

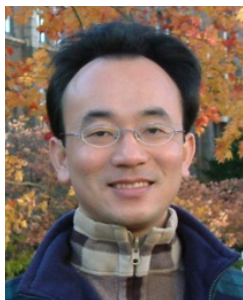


Text Mining and Natural Language Processing (文字探勘與自然語言處理)

Time: 2017/01/23 (Mon) (14:00-17:00)

Place: 國立臺北護理健康大學 城區部 (台北市內江街89號) C302

Host: 祝國忠 院長 (健康科技學院院長)



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/>

2017-01-23





戴敏育 博士 (Min-Yuh Day, Ph.D.)

淡江大學資管系專任助理教授

中央研究院資訊科學研究所訪問學人

國立台灣大學資訊管理博士

Publications Co-Chairs, IEEE/ACM International Conference on
Advances in Social Networks Analysis and Mining (ASONAM 2013-)

Program Co-Chair, IEEE International Workshop on
Empirical Methods for Recognizing Inference in Text (IEEE EM-RITE 2012-)

Workshop Chair, The IEEE International Conference on
Information Reuse and Integration (IEEE IRI)



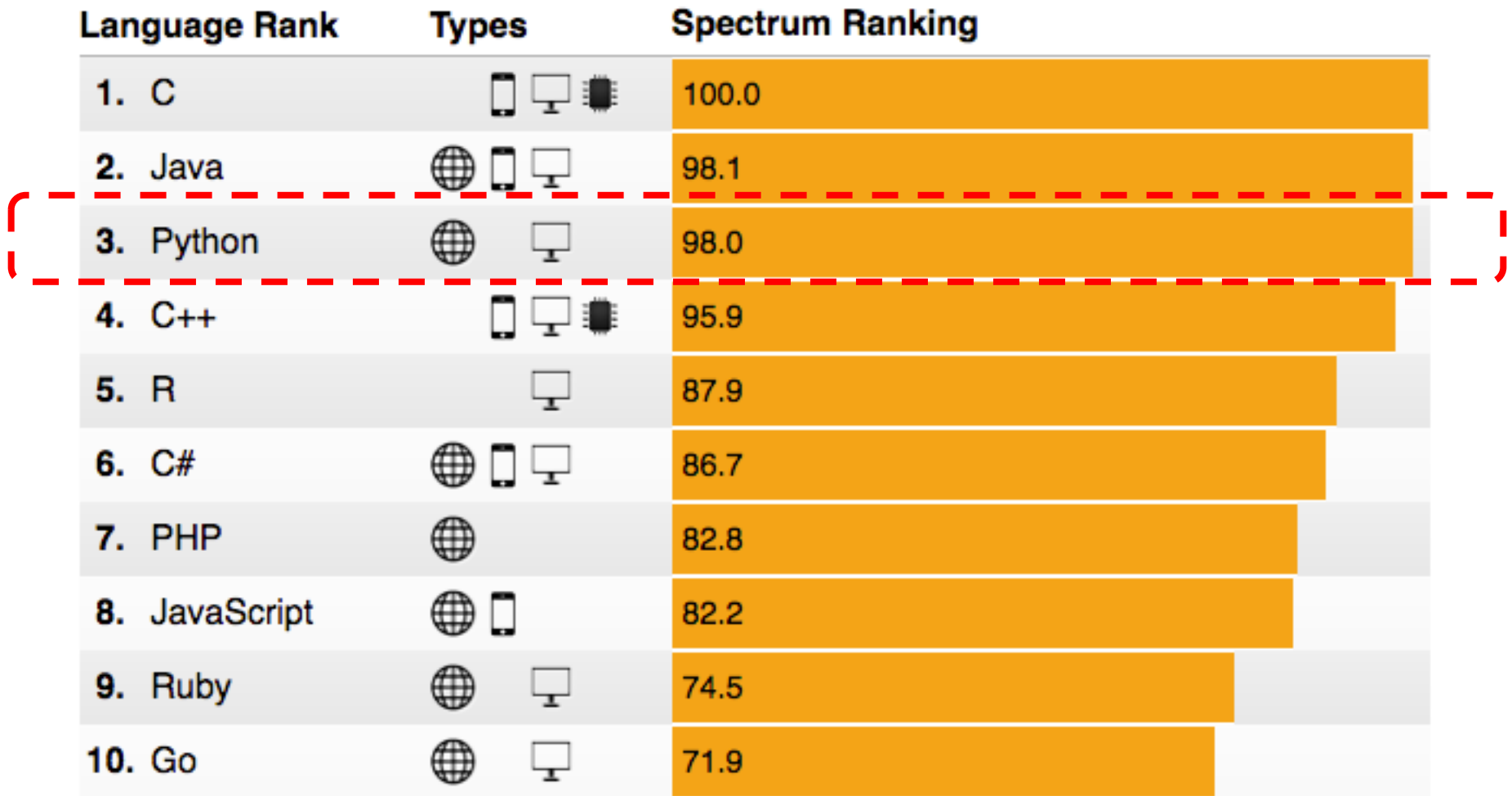
Text Mining (TM)

Natural Language Processing (NLP)

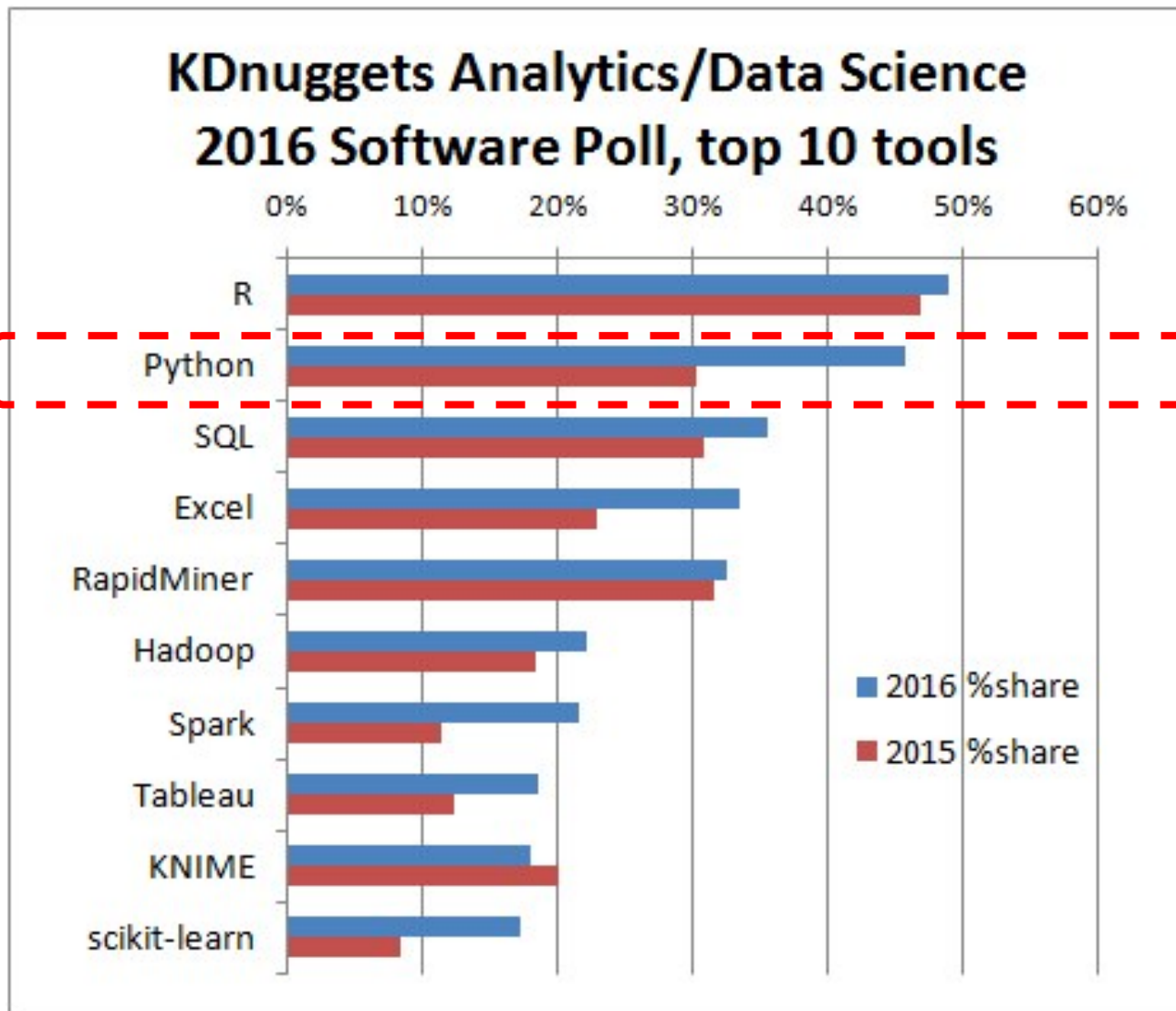
Outline

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

Python for Big Data Analytics



Python: Analytics and Data Science Software



Python

Python

PSF

Docs

PyPI

Jobs

Community



GO

Socialize

Sign In

About

Downloads

Documentation

Community

Success Stories

News

Events

```
# Python 3: List comprehensions
>>> fruits = ['Banana', 'Apple', 'Lime']
>>> loud_fruits = [fruit.upper() for fruit in fruits]
>>> print(loud_fruits)
['BANANA', 'APPLE', 'LIME']

# List and the enumerate function
>>> list(enumerate(fruits))
[(0, 'Banana'), (1, 'Apple'), (2, 'Lime')]
```



Compound Data Types

Lists (known as arrays in other languages) are one of the compound data types that Python understands. Lists can be indexed, sliced and manipulated with other built-in functions. [More about lists in Python 3](#)

1

2

3

4

5

Python is a programming language that lets you work quickly and integrate systems more effectively. [>>> Learn More](#)

Get Started

Download

Docs

Jobs

<https://www.python.org/>

Python is an
interpreted,
object-oriented,
high-level
programming language
with
dynamic semantics.

Anaconda

CONTINUUM[®]
ANALYTICS

ANACONDA

COMMUNITY

SERVICES

SOLUTIONS

ABOUT

RESOURCES

LOG IN SUPPORT CONTACT

ANACONDA GIVES
SUPERPOWERS TO
PEOPLE WHO CHANGE
THE WORLD



ANACONDA[®]

Modern open source analytics platform powered
by Python

DOWNLOAD FOR FREE

ANACONDA NOW AVAILABLE FOR CLOUDERA CDH

WHY YOU'LL LOVE ANACONDA

Making it easy to install, intuitive to discover, quick to analyze, simple to collaborate, and accessible to all.

**Committed to Open
Source. Now and
forever.**

**Tested and certified
packages to cover
your back.**

**Explore and visualize
complex data easily.**

**All the analytics you
ever wanted and
more.**

<https://www.continuum.io/>

import nltk nltk.download()

The screenshot shows a Jupyter Notebook interface with a browser window at localhost:8888/notebooks/TextMiningNLP.ipynb. The notebook has a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar. The code cell contains the following Python code:

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

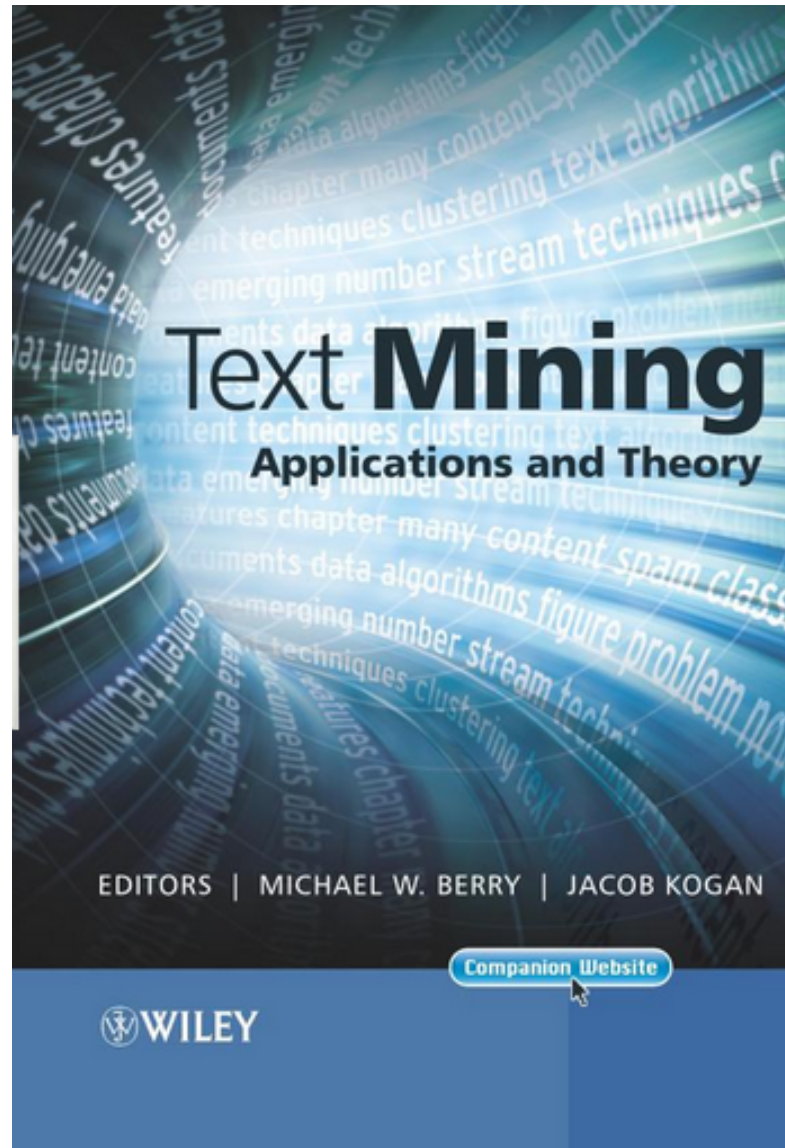
The NLTK Downloader window is open, displaying a table of available packages. The table has columns for Identifier, Name, Size, and Status. The 'All Packages' tab is selected, showing three entries: 'all', 'all-corpora', and 'book'. All three are listed as 'not installed'.

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

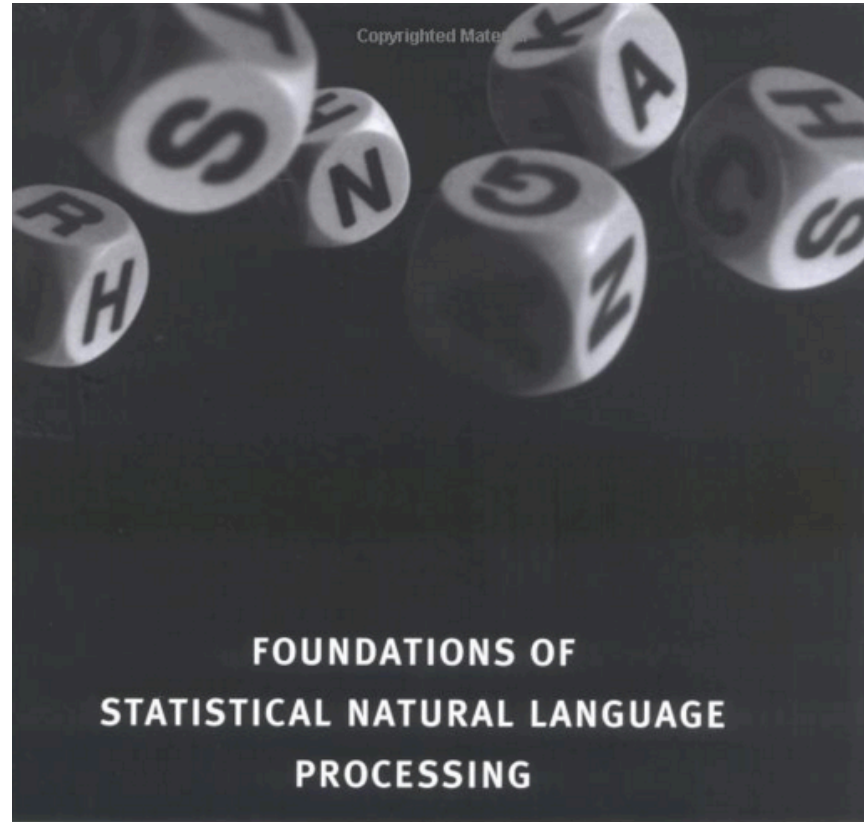
At the bottom of the window, there are fields for 'Server Index' (http://www.nltk.org/nltk_data/) and 'Download Directory' (/Users/imyday/nltk_data), along with 'Download' and 'Refresh' buttons.

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

Text Mining

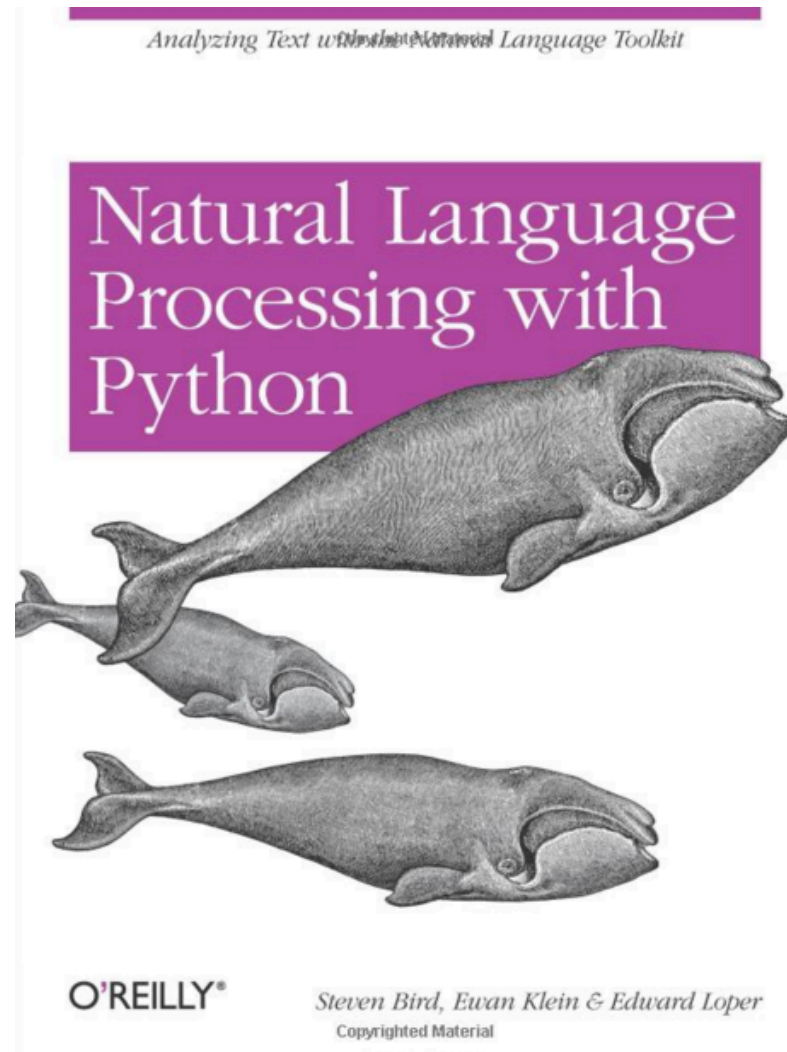


Christopher D. Manning and Hinrich Schütze (1999),
**Foundations of
Statistical Natural Language Processing,**
The MIT Press



**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/



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/. 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](#)

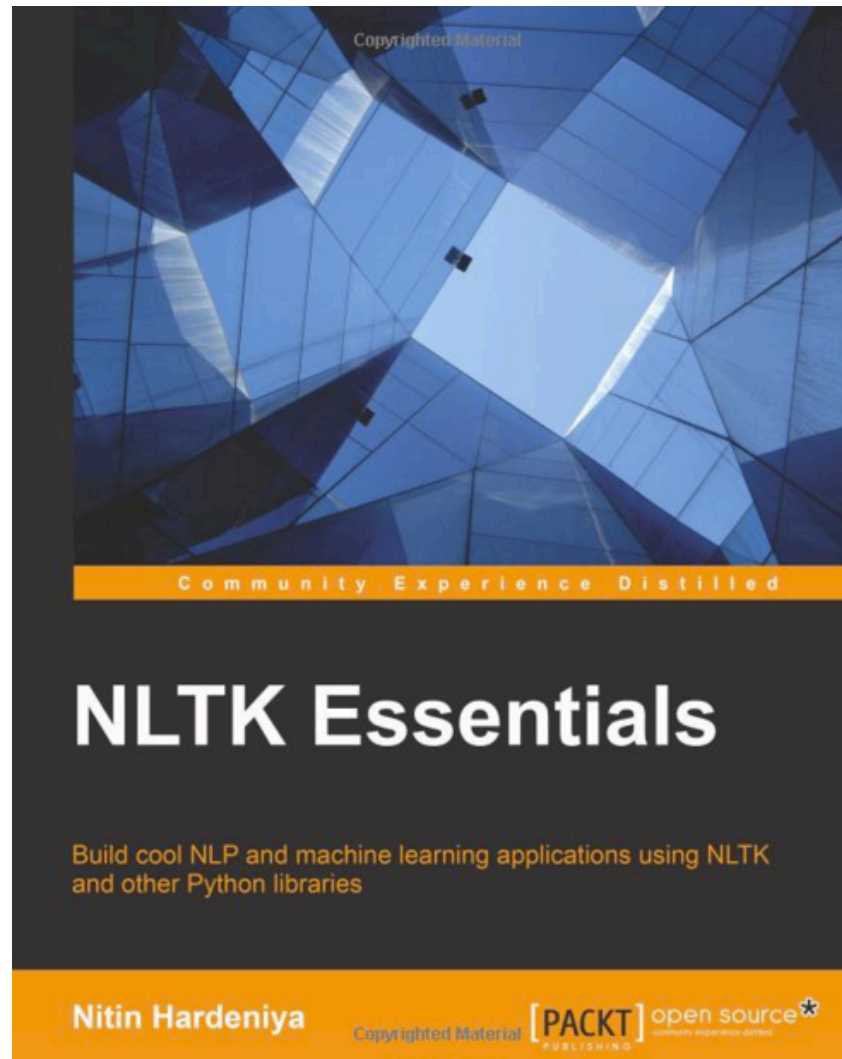
[Bibliography](#)

[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



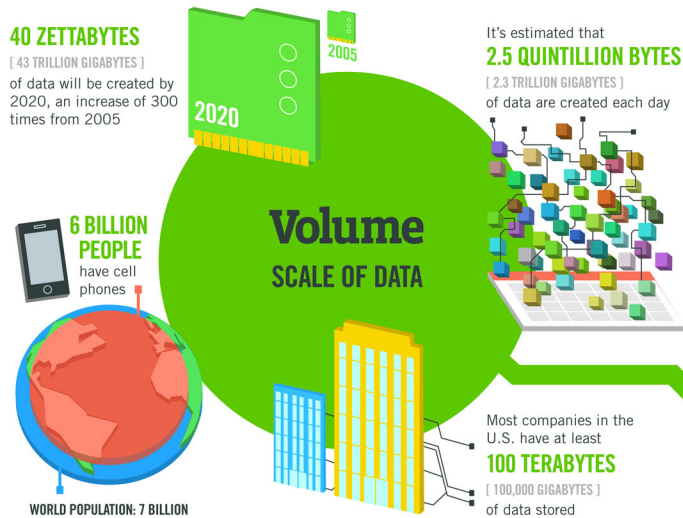
<http://www.amazon.com/NLTK-Essentials-Nitin-Hardeniya/dp/1784396907>

Text Mining **(text data mining)**

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

Big Data Analytics

Big Data 4 V



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT
are shared on Facebook every month



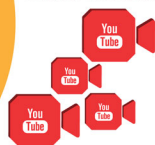
Variety

DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users



The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
— almost 2.5 connections per person on earth



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



27% OF RESPONDENTS

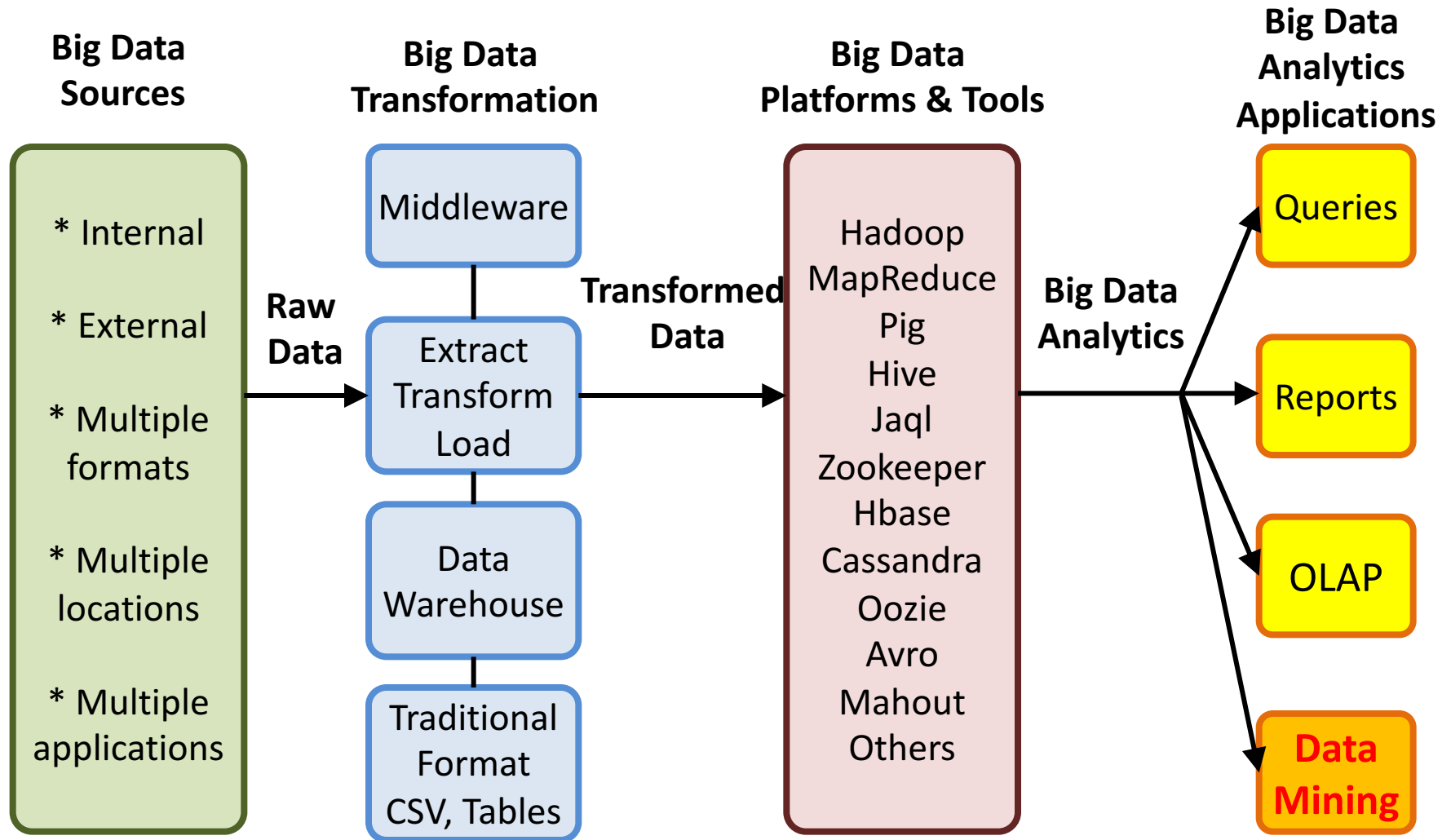
in one survey were unsure of how much of their data was inaccurate

Veracity

UNCERTAINTY OF DATA

Value

Architecture of Big Data Analytics

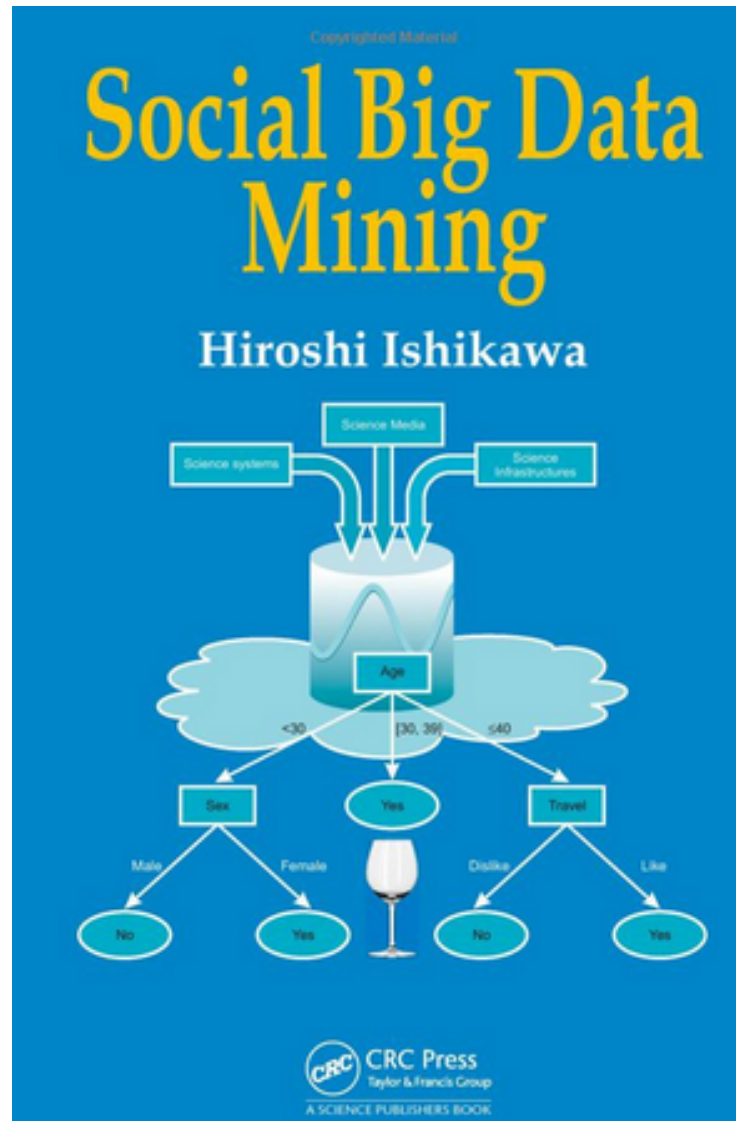


Architecture of Big Data Analytics



Social Big Data Mining

(Hiroshi Ishikawa, 2015)

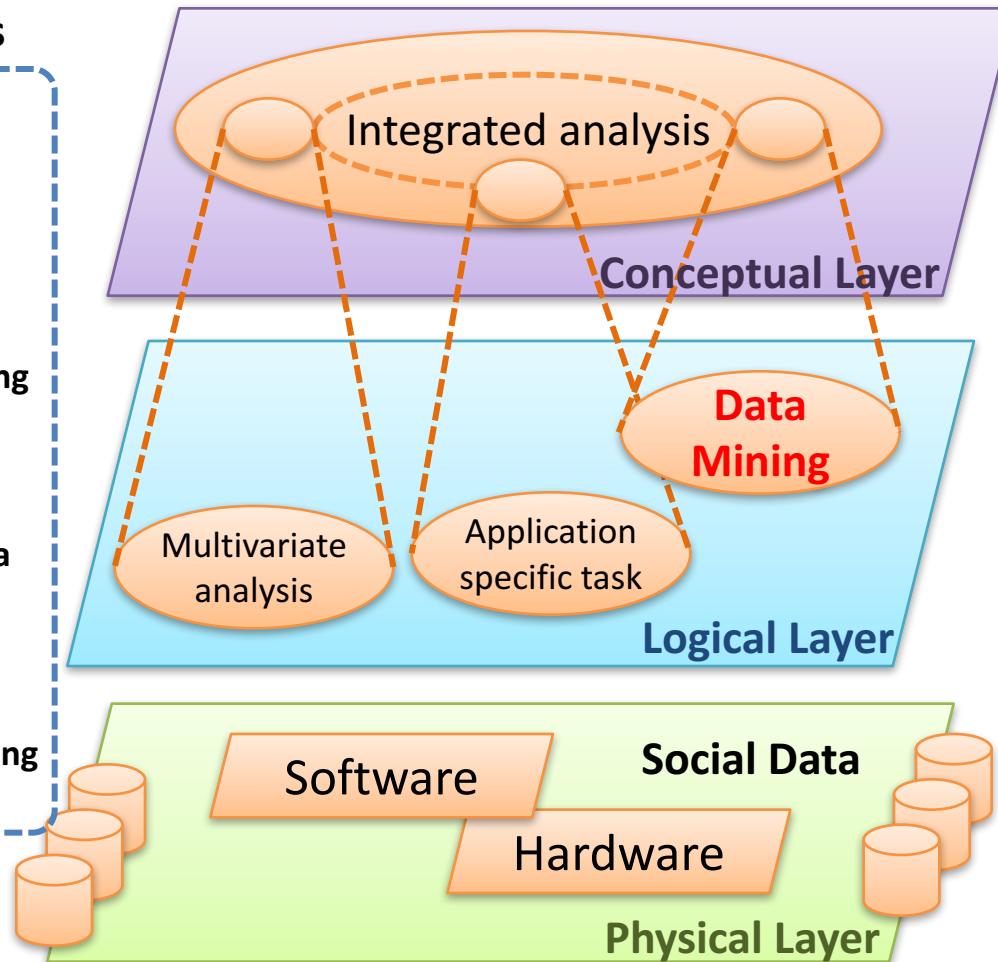


Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Enabling Technologies

- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing



Analysts

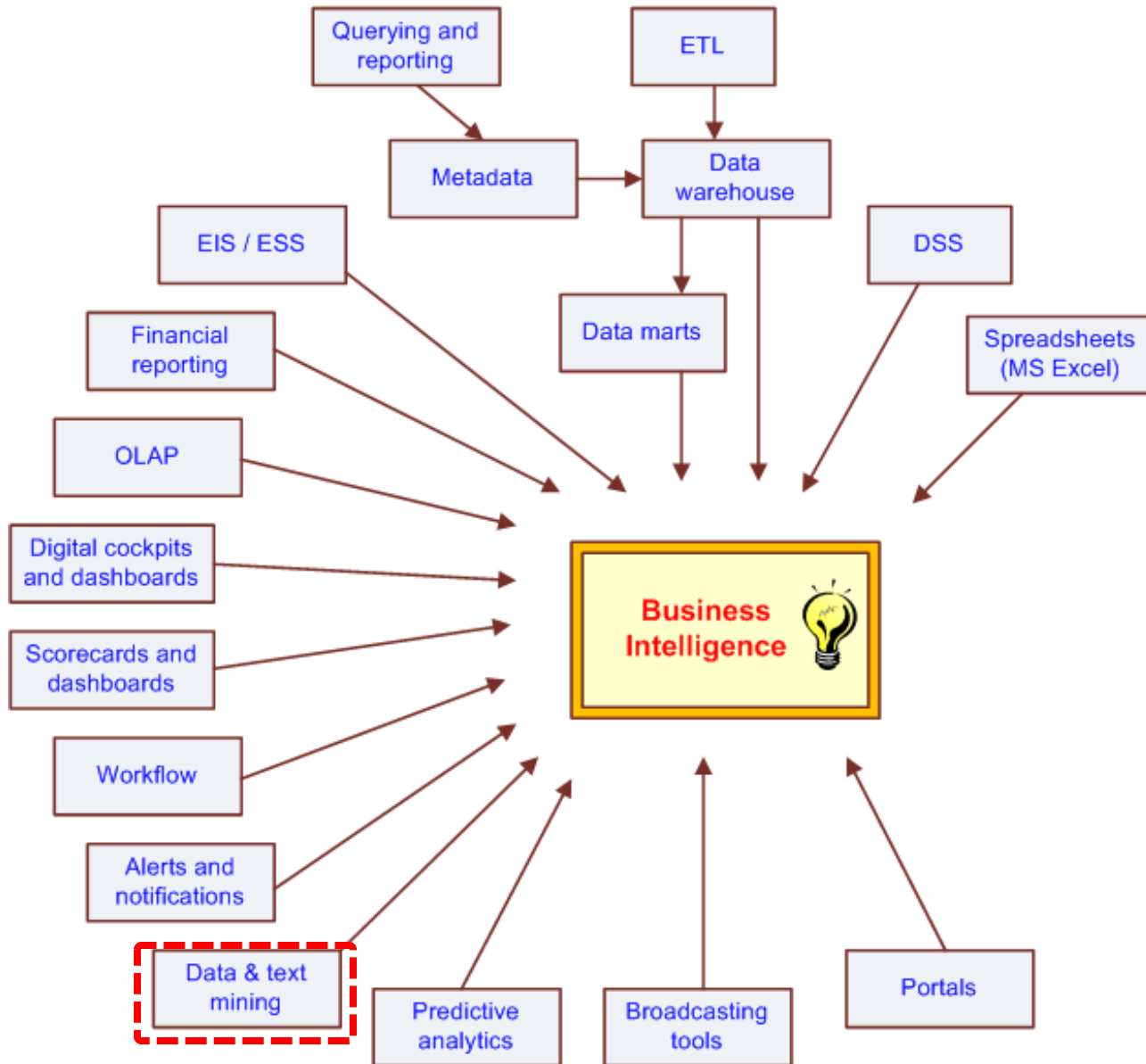
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Deep Learning

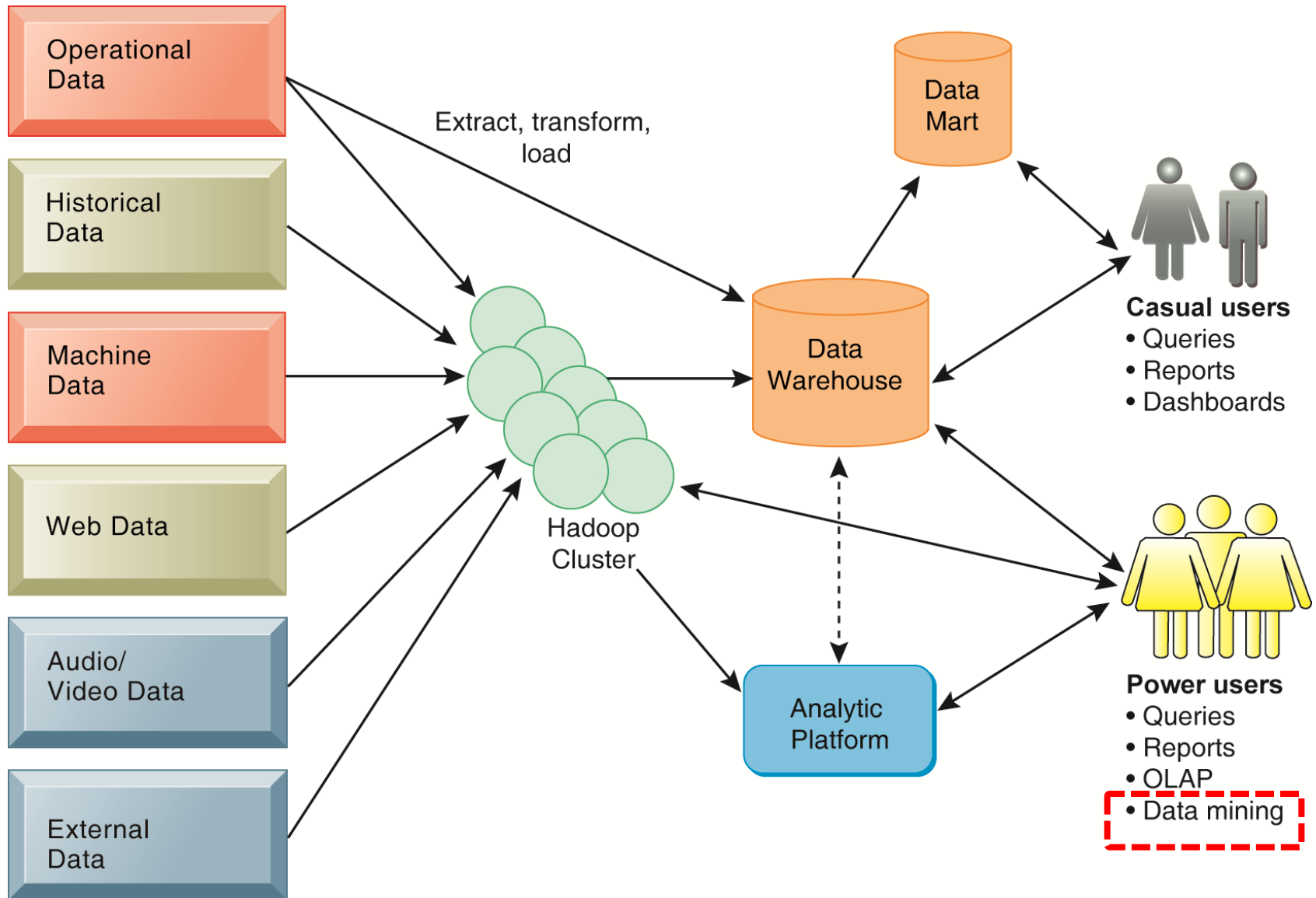
Intelligence from Big Data



The Evolution of BI Capabilities



Business Intelligence (BI) Infrastructure



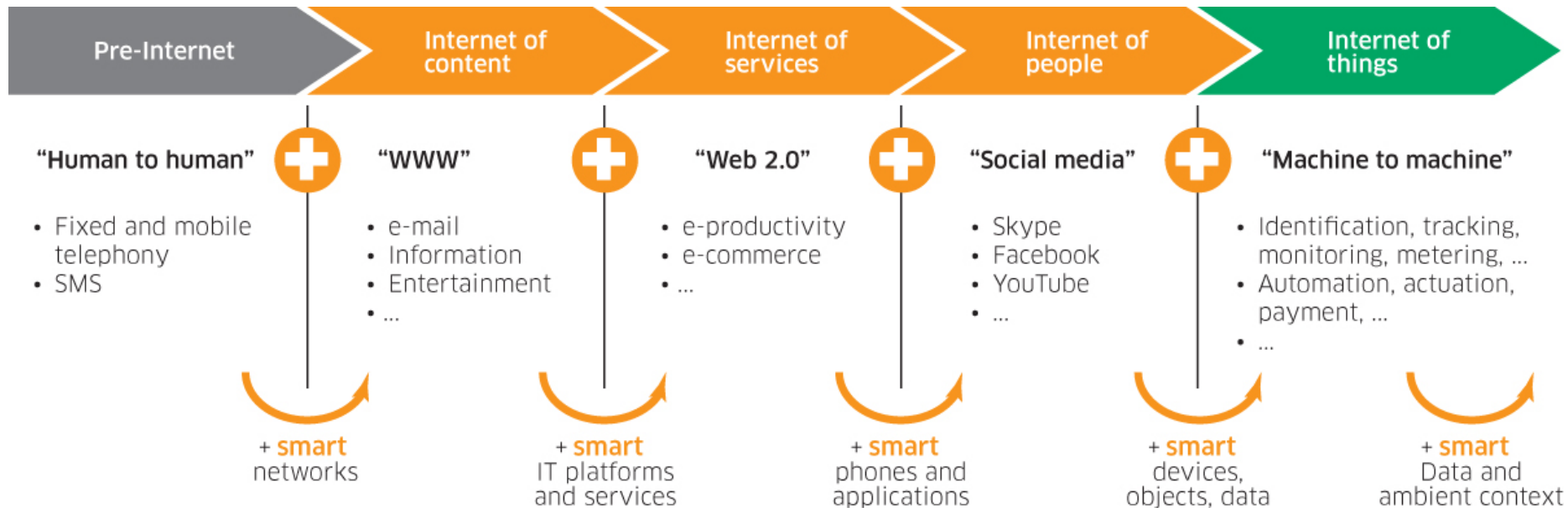
Social Media



Internet Evolution

Internet of People (IoP): Social Media

Internet of Things (IoT): Machine to Machine



Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015

<http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/>

Emotions



Love

Anger

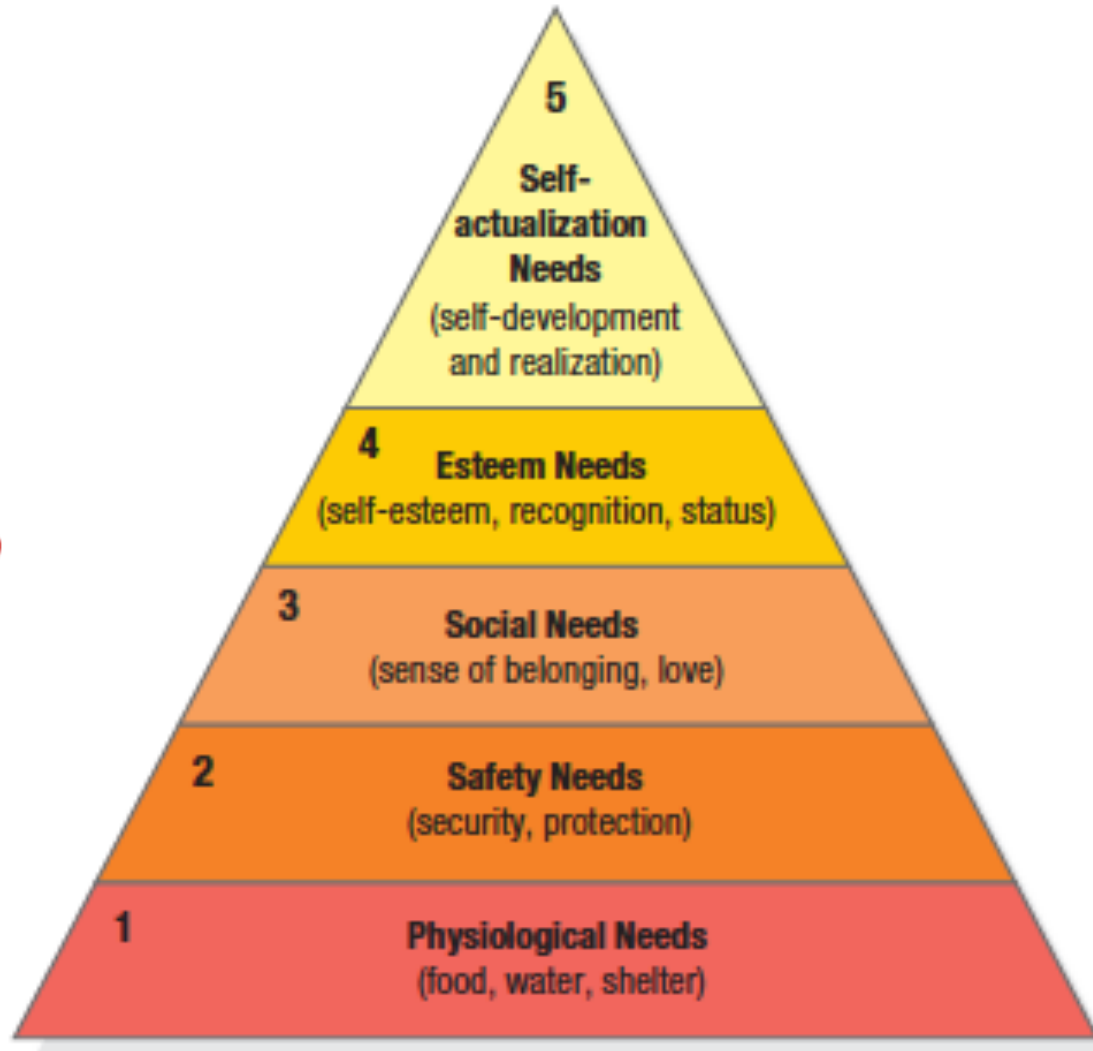
Joy

Sadness

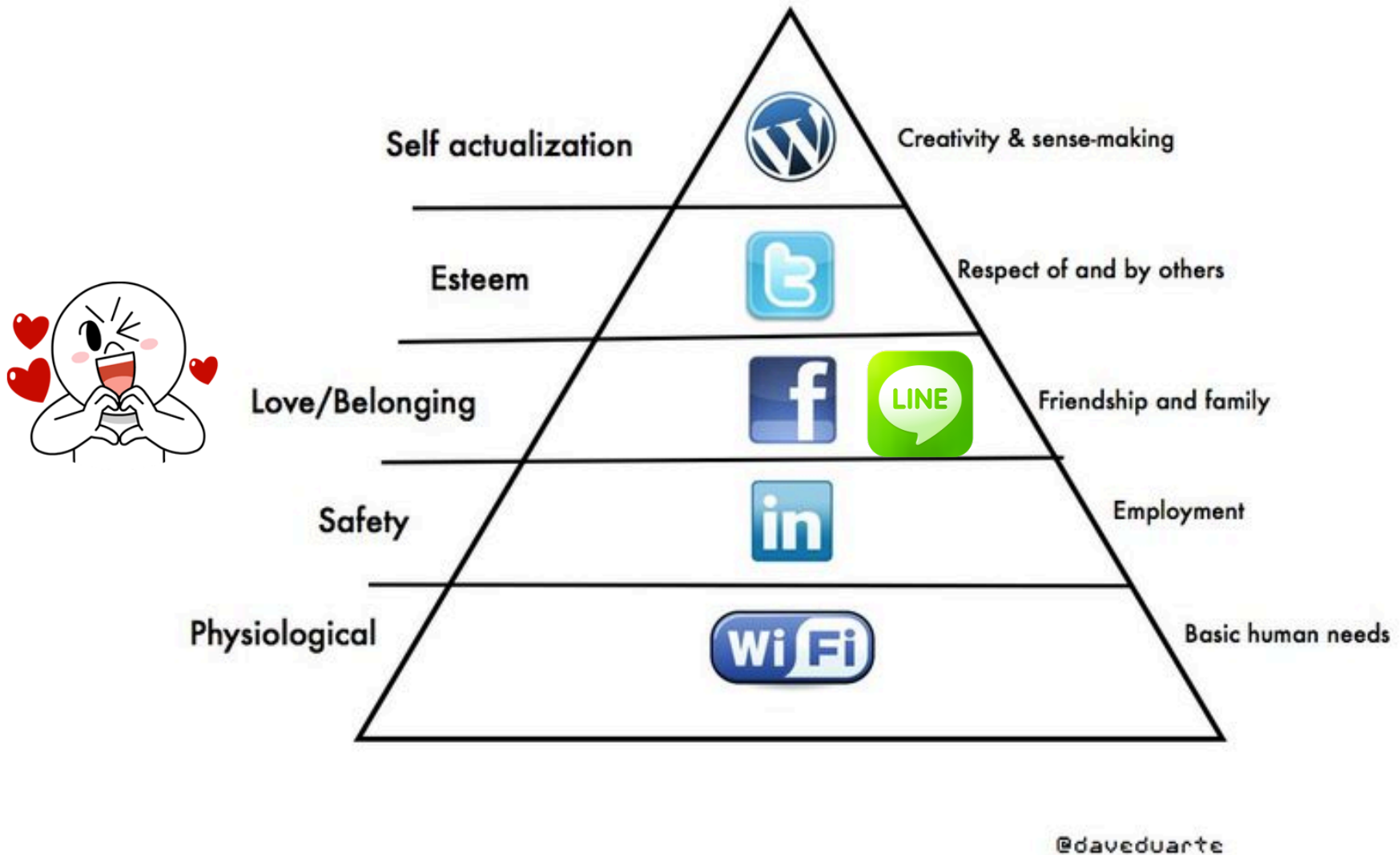
Surprise

Fear

Maslow's Hierarchy of Needs

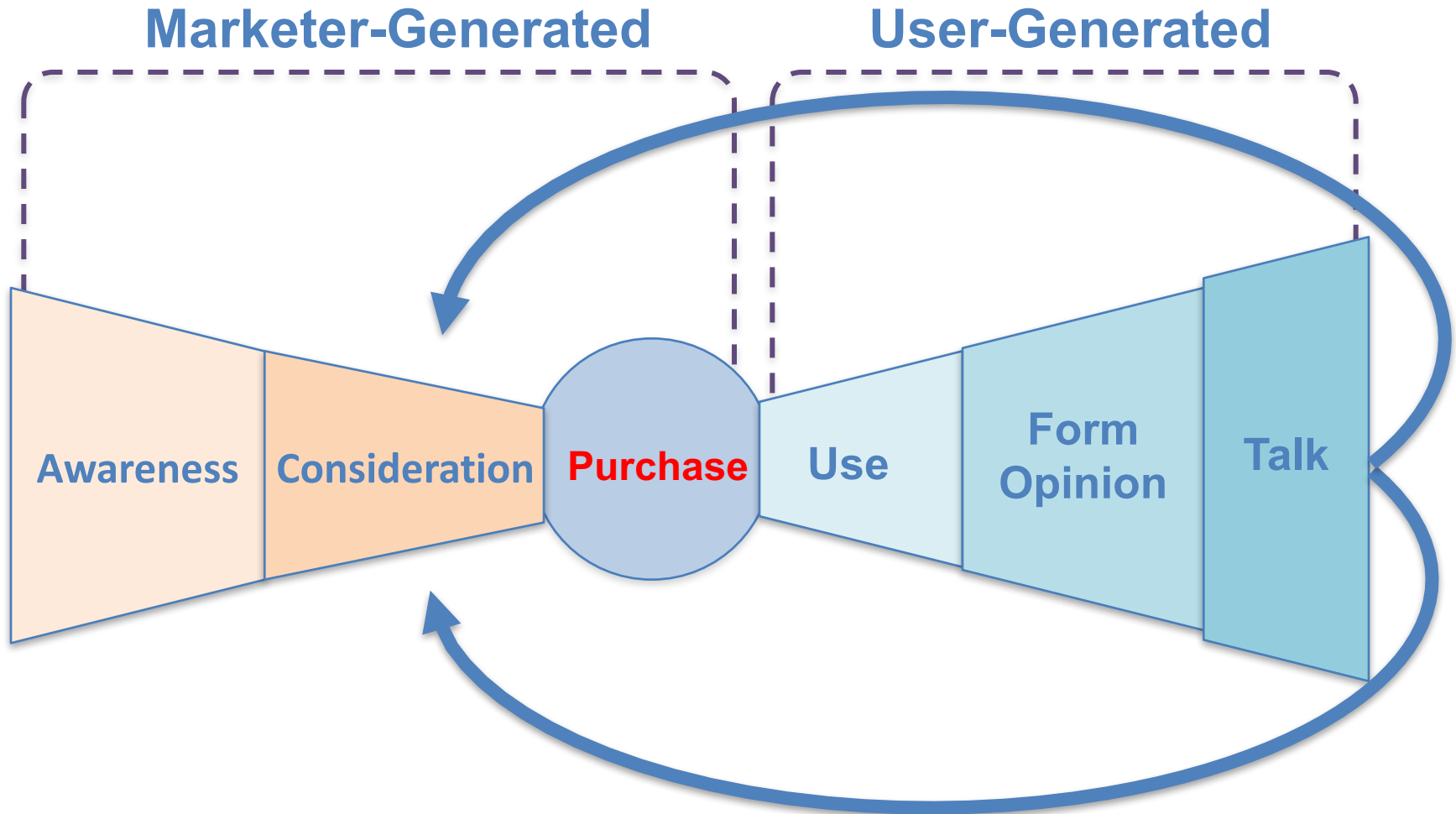


Social Media Hierarchy of Needs

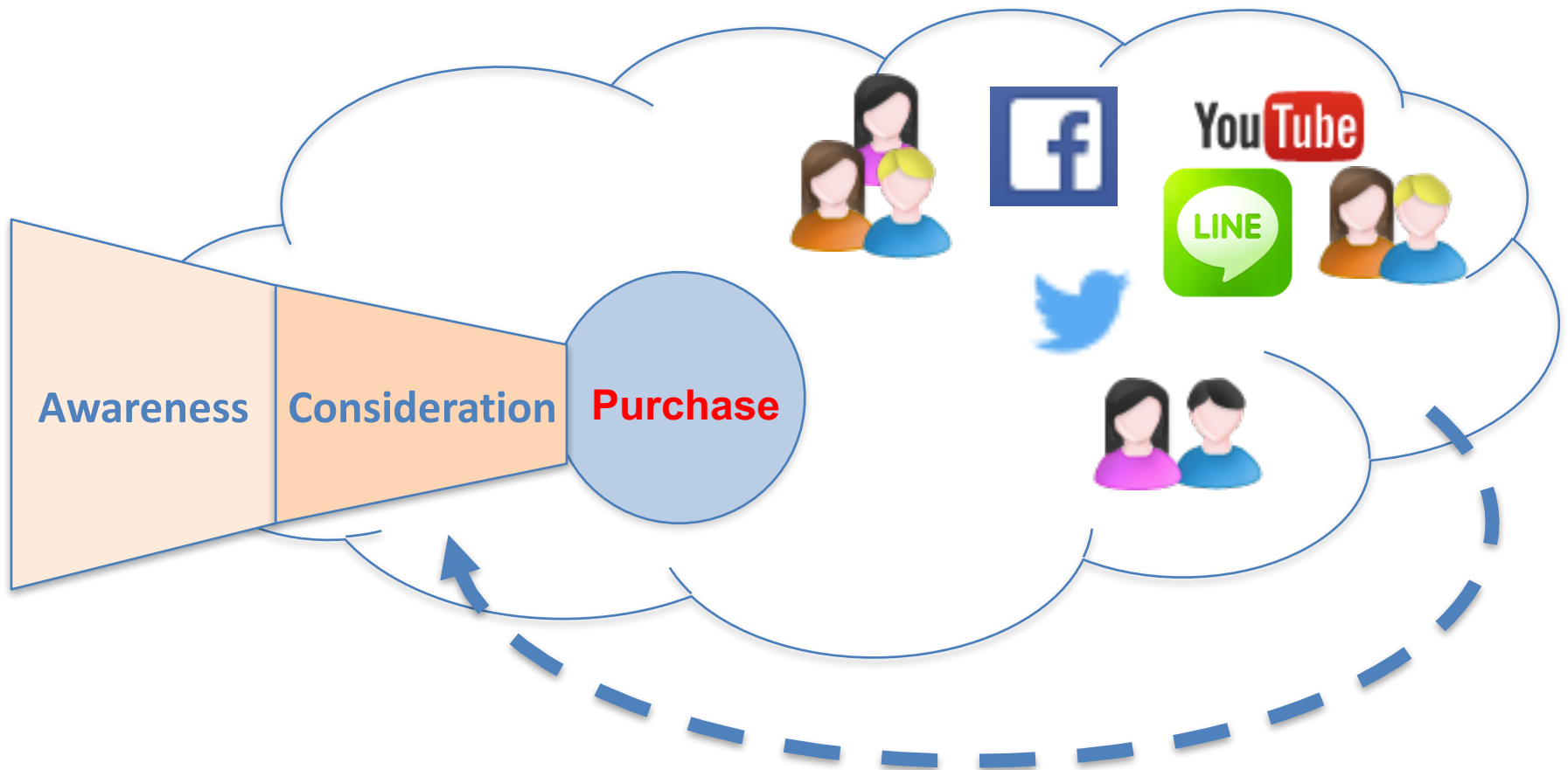


The Social Feedback Cycle

Consumer Behavior on Social Media



The New Customer Influence Path





Example of Opinion: review segment on iPhone



“I bought an iPhone a few days ago.

It was such a nice phone.

The touch screen was really cool.

The voice quality was clear too.

However, my mother was mad with me as I did not tell her before I bought it.

She also thought the phone was too expensive, and wanted me to return it to the shop. ... ”

Example of Opinion: review segment on iPhone

“(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

(4) The voice quality was **clear** too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”



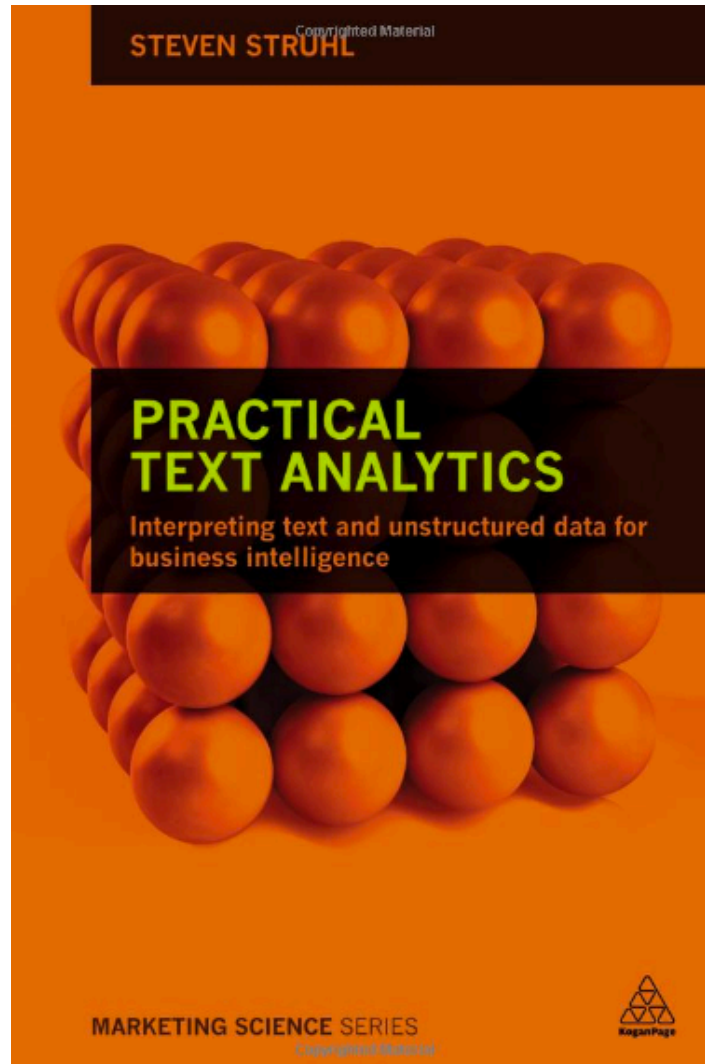
**+Positive
Opinion**



**-Negative
Opinion**

Text Mining Technologies

**Steven Struhl (2015),
Practical Text Analytics:
Interpreting Text and Unstructured Data for Business Intelligence
(Marketing Science), Kogan Page**



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

Text mining

Text Data Mining

Intelligent Text Analysis

Knowledge-Discovery in Text (KDT)

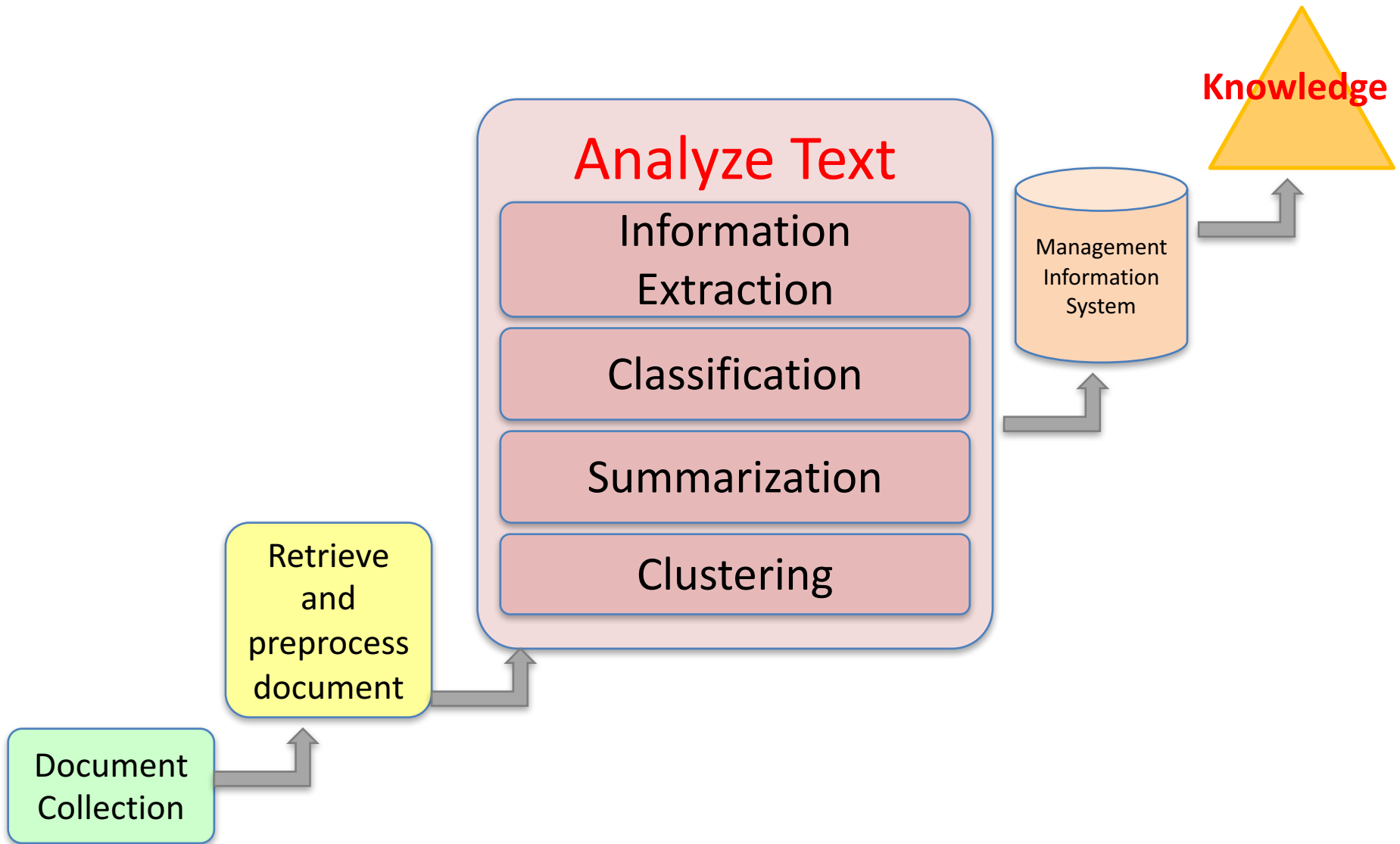
Text Mining:
the process of extracting
interesting and non-trivial
information and knowledge
from unstructured text.

Text Mining:
discovery by computer of
new, previously
unknown information,
by automatically
extracting information
from different written resources.

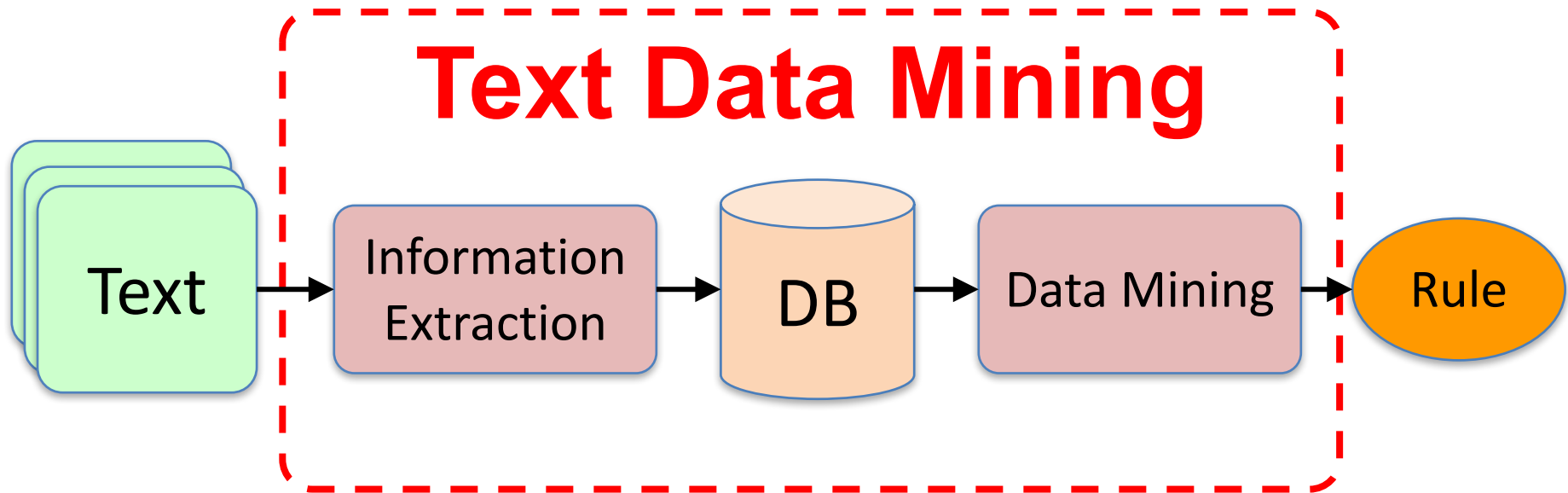
Text Mining (TM)

**Natural Language Processing
(NLP)**

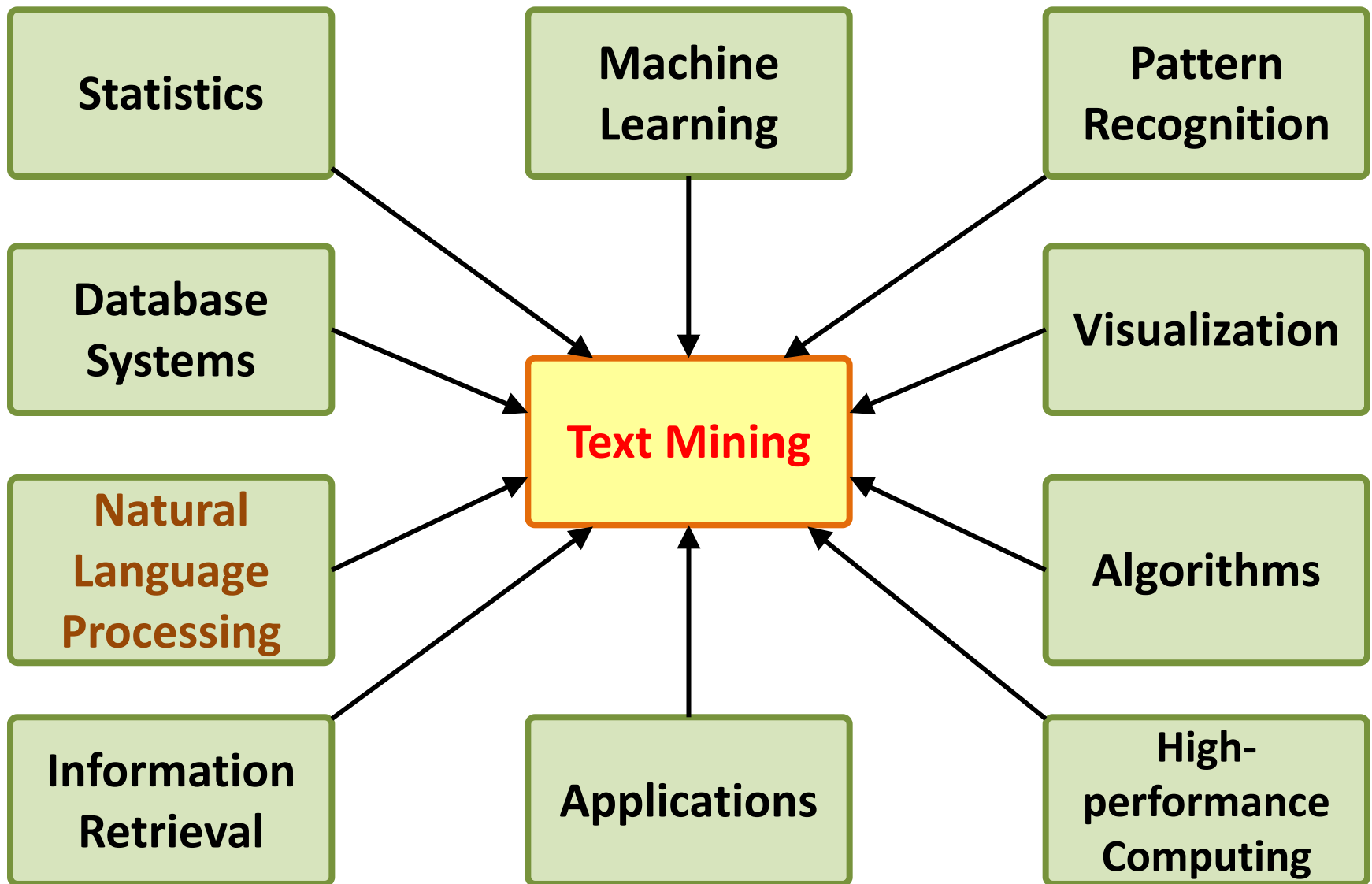
An example of Text Mining



Overview of Information Extraction based Text Mining Framework



Text Mining Technologies



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



Data Mining:

Core **Analytics** Process

The **KDD** Process for
Extracting Useful **Knowledge**
from Volumes of **Data**

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

The **KDD Process** for Extracting Useful **Knowledge** from Volumes of **Data**.

Communications of the ACM, 39(11), 27-34.

Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information.

The KDD Process for Extracting Useful Knowledge from Volumes of Data


AS WE MARCH INTO THE AGE of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls,

and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses.

Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer

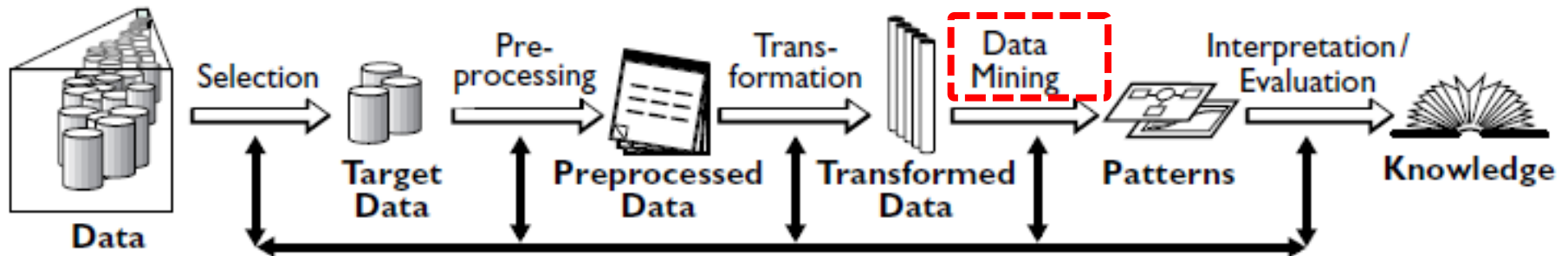
Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth



Data Mining

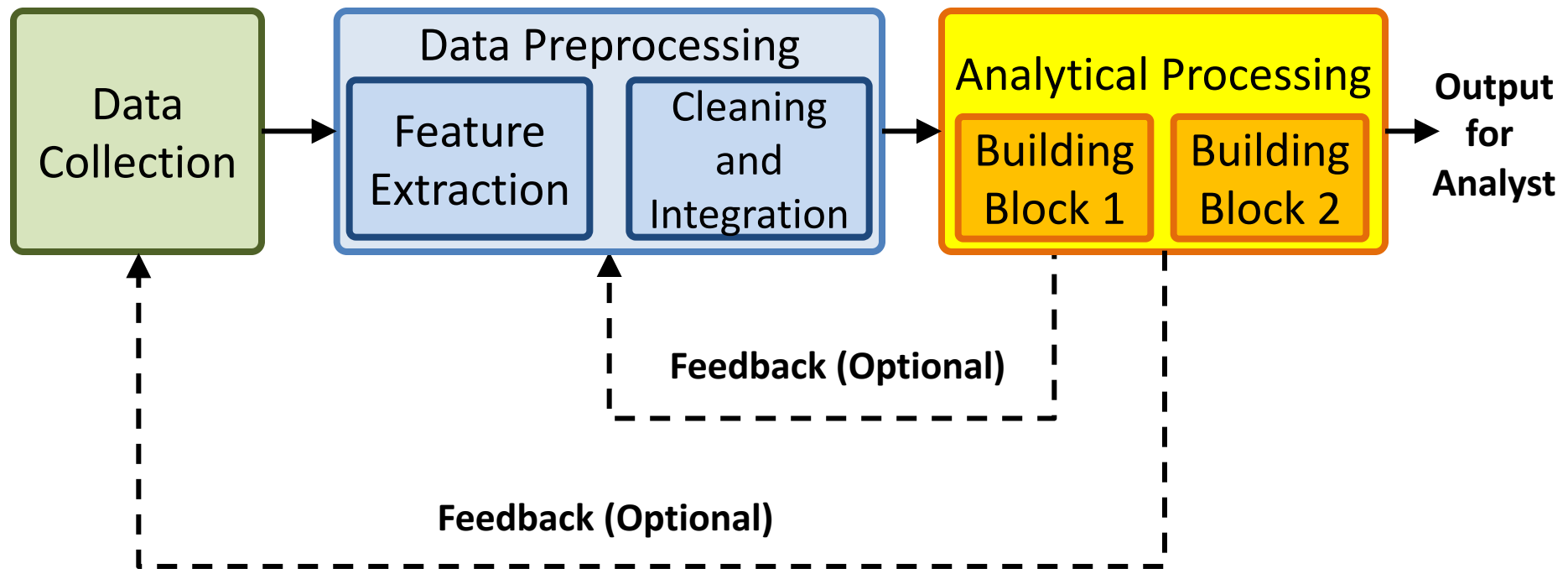
Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)



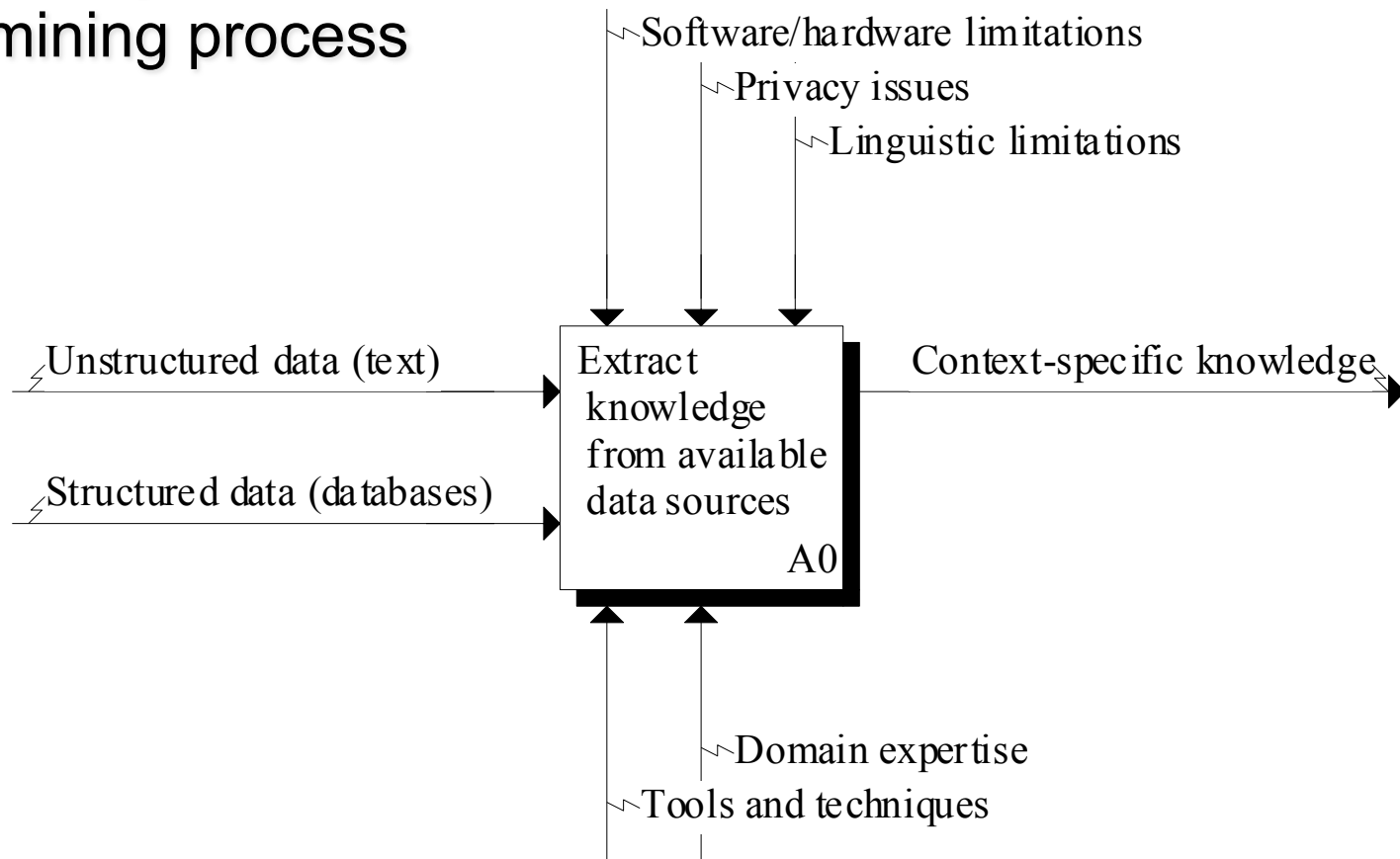
Data Mining Processing Pipeline

(Charu Aggarwal, 2015)

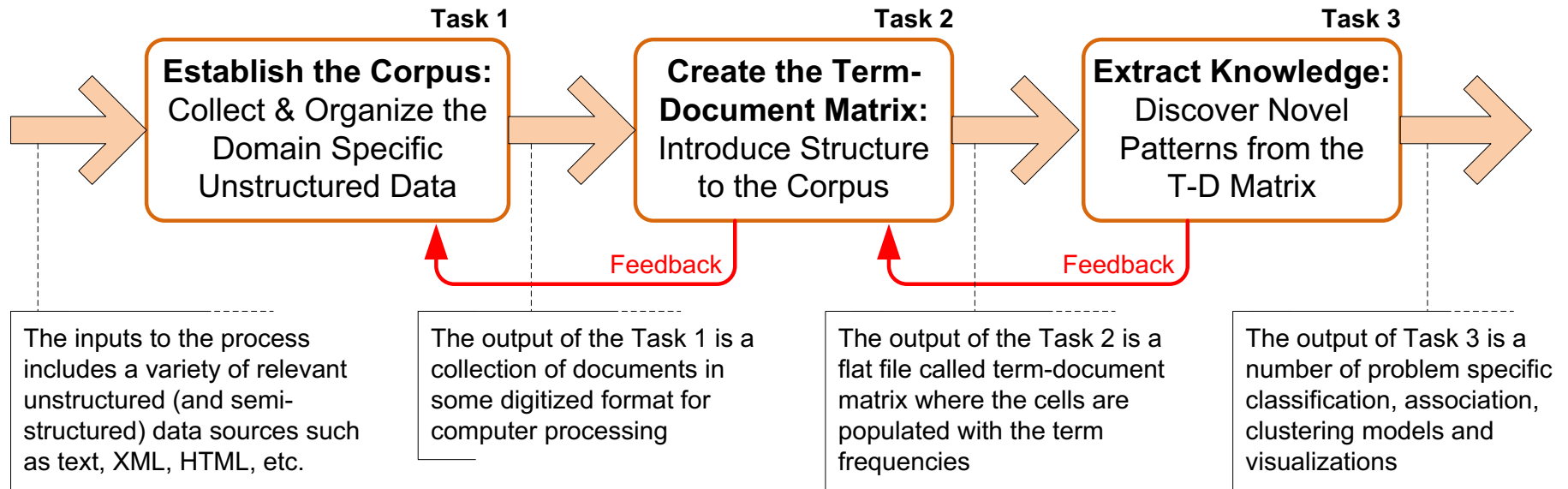


Text Mining Process

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

<div>Terms</div> <div>Documents</div>	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-Documents 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

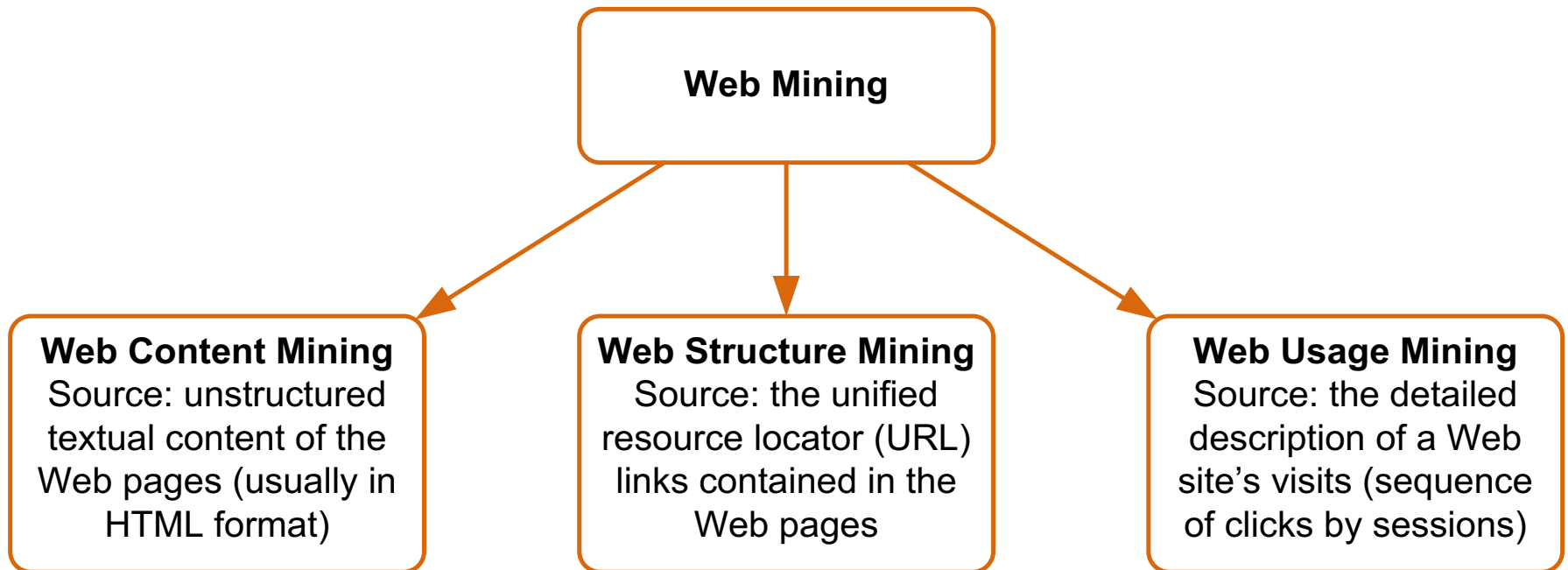
- **Step 2:** Create the Term-by-Document Matrix (TDM), cont.
 - TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
 - Manual - a domain expert goes through it
 - Eliminate terms with very few occurrences in very few documents (?)
 - Transform the matrix using singular value decomposition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

Web Mining

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



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)

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](#)

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 28270134 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](#)

[隱私權聲明](#) | [版權聲明](#)



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

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

[🏠](#)
[即時](#)
[要聞](#)
[娛樂](#)
[運動](#)
[全球](#)
[社會](#)
[產經](#)
[股市](#)
[健康](#)
[生活](#)
[文教](#)
[評論](#)
[地方](#)
[兩岸](#)
[國際](#)

莎士比亞在淡江 遇見賽萬提斯

莎士比亞在淡江 遇見賽萬提斯

2016-04-26 02:27 聯合報 記者徐葳倫／淡水報導

分享4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作者賽萬提斯逝世之日。英專起家的淡江大學舉辦「當莎士比亞遇見賽萬提斯」活動，規畫主題書展、彩繪活動，並添購新書，拉近學生與經典文學的距離。

首波登場的「主題書展」，展出2大文豪經典作品的原著、各種譯本以及DVD、電子書等數位化資料，校方也添購許多新書，吸引學生「搶鮮」閱讀經典名作。現場還規畫「彩繪大師」，讓學生發揮創意，畫出五彩繽紛的莎士比亞和賽萬提斯人像。

英語系四年級學生陳彥伶說，讀英語系接觸莎士比亞作品，但過去沒有舉辦書展時，這些作品都放在圖書館8樓，現在搬到1樓大廳陳列，不僅有很多莎士比亞、賽萬提斯的經典新書，還可藉由電子書、電影理解兩位作家，是以前沒有過的體驗。

英語系四年級學生鄭少淮表示，莎士比亞的「馬克白」、「羅密歐與茱麗葉」都已經讀過很多次，從經典文學中理解不同城市、國家的文化。

日文系學生賴喬郁說，原本只是喜歡塗鴉才來參加活動，後來才知道畫的是2個大文豪，接觸他們的作品，文學經典「原來離我這麼近」。

淡江大學外語學院院長陳小雀表示，莎士比亞的「to be, or not to be; that is the question」，賽萬提斯的「看得越多，行得越遠；書讀得越多，知識就越廣博」，都是來自文學的名言，校方希望用最簡單的方式，讓學生知道「文學不難」，就在你我身邊。



淡江大學舉辦「當莎士比亞遇見賽萬提斯」系列活動，讓師生幫莎士比亞、賽萬提斯著色，畫出五彩繽紛的「文學大師」。記者徐威倫／攝影

4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作

<http://udn.com/news/story/7323/1653437-%E8%8E%E5%A3%AB%E6%AF%94%E4%BA%9E%E5%9C%A8%E6%B7%A1%E6%B1%9F-0%E8%8A%87%E8%A8%8D%E8%BD%E8%8B%A8%E8%85%88%E8%88%A5>

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.

自 2014/01/06 起，本斷詞系統已經處理過 28270134 篇文章

送出

清除

莎士比亞在淡江 遇見賽萬提斯
2016-04-26 02:27 聯合報 記者徐葳倫 / 淡水報導

分享4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作者賽萬提斯逝世之日。英專起家的淡江大學舉辦「當莎士比亞遇見賽萬提斯」活動，規畫主題書展、彩繪活動，並添購新書，拉近學生與經典文學的距離。

首波登場的「主題書展」，展出2大文豪經典作品的原著、各種譯本以及DVD、電子書等數位化資料，校方也添購許多新書，吸引學生「搶鮮」閱讀經典名作。現場還規畫「彩繪大師」，讓學生發揮創意，畫出五彩繽紛的莎士比亞和賽萬提斯人像。英語系四年級學生陳彥伶說，讀英語系接觸莎士比亞作品，但過去沒有舉辦書展時，這些作品都放在圖書館8樓，現在搬到1樓大廳陳列，不僅有很多莎士比亞、賽萬提斯的經典新書，還可藉由電子書、電影理解兩位作家，是以前沒有過的體驗。

英語系四年級學生鄭少淮表示，莎士比亞的「馬克白」、「羅密歐與茱麗葉」都已經讀過很多次，從經典文學中理解不同城市、國家的文化。

日文系學生賴喬郁說，原本只是喜歡塗鴉才來參加活動，後來才知道畫的是2個大文豪，接觸他們的作品，文學經典「原來離我這麼近」。

淡江大學外語學院院長陳小雀表示，莎士比亞的「to be, or not to be; that is the question」，賽萬提斯的「看得越多，行得越遠；書讀得越多，知識就越廣博」，都是來自文學的名言，校方希望用最簡單的方式，讓學生知道「文學不難」，就在你我身邊。

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.

莎士比亞(Nb) 在(P) 淡江(Nb) 遇見(VC) 賽萬提(Nb) 斯(Nep) 2016(Neu) -(FW) 04(Neu) -(FW) 2602(Neu) :(COLONCATEGORY) 27(Neu) 聯合報(Nb) 記者(Na) 徐葳倫(Nb) 淡水(Nc) 報導(Na) 分享(VJ) 4月(Nd) 23日(Nd) 是(SHI) 「(PARENTHESISCATEGORY) 也(D) 是(SHI) 英國(Nc) 大(VH) 文豪(Na) 莎士比亞(Nb) 的(DE) 生日(Na) 與(Caa) 忌日(Na) ,(COMMACATEGORY) 及(Caa) 「(PARENTHESISCATEGORY) 唐吉訶德(Nb) 」(PARENTHESISCATEGORY) 作者(Na) 賽萬提(Nb) 斯(Nep) 逝世(VH) 之(DE) 日(Na) 英(Nc) 專(D) 起家(VA) 的(DE) 淡江(Nb) 大學(Nc) 舉辦(VC) 「(PARENTHESISCATEGORY) 當(P) 莎士比亞(Nb) 遇見(VC) 賽萬提(Nb) 規畫(VC) 主題(Na) 書展(Na) 、(PAUSECATEGORY) 彩繪(VC) 活動(Na) ,(COMMACATEGORY) 並(Cbb) 添購(VC) 新書(Na) ,(COMMACATEGORY) 拉近(VC) 學生(Na) 與(Caa) 經典(Na) 文學(Na) 的(DE) 距離(Na) 。(PERIODCATEGORY) 首(Nes) 波(Nf) 登場(VA) 的(T) 「(PARENTHESISCATEGORY) 主題(Na) 書展(Na) 」(PARENTHESISCATEGORY) ,(COMMACATEGORY) 展出(VC) 2(Neu) 大(VH) 文豪(Na) 經典(Na) 作品(Na) 的(DE) 原著(Na) 、(PAUSECATEGORY) 各(Nes) 種(Nf) 譯本(Na) 以及(Caa) 校方(Na) 也(D) 添購(VC) 許多(Nega) 新書(Na) ,(COMMACATEGORY) 吸引(VJ) 學生(Na) 「(PARENTHESISCATEGORY) 搶鮮(Na) 」(PARENTHESISCATEGORY) 閱讀(VC) 經典(Na) 名作(Na) 。(PERIODCATEGORY) 現場(Nc) 還(D) 規畫(VC) 「(PARENTHESISCATEGORY) 彩繪(VC) 大師(Na) 」(PARENTHESISCATEGORY) ,(COMMACATEGORY) 讓(VL) 學生(Na) 發揮(VJ) 創意(Na) ,(COMMACATEGORY) 畫出(VC) 五彩繽紛(VH) 的(DE) 莎士比亞(Nb) 和(Caa) 賽萬提(Nb) 斯人(Na) 像(VG) 。(PERIODCATEGORY) 英語系(Nc) 四年級(Na) 學生(Na) 陳彥伶(Nb) 說(VE) ,(COMMACATEGORY) 讀(VC) 英語系(Nc) 接觸(VC) 莎士比亞(Nb) 作品(Na) ,(COMMACATEGORY) 但(Cbb) 過去(Nd) 沒有(D) 舉辦(VC) 書展(Na) 時(Ng) ,(COMMACATEGORY) 這些(Nega) 作品(Na) 都(D) 放(VC) 在(P) 圖書館(Nc) 8樓(Nc) ,(COMMACATEGORY)

CKIP 中研院中文斷詞系統

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

莎士比亞在淡江 遇見賽萬提斯

2016-04-26 02:27 聯合報 記者徐葳倫／淡水報導

分享4月23日是「世界閱讀日」，也是英國大文豪莎士比亞的生日與忌日，及「唐吉訶德」作者賽萬提斯逝世之日。英專起家的淡江大學舉辦「當莎士比亞遇見賽萬提斯」活動，規畫主題書展、彩繪活動，並添購新書，拉近學生與經典文學的距離。

莎士比亞(Nb) 在(P) 淡江(Nb) 遇見(VC) 賽萬提(Nb) 斯(Nep) 2016(Neu) -(FW) 04
(Neu) -(FW) 2602(Neu) :(COLONCATEGORY)
27(Neu) 聯合報(Nb) 記者(Na) 徐葳倫(Nb) 淡水(Nc) 報導(Na) 分享(VJ) 4月(Nd) 23日
(Nd) 是(SHI) 「(PARENTHESISCATEGORY) 世界(Nc) 閱讀日(Na) 」
(PARENTHESISCATEGORY) ，(COMMACATEGORY)
也(D) 是(SHI) 英國(Nc) 大(VH) 文豪(Na) 莎士比亞(Nb) 的(DE) 生日(Na) 與(Caa) 忌日
(Na) ，(COMMACATEGORY)
及(Caa) 「(PARENTHESISCATEGORY) 唐吉訶德(Nb) 」(PARENTHESISCATEGORY) 作者
(Na) 賽萬提(Nb) 斯(Nep) 逝世(VH) 之(DE) 日(Na) 。(PERIODCATEGORY)
英(Nc) 專(D) 起家(VA) 的(DE) 淡江(Nb) 大學(Nc) 舉辦(VC) 「
(PARENTHESISCATEGORY) 當(P) 莎士比亞(Nb) 遇見(VC) 賽萬提(Nb) 斯(Nep) 」
(PARENTHESISCATEGORY) 活動(Na) ，(COMMACATEGORY)
規畫(VC) 主題(Na) 書展(Na) 、(PAUSECATEGORY) 彩繪(VC) 活動(Na) ，
(COMMACATEGORY)
並(Cbb) 添購(VC) 新書(Na) ，(COMMACATEGORY)
拉近(VC) 學生(Na) 與(Caa) 經典(Na) 文學(Na) 的(DE) 距離(Na) 。(PERIODCATEGORY)



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 ↕

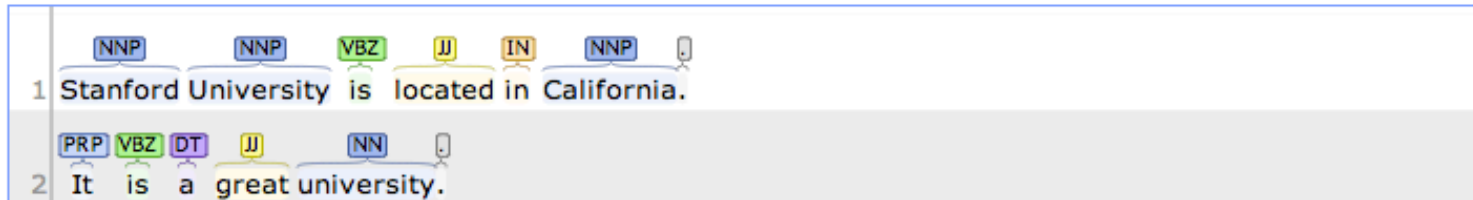
Please enter your text here:

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

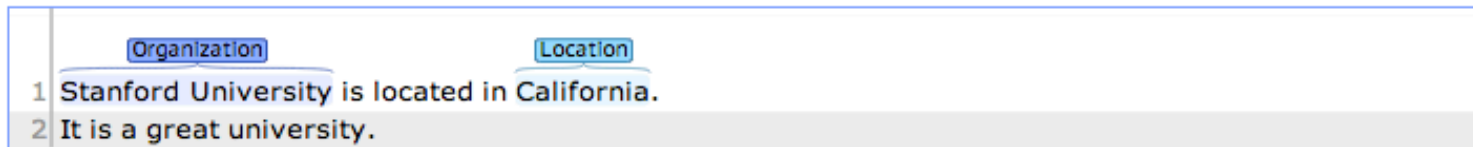
Submit

Clear

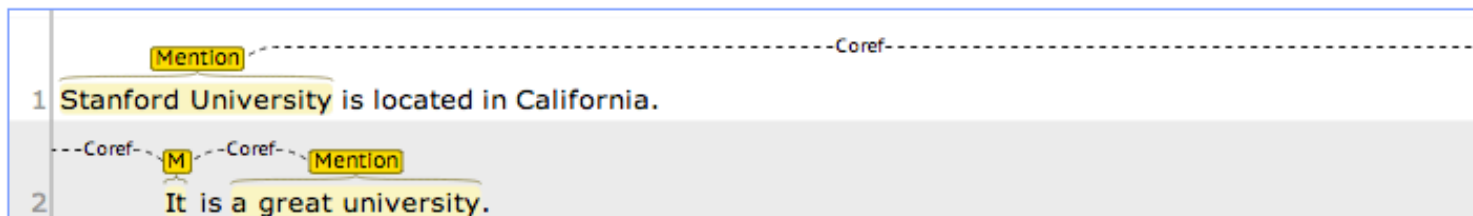
Part-of-Speech:



Named Entity Recognition:



Coreference:

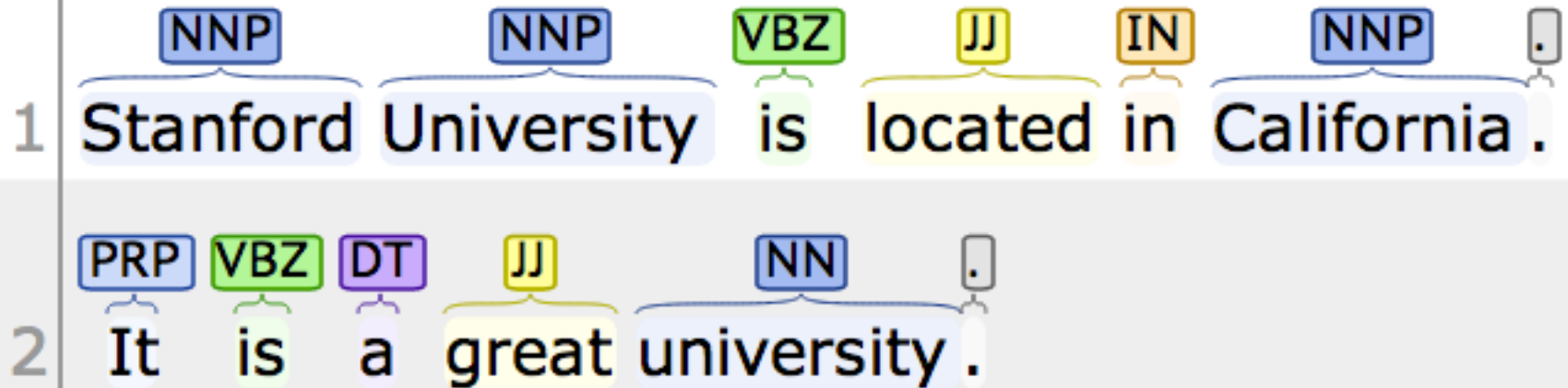


Stanford CoreNLP

<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:

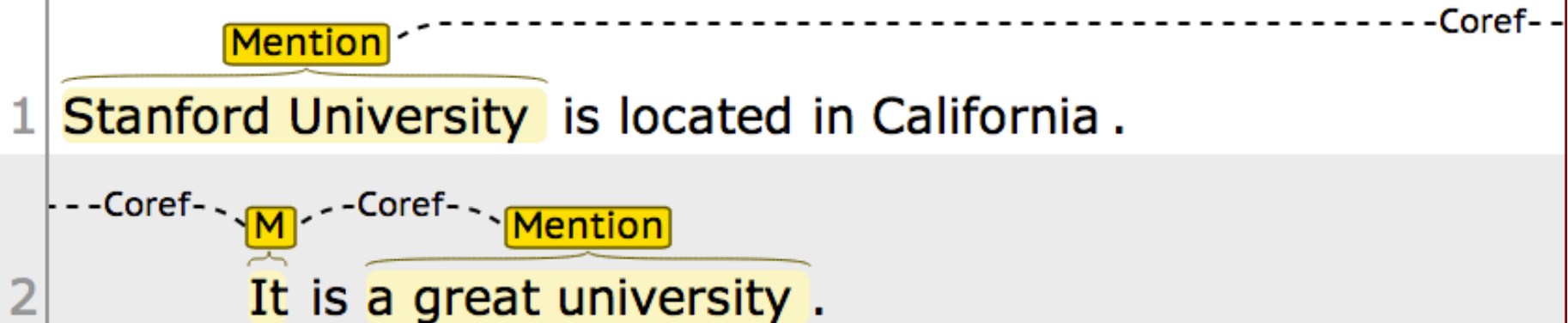
	Organization		Location
1	Stanford University	is located in	California .
2	It is a great university .		

Stanford CoreNLP

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

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

Coreference:

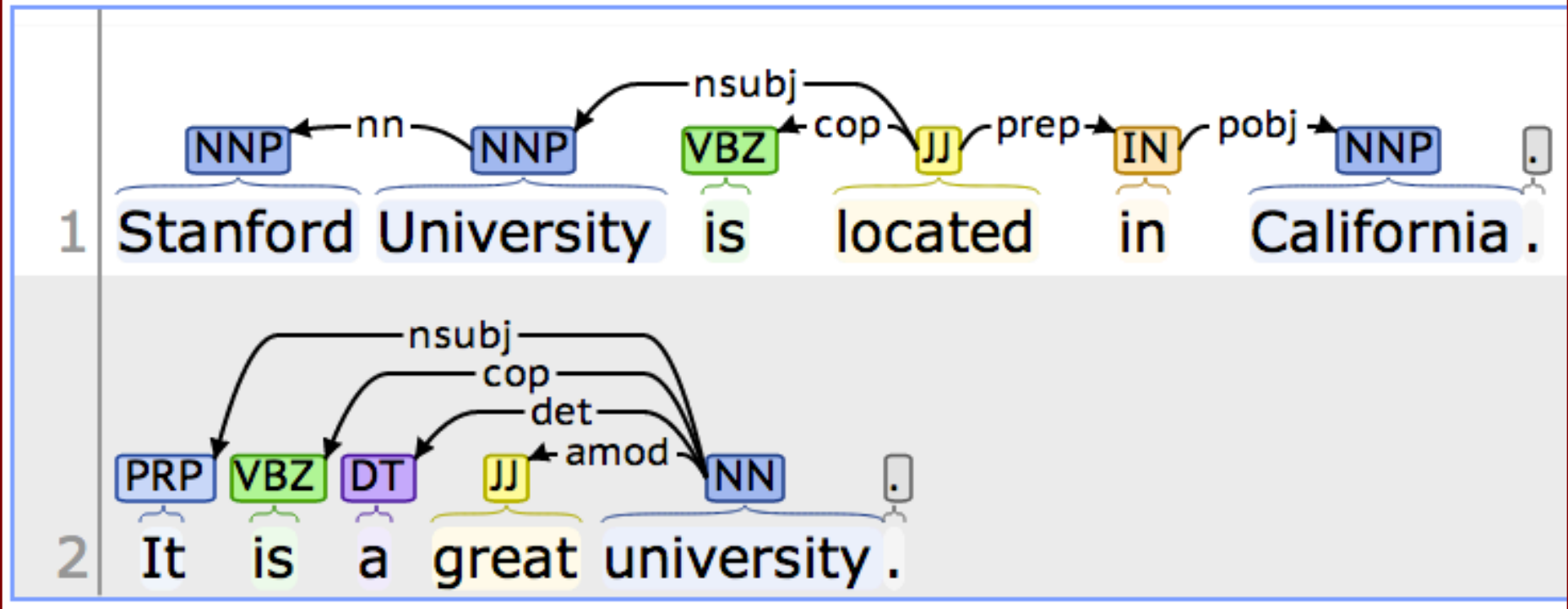


Stanford CoreNLP

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

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

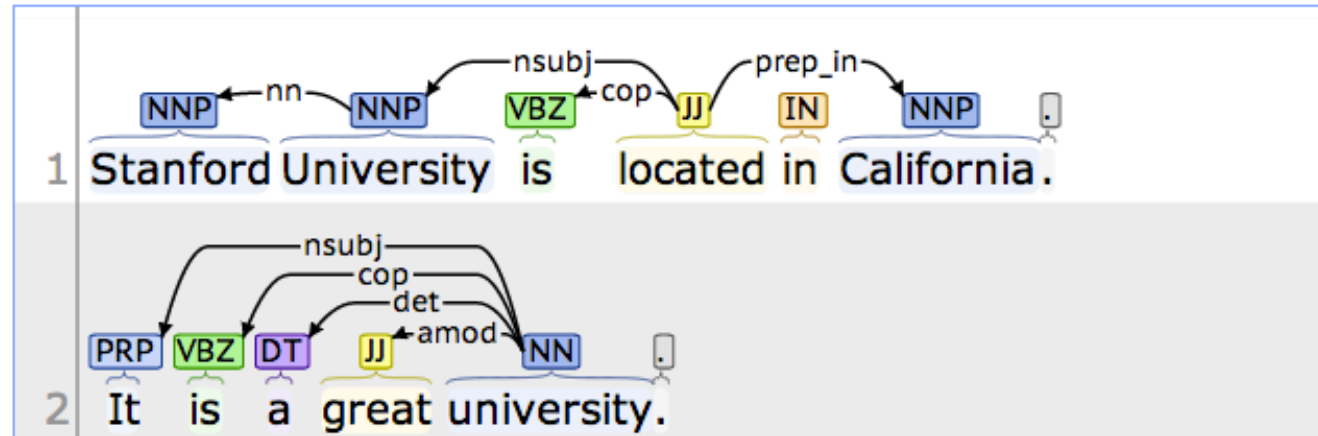
Basic dependencies:



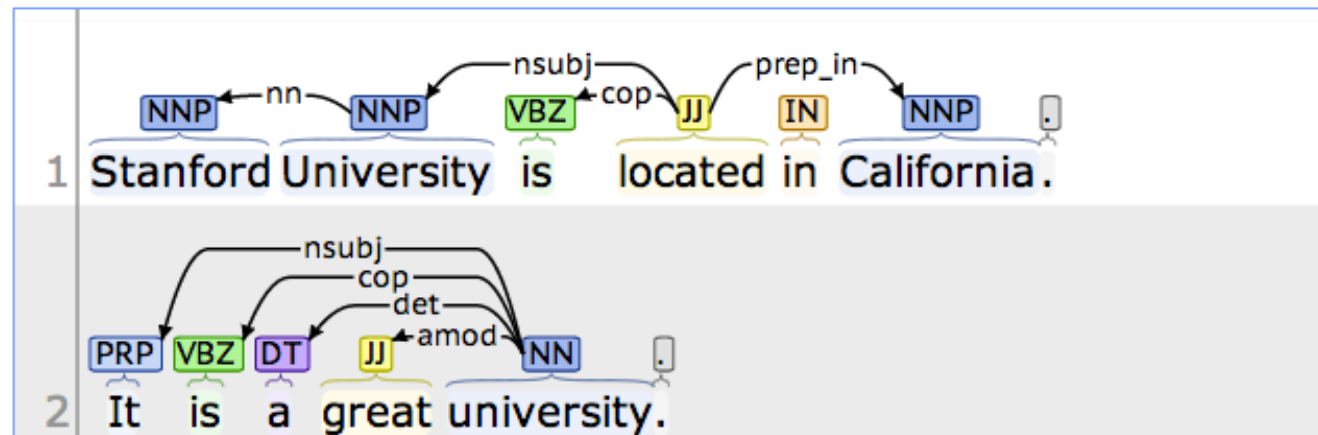
Stanford CoreNLP

<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

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 #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 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)

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>.

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

Copyright © 2019, Stanford University. All Rights Reserved.

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

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** (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

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** (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

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)

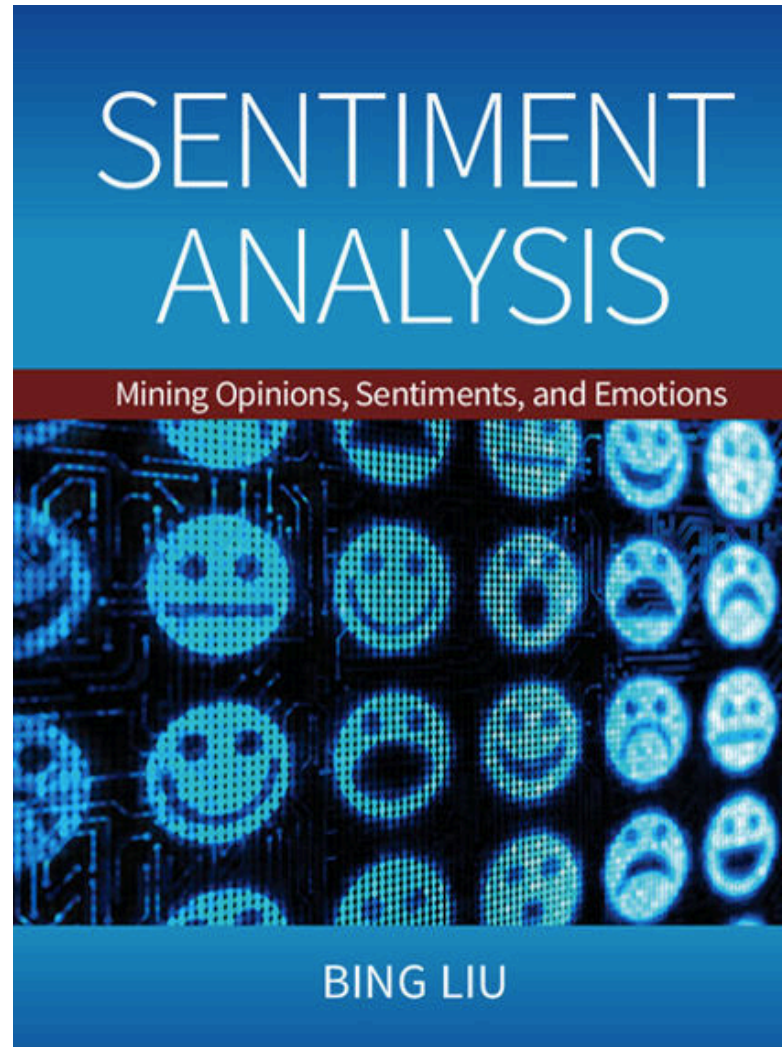
授權金額：免費授權學術使用。

授權日期：2011.05.03。

HIT IR: <http://ir.hit.edu.cn/>

Architectures of Sentiment Analytics

Bing Liu (2015),
Sentiment Analysis:
Mining Opinions, Sentiments, and Emotions,
Cambridge University Press



Sentiment Analysis and Opinion Mining

- Computational study of
opinions,
sentiments,
subjectivity,
evaluations,
attitudes,
appraisal,
affects,
views,
emotions,
ets., expressed in text.
 - Reviews, blogs, discussions, news, comments, feedback, or any other documents

Research Area of Opinion Mining

- Many names and tasks with difference objective and models
 - Sentiment analysis
 - Opinion mining
 - Sentiment mining
 - Subjectivity analysis
 - Affect analysis
 - Emotion detection
 - Opinion spam detection

Example of Opinion: review segment on iPhone

“(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

(4) The voice quality was **clear** too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

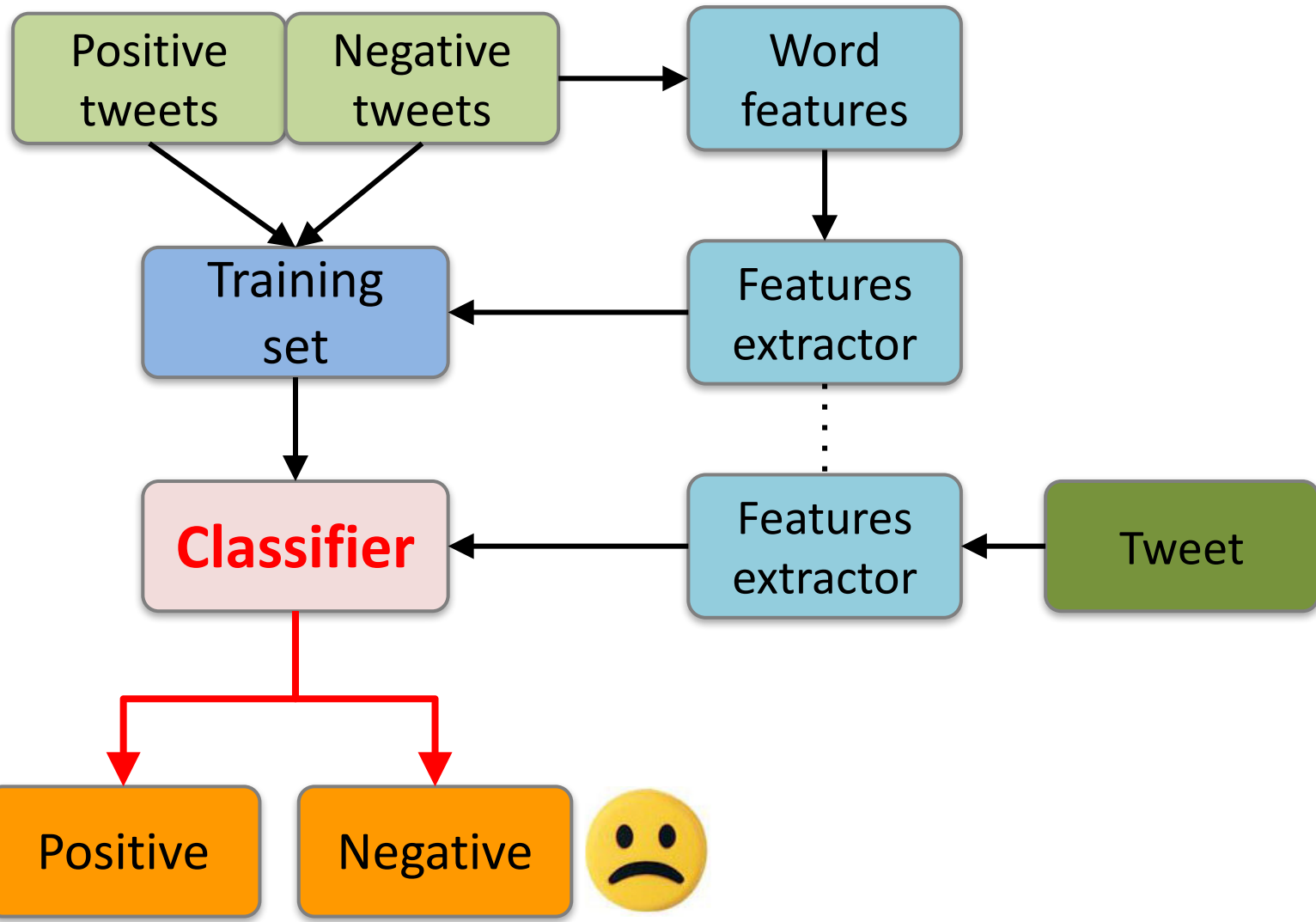


**+Positive
Opinion**

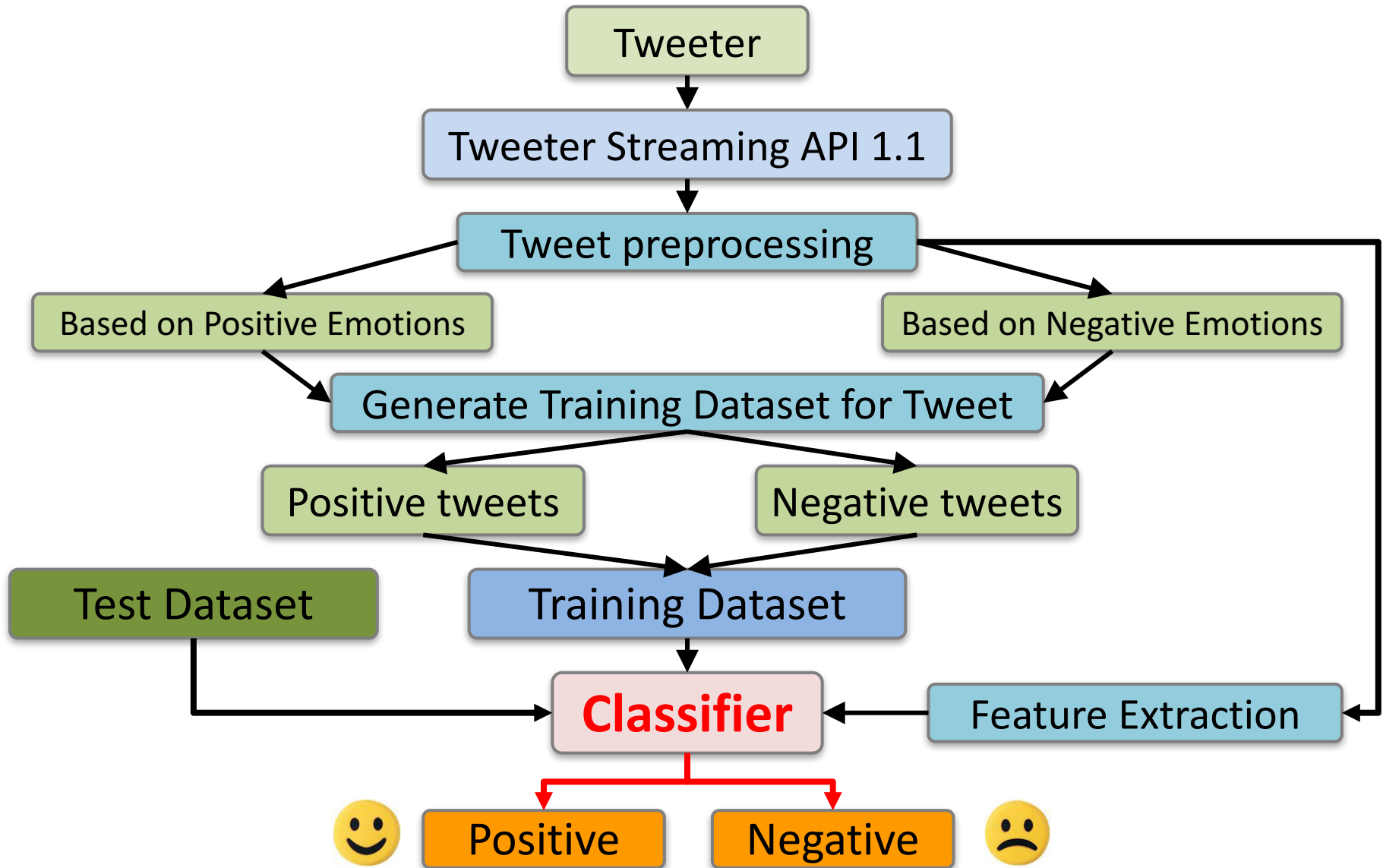


**-Negative
Opinion**

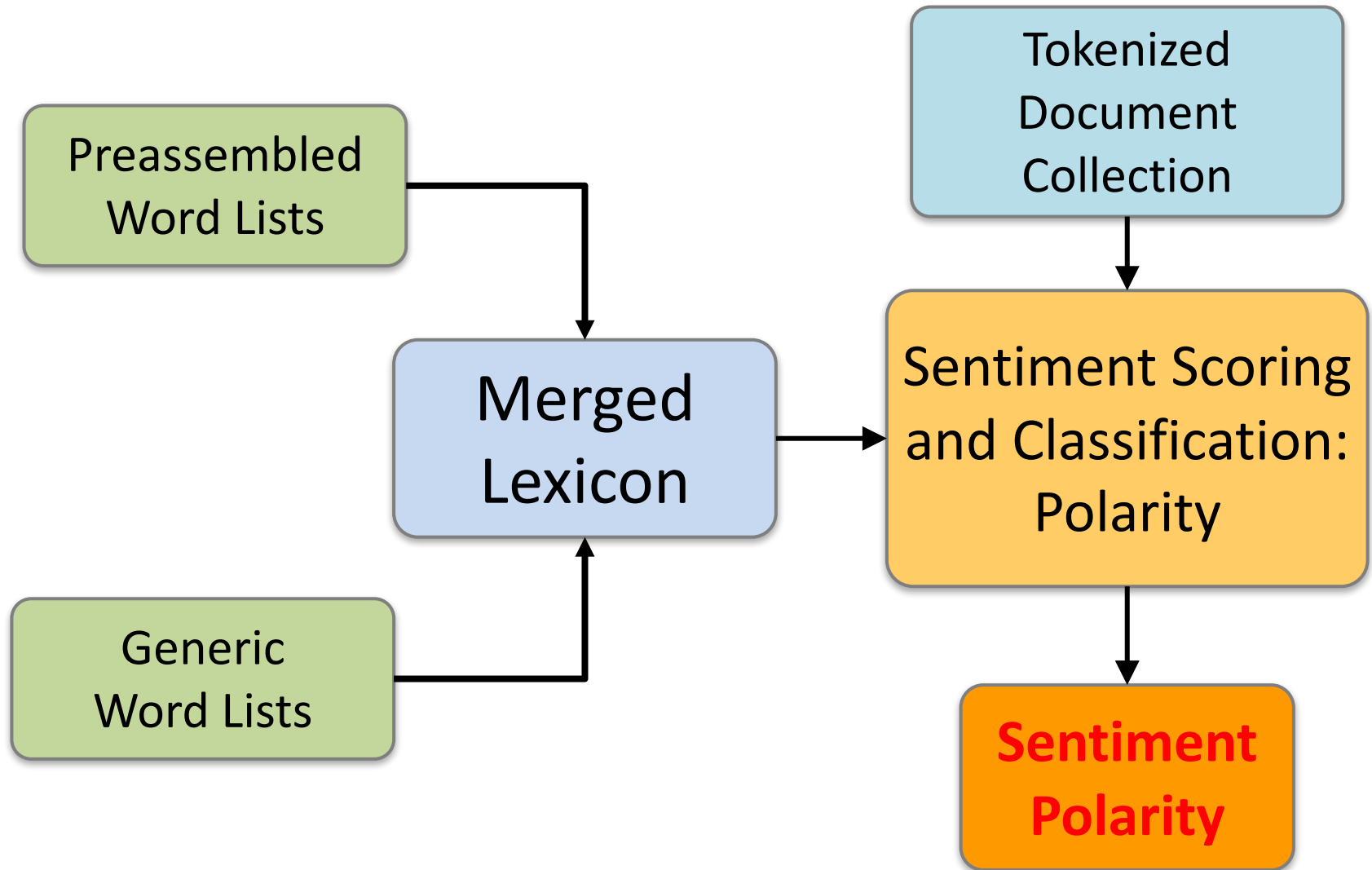
Sentiment Analysis Architecture



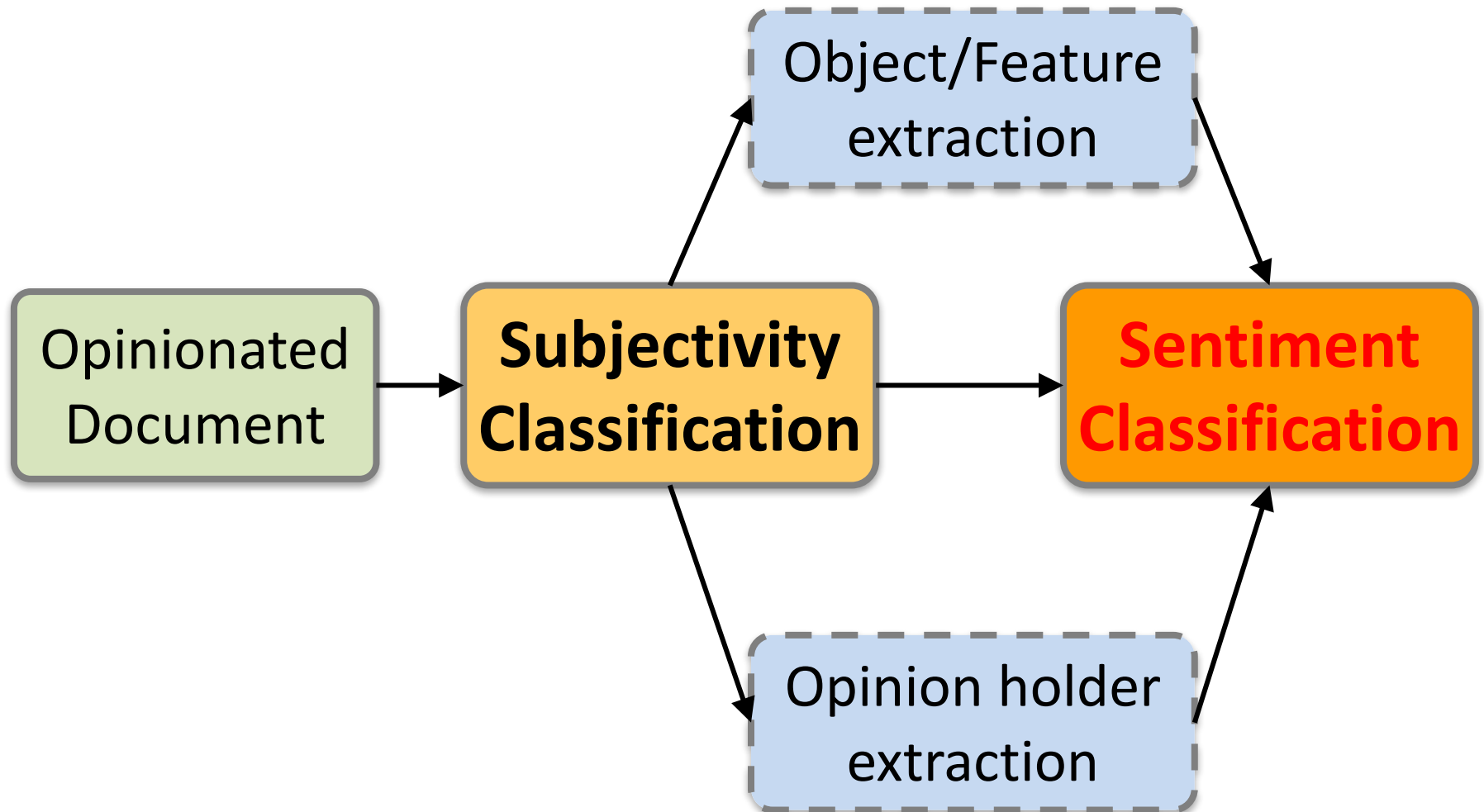
Sentiment Classification Based on Emoticons



Lexicon-Based Model



Sentiment Analysis Tasks



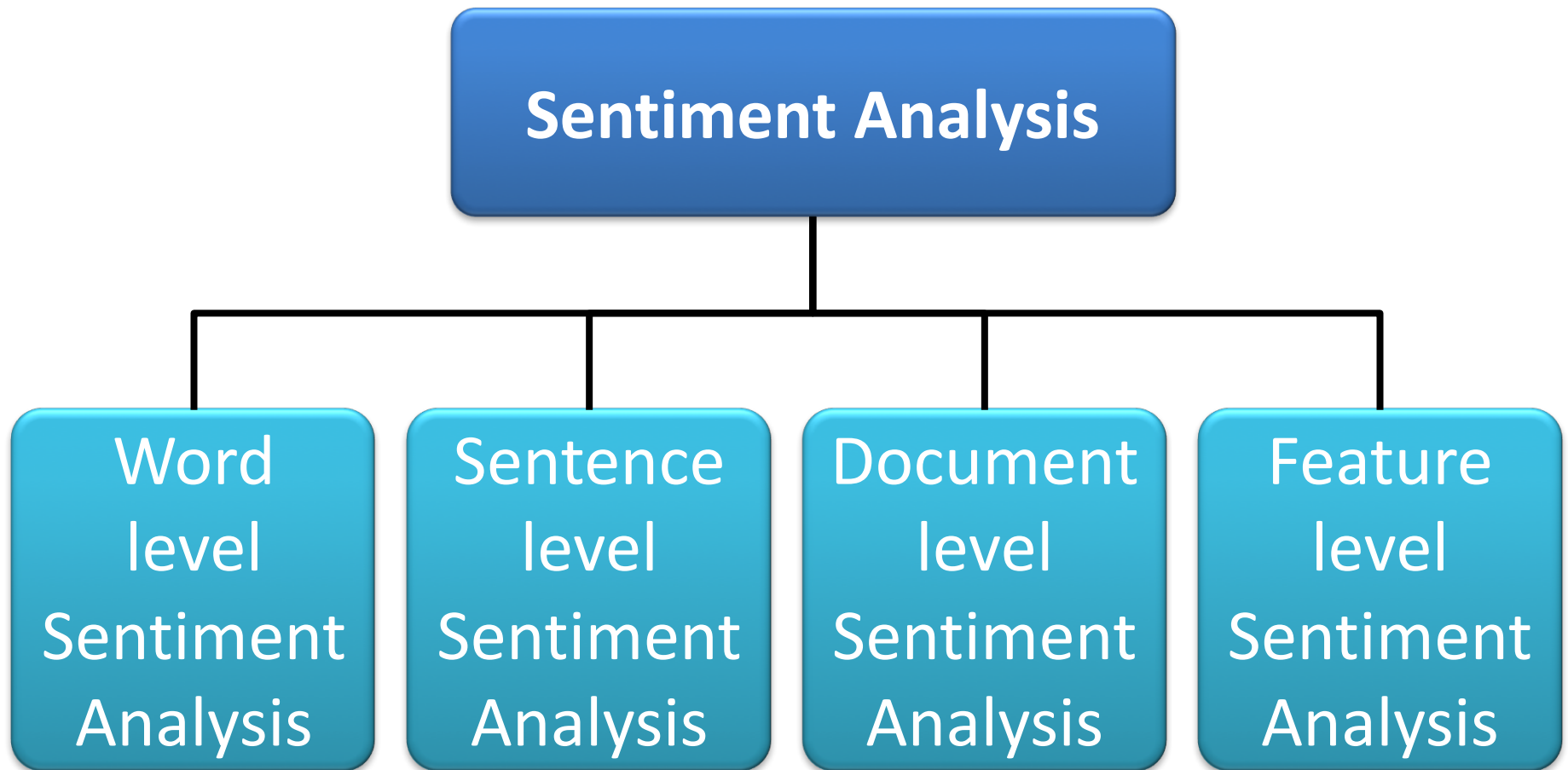
Sentiment Analysis

vs.

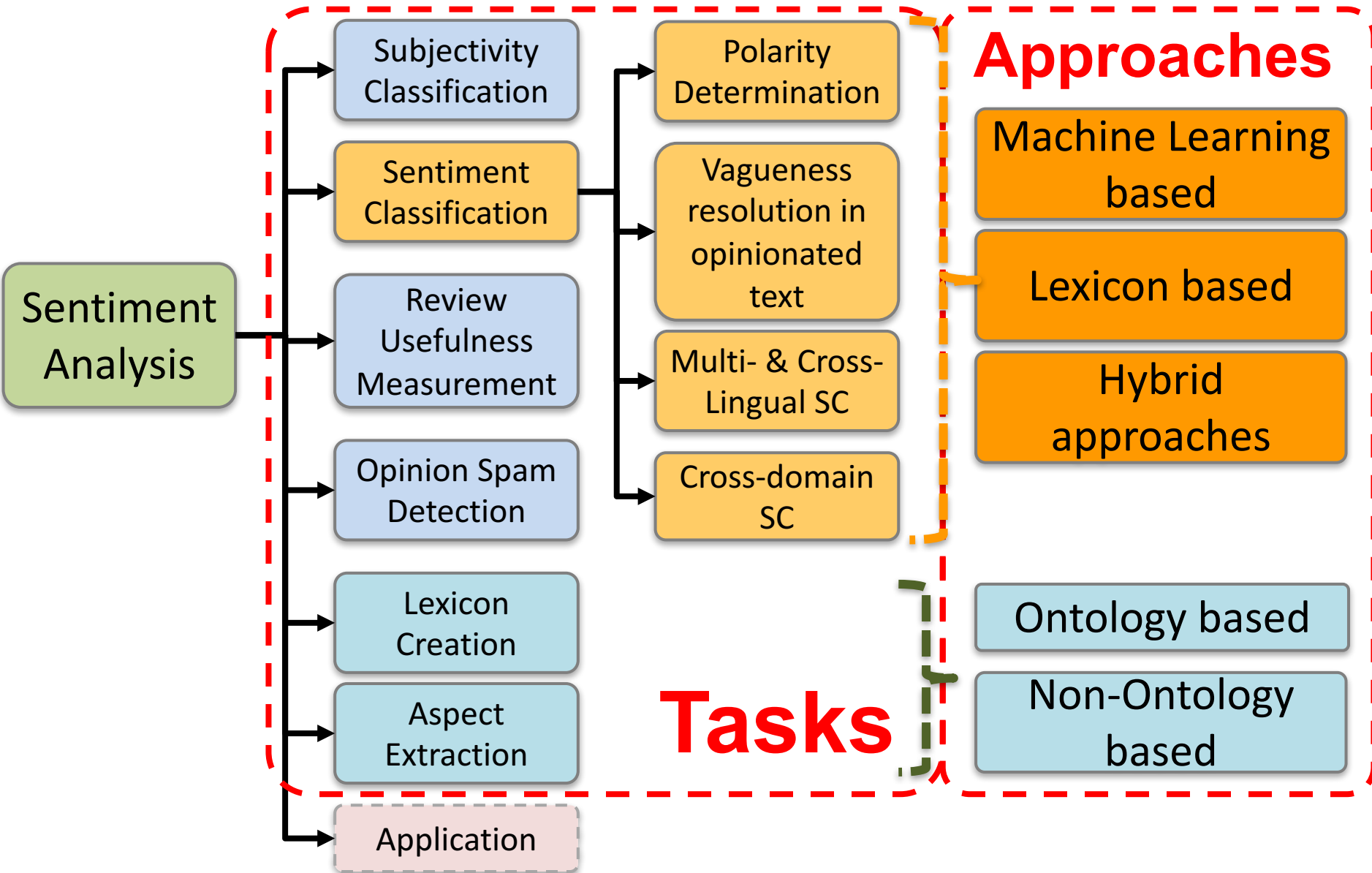
Subjectivity Analysis

Sentiment Analysis	Subjectivity Analysis
Positive	Subjective
Negative	
Neutral	Objective

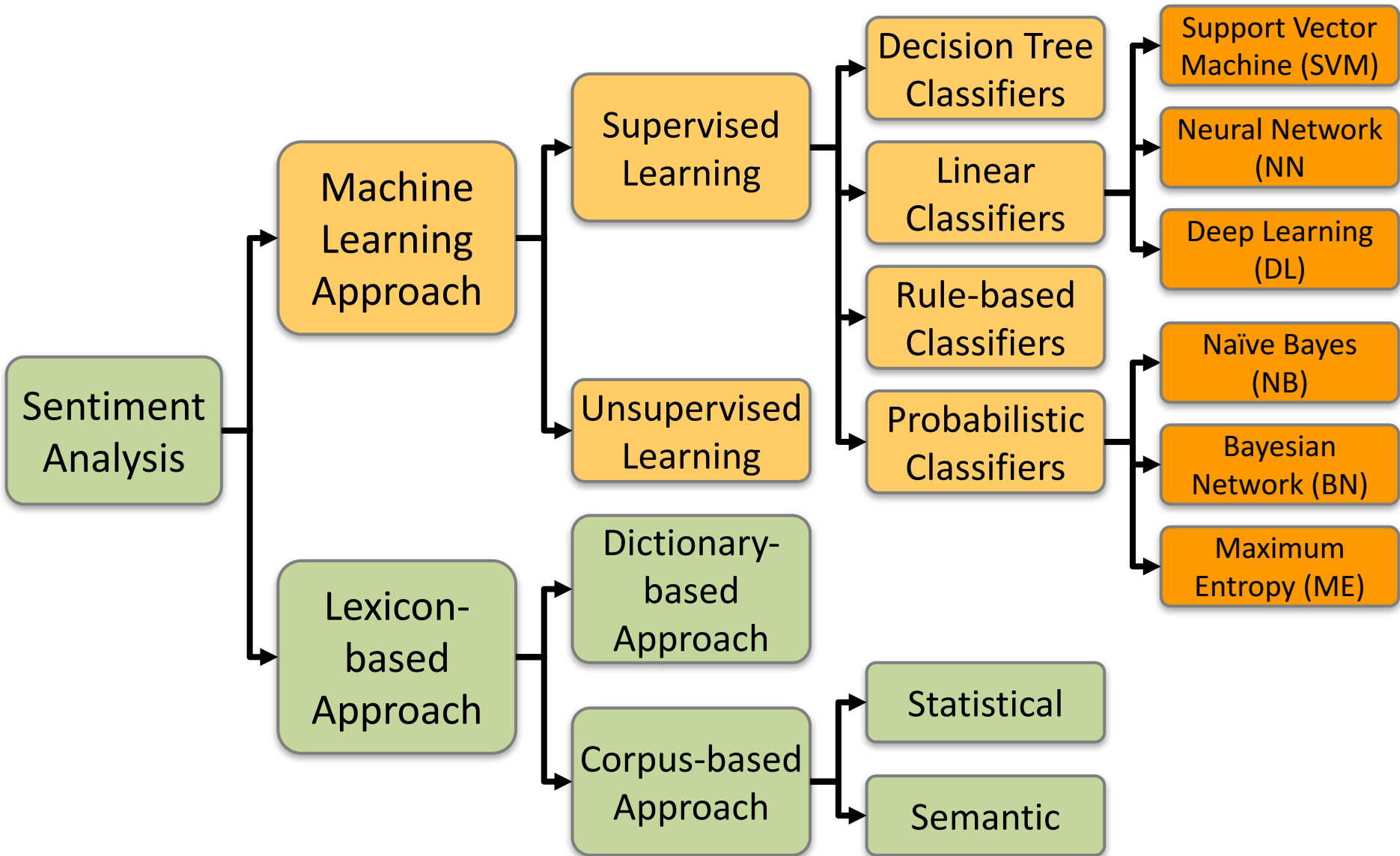
Levels of Sentiment Analysis



Sentiment Analysis



Sentiment Classification Techniques



Example of SentiWordNet

POS	ID	PosScore		NegScore		SynsetTerms	Gloss
a	00217728	0.75	0		beautiful#1	delighting the senses or exciting intellectual or emotional admiration; "a beautiful child"; "beautiful country"; "a beautiful painting"; "a beautiful theory"; "a beautiful party"	
a	00227507	0.75	0		best#1	(superlative of `good') having the most positive qualities; "the best film of the year"; "the best solution"; "the best time for planting"; "wore his best suit"	
r	00042614	0	0.625		unhappily#2	sadly#1	in an unfortunate way; "sadly he died before he could see his grandchild"
r	00093270	0	0.875		woefully#1	sadly#3	lamentably#1 deplorably#1 in an unfortunate or deplorable manner; "he was sadly neglected"; "it was woefully inadequate"
r	00404501	0	0.25		sadly#2		with sadness; in a sad manner; "``She died last night,' he said sadly"

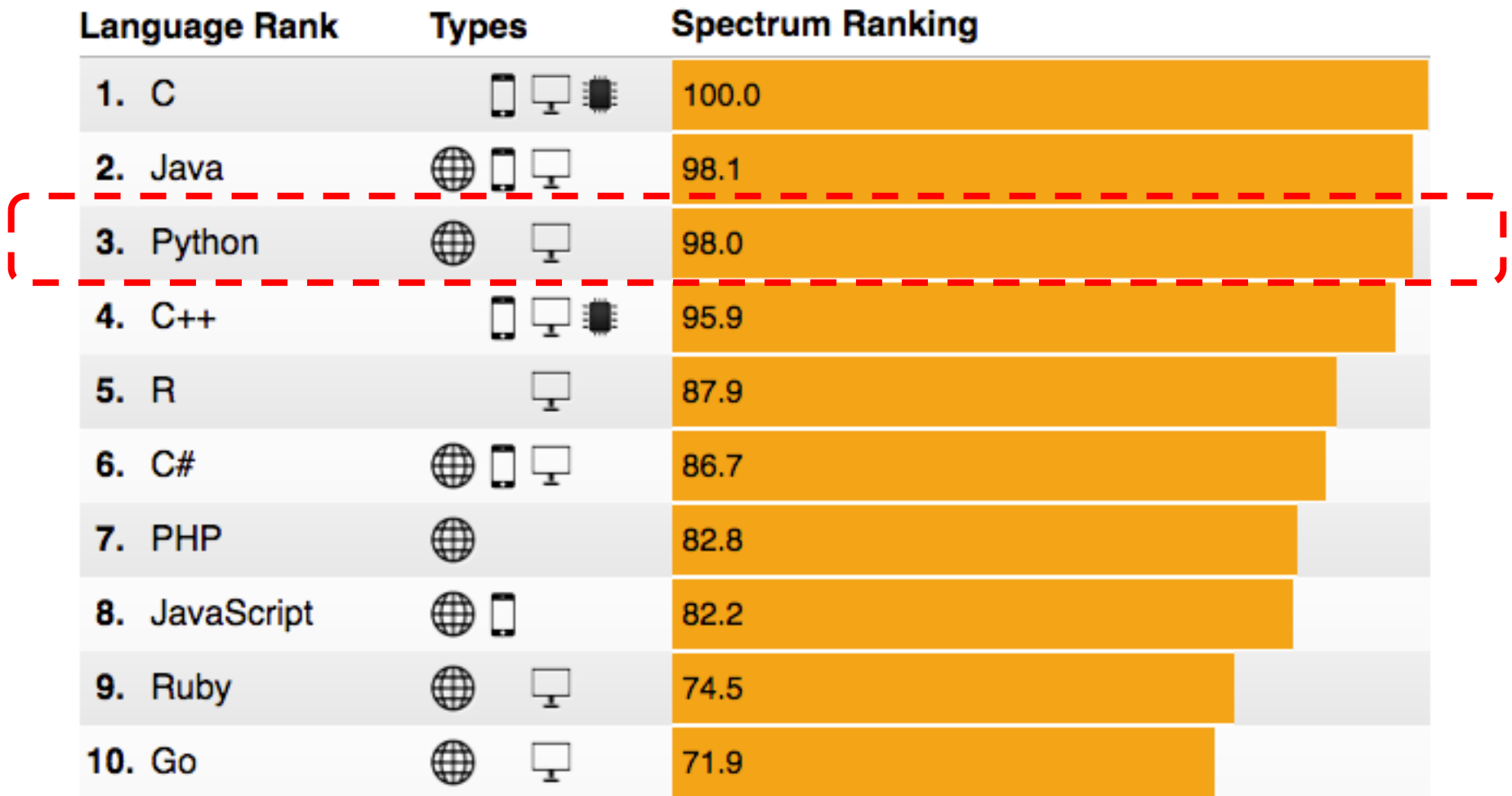
Evaluation of Text Mining and Sentiment Analysis

- Evaluation of Information Retrieval
- Evaluation of Classification Model (Prediction)
 - Accuracy
 - Precision
 - Recall
 - F-score

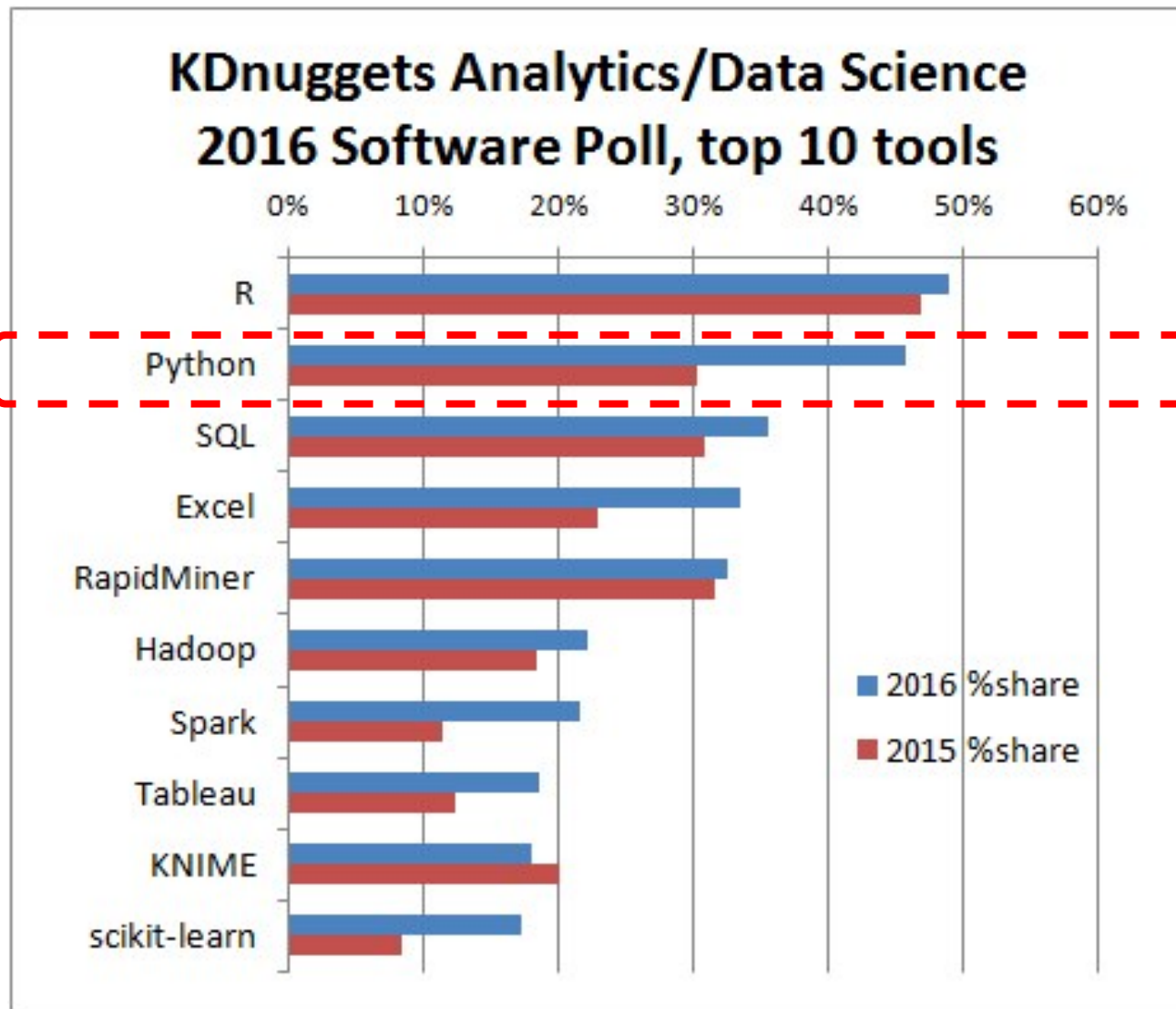
Natural Language Processing with NLTK in Python



Python for Big Data Analytics



Python: Analytics and Data Science Software



Python

Python

PSF

Docs

PyPI

Jobs

Community



GO

Socialize

Sign In

About

Downloads

Documentation

Community

Success Stories

News

Events

```
# Python 3: List comprehensions
>>> fruits = ['Banana', 'Apple', 'Lime']
>>> loud_fruits = [fruit.upper() for fruit in fruits]
>>> print(loud_fruits)
['BANANA', 'APPLE', 'LIME']

# List and the enumerate function
>>> list(enumerate(fruits))
[(0, 'Banana'), (1, 'Apple'), (2, 'Lime')]
```



Compound Data Types

Lists (known as arrays in other languages) are one of the compound data types that Python understands. Lists can be indexed, sliced and manipulated with other built-in functions. [More about lists in Python 3](#)

1

2

3

4

5

Python is a programming language that lets you work quickly and integrate systems more effectively. [>>> Learn More](#)

Get Started

Download

Docs

Jobs

<https://www.python.org/>

Python is an
interpreted,
object-oriented,
high-level
programming language
with
dynamic semantics.

Anaconda

CONTINUUM[®]
ANALYTICS

ANACONDA

COMMUNITY

SERVICES

SOLUTIONS

ABOUT

RESOURCES

LOG IN SUPPORT CONTACT

ANACONDA GIVES
SUPERPOWERS TO
PEOPLE WHO CHANGE
THE WORLD



ANACONDA[®]

Modern open source analytics platform powered
by Python

DOWNLOAD FOR FREE

ANACONDA NOW AVAILABLE FOR CLOUDERA CDH

WHY YOU'LL LOVE ANACONDA

Making it easy to install, intuitive to discover, quick to analyze, simple to collaborate, and accessible to all.

**Committed to Open
Source. Now and
forever.**

**Tested and certified
packages to cover
your back.**

**Explore and visualize
complex data easily.**

**All the analytics you
ever wanted and
more.**

<https://www.continuum.io/>

Download Anaconda

DOWNLOAD ANACONDA NOW!

Jump to: [Windows](#)  [Linux](#)

Get Superpowers with Anaconda

Anaconda is a completely free Python distribution (including for commercial use and redistribution). It includes more than 400 of the most popular Python packages for science, math, engineering, and data analysis. See the packages included with Anaconda and [the Anaconda changelog](#).

Which version should I download and install?

Because Anaconda includes installers for Python 2.7 and 3.5, either is fine. Using either version, you can use Python 3.4 with the conda command. You can create a 3.5 environment with the conda command if you've downloaded 2.7 — and vice versa.

If you don't have time or disk space for the entire distribution, try [Miniconda](#), which contains only conda and Python. Then install just the individual packages you want through the conda command.



Download Anaconda Python 3.5

Anaconda for OS X

PYTHON 2.7	PYTHON 3.5
<div>Mac OS X 64-bit Graphical Installer</div> <div>274M (OS X 10.7 or higher)</div>	<div>Mac OS X 64-bit Graphical Installer</div> <div>267M (OS X 10.7 or higher)</div>
<div>Mac OS X 64-bit Command-Line installer</div> <div>239M (OS X 10.7 or higher)</div>	<div>Mac OS X 64-bit Command-Line installer</div> <div>233M (OS X 10.7 or higher)</div>

OS X Anaconda Installation

Choose either the graphical installer or the command line installer for OS X.

Graphical Installer:

1. Download the graphical installer.
2. Double-click the downloaded .pkg file and follow the instructions.

OS X Anaconda Installation

OS X Anaconda Installation

Choose either the graphical installer or the command line installer for OS X.

Graphical Installer:

1. Download the graphical installer.
2. Double-click the downloaded .pkg file and follow the instructions.

Command Line Installer:

1. Download the command line installer.
2. In your terminal window, type one of the below and follow the instructions:

Python 2.7:

```
bash Anaconda2-2.5.0-MacOSX-x86_64.sh
```

Python 3.5:

```
bash Anaconda3-2.5.0-MacOSX-x86_64.sh
```

NOTE: Include the "bash" command even if you are not using the bash shell.

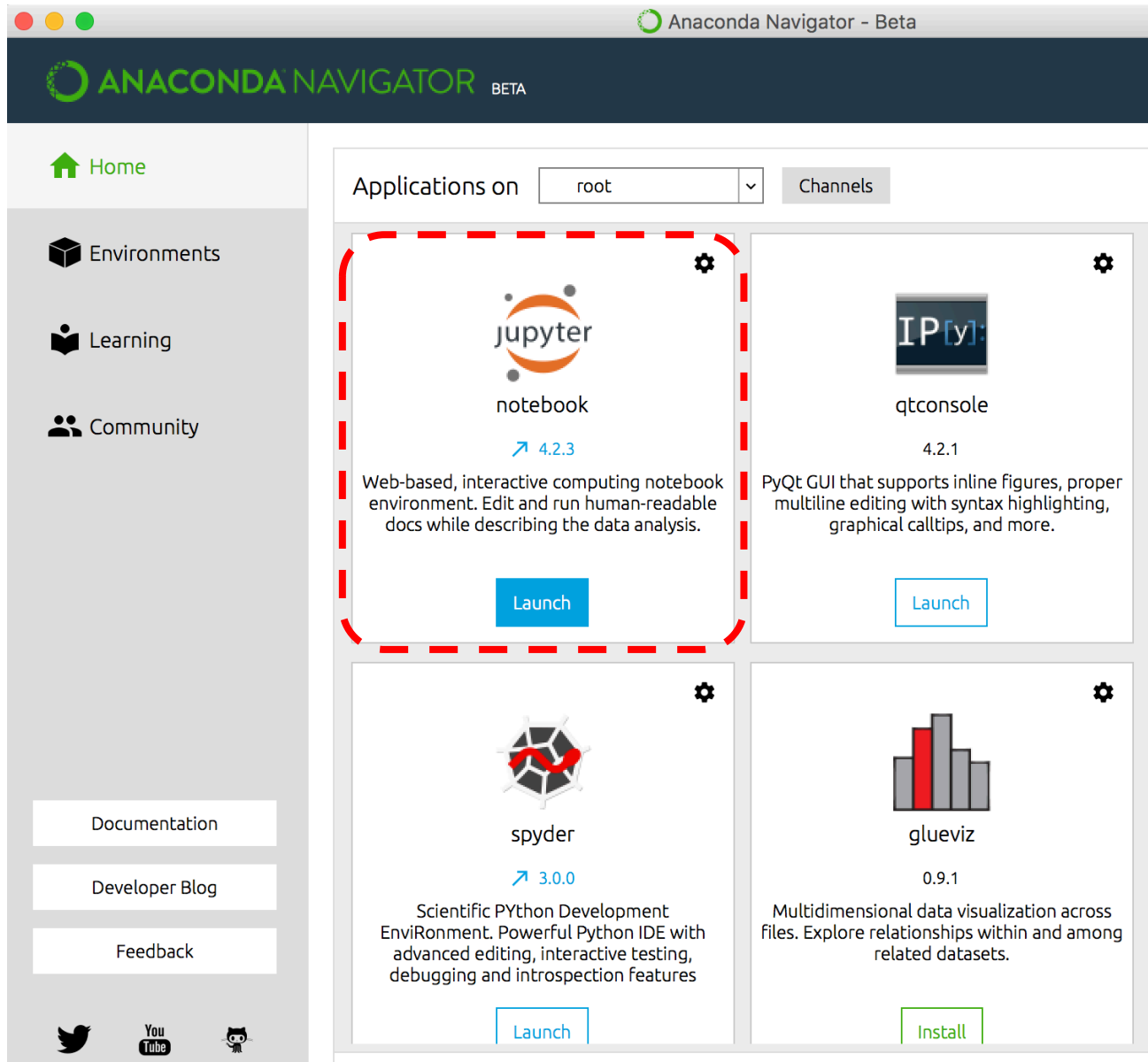
3. Optional: [Verify data integrity with MD5.](#)

<https://www.continuum.io/downloads>

Anaconda-Navigator



Jupyter notebook



Python versions (py2 and py3)

- Python 0.9.0 released in 1991 (first release)
- Python 1.0 released in 1994
- Python 2.0 released in 2000
- Python 2.6 released in 2008
- **Python 2.7 released in 2010**
- Python 3.0 released in 2008
- Python 3.3 released in 2010
- Python 3.4 released in 2014
- Python 3.5 released in 2015

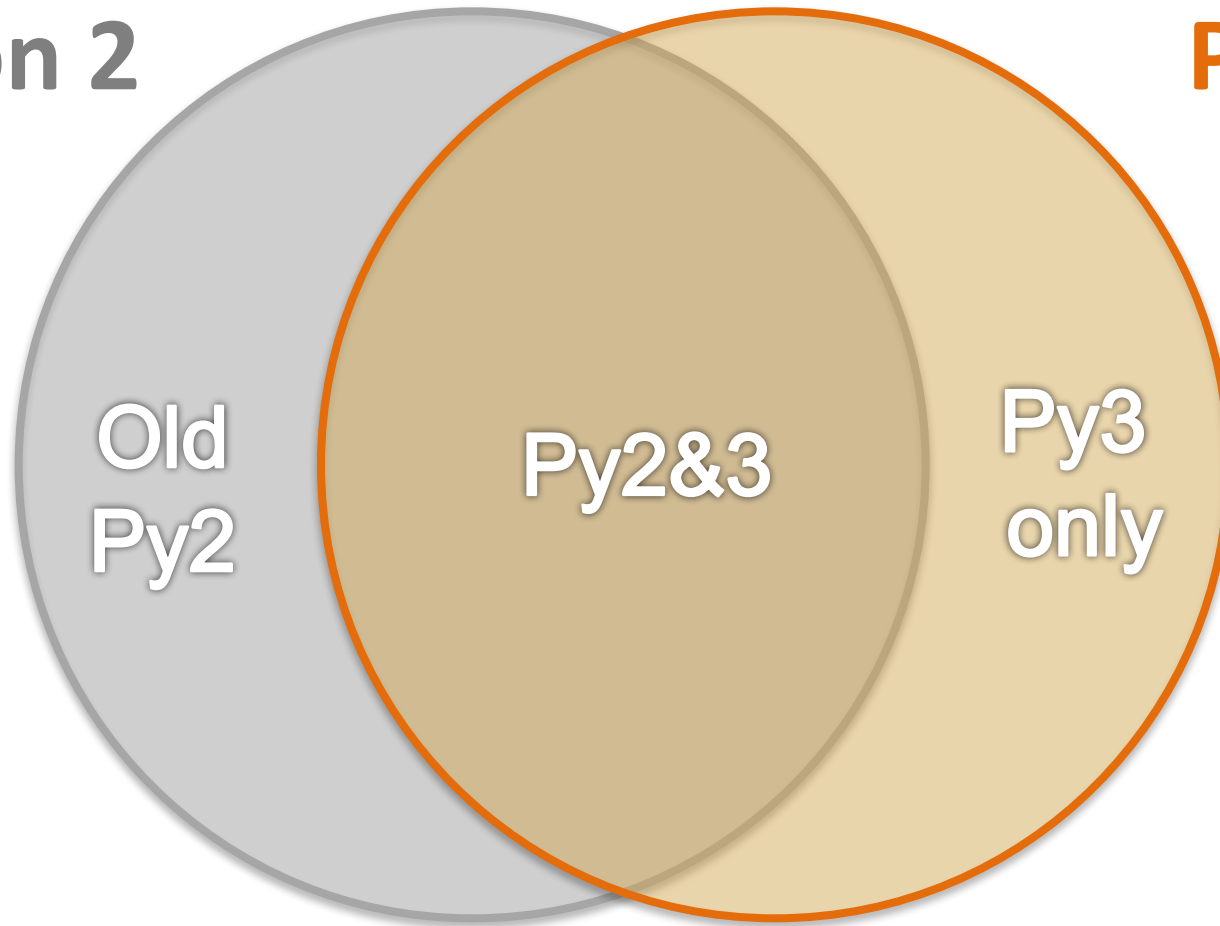
Python (Python 2.7 & Python 3.5)



Standard Syntax

Python 2

Python 3



Source: PyCon Australia (2014), Writing Python 2/3 compatible code by Edward Schofield

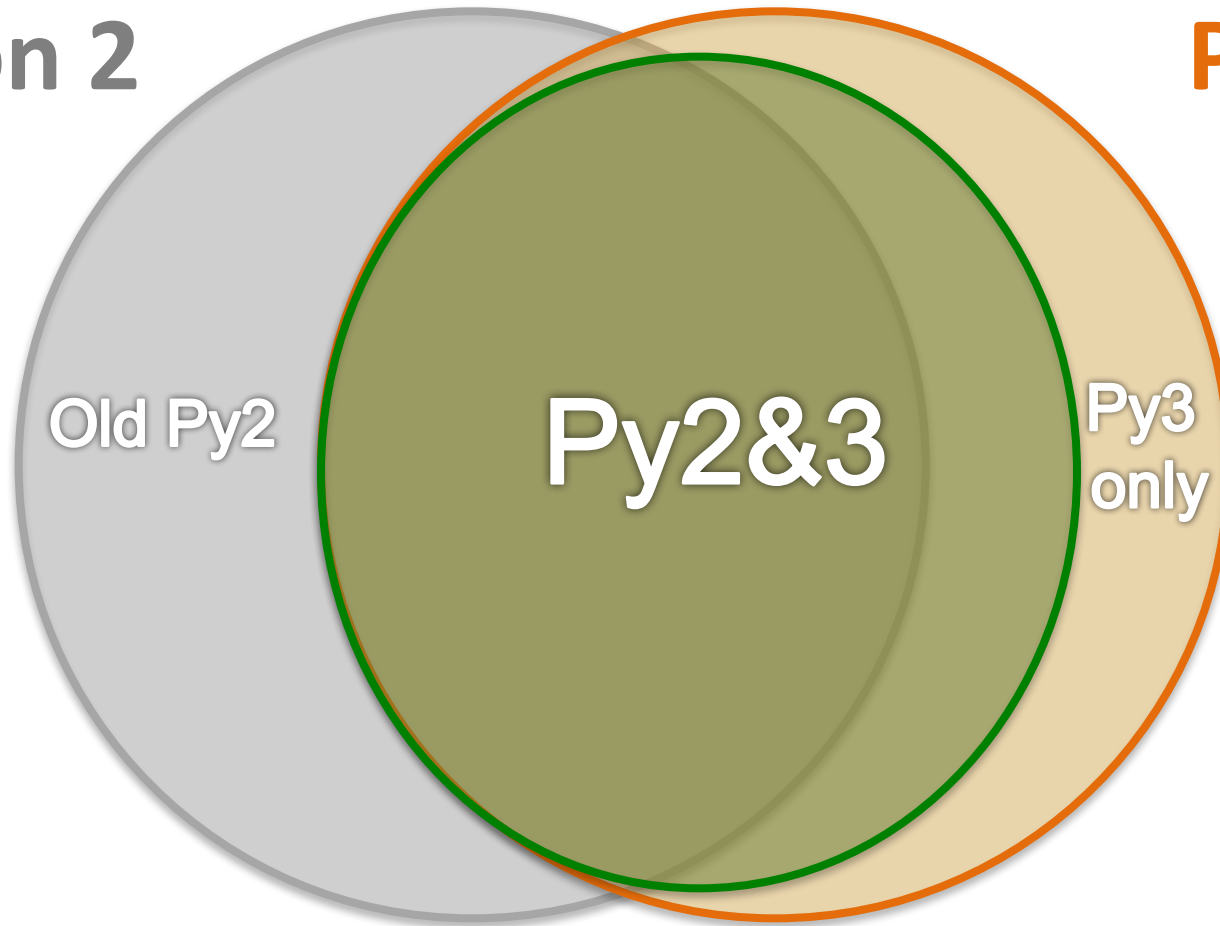
<https://www.youtube.com/watch?v=KOqk8j11aAI>

```
from __future__ import ...
```



Python 2

Python 3



Source: PyCon Australia (2014), Writing Python 2/3 compatible code by Edward Schofield

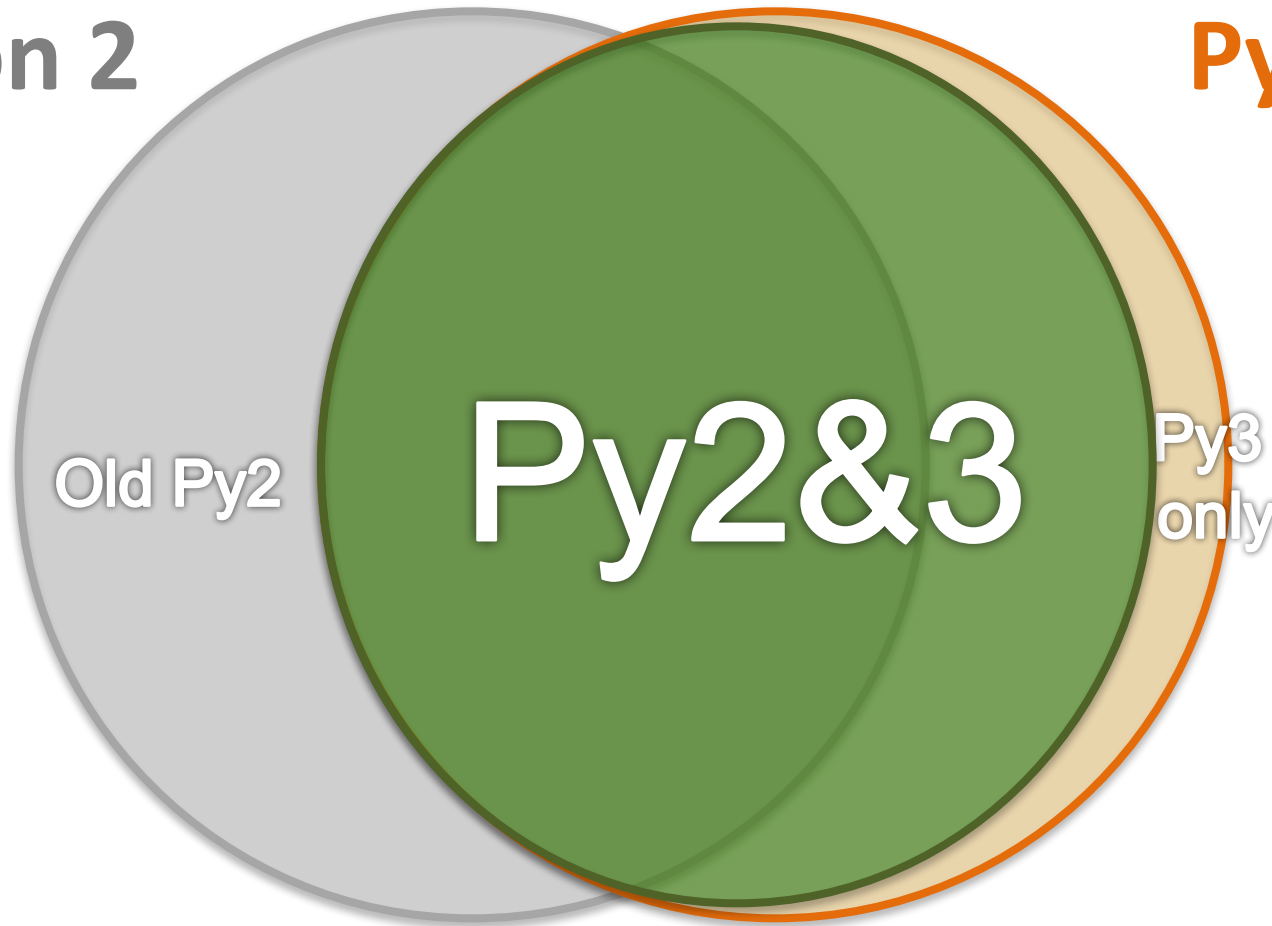
<https://www.youtube.com/watch?v=KOqk8j11aAI>

```
from future.builtins import *
```



Python 2

Python 3



Source: PyCon Australia (2014), Writing Python 2/3 compatible code by Edward Schofield

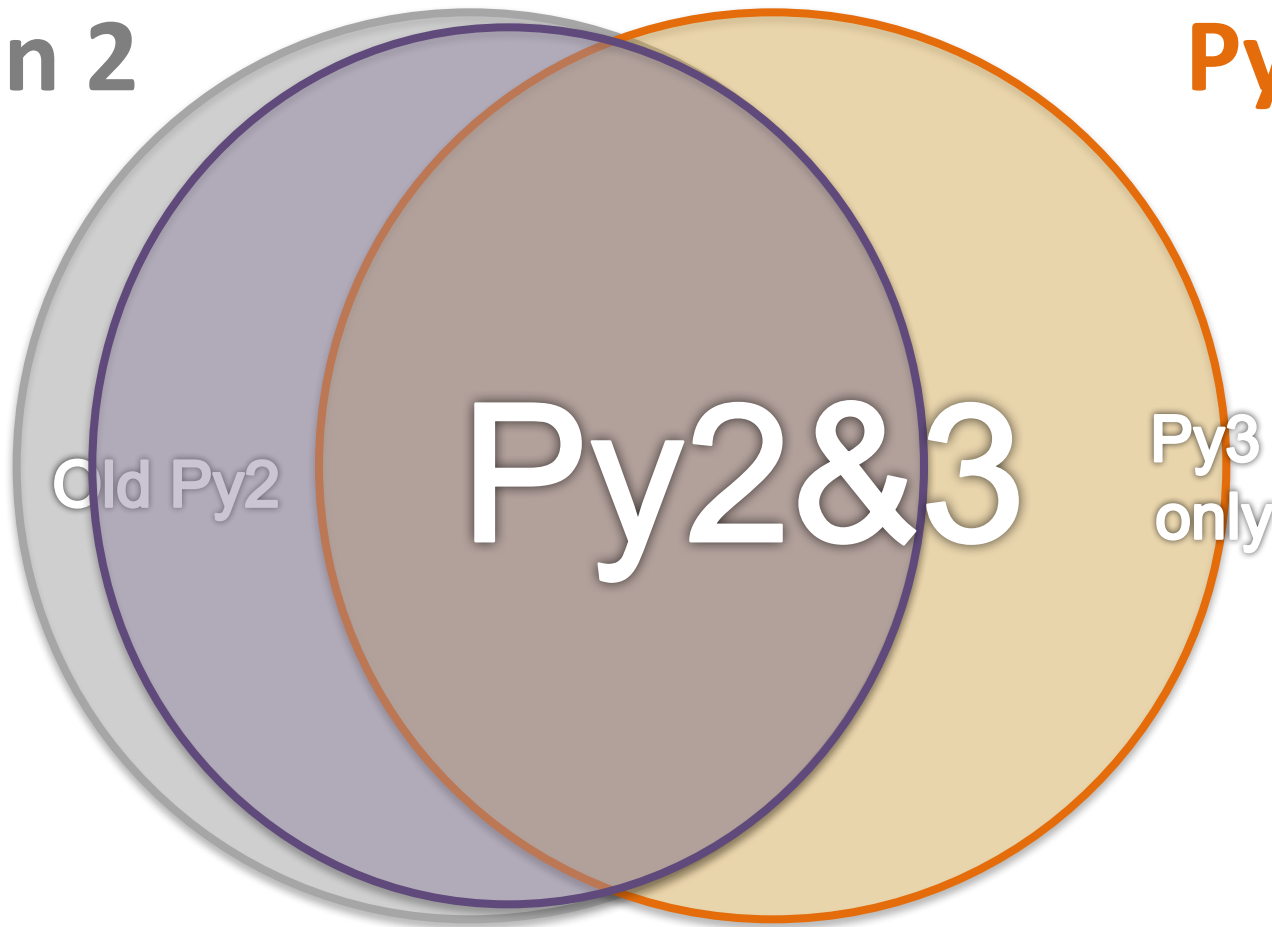
<https://www.youtube.com/watch?v=KOqk8j11aAI>

```
from past.builtins import *
```



Python 2

Python 3



Source: PyCon Australia (2014), Writing Python 2/3 compatible code by Edward Schofield

<https://www.youtube.com/watch?v=KOqk8j11aAI>

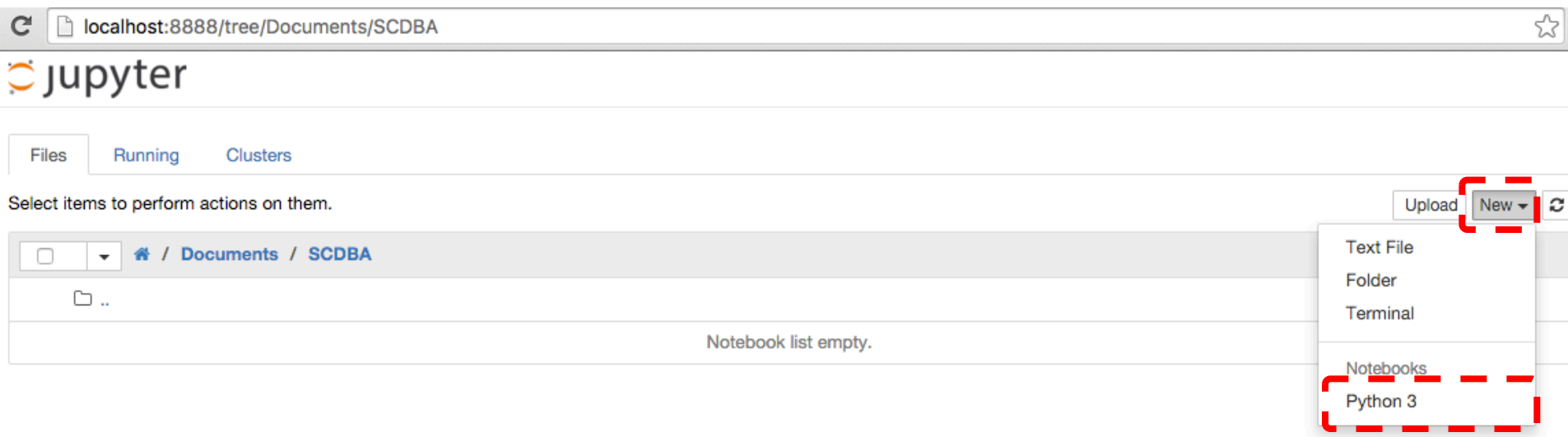
ipython notebook

ipython notebook

```
imyday — python • python.app //anaconda/bin/ipython notebook — 80x24
[iMydaytekiMacBook-Pro:~ imyday$ ipython notebook]
[I 14:26:49.944 NotebookApp] Serving notebooks from local directory: /Users/imyday
ay
[I 14:26:49.944 NotebookApp] 0 active kernels
[I 14:26:49.944 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
[I 14:26:49.944 NotebookApp] Use Control-C to stop this server and shut down all
kernels (twice to skip confirmation).
[W 14:26:56.639 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c
960677/channels?session_id=265FB16817FB4AB79202F6D3C3BDB0E6 (::1): Kernel does n
ot exist: a87ab95b-6d6e-44d3-aaa7-c1901c960677
[W 14:26:56.663 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c
960677/channels?session_id=265FB16817FB4AB79202F6D3C3BDB0E6 (::1) 95.43ms refere
r=None
[W 14:26:56.681 NotebookApp] 404 GET /api/kernels/b7fae9a6-d77b-4ead-832c-c070b1
8d642b/channels?session_id=EF4C761633E541C88568CDBCDE1091B7 (::1): Kernel does n
ot exist: b7fae9a6-d77b-4ead-832c-c070b18d642b
[W 14:26:56.683 NotebookApp] 404 GET /api/kernels/b7fae9a6-d77b-4ead-832c-c070b1
8d642b/channels?session_id=EF4C761633E541C88568CDBCDE1091B7 (::1) 6.62ms referer
=None
[W 14:27:29.595 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c
960677/channels?session_id=265FB16817FB4AB79202F6D3C3BDB0E6 (::1): Kernel does n
ot exist: a87ab95b-6d6e-44d3-aaa7-c1901c960677
[W 14:27:29.631 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c
```

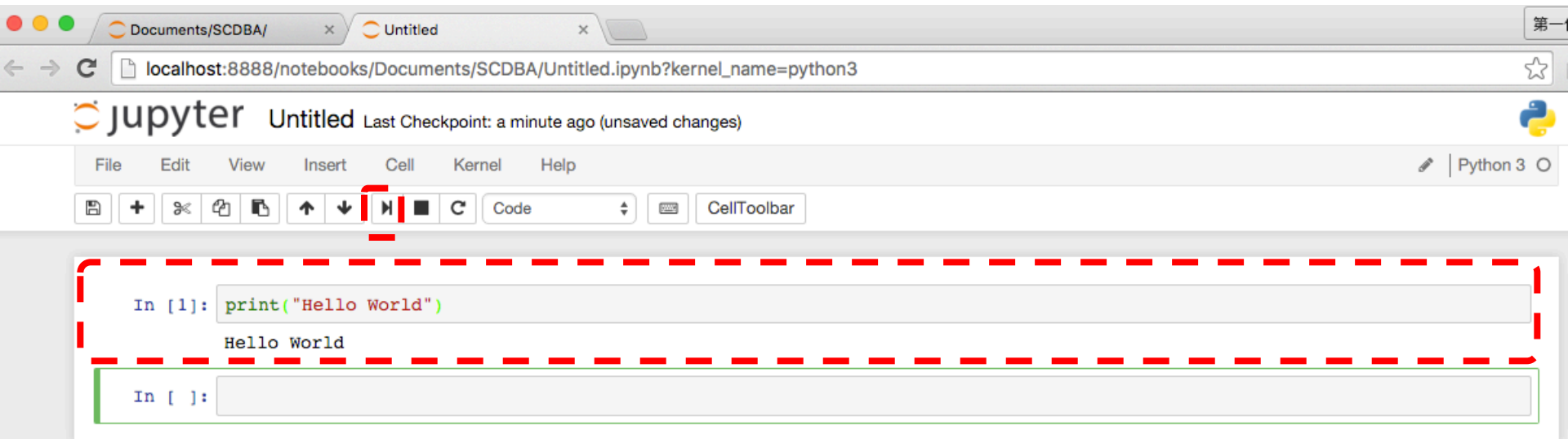
jupyter notebook

Python 3



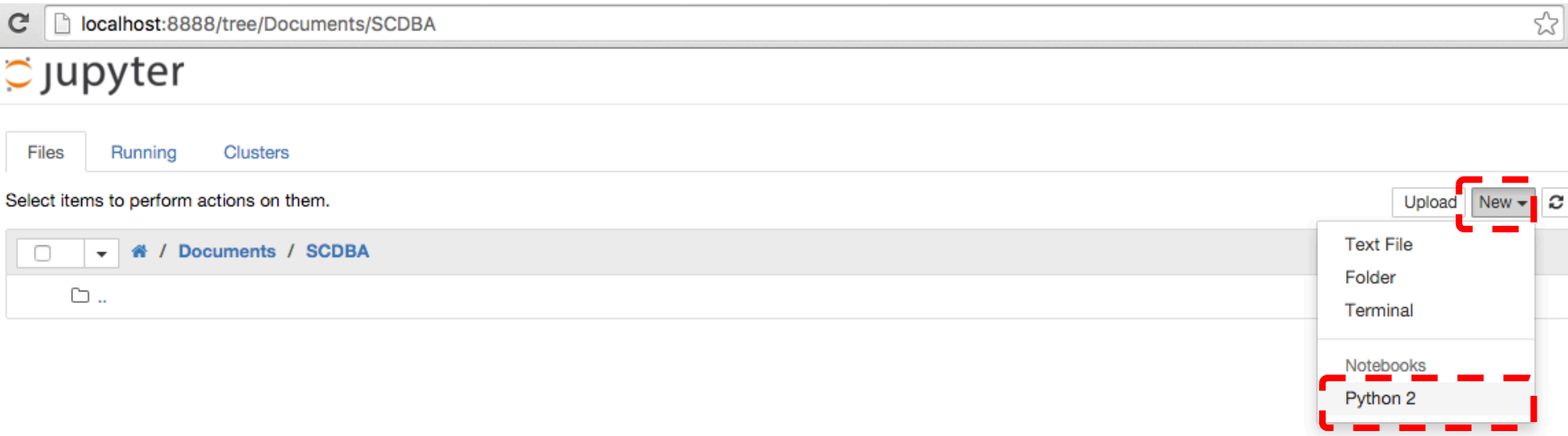
jupyter notebook

Python 3



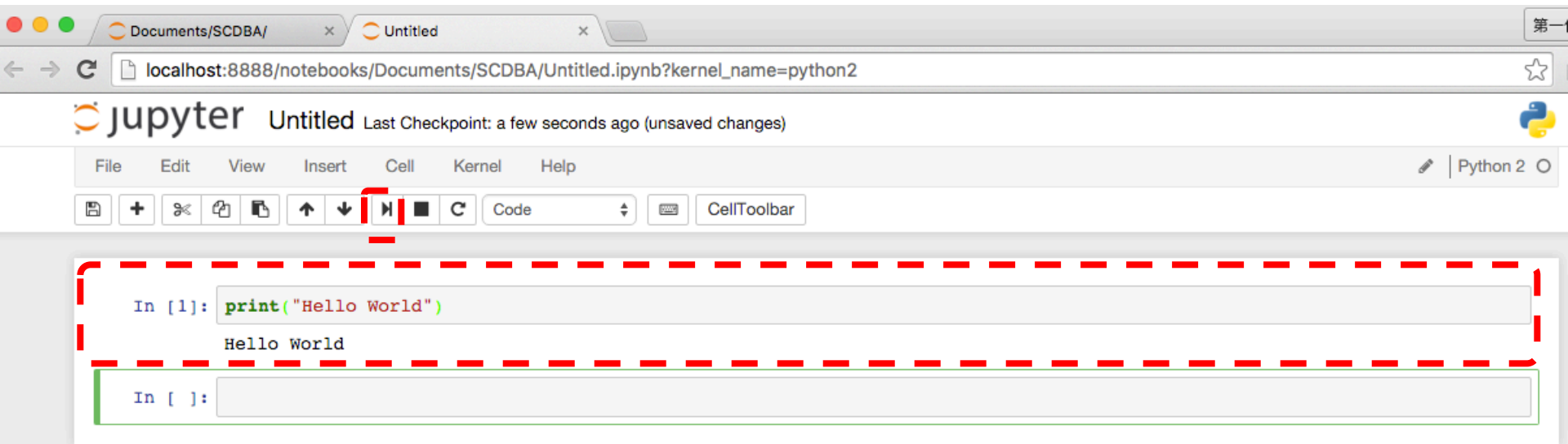
jupyter notebook

Python 2



jupyter notebook


Python 2



ipython notebook


jupyter notebook

← → ↺ localhost:8888/tree ☆

 jupyter

Files Running Clusters

Select items to perform actions on them. Upload New ▾ ↻

<input type="checkbox"/>	<input type="button" value="v"/>	
<input type="checkbox"/>		AndroidStudioProjects
<input type="checkbox"/>		app
<input type="checkbox"/>		Applications
<input type="checkbox"/>		AppsPro
<input type="checkbox"/>		bin
<input type="checkbox"/>		Desktop
<input type="checkbox"/>		Development
<input type="checkbox"/>		Documents
<input type="checkbox"/>		Downloads
<input type="checkbox"/>		Dropbox
<input type="checkbox"/>		imtkuapp5
<input type="checkbox"/>		jEdit
<input type="checkbox"/>		man
<input type="checkbox"/>		Movies
<input type="checkbox"/>		Music
<input type="checkbox"/>		OneDrive
<input type="checkbox"/>		Pictures

jupyter notebook

The screenshot shows the Jupyter Notebook web interface in a browser window. The address bar shows `localhost:8888/tree`. The interface has tabs for **Files**, **Running**, and **Clusters**. Below the tabs, it says "Select items to perform actions on them." There is a list of files and folders, each with a checkbox and a folder icon. The files are: `AndroidStudioProjects`, `app`, `Applications`, `AppsPro`, `bin`, `Desktop`, `Development`, `Documents`, `Downloads`, `Dropbox`, `imtkuapp5`, `jEdit`, `man`, `Movies`, `Music`, `OneDrive`, and `Pictures`. On the right side, 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 **Python 2** option is highlighted. Below the dropdown menu, there is a tooltip that says "Create a new notebook with Python 2". At the bottom left, there is a status bar showing `localhost:8888/tree#`.

Home x

localhost:8888/tree

jupyter

Files Running Clusters

Select items to perform actions on them.

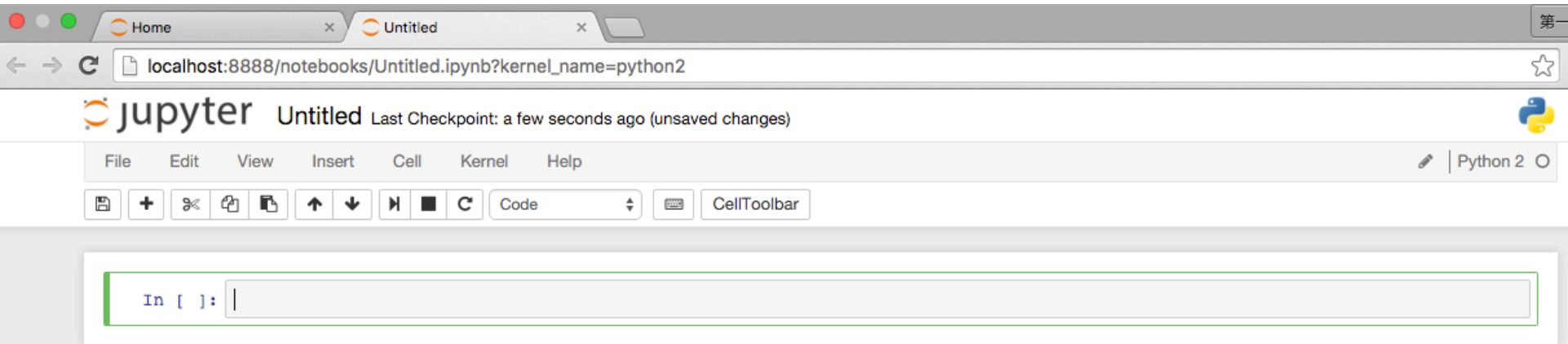
Upload New

Text File
Folder
Terminal
Notebooks
Python 2

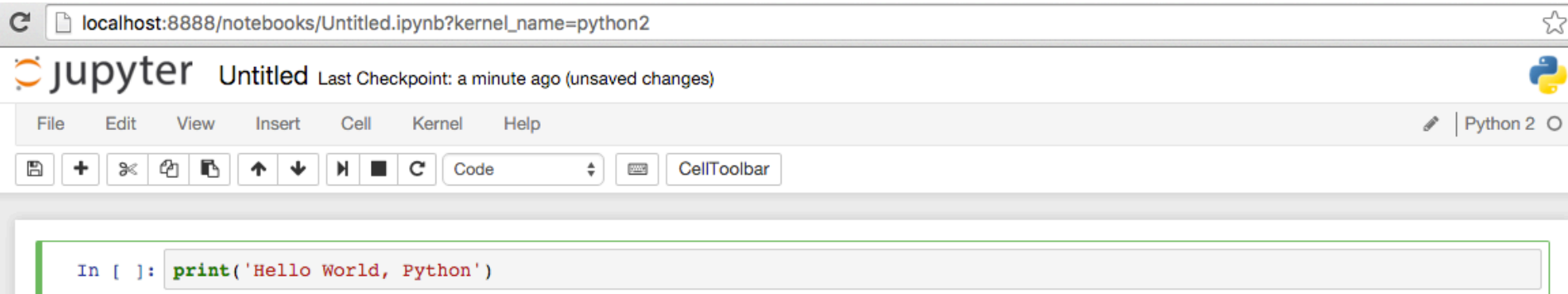
Create a new notebook with Python 2

localhost:8888/tree#

jupyter notebook



```
print('Hello World, Python')
```



The screenshot shows a web browser window with the address bar displaying `localhost:8888/notebooks/Untitled.ipynb?kernel_name=python2`. The Jupyter interface includes a header with the Jupyter logo, the text "Untitled", and a status message "Last Checkpoint: a minute ago (unsaved changes)". Below the header is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, and Help. To the right of the menu bar is a toolbar with icons for saving, adding, deleting, and other actions, along with a dropdown menu currently set to "Code". The main area of the notebook contains a single code cell with the prompt `In []:` followed by the code `print('Hello World, Python')`. The code is syntax-highlighted, with `print` in green, the string in red, and the parentheses and comma in black.

```
print('Hello World, Python')
```

localhost:8888/notebooks/Untitled.ipynb?kernel_name=python2

jupyter Untitled Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python 2

Code CellToolbar

```
In [1]: print('Hello World, Python')
```

Hello World, Python

```
In [ ]:
```

print

```
# Python 2 only:  
print 'Hello'
```

```
# Python 2 and 3:  
print('Hello')
```

```
# Python 2 only:  
print 'Hello', 'Guido'
```

```
# Python 2 and 3:  
from __future__ import print_function #(at top of module)  
  
print('Hello', 'Guido')
```

Writing Python 2-3 compatible code

Essential syntax differences

print

```
# Python 2 only:  
print 'Hello'
```

```
# Python 2 and 3:  
print('Hello')
```

To print multiple strings, import `print_function` to prevent Py2 from interpreting it as a tuple:

```
# Python 2 only:  
print 'Hello', 'Guido'
```

```
# Python 2 and 3:  
from __future__ import print_function    # (at top of module)  
  
print('Hello', 'Guido')
```

Unicode (text) string literals

```
# Python 2 only  
s1 = 'The Zen of Python'  
s2 = u'きたないのよりきれいな方がいい\n'
```

```
# Python 2 and 3  
s1 = u'The Zen of Python'  
s2 = u'きたないのよりきれいな方がいい\n'
```

Unicode (text) string literals

```
# Python 2 and 3
from __future__ import unicode_literals # at top of module

s1 = 'The Zen of Python'
s2 = 'きたないのよりきれいな方がいい\n'
```



Text input and output

```
print("Hello World")
```

```
print("Hello World\nThis is a message")
```

```
x = 3  
print(x)
```

```
x = 2  
y = 3  
print(x, ' ', y)
```

```
name = input("Enter a name: ")
```

```
x = int(input("What is x? "))
```

```
x = float(input("Write a number"))
```

Variables

```
x = 2
```

```
price = 2.5
```

```
word = 'Hello'
```

```
word = 'Hello'
```

```
word = "Hello"
```

```
word = '''Hello'''
```

```
x = 2
```

```
x = x + 1
```

```
x = 5
```

Python Basic Operators

```
print('7 + 2 =', 7 + 2)
print('7 - 2 =', 7 - 2)
print('7 * 2 =', 7 * 2)
print('7 / 2 =', 7 / 2)
print('7 // 2 =', 7 // 2)
print('7 % 2 =', 7 % 2)
print('7 ** 2 =', 7 ** 2)
```

```
print('7 + 2 =', 7 + 2)
print('7 - 2 =', 7 - 2)
print('7 * 2 =', 7 * 2)
print('7 / 2 =', 7 / 2)
print('7 // 2 =', 7 // 2)
print('7 % 2 =', 7 % 2)
print('7 ** 2 =', 7 ** 2)
```

```
7 + 2 = 9
7 - 2 = 5
7 * 2 = 14
7 / 2 = 3.5
7 // 2 = 3
7 % 2 = 1
7 ** 2 = 49
```

BMI Calculator in Python

```
height_cm = float(input("Enter your height in cm: "))  
weight_kg = float(input("Enter your weight in kg: "))  
  
height_m = height_cm/100  
BMI = (weight_kg/(height_m**2))  
  
print("Your BMI is: " + str(round(BMI,1)))
```

If statements

> greater than
< smaller than
== equals
!= is not

```
score = 80
if score >=60 :
    print( "Pass" )
else:
    print( "Fail" )
```

For loops

```
for i in range(1,11):  
    print(i)
```

```
1  
2  
3  
4  
5  
6  
7  
8  
9  
10
```

For loops

```
for i in range(1,10):  
    for j in range(1,10):  
        print(i, ' * ', j, ' = ', i*j)
```

```
9 * 1 = 9  
9 * 2 = 18  
9 * 3 = 27  
9 * 4 = 36  
9 * 5 = 45  
9 * 6 = 54  
9 * 7 = 63  
9 * 8 = 72  
9 * 9 = 81
```

Functions

```
def convertCMtoM(xcm):  
    m = xcm/100  
    return m
```

```
cm = 180  
m = convertCMtoM(cm)  
print(str(m))
```

1.8

Lists

```
x = [60, 70, 80, 90]  
print(len(x))  
print(x[0])  
print(x[1])  
print(x[-1])
```

60

70

90

Tuples

A **tuple** in Python is a collection that **cannot be modified**.

A tuple is defined using **parenthesis**.

```
x = (10, 20, 30, 40, 50)
```

```
print(x[0])
```

```
print(x[1])
```

```
print(x[2])
```

```
print(x[-1])
```

10

20

30

50

Python Ecosystem

Python Ecosystem

import math

```
x = log(1)
print(x)
```

```
-----
NameError                                Traceback (most recent call last)
<ipython-input-64-55d85b4998db> in <module>()
----> 1 x = log(1)
      2 print(x)

NameError: name 'log' is not defined
```

math.log?

```
import math
x = math.log(1)
print(x)
```

0.0

```
math.log(8, 2)|
```

3.0

```
Docstring:
log(x[, base])
```

```
Return the logarithm of x to the given base.
If the base not specified, returns the natural logarithm (base e) of x.
Type:      builtin_function_or_method
```

NumPy

- NumPy provides a **multidimensional array** object to store homogenous or heterogeneous data; it also provides **optimized functions/methods** to operate on this array object.

NumPy

```
v = range(1, 6)  
print(v)  
2 * v  
  
import numpy as np  
v = np.arange(1, 6)  
  
v  
  
2 * v
```

```
v = range (1, 6)
print(v)
```

```
[1, 2, 3, 4, 5]
```

```
2 * v
```

```
[1, 2, 3, 4, 5, 1, 2, 3, 4, 5]
```

```
import numpy as np
v = np.arange(1, 6)
v
```

```
array([1, 2, 3, 4, 5])
```

```
2 * v
```

```
array([ 2,  4,  6,  8, 10])
```

Compatible Python 2 and Python 3 Code

- `print()`
- Exceptions
- Division
- Unicode strings
- Bad imports

Compatible Python 2 and Python 3 Code

```
print()  
print("This works in py2 and py3")  
  
from __future__ import print_function  
print("Hello", "World")
```

File IO with open()

```
# Python 2 only
f = open('myfile.txt')
data = f.read()                # as a byte string
text = data.decode('utf-8')
```



```
# Python 2 and 3: alternative 1
from io import open
f = open('myfile.txt', 'rb')
data = f.read()                # as bytes
text = data.decode('utf-8')    # unicode, not bytes
```



```
# Python 2 and 3: alternative 2
from io import open
f = open('myfile.txt', encoding='utf-8')
text = f.read()                # unicode, not bytes
```

Six: Python 2 and 3 Compatibility Library

[←](#) [→](#) [↺](#) [https://pythonhosted.org/six/](#) [☆](#) [🌐](#) [☰](#)

six 1.10.0 documentation » [modules](#) | [index](#)

Table Of Contents

- Six: Python 2 and 3 Compatibility Library
 - Indices and tables
 - Package contents
 - Constants
 - Object model compatibility
 - Syntax compatibility
 - Binary and text data
 - unittest assertions
 - Renamed modules and attributes compatibility
 - urllib parse
 - urllib error
 - urllib request
 - urllib response
 - Advanced - Customizing renames

This Page

[Show Source](#)

Quick search

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

Six: Python 2 and 3 Compatibility Library

Six provides simple utilities for wrapping over differences between Python 2 and Python 3. It is intended to support codebases that work on both Python 2 and 3 without modification. six consists of only one Python file, so it is painless to copy into a project.

Six can be downloaded on [PyPi](#). Its bug tracker and code hosting is on [BitBucket](#).

The name, “six”, comes from the fact that 2*3 equals 6. Why not addition? Multiplication is more powerful, and, anyway, “five” has already been snatched away by the (admittedly now moribund) Zope Five project.

Indices and tables

- [Index](#)
- [Search Page](#)

Package contents

six.PY2
A boolean indicating if the code is running on Python 2.

six.PY3
A boolean indicating if the code is running on Python 3.

Constants

Six provides constants that may differ between Python versions. Ones ending `_types` are mostly useful as the second argument to `isinstance` or `issubclass`.

six.class_types
Possible class types. In Python 2, this encompasses old-style and new-style classes. In Python 3, this is just new-styles.

Conda Test Drive

← → ↺

conda.pydata.org/docs/test-drive.html

☆

🏠 Conda

Search docs

Get started

Intro to conda

Download conda

Installation

Test drive

Conda cheat sheet

Using conda

Building packages

Help & reference

Get involved

Docs » Get started » Test drive

Edit on GitHub

Test drive

To start the conda 30-minute test drive, you should have already followed our 2-minute *Quick install* guide to download, install and update Miniconda, OR have downloaded, installed and updated Anaconda or Miniconda on your own.

NOTE: After installing, be sure you have closed and then re-opened the terminal window so the changes can take effect.

Conda test drive milestones:

1. **USING CONDA.** First we will verify that you have installed Anaconda or Miniconda, and check that it is updated to the current version. 3 min.
2. **MANAGING ENVIRONMENTS.** Next we will play with environments by creating a few environments, so you can learn to move easily between the environments. We will also verify which environment you are in, and make an exact copy of an environment as a backup. 10 min.
3. **MANAGING PYTHON.** Then we will check to see which versions of Python are available to install, install another version of Python, and switch between versions. 4 min.
4. **MANAGING PACKAGES.** We play with packages. We will a) list packages installed on your computer, b) see a list of available packages, and c) install and remove some packages using conda install. For packages not available using conda install, we will d) search on Anaconda.org. For packages that are in neither location, we'll e) install a package with the pip package manager. We will also install a free 30 day trial of Continuum's commercial package IOPro. 10 min.
5. **REMOVING PACKAGES, ENVIRONMENTS, OR CONDA.** We'll end the test drive by removing

<http://conda.pydata.org/docs/test-drive.html>

Managing Conda and Anaconda

Managing conda and anaconda

conda info

Verify conda is installed, check version #

conda update conda

Update conda package and environment manager to current version

conda update anaconda

Update the anaconda meta package (the library of packages ready to install with **conda** command)

Managing environments

Managing environments

conda info --envs or **conda info -e** Get a list of all my environments, active environment shown with *

conda create --name snowflakes biopython Create an environment and install program(s)

or

conda create -n snowflakes biopython

***TIP:** To avoid dependency conflicts, install all programs in the environment (snowflakes) at the same time.*

***TIP:** Environments install by default into the envs directory in your conda directory. You can specify a different path; see **conda create --help** for details.*

source activate snowflakes (Linux, Mac)

Activate the new environment to use it

activate snowflakes (Windows)

***TIP:** Activate prepends the path to the snowflakes environment.*

conda create -n bunnies python=3.4 astroid Create a new environment, specify Python version

conda create -n flowers --clone snowflakes Make exact copy of an environment

conda remove -n flowers --all

Delete an environment

conda env export > puppies.yml

Save current environment to a file

conda env create -f puppies.yml

Load environment from a file

Managing Python

Managing Python

conda search --full-name python
or
conda search -f python

Check versions of Python available to install

conda create -n snakes python=3.4

Install different version of Python in new environment

source activate snakes (*Linux, Mac*)
activate snakes (*Windows*)

Switch to the new environment that has a different version of Python

TIP: *Activate prepends the path to the snakes environment.*

Managing Packages in Python

Managing packages, including Python

conda list

View list of packages and versions installed in active environment

conda search beautiful-soup

Search for a package to see if it is available to conda install

conda install -n bunnies beautiful-soup Install a new package

NOTE: If you do not include the name of the new environment (**-n bunnies**) it will install in the current active environment.

TIP: To view list of all packages available through **conda install**, visit <http://docs.continuum.io/anaconda/pkg-docs.html>.

conda update beautiful-soup

Update a package in the current environment

conda search --override-channels -c pandas bottleneck Search for a package in a specific location (i.e. the pandas channel on Anaconda.org)

NOTE: Or go to Anaconda.org in the browser and search by package name. This will show the specific channel (owner) through which it is available.

conda install -c pandas bottleneck Install a package from a specific channel

conda search --override-channels -c defaults beautiful-soup Search for a package to see if it is available from the Anaconda repository

source activate bunnies (Linux, Mac)
activate bunnies (Windows)
pip install see

Activate the environment where you want to install a package and install it with pip (included with Anaconda and Miniconda)

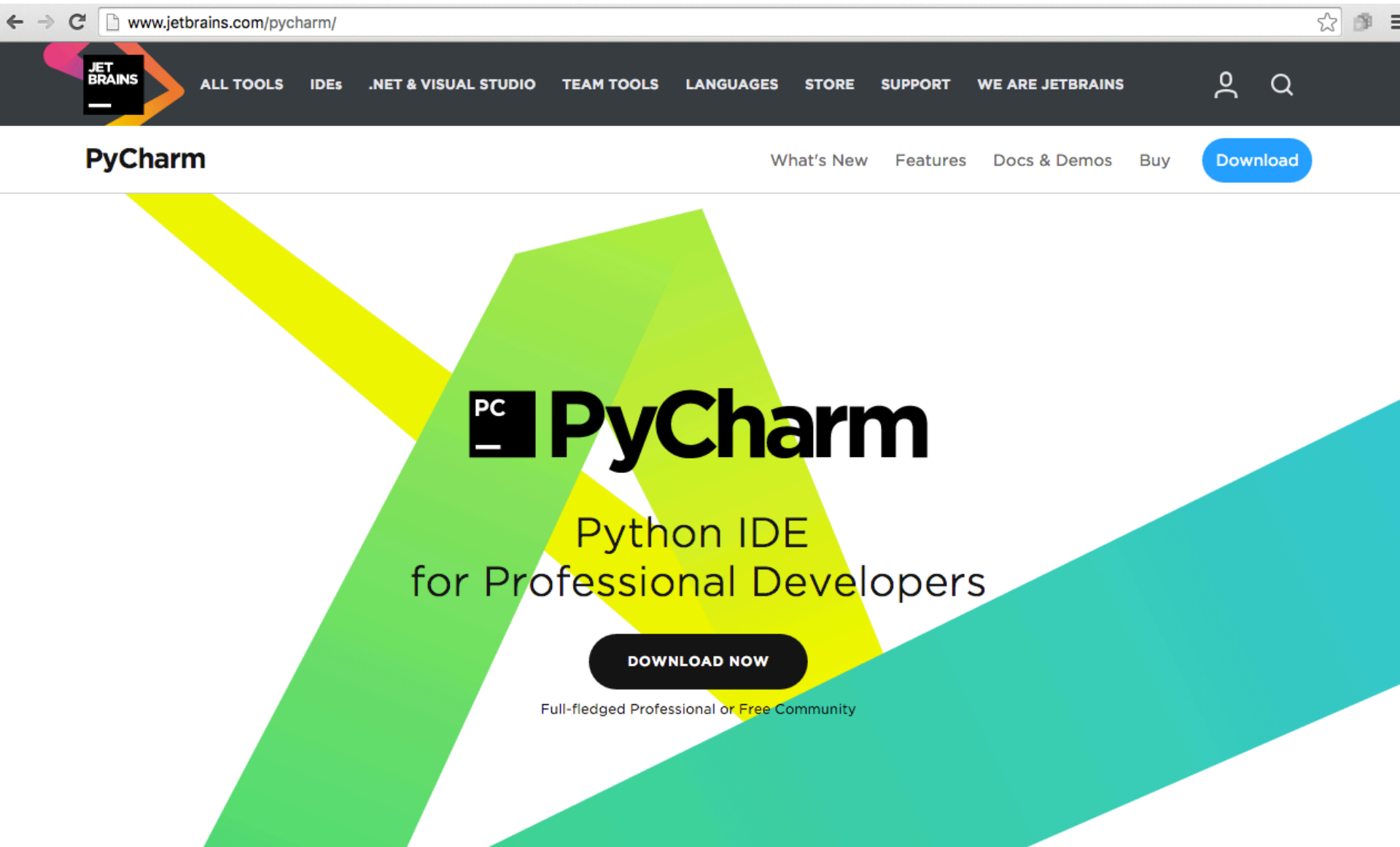
conda install iopro accelerate

Install commercial Continuum packages

conda skeleton pypi pyinstrument
conda build pyinstrument

Build a Conda package from a Python Package Index (PyPI) Package

PyCharm: Python IDE



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.)

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 screenshot shows the JupyterLab web interface in a browser. The address bar displays 'localhost:8888/tree#'. The Jupyter logo is visible at the top left. Below the logo, there are tabs for 'Files', 'Running', and 'Clusters'. The 'Files' tab is active, showing a file tree. A message 'Select items to perform actions on them.' is displayed above the file list. The file list contains several folders: Applications, Desktop, Development, Documents, and Downloads. On the right side of the file list, there are buttons for 'Upload', 'New', and a refresh icon. The 'New' button is highlighted with a red box, and its dropdown menu is open, showing options: 'Text File', 'Folder', 'Terminal', 'Notebooks', and 'Python 2'. The 'Terminal' option is 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
astropy                  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-webrtc	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
pycparser	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')`



<code>_Quotient</code>	<code>COOKIELIB</code>	<code>nis</code>	<code>cadnanny</code>
<code>_Qt</code>	<code>copy</code>	<code>nlTK</code>	<code>tarfile</code>
<code>_Res</code>	<code>copy_reg</code>	<code>nntplib</code>	<code>telnetlib</code>
<code>_Scrap</code>	<code>copyreg</code>	<code>nose</code>	<code>tempfile</code>
<code>_Snd</code>	<code>crypt</code>	<code>notebook</code>	<code>terminado</code>
<code>_TE</code>	<code>cryptography</code>	<code>ntpath</code>	<code>terminalcommand</code>
<code>_Win</code>	<code>csv</code>	<code>nturl2path</code>	<code>termios</code>
<code>__builtin__</code>	<code>ctypes</code>	<code>numba</code>	<code>test_path</code>
<code>__future__</code>	<code>curl</code>	<code>numbers</code>	<code>test_pycosat</code>
<code>_abcoll</code>	<code>curses</code>	<code>numexpr</code>	<code>tests</code>
<code>_ast</code>	<code>cycler</code>	<code>numpy</code>	<code>textwrap</code>
<code>_bisect</code>	<code>cython</code>	<code>odo</code>	<code>this</code>
<code>_builtinSuites</code>	<code>cythonmagic</code>	<code>opcode</code>	<code>thread</code>
<code>_cffi_backend</code>	<code>cytoolz</code>	<code>openpyxl</code>	<code>threading</code>
<code>_codecs</code>	<code>datashape</code>	<code>operator</code>	<code>time</code>
<code>_codecs_cn</code>	<code>datetime</code>	<code>optparse</code>	<code>timeit</code>
<code>_codecs_hk</code>	<code>dateutil</code>	<code>os</code>	<code>tkColorChooser</code>
<code>_codecs_iso2022</code>	<code>dbhash</code>	<code>os2emxpath</code>	<code>tkCommonDialog</code>
<code>_codecs_jp</code>	<code>dbm</code>	<code>osax</code>	<code>tkFileDialog</code>
<code>_codecs_kr</code>	<code>decimal</code>	<code>pandas</code>	<code>tkFont</code>
<code>_codecs_tw</code>	<code>decorator</code>	<code>parser</code>	<code>tkMessageBox</code>

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 browser window at localhost:8888/notebooks/TextMiningNLP.ipynb. The notebook title is 'TextMiningNLP' and it shows 'Last Checkpoint: an hour ago (autosaved)'. The menu bar includes File, Edit, View, Insert, Cell, Kernel, and Help. The toolbar shows icons for saving, adding cells, undo, redo, and running code. The code cell contains the following Python code:

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

Below the code cell, the 'NLTK Downloader' window is open. It has tabs for Collections, Corpora, Models, and All Packages. The 'All Packages' tab is selected, showing a table with the following data:

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 buttons for 'Download' and 'Refresh'. Below these buttons, the 'Server Index' is set to http://www.nltk.org/nltk_data/ and the 'Download Directory' is set to `/Users/imyday/nltk_data`.

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
```

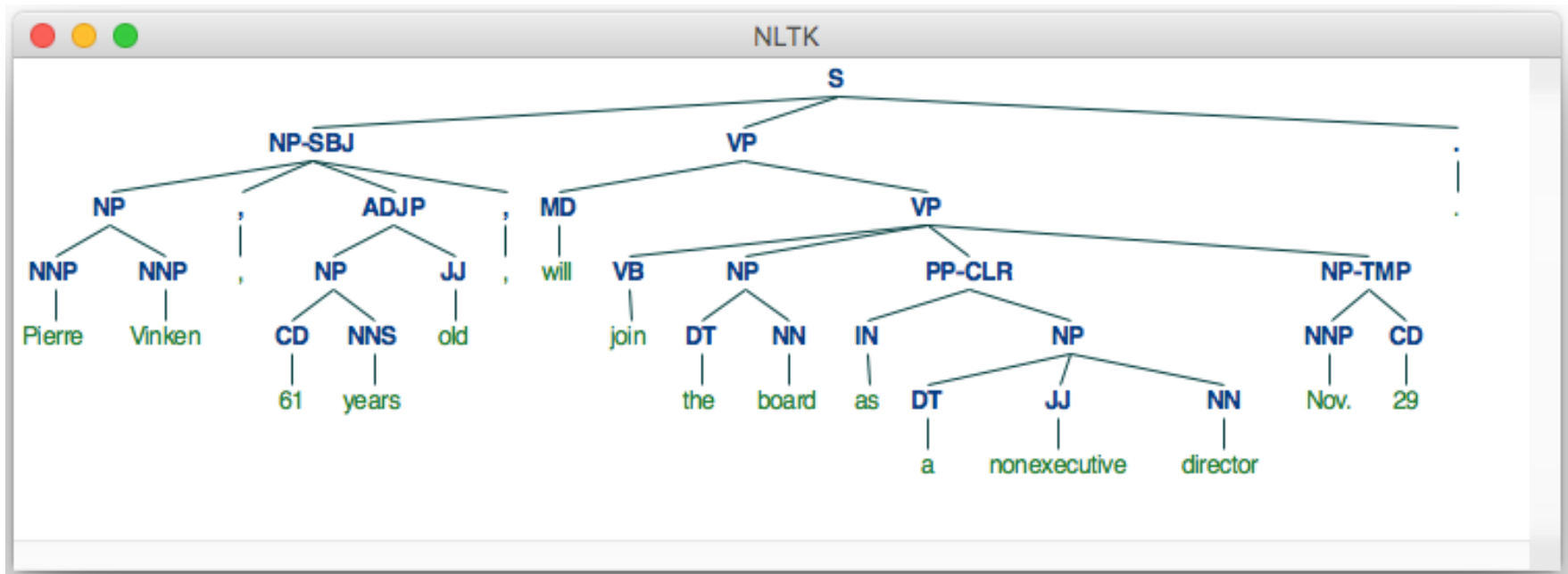
```
entities = nltk.chunk.ne_chunk(tagged)
entities
```

```
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
```

Python Jieba “结巴” 中文分词

GitHub, Inc. [US] <https://github.com/fxsjy/jieba>



Personal Open source Business Explore

Pricing Blog Support

This repository

Search

Sign in

Sign up

fxsjy / jieba

Watch

761

★ Star

7,187

🍴 Fork

2,252

Code

Issues 226

Pull requests 14

Projects 0

Wiki

Pulse

Graphs

结巴中文分词

485 commits

2 branches

23 releases

31 contributors

MIT

Branch: master

New pull request

Find file

Clone or download



fxsjy committed on GitHub Merge pull request #382 from huntzhan/master

Latest commit 8ba26cf on Aug 5, 2016

extra_dict	update to v0.33	2 years ago
jieba	Bugfix for HMM=False in parallelism.	6 months ago
test	Bugfix for HMM=False in parallelism.	6 months ago
.gitattributes	first commit	4 years ago
.gitignore	update jieba3k	2 years ago
Changelog	version change 0.38	a year ago
LICENSE	add a license file	4 years ago
MANIFEST.in	include Changelog & README.md in the distribution package	4 years ago
README.md	Update README.md	8 months ago

<https://github.com/fxsjy/jieba>

Python Jieba “结巴” 中文分词

```
import jieba
import jieba.posseg as pseg
sentence = "銀行產業正在改變，金融機構欲挖角科技人才"
words = jieba.cut(sentence)
print(sentence)
print(" ".join(words))
wordspos = pseg.cut(sentence)
result = ''
for word, pos in wordspos:
    print(word + ' (' + pos + ')')
    result = result + ' ' + word + ' (' + pos + ')'
print(result.strip())
```

import jieba

words = jieba.cut(sentence)

```
import jieba
import jieba.posseg as pseg
sentence = "銀行產業正在改變，金融機構欲挖角科技人才"
words = jieba.cut(sentence)
print(sentence)
print(" ".join(words))    #銀行 產業 正在 改變 ， 金融 機構 欲 挖角 科技人才

wordspos = pseg.cut(sentence)
result = ''
for word, pos in wordspos:
    print(word + ' (' + pos + ')')
    result = result + ' ' + word + ' (' + pos + ') '
print(result.strip())    #銀行(n) 產業(n) 正在(t) 改變(v) ， (x) 金融(n) 機構(n) 欲(d) 挖角(n) 科技人才(n)
```

銀行產業正在改變，金融機構欲挖角科技人才

銀行 產業 正在 改變 ， 金融 機構 欲 挖角 科技人才

銀行 (n)

產業 (n)

正在 (t)

改變 (v)

， (x)

金融 (n)

機構 (n)

欲 (d)

挖角 (n)

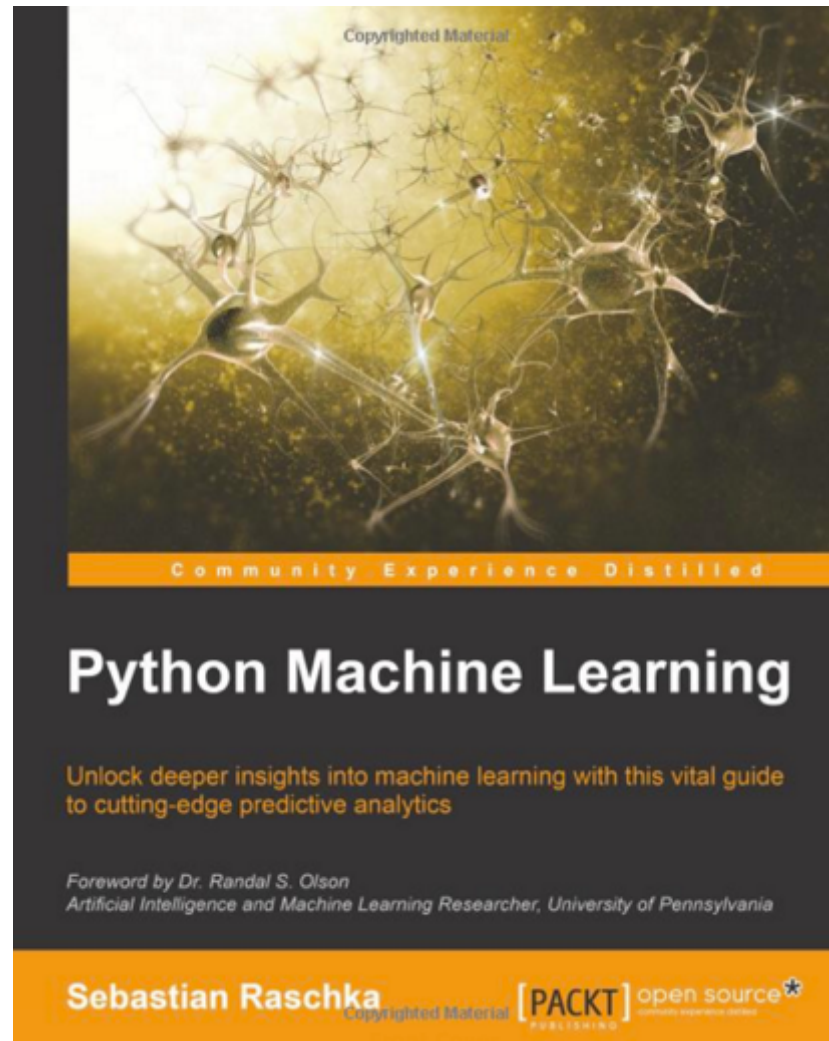
科技人才 (n)

銀行(n) 產業(n) 正在(t) 改變(v) ， (x) 金融(n) 機構(n) 欲(d) 挖角(n) 科技人才(n)

Python Jieba “结巴” 中文分词

- <https://github.com/fxsjy/jieba>
- `jieba.set_dictionary('data/dict.txt.big')`
 - `#/anaconda/lib/python3.5/site-packages/jieba`
 - `dict.txt` (5.4MB)(349,046)
 - `dict.txt.big.txt` (8.6MB)(584,429)
 - `dict.txt.small.txt` (1.6MB)(109,750)
 - `dict.tw.txt` (4.2MB)(308,431)
- https://github.com/ldkrshi/jieba-zh_TW
 - 结巴中文斷詞台灣繁體版本

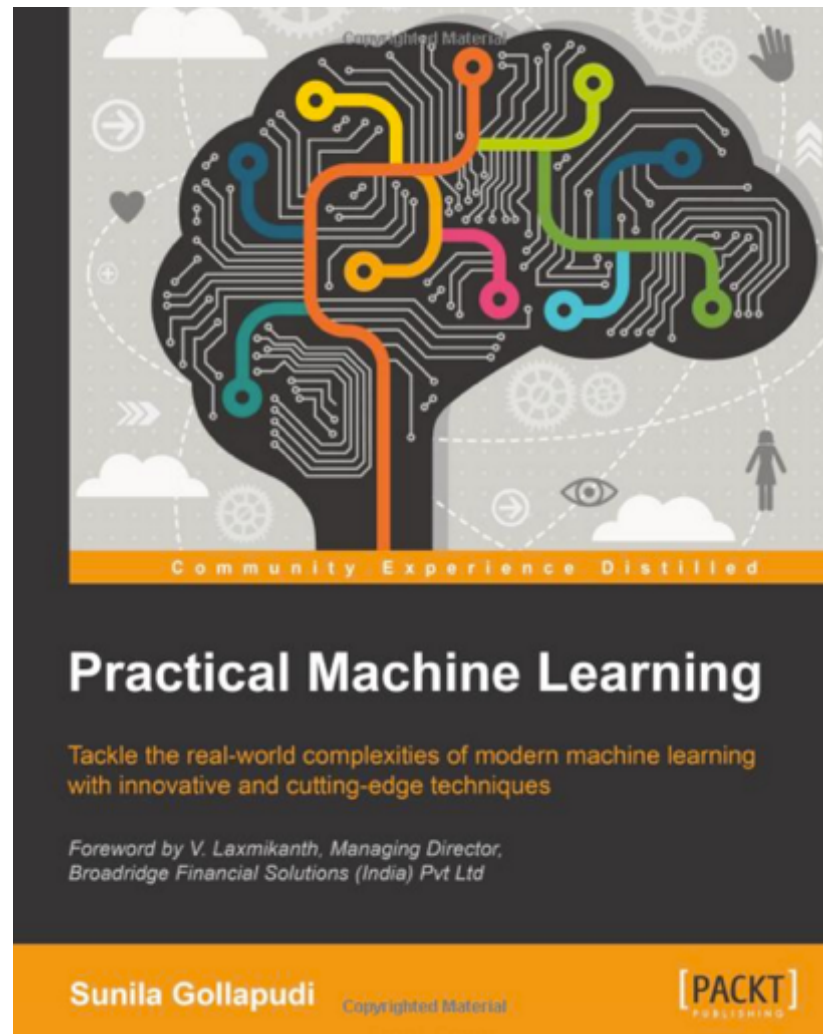
Sebastian Raschka (2015),
Python Machine Learning,
Packt Publishing



Sunila Gollapudi (2016),

Practical Machine Learning,

Packt Publishing



Machine Learning Models

Deep Learning

Kernel

Association rules

Ensemble

Decision tree

Dimensionality reduction

Clustering

Regression Analysis

Bayesian

Instance based

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/
- 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



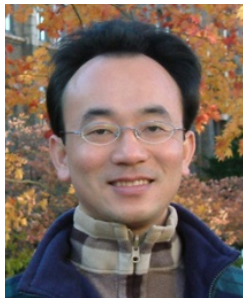
Q & A

Text Mining and Natural Language Processing (文字探勘與自然語言處理)

Time: 2017/01/23 (Mon) (14:00-17:00)

Place: 國立臺北護理健康大學 城區部 (台北市內江街89號) C302

Host: 祝國忠 院長 (健康科技學院院長)



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/>

2017-01-23

