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
Assistant Professor
Dept. of Information Management, Tamkang University

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

<table>
<thead>
<tr>
<th>Language Rank</th>
<th>Types</th>
<th>Spectrum Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. C</td>
<td>📱💻📱</td>
<td>100.0</td>
</tr>
<tr>
<td>2. Java</td>
<td>🌐📱💻</td>
<td>98.1</td>
</tr>
<tr>
<td>3. Python</td>
<td>🌐💻</td>
<td>98.0</td>
</tr>
<tr>
<td>4. C++</td>
<td>📱💻📱</td>
<td>95.9</td>
</tr>
<tr>
<td>5. R</td>
<td>📱💻</td>
<td>87.9</td>
</tr>
<tr>
<td>6. C#</td>
<td>🌐📱💻</td>
<td>86.7</td>
</tr>
<tr>
<td>7. PHP</td>
<td>🌐💻</td>
<td>82.8</td>
</tr>
<tr>
<td>8. JavaScript</td>
<td>🌐📱💻</td>
<td>82.2</td>
</tr>
<tr>
<td>9. Ruby</td>
<td>🌐💻</td>
<td>74.5</td>
</tr>
<tr>
<td>10. Go</td>
<td>🌐💻</td>
<td>71.9</td>
</tr>
</tbody>
</table>

Python: Analytics and Data Science Software

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

# List and the enumerate function
```python
>>> 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

Python is a programming language that lets you work quickly and integrate systems more effectively. Learn More
Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.

Source: https://www.python.org/doc/essays/blurb/
ANAconda

ANAconda GIVES SUPERPOWERS TO PEOPLE WHO CHANGE THE WORLD

Modern open source analytics platform powered by Python

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

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

http://www.amazon.com/Natural-Language-Processing-Python-Steven/dp/0596516495
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

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

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Volume

Scale of Data

40 ZETTAYES (43 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005

6 billion people have cell phones

2020

WORLD POPULATION: 7 BILLION

6.5 quintillion bytes (2.3 trillion gigabytes) of data are created each day

2005

Volume

Most companies in the U.S. have at least 100 terabytes (10,000 gigabytes) of data stored

Velocity

Analysis of Streaming Data

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

By 2016, it is projected there will be 18.9 billion network connections -- almost 2.5 connections per person on earth

Variety

Different Forms of Data

As of 2011, the global size of data in healthcare was estimated to be 150 Exabytes (160 billion gigabytes)

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

30 billion pieces of content are shared on Facebook every month

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

Veracity

Uncertainty of Data

1 in 3 business leaders don't trust the information they use to make decisions

27% of respondents in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around $3.1 trillion a year

Source: https://www-01.ibm.com/software/data/bigdata/
Value
Architecture of Big Data Analytics

**Big Data Sources**
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

**Big Data Transformation**
- Middleware
- Extract Transform Load
- Data Warehouse
- Traditional Format CSV, Tables

**Big Data Platforms & Tools**
- Hadoop
- MapReduce
- Pig
- Hive
- Jaql
- Zookeeper
- Hbase
- Cassandra
- Oozie
- Avro
- Mahout
- Others

**Big Data Analytics Applications**
- Queries
- Reports
- OLAP
- Data Mining

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Data Mining

Big Data Analytics Applications

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools

Big Data Analytics Applications
- Queries
- Reports
- OLAP
- Data Mining

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
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

Conceptual Layer
- Integrated analysis

Logical Layer
- Multivariate analysis
- Application specific task
- Data Mining

Physical Layer
- Software
- Social Data
- Hardware

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Deep Learning
Intelligence from Big Data

Source: https://www.vlab.org/events/deep-learning/
The Evolution of BI Capabilities

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

- Operational Data
- Historical Data
- Machine Data
- Web Data
- Audio/Video Data
- External Data

Data Mart

Casual users
- Queries
- Reports
- Dashboards

Power users
- Queries
- Reports
- OLAP
- Data mining

Hadoop Cluster

Data Warehouse

Operational Data

Extract, transform, load

Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine

Emotions

- Love
- Joy
- Surprise
- Anger
- Sadness
- Fear

Maslow’s Hierarchy of Needs

1. Physiological Needs (food, water, shelter)
2. Safety Needs (security, protection)
3. Social Needs (sense of belonging, love)
4. Esteem Needs (self-esteem, recognition, status)
5. Self-actualization Needs (self-development and realization)

Social Media Hierarchy of Needs

Source: http://www.pinterest.com/pin/18647785930903585/
The Social Feedback Cycle
Consumer Behavior on Social Media

Marketer-Generated

User-Generated

Awareness
Consideration
Purchase
Use
Form Opinion
Talk

Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement
The New Customer Influence Path

Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement
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 ____ a few days ago.
(2) It was such a ____ phone.
(3) The ____ screen was really ____.
(4) The ____ quality was ____ 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 ____ and wanted me to return it to the shop. ...”

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

http://www.amazon.com/Practical-Text-Analytics-Interpreting-Unstructured/dp/0749474017
Text Mining Concepts

• 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)

• Unstructured corporate data is doubling in size every 18 months

• Tapping into these information sources is not an option, but a need to stay competitive

• Answer: text mining
  – A semi-automated process of extracting knowledge from unstructured data sources
  – a.k.a. text data mining or knowledge discovery in textual databases

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

1. Document Collection
2. Retrieve and preprocess document
3. Analyze Text
   - Information Extraction
   - Classification
   - Summarization
   - Clustering
4. Knowledge Management Information System

Overview of Information Extraction based Text Mining Framework

Text Data Mining

Text Mining Technologies

- Statistics
- Database Systems
- Natural Language Processing
- Information Retrieval
- Machine Learning
- Pattern Recognition
- Visualization
- Algorithms
- High-performance Computing
- Applications

Adapted from: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier
Data Mining versus Text Mining

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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining:
Core Analytics Process

The KDD Process for Extracting Useful Knowledge from Volumes of Data

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

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

Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)

Data Mining Processing Pipeline
(Charu Aggarwal, 2015)

Data Collection → Data Preprocessing:
- Feature Extraction
- Cleaning and Integration

Analytical Processing:
- Building Block 1
- Building Block 2

Output for Analyst

Feedback (Optional)

Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer
**Text Mining Process**

Context diagram for the text mining process

- Unstructured data (text)
- Structured data (databases)
- Extract knowledge from available data sources
- Context-specific knowledge
- Domain expertise
- Tools and techniques
- Software/hardware limitations
- Privacy issues
- Linguistic limitations

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

**Task 1: Establish the Corpus:** Collect & Organize the Domain Specific Unstructured Data
- The inputs to the process includes a variety of relevant unstructured (and semi-structured) data sources such as text, XML, HTML, etc.

**Task 2: Create the Term-Document Matrix:** Introduce Structure to the Corpus
- The output of Task 1 is a collection of documents in some digitized format for computer processing.
- The output of Task 2 is a flat file called term-document matrix where the cells are populated with the term frequencies.

**Task 3: Extract Knowledge:** Discover Novel Patterns from the T-D Matrix
- The output of Task 3 is a number of problem specific classification, association, clustering models and visualizations.

The three-step text mining process

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

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

**• Step 2: Create the Term–by–Document Matrix**

| Documents   | Terms                  | 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                  |                      |             |     |      |
| ...         |                        |                 |                    |                      |             |     |      |

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

• Syntax versus semantics based text mining

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

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

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

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

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

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

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

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

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

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

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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
歐巴馬是美國的一位總統
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。
自 2014/01/06 起，本斷詞系統已經處理過 28270134 篇文章

莎士比亞在淡江 遇見賽萬提斯
2016-04-26 02:27 聯合報 記者徐筱倫 / 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「唐吉珂德」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良說，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雋表示，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。

日文系學生賴嘉郁說，原本只是喜歡塗鴉才來參加活動，後來才知道是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；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/
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。
The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

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

Supported software distributions

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

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

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

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

**Stanford POS Tagger**
A maximum-entropy (CMM) part-of-speech (POS) tagger for English,
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford CoreNLP
http://nlp.stanford.edu:8080/corenlp/process

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

Coreference:

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

1 Stanford University is located in California.

2 It is a great university.

Collapsed CC-processed dependencies:

1 Stanford University is located in California.

2 It is a great university.

Visualisation provided using the brat visualisation/annotation software. Copyright © 2011, Stanford University, All Rights Reserved.

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Lemma</th>
<th>Char begin</th>
<th>Char end</th>
<th>POS</th>
<th>NER</th>
<th>Normalized NER</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>Stanford</td>
<td>0</td>
<td>8</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>University</td>
<td>University</td>
<td>9</td>
<td>19</td>
<td>NNP</td>
<td>ORGANIZATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>is</td>
<td>be</td>
<td>20</td>
<td>22</td>
<td>VBZ</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>located</td>
<td>located</td>
<td>23</td>
<td>30</td>
<td>JJ</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>in</td>
<td>in</td>
<td>31</td>
<td>33</td>
<td>IN</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>California</td>
<td>California</td>
<td>34</td>
<td>44</td>
<td>NNP</td>
<td>LOCATION</td>
<td></td>
<td>PER0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>44</td>
<td>45</td>
<td>.</td>
<td>O</td>
<td></td>
<td>PER0</td>
</tr>
</tbody>
</table>

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

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

Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Stanford University is located in California. It is a great university.
Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

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

That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares.

Related: Gates reclaims title of world's richest billionaire

Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.

It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation.

The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

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

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

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

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET
Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft over the past two days. <ORGANIZATION>Microsoft</ORGANIZATION> (MSFT) brought 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: english.muc.7class.distsim.crf.ser.gz
Output Format: xml
Preserve Spacing: yes

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.
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) —

Copyright © 2011, Stanford University, All Rights Reserved.
Stanford Named Entity Tagger (NER)

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

Stanford Named Entity Tagger

Classifier: english.conll.4class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

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

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

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

Potential tags:

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

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

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

Potential tags:
- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

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

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 reclaimed title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.
Bill Gates no longer Microsoft's biggest shareholder By Patrick Sheridan @CNNTech May 2, 2014 5:46 PM ET NEW YORK - For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft, bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer, who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft'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.
自然語言處理與資訊檢索研究資源
http://mail.tku.edu.tw/myday/resources/

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

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

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

4. 哈工大資訊檢索研究中心(HIT-CIR)語言技術平臺
語料資源
哈工大資訊檢索研究中心漢語依存樹庫〔HIT-CIR Chinese Dependency Treebank〕
哈工大資訊檢索研究中心同義詞詞林擴展版〔HIT-CIR Tongyici Cilin (Extended)〕
語言處理模組
斷句 (SplitSentence: Sentence Splitting)
詞法分析 (IRLAS: Lexical Analysis System)
基於SVMTool的詞性標注 (PosTag: Part-of-speech Tagging)
命名實體識別 (NER: Named Entity Recognition)
基於動態局部優化的依存句法分析 (Parser: Dependency Parsing)
基於圖的依存句法分析 (GParser: Graph-based DP)
全文詞義消歧 (WSD: Word Sense Disambiguation)
淺層語義標注模組 (SRL: hallow Semantics Labeling)
資料表示
語言技術置標語言 (LTML: Language Technology Markup Language)
視覺化工具
LTML視覺化XSL

授權單位：哈工大資訊檢索研究中心(HIT-CIR)
授權金額：免費授權學術使用。
授權日期：2011.05.03。
HIT IR: http://ir.hit.edu.cn/
Architectures of Sentiment Analytics
Bing Liu (2015), Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press

http://www.amazon.com/Sentiment-Analysis-Opinions-Sentiments-Emotions/dp/1107017890
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

“(1) I bought an iPhone a few days ago.
(2) It was such a nice phone.
(3) The touchscreen 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. ...”
Sentiment Analysis Architecture

Sentiment Classification Based on Emoticons

Lexicon-Based Model

Preassembled Word Lists → Merged Lexicon

Generic Word Lists → Merged Lexicon

Tokenized Document Collection → Sentiment Scoring and Classification: Polarity

Sentiment Polarity

Sentiment Analysis Tasks

- Opinionated Document
- Subjectivity Classification
- Sentiment Classification
- Opinion holder extraction
- Object/Feature extraction

### Sentiment Analysis vs. Subjectivity Analysis

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Subjectivity Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Subjective</td>
</tr>
<tr>
<td>Negative</td>
<td>Objective</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
</tr>
</tbody>
</table>
Levels of Sentiment Analysis

1. Word level Sentiment Analysis
2. Sentence level Sentiment Analysis
3. Document level Sentiment Analysis
4. Feature level Sentiment Analysis

Sentiment Analysis

Tasks

- Subjectivity Classification
- Sentiment Classification
- Review Usefulness Measurement
- Opinion Spam Detection
- Lexicon Creation
- Aspect Extraction
- Polarity Determination
- Vagueness resolution in opinionated text
- Multi- & Cross-Lingual SC
- Cross-domain SC

Approaches

- Machine Learning based
- Lexicon based
- Hybrid approaches
- Ontology based
- Non-Ontology based

Sentiment Classification Techniques

Sentiment Analysis
- Machine Learning Approach
  - Supervised Learning
    - Decision Tree Classifiers
    - Linear Classifiers
    - Rule-based Classifiers
  - Unsupervised Learning
    - Probabilistic Classifiers
  - Dictionary-based Approach
    - Statistical
    - Semantic
  - Corpus-based Approach
    - Support Vector Machine (SVM)
    - Neural Network (NN)
    - Deep Learning (DL)
    - Naïve Bayes (NB)
    - Bayesian Network (BN)
    - Maximum Entropy (ME)

### Example of SentiWordNet

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

<table>
<thead>
<tr>
<th>Language Rank</th>
<th>Types</th>
<th>Spectrum Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. C</td>
<td>📡📱💻💻</td>
<td>100.0</td>
</tr>
<tr>
<td>2. Java</td>
<td>🌍📱💻💻</td>
<td>98.1</td>
</tr>
<tr>
<td>3. Python</td>
<td>🌍📱💻💻</td>
<td>98.0</td>
</tr>
<tr>
<td>4. C++</td>
<td>📡📱💻💻</td>
<td>95.9</td>
</tr>
<tr>
<td>5. R</td>
<td>📡💻</td>
<td>87.9</td>
</tr>
<tr>
<td>6. C#</td>
<td>🌍📱💻</td>
<td>86.7</td>
</tr>
<tr>
<td>7. PHP</td>
<td>🌍💻</td>
<td>82.8</td>
</tr>
<tr>
<td>8. JavaScript</td>
<td>🌍📱💻</td>
<td>82.2</td>
</tr>
<tr>
<td>9. Ruby</td>
<td>🌍💻</td>
<td>74.5</td>
</tr>
<tr>
<td>10. Go</td>
<td>🌍💻</td>
<td>71.9</td>
</tr>
</tbody>
</table>

Python: Analytics and Data Science Software

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

# List and the enumerate function
```python
>>> 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

Python is a programming language that lets you work quickly and integrate systems more effectively. Learn More

https://www.python.org/
Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.

Source: https://www.python.org/doc/essays/blurb/
Anaconda

ANAconda gives SUPERpowers to people who change the world

Modern open source analytics platform powered by Python

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 | OS X | 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.

https://www.continuum.io/downloads
Download Anaconda Python 3.5

Anaconda for OS X

<table>
<thead>
<tr>
<th>PYTHON 2.7</th>
<th>PYTHON 3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Mac OS X 64-bit Graphical Installer" /></td>
<td><img src="https://via.placeholder.com/150" alt="Mac OS X 64-bit Graphical Installer" /></td>
</tr>
<tr>
<td>274M (OS X 10.7 or higher)</td>
<td>267M (OS X 10.7 or higher)</td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Mac OS X 64-bit Command-Line Installer" /></td>
<td><img src="https://via.placeholder.com/150" alt="Mac OS X 64-bit Command-Line Installer" /></td>
</tr>
<tr>
<td>239M (OS X 10.7 or higher)</td>
<td>233M (OS X 10.7 or higher)</td>
</tr>
</tbody>
</table>

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.

https://www.continuum.io/downloads
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

Source: Yves Hilpisch (2014), Python for Finance: Analyze Big Financial Data, O'Reilly
Python (Python 2.7 & Python 3.5) Standard Syntax

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

Old Py2

Py2&3

Py3 only

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 *
from past.builtins import *

Python 2

Python 3

Py2&3

Old Py2

Py3 only

Source: PyCon Australia (2014), Writing Python 2/3 compatible code by Edward Schofield
https://www.youtube.com/watch?v=KOqk8j11aAI
ipython notebook
ipython notebook

```
[iMydaytekiMacBook-Pro:~ imyday]$ ipython notebook
[I 14:26:49.944 NotebookApp] 0 active kernels
[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-c1901c960677/channels?session_id=265FB16817FB4A79202F6D3C3BDB0E6 (::1): Kernel does not exist: a87ab95b-6d6e-44d3-aaa7-c1901c960677
[W 14:26:56.663 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c960677/channels?session_id=265FB16817FB4A79202F6D3C3BDB0E6 (::1) 95.43ms referer=None
[W 14:26:56.681 NotebookApp] 404 GET /api/kernels/b7fae9a6-d77b-4ead-832c-c070b18d642b/channels?session_id=EF4C761633E541C88568CDBCDE1091B7 (::1): Kernel does not exist: b7fae9a6-d77b-4ead-832c-c070b18d642b
[W 14:26:56.683 NotebookApp] 404 GET /api/kernels/b7fae9a6-d77b-4ead-832c-c070b18d642b/channels?session_id=EF4C761633E541C88568CDBCDE1091B7 (::1) 6.62ms referer=None
[W 14:27:29.595 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c960677/channels?session_id=265FB16817FB4A79202F6D3C3BDB0E6 (::1): Kernel does not exist: a87ab95b-6d6e-44d3-aaa7-c1901c960677
[W 14:27:29.631 NotebookApp] 404 GET /api/kernels/a87ab95b-6d6e-44d3-aaa7-c1901c960677
```
jupyter notebook

Python 3
jupyter notebook
Python 3

In [1]: print("Hello World")
Hello World

In [ ]:
jupyter notebook
Python 2
jupyter notebook
Python 2

In [1]: print("Hello World")

Hello World
ipython notebook
jupyter notebook
jupyter notebook
print('Hello World, Python')
print('Hello World, Python')
# 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')

http://python-future.org/compatible_idioms.html
Writing Python 2-3 compatible code

Essential syntax differences

```python
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')
```

http://python-future.org/compatible_idioms.html
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

# 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

http://python-future.org/compatible_idioms.html
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
'
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"))

Source: http://pythonprogramminglanguage.com/text-input-and-output/
Variables

\[
x = 2
\]
\[
\text{price} = 2.5
\]
\[
\text{word} = 'Hello'
\]
\[
\text{word} = 'Hello'
\]
\[
\text{word} = "Hello"
\]
\[
\text{word} = '''Hello'''
\]
\[
x = 2
\]
\[
x = x + 1
\]
\[
x = 5
\]
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)
BMI Calculator in Python

```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

```python
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

Source: http://pythonprogramminglanguage.com/
For loops

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

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

1.8
Lists

```python
x = [60, 70, 80, 90]
print(len(x))
print(x[0])
print(x[1])
print(x[2])
print(x[-1])
```

60
70
90
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])  # 10
print(x[1])  # 20
print(x[2])  # 30
print(x[-1]) # 50
```
Python Ecosystem
**Python Ecosystem**

```python
import math

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

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

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

```
0.0
```

**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`

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

```
3.0
```
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.
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

Source: DrapsTV (2016), Compatible Python 2 & 3 Code
https://www.youtube.com/watch?v=5Pwc-Rd4qJA
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

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 in `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.

https://pythonhosted.org/six/
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...
## Managing Conda and Anaconda

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>conda info</td>
<td>Verify conda is installed, check version #</td>
</tr>
<tr>
<td>conda update conda</td>
<td>Update conda package and environment manager to current version</td>
</tr>
<tr>
<td>conda update anaconda</td>
<td>Update the anaconda meta package (the library of packages ready to install with conda command)</td>
</tr>
</tbody>
</table>

## 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 snowflakes (Windows)**

Activate the new environment to use it

*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**

**conda env create -f puppies.yml**

Save current envirinment to a file

Load environment fromm a file

http://conda.pydata.org/docs/_downloads/conda-cheatsheet.pdf
# Managing Python

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>conda search --full-name python</code> or <code>conda search -f python</code></td>
<td>Check versions of Python available to install</td>
</tr>
<tr>
<td><code>conda create -n snakes python=3.4</code></td>
<td>Install different version of Python in new environment</td>
</tr>
<tr>
<td><code>source activate snakes (Linux, Mac)</code></td>
<td>Switch to the new environment that has a different version of Python</td>
</tr>
<tr>
<td><code>activate snakes (Windows)</code></td>
<td></td>
</tr>
</tbody>
</table>

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

# Managing Packages, Including Python

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>conda list</strong></td>
<td>View list of packages and versions installed in active environment</td>
</tr>
<tr>
<td><strong>conda search beautiful-soup</strong></td>
<td>Search for a package to see if it is available to conda install</td>
</tr>
</tbody>
</table>
| **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** | Activate the environment where you want to install a package and install it with pip (included with Anaconda and Miniconda) |
| **activate bunnies** | (Windows) |
| **pip install see** | Install commercial Continuum packages |
| **conda skeleton pypi pyinstrument** | Build a Conda package from a Python Package Index (PyPI) Package |

PyCharm: Python IDE

http://www.jetbrains.com/pycharm/
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:

```python
>>> import nltk

http://www.nltk.org/
```
jupyter notebook
Jupyter New Terminal
```bash
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
decimal                     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
```
<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
<th>Python Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>nltk</td>
<td>3.1</td>
<td>py27_0</td>
</tr>
<tr>
<td>node-webkit</td>
<td>0.10.1</td>
<td></td>
</tr>
<tr>
<td>nose</td>
<td>1.3.7</td>
<td>py27_0</td>
</tr>
<tr>
<td>notebook</td>
<td>4.1.0</td>
<td>py27_0</td>
</tr>
<tr>
<td>numba</td>
<td>0.23.1</td>
<td>np110py27_0</td>
</tr>
<tr>
<td>numexpr</td>
<td>2.4.6</td>
<td>np110py27_1</td>
</tr>
<tr>
<td>numpy</td>
<td>1.10.4</td>
<td>py27_0</td>
</tr>
<tr>
<td>odo</td>
<td>0.4.0</td>
<td>py27_0</td>
</tr>
<tr>
<td>openpyxl</td>
<td>2.3.2</td>
<td>py27_0</td>
</tr>
<tr>
<td>openssl</td>
<td>1.0.2g</td>
<td></td>
</tr>
<tr>
<td>pandas</td>
<td>0.18.0</td>
<td>np110py27_0</td>
</tr>
<tr>
<td>path.py</td>
<td>8.1.2</td>
<td>py27_1</td>
</tr>
<tr>
<td>patsy</td>
<td>0.4.0</td>
<td>np110py27_0</td>
</tr>
<tr>
<td>pep8</td>
<td>1.7.0</td>
<td>py27_0</td>
</tr>
<tr>
<td>pexpect</td>
<td>3.3</td>
<td>py27_0</td>
</tr>
<tr>
<td>pickleshare</td>
<td>0.5</td>
<td>py27_0</td>
</tr>
<tr>
<td>pillow</td>
<td>3.1.0</td>
<td>py27_0</td>
</tr>
<tr>
<td>pip</td>
<td>8.1.0</td>
<td>py27_0</td>
</tr>
<tr>
<td>ply</td>
<td>3.8</td>
<td>py27_0</td>
</tr>
<tr>
<td>psutil</td>
<td>3.4.2</td>
<td>py27_0</td>
</tr>
<tr>
<td>ptyprocess</td>
<td>0.5</td>
<td>py27_0</td>
</tr>
<tr>
<td>py</td>
<td>1.4.31</td>
<td>py27_0</td>
</tr>
<tr>
<td>pyasn1</td>
<td>0.1.9</td>
<td>py27_0</td>
</tr>
<tr>
<td>pyaudio</td>
<td>0.2.7</td>
<td>py27_0</td>
</tr>
<tr>
<td>pycosat</td>
<td>0.6.1</td>
<td>py27_0</td>
</tr>
<tr>
<td>pycparser</td>
<td>2.14</td>
<td>py27_0</td>
</tr>
<tr>
<td>pycrypto</td>
<td>2.6.1</td>
<td>py27_0</td>
</tr>
<tr>
<td>pycurl</td>
<td>7.19.5.3</td>
<td>py27_0</td>
</tr>
<tr>
<td>pyflakes</td>
<td>1.0.0</td>
<td>py27_0</td>
</tr>
</tbody>
</table>
```python
help('modules')
```

```
In [2]: help('modules')

<table>
<thead>
<tr>
<th>Module</th>
<th>Module</th>
<th>Module</th>
<th>Module</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>builtins</strong></td>
<td><strong>future</strong></td>
<td><strong>builtin</strong></td>
<td><strong>codecs</strong></td>
</tr>
<tr>
<td>abcoll</td>
<td>ast</td>
<td>bisect</td>
<td>builtinsSuites</td>
</tr>
<tr>
<td>ast</td>
<td>cycler</td>
<td>cython</td>
<td>cpythonmagics</td>
</tr>
<tr>
<td>_codecs</td>
<td>_codecs_cn</td>
<td>_codecs_hk</td>
<td>_codecs_iso2022</td>
</tr>
<tr>
<td>_codecs_iso2022</td>
<td>_codecs_jp</td>
<td>_codecs_kr</td>
<td>_codecs_tw</td>
</tr>
<tr>
<td>codecs</td>
<td>codecs_cn</td>
<td>codecs_hk</td>
<td>codecs_iso2022</td>
</tr>
<tr>
<td>codecs_iso2022</td>
<td>dbhash</td>
<td>dateutil</td>
<td>os</td>
</tr>
<tr>
<td>_codecs_jp</td>
<td>dbm</td>
<td>os2emxpath</td>
<td>tkColorChooser</td>
</tr>
<tr>
<td>codecs_kr</td>
<td>decimal</td>
<td>osax</td>
<td>tkCommonDialog</td>
</tr>
<tr>
<td>_codecs_tw</td>
<td>decorator</td>
<td>pandas</td>
<td>tkFileDialog</td>
</tr>
<tr>
<td>codecs</td>
<td></td>
<td>parser</td>
<td>tkFont</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tkMessageBox</td>
</tr>
</tbody>
</table>
```

`tabnanny`  `ntplib`  `telnetlib`  `terminado`  `terminalcommand`  `termios`  `numba`  `test_path`  `test_pycosat`  `tests`  `numpy`  `textwrap`  `this`  `thread`  `threading`  `time`  `timeit`  `tkColorChooser`  `tkCommonDialog`  `tkFileDialog`  `tkFont`  `tkMessageBox`
import nltk
import nltk
nltk.download()
import nltk
nltk.download()
import nltk
nltk.download()
nltk_data

- chunkers
- corpora
- grammars
- help
- models
- stemmers
- taggers
- tokenizers
At eight o'clock on Thursday morning Arthur didn't feel very good.
At eight o'clock on Thursday morning Arthur didn't feel very good.
import nltk
sentence = "At eight o'clock on Thursday morning Arthur didn't feel very good."
tokens = nltk.word_tokenize(sentence)
tokens
print(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', 'very', 'good', '.']

Source: http://www.nltk.org/
tagged = nltk.pos_tag(tokens)
tagged[0:6]
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'),
        ('.', '.')]
At eight o'clock on Thursday morning
Arthur didn't feel very good.
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'), (',', '.')]}

Source: [http://www.nltk.org/](http://www.nltk.org/)
from nltk.corpus import treebank
t = treebank.parsed_sents('wsj_0001.mrg')[0]
t.draw()
(S
  (NP-SBJ (NP (NNP Pierre) (NNP Vinken))
  (, ,))
  (ADJP
  (NP (CD 61) (NNS years))
  (JJ old))
  (, ,))
  (VP (MD will)
  (VP (VB join)
  (NP (DT the) (NN board))
  (PP-CLR (IN as)
  (NP (DT a) (JJ nonexecutive) (NNN director)))
  (NP-TMP (NNP Nov.) (CD 29))))
  (.. .))

(S
  (NP-SBJ (NNP Mr.) (NNP Vinken))
  (VP (VBZ is)
  (NP-PRD
  (NP (NN chairman))
  (PP (IN of)
  (NP
  (NP (NNP Elsevier) (NNP N.V.))
  (, ,)
  (NP (DT the) (NNP Dutch) (VBG publishing) (NN group)))))))
  (.. .))
Python Jieba “结巴”中文分词

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

fxsjy / jieba

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

- extra_dict
  update to v0.33

- jieba
  Bugfix for HMM=False in parallelism.

- test
  Bugfix for HMM=False in parallelism.

- .gitattributes
  first commit

- .gitignore
  update jieba3k

- Changelog
  version change 0.38

- LICENSE
  add a license file

- MANIFEST.in
  include Changelog & README.md in the distribution package

- README.md
  Update README.md

https://github.com/fxsjy/jieba
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.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)
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/ldkrsi/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

Source: Sunila Gollapudi (2016), Practical Machine Learning, Packt Publishing
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


• 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


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

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