagged = nltk.pos_tag(tokens)
tagged[0:6]
```

```
In [3]: tagged = nltk.pos_tag(tokens)
tagged[0:6]
```

```
Out[3]: [('At', 'IN'),
          ('eight', 'CD'),
          ("o'clock", 'NN'),
          ('on', 'IN'),
          ('Thursday', 'NNP'),
          ('morning', 'NN')]
```

# tagged

```
In [4]: tagged
```

```
Out[4]: [('At', 'IN'),  
         ('eight', 'CD'),  
         ("o'clock", 'NN'),  
         ('on', 'IN'),  
         ('Thursday', 'NNP'),  
         ('morning', 'NN'),  
         ('Arthur', 'NNP'),  
         ('did', 'VBD'),  
         ("n't", 'RB'),  
         ('feel', 'VB'),  
         ('very', 'RB'),  
         ('good', 'JJ'),  
         ('.', '.')] ]
```

# print (tagged)

In [5]: `print(tagged)`

```
[('At', 'IN'), ('eight', 'CD'), ('o'clock', 'NN'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'), ('Arthur', 'NNP'), ('did', 'VBD'), ('n't', 'RB'), ('feel', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')] ]
```

```
[('At', 'IN'), ('eight', 'CD'), ('o'clock', 'NN'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'), ('Arthur', 'NNP'), ('did', 'VBD'), ('n't', 'RB'), ('feel', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')] ]
```

**At eight o'clock on Thursday morning  
Arthur didn't feel very good.**

```
entities = nltk.chunk.ne_chunk(tagged)
entities
```

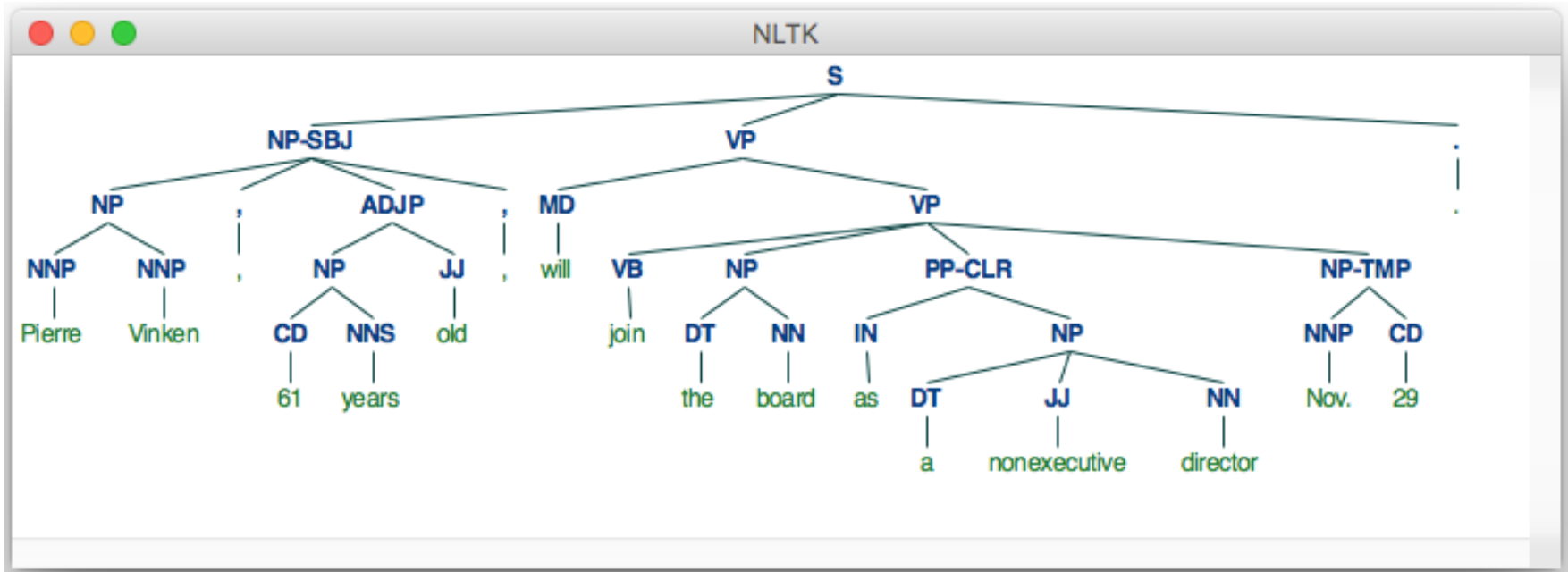
```
entities = nltk.chunk.ne_chunk(tagged)
entities
```

```
Tree('S', [('At', 'IN'), ('eight', 'CD'), ("o'clock", 'NN'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'), Tree('PERSON', [('Arthur', 'NNP'])], ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')])
```

```
Tree('S', [('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'), Tree('PERSON', [('Arthur', 'NNP'])], ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')])
```

```
from nltk.corpus import treebank
t = treebank.parsed_sents('wsj_0001.mrg')[0]
t.draw()
```

```
from nltk.corpus import treebank
t = treebank.parsed_sents('wsj_0001.mrg')[0]
t.draw()
```





# wsj\_0001.mrg



wsj\_0001.mrg



wsj\_0002.mrg



wsj\_0003.mrg



wsj\_0004.mrg



wsj\_0005.mrg



wsj\_0006.mrg



wsj\_0007.mrg



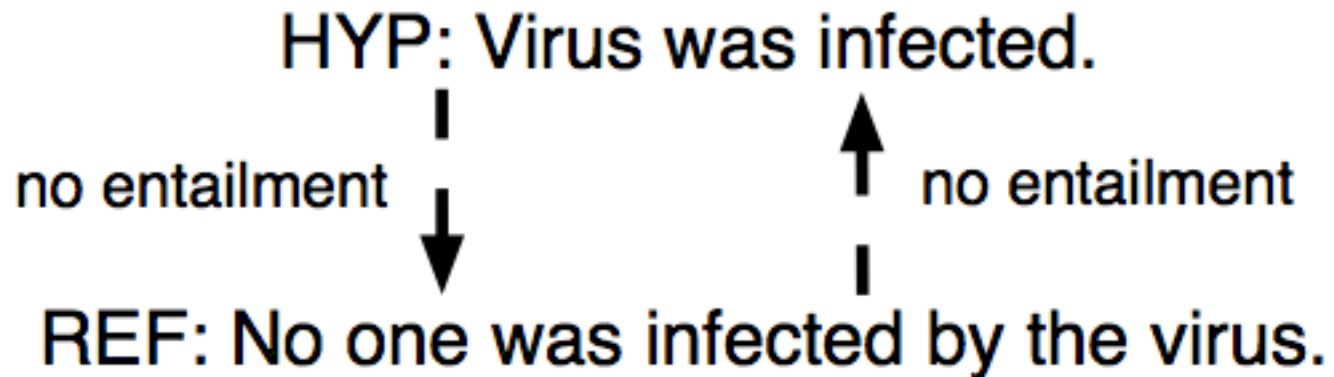
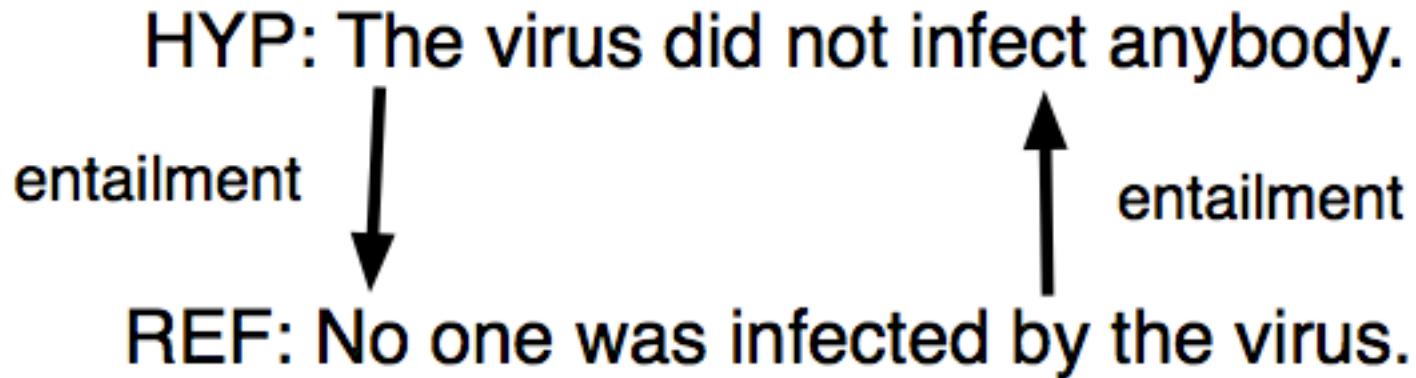
wsj\_0008.mrg

Macintosh HD > Users > imyday > nltk\_data > corpora > treebank > combined > wsj\_0001.mrg

# wsj\_0001.mrg

```
wsj_0001.mrg  x
1
2 ( (S
3   (NP-SBJ
4     (NP (NNP Pierre) (NNP Vinken) )
5     (, ,)
6     (ADJP
7       (NP (CD 61) (NNS years) )
8       (JJ old) )
9     (, ,) )
10  (VP (MD will)
11     (VP (VB join)
12       (NP (DT the) (NN board) )
13       (PP-CLR (IN as)
14         (NP (DT a) (JJ nonexecutive) (NN director) ))
15       (NP-TMP (NNP Nov.) (CD 29) )))
16  (. .) ))
17 ( (S
18   (NP-SBJ (NNP Mr.) (NNP Vinken) )
19   (VP (VBZ is)
20     (NP-PRD
21       (NP (NN chairman) )
22       (PP (IN of)
23         (NP
24           (NP (NNP Elsevier) (NNP N.V.) )
25           (, ,)
26           (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
27   (. .) ))
28
```

# 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)
- Natural Language Processing with NLTK in Python

# References

- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.
- Steven Bird, Ewan Klein and Edward Loper, Natural Language Processing with Python, 2009, O'Reilly Media, <http://www.nltk.org/book/> , [http://www.nltk.org/book\\_1ed/](http://www.nltk.org/book_1ed/)
- Nitin Hardeniya, NLTK Essentials, 2015, Packt Publishing
- 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/>
- Christopher D. Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, 1999, The MIT Press
- Text Mining, [http://en.wikipedia.org/wiki/Text\\_mining](http://en.wikipedia.org/wiki/Text_mining)