NLTK: Natural Language Toolkit Overview and Application

Jimmy Lai
jimmy.lai@oi-sys.com
Software Engineer @ Oxygen Intelligence
2012/06/09

2012 TAIWAN PYCON

聚寶評+

Outline

1. An application based on NLP: 聚寶評
2. Introduction to Natural Language Processing
3. Brief History of NLTK
4. NLTK

聚寶評 www.ezpao.com

聚寶評 www.ezpao.com

聚寶評 找美食 搜店家

新竹市 | 火鍋 | 找店家 | 地圖搜尋

選地點：縣市 > 鄉鎮市區

縮小搜尋範圍 (重設)：

本次搜尋：在 新竹市 找 火鍋 的店家

約有 93 筆結果

語意分析搜尋引擎

A 太和殿鴛鴦麻辣火鍋 300 新竹市經國路二段99號

網友分享標籤：綜合鍋類、連鎖店、麻辣火鍋
網友分享菜：鴛鴦麻辣火鍋、綜合菇盤、手工花枝漿 ... 共9道

86%
短評共64筆

個人頭像：網路上也是好評不斷 還有人說比鼎王更好吃 看的我整個熱血沸騰啦! ... 太和殿鴛鴦麻辣火鍋 2010/07/30

經驗分享： 7 0 0 選 0 歇 0

B 聚北海道昆布鍋 新竹市中正路2號5樓

網友分享標籤：綜合鍋、謝師宴、大型團體聚會、火鍋
網友分享菜：梅子醋、昆布烏龍麵、精緻昆布鍋套餐 ... 共15道

96%
短評共595筆

wensuping：嚴選套餐甜品～金字塔紅豆冰堡，雖然天氣冷，但我還是堅持點這道～好食。 ... 聚-北海道昆布鍋 2011/01/06

經驗分享： 21 0 3 選 0 歇 0

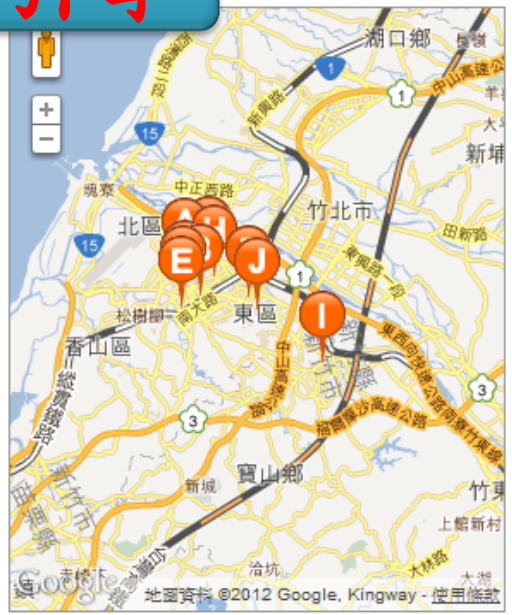
C 台北林季麻辣火鍋連鎖店 新竹市民生路266號

網友分享標籤：綜合鍋類、麻辣火鍋、吃到飽
網友分享菜：麻辣鍋、鴛鴦鍋、酸白菜鍋、麻辣火鍋 ... 共15道

91%
短評共79筆

hisldudi：肉片 / 火鍋餃類 / 排骨酥 / 蛤蜊 /鴨血豆腐 / 香菇丸 這些都挺推薦的。 ... [推薦][新竹市][麻辣火鍋] 林季麻辣火鍋 (新竹民生店) 2010/03/06

經驗分享： 9 0 0 選 0 歇 0



搜尋地點：新竹市

店家排行榜

<返回 本次搜尋結果

聚北海道昆布鍋

新竹市中正路2號5樓 MAP

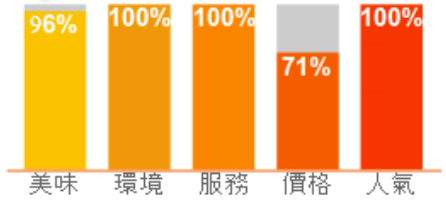
電話：03-5260688
營業時間：
消費價位：一般價位

本店提供的服務： 編輯

經驗分享：21 0 3 遷 0 歇 0

網友分享標籤：綜合鍋、謝師宴、大型團體聚會、節日聚會、綜合鍋類、家庭聚會
網友分享菜：梅子醋(16)、 昆布烏龍麵(16)、 精緻昆布鍋套餐(13)、 北海道金橙奶酪(11)、 套餐(11)、 蟹肉蒟蒻球(10)、 三色魚麵(10)、 北海道黃金稀飯(10)、 北海道昆布鍋套餐(9)、 精燉桂花赤豆饌(9)、 金字塔紅豆冰堡(9)、 嫩肩牛小排(9)、 豬雞雙拼主餐(8)、 昆布滋養湯(8)、 昆布麻辣湯(8)

評論主題分析



短評共595筆

網友分享菜分析

短評共 595 筆　　搜尋文章內容：請輸入關鍵字

全部主題	全部	內容	全部時間	資料來源
美味		"聚"北海道昆布鍋 – in新竹 **cqovkrpe**：↑手工創意單點：蝦仁福袋and山藥玉子燒雙拼 ↑服務人員建議點雙拼的可以吃到兩種不同的東西 山藥玉子燒口感很Q ...	2009/12/17 很久以前	隨意窩
魚頭海鮮主餐		新竹-聚北海道昆布鍋 **阿叮**：隔壁桌的秀華點了季節魚頭海鮮主餐 端來是大大的一隻魚頭和一些海鮮 看起來不錯吃旁邊還有蝦罵和花枝 ...	2010/12/07 一年以前	蕃薯藤
美味		[家庭聚餐]生日聚-北海道昆布鍋 **alice**：嫩肩牛小排是〔北海道嚴選套餐〕才有的主餐，它真的也沒讓我失望。均勻的油花分在的肉片上，放到鍋中涮幾下就熟了，入口軟嫩，滋味好呀！爸媽主餐選雞肉，雞肉也很不錯；而豬肉是雪花豬，有特別的口感，也很不賴唷～ ...	2010/12/05 一年以前	新浪部落
北海道昆布鍋		[食-新竹]〞聚〞在一起吃火鍋 **aorange**：北海道昆布鍋清清爽爽的，昆布滋養鍋有中藥的味道， ...	2009/10/28 很久以前	愛評網

正評/負評分析

手機版

m.ezpao.com

網友分享菜+評論分析

網友分享菜+正負評分析

Natural Language Processing (NLP)

- 語音識別（Speech recognition）
- 詞性標註（Part-of-speech tagging）
- 句法分析（Parsing）
- 自然語言生成(Natural language generation)
- 文本分類（Text classification）
- 信息抽取（Information extraction）
- 機器翻譯（Machine translation）
- 文字蘊涵（Textual entailment）

via Wikipedia

NLTK: Natural Language Toolkit

- http://www.nltk.org/
- Author: Steven Bird, Edward Loper, Ewan Klein
- Originally developed for class student has background either in computer science or linguistics.
- Currently:
  - Education: over 100 courses in 23 countries.
  - Research: over 250 papers cites NLTK.

Outline

1. An application based on NLP: 聚寶評
2. Introduction to Natural Language Processing
3. Brief History of NLTK
4. NLTK:
   a. Installation
   b. Annotated Text Corpora
   c. Text Tokenization, Normalization, Analysis, Distribution Analysis
   d. Part-of-speech Tagging
   e. Text Classification
   f. Named Entity Recognition

Install NLTK

Python 2.6+

pip install numpy
pip install matplotlib
pip install nltk

Annotated Text Corpora (1/3)

nltk.corpus

- Corpus: 語料庫，含有某種結構化標記的資料集合，可能包含多種語言。
- 例如：
  - stopwords: 常見字字典
  - sinica_treebank: 中文語句結構標記語料庫
  - brown: 包含15種分類及詞性標記的英語語料庫
  - wordnet: 包含詞性、同義反義的英語字典

Annotated Text Corpora (2/3)

nltk.corpus

```
import nltk
#download corpus on demand
nltk.download()

#stopwords
nltk.corpus.stopwords.words()
nltk.corpus.stopwords.words('english')
nltk.corpus.stopwords.words('french')

some_english_stopwords = ['most', 'me',
'below', 'when', 'which', 'what', 'of', 'it',
'very', 'our']
```

```
#Chinese treebank
nltk.corpus.sinica_treebank

#Examples
(嘉珍, Nba)
(和, Caa)
(我, Nhaa)
(住在, VC1)
(同一條, DM)
(巷子, Nab)

(我們, Nhaa)
(是, V_11)
(鄰居, Nab)

(也, Dbb)
(是, V_11)
(同班, Nv3)
(同學, Nab)
```

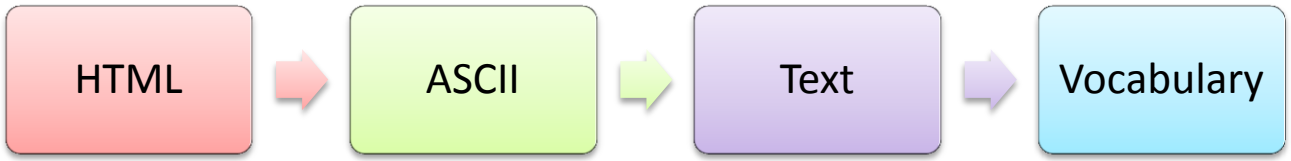
Annotated Text Corpora (3/3)

nltk.corpus

ACE Named Entity Chunker (Maximum entropy)
Australian Broadcasting Commission 2006
Alpino Dutch Treebank
BioCreAtIvE (Critical Assessment of Information Extraction Systems in Biology)
Brown Corpus
Brown Corpus (TEI XML Version)
CESS-CAT Treebank
CESS-ESP Treebank
Chat-80 Data Files
City Database
The Carnegie Mellon Pronouncing Dictionary (0.6)
ComTrans Corpus Sample
CONLL 2000 Chunking Corpus
CONLL 2002 Named Entity Recognition Corpus
Dependency Treebanks from CoNLL 2007 (Catalan and Basque Subset)
Dependency Parsed Treebank
Sample European Parliament Proceedings Parallel Corpus
Portuguese Treebank
Gazeteer Lists
Genesis Corpus
Project Gutenberg Selections
NIST IE-ER DATA SAMPLE
C-Span Inaugural Address Corpus
Indian Language POS-Tagged Corpus
JEITA Public Morphologically Tagged Corpus (in ChaSen format)
PC-KIMMO Data Files
KNB Corpus (Annotated blog corpus)
Language Id Corpus
Lin's Dependency Thesaurus
MAC-MORPHO: Brazilian Portuguese news text with part-of-speech tags
Machado de Assis -- Obra Completa
Sentiment Polarity Dataset Version 2.0
Names Corpus, Version 1.3 (1994-03-29)
NomBank Corpus 1.0
NPS Chat
Paradigm Corpus

Text Tokenization
nltk.tokenize

Web Text Processing Flow




```
from urllib import urlopen
html = urlopen(url).read()
raw = nltk.clean_html(html)
sents = nltk.sent_tokenize(raw)
tokens = []                                  # wordpunct_tokenize: ['3', '.', '33']
for sent in sents:                           # word_tokenize: ['3.33']
    tokens.extend(nltk.word_tokenize(sent))
text = nltk.Text(tokens)
words = [word.lower() for word in text]
vocab = sorted(set(words))
```

Text Normalization (1/2)

nltk.stem

- Stem: 將單字(現在式、過去式、單複數)還原成原型。可以將不同形式的單字歸類為同一個單字。
- 著名演算法：Porter Stemmer

Text Normalization (2/2)
nltk.stem

```
In [1]: words = nltk.corpus.conll2000.words()[:20]
In [2]: print words
['Confidence', 'in', 'the', 'pound', 'is', 'widely', 'expected', 'to', 'take', 'another',
'sharp', 'dive', 'if', 'trade', 'figures', 'for', 'September', ',', 'due', 'for']

In [3]: stemmer = nltk.stem.PorterStemmer()
In [4]: print [stemmer.stem(word) for word in words]
['Confid', 'in', 'the', 'pound', 'is', 'wide', 'expect', 'to', 'take', 'anoth',
'sharp', 'dive', 'if', 'trade', 'figur', 'for', 'Septemb', ',', 'due', 'for']

In [5]: stemmer = nltk.stem.LancasterStemmer()
In [6]: print [stemmer.stem(word) for word in words]
['confid', 'in', 'the', 'pound', 'is', 'wid', 'expect', 'to', 'tak', 'anoth',
'sharp', 'div', 'if', 'trad', 'fig', 'for', 'septemb', ',', 'due', 'for']

In [7]: stemmer = nltk.stem.WordNetLemmatizer()
In [8]: print [stemmer.lemmatize(word) for word in words]
['Confidence', 'in', 'the', 'pound', 'is', 'widely', 'expected', 'to', 'take', 'another',
'sharp', 'dive', 'if', 'trade', 'figure', 'for', 'September', ',', 'due', 'for']
```

Text Analysis

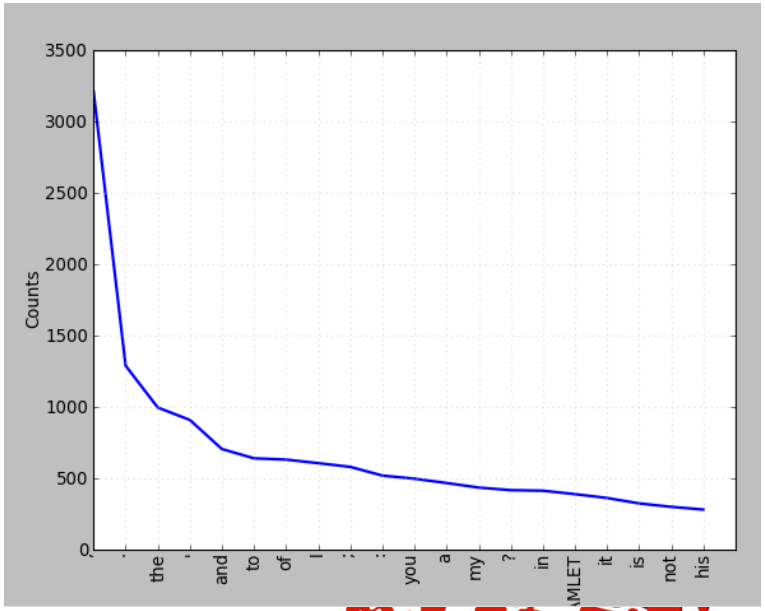
nltk.text

```
#using sinica treebank dataset
In [1]: text = nltk.Text(nltk.corpus.sinica_treebank.words())

#common context
In [2]: text.common_contexts(['是', '在'])
一直_我 也_一 也_台北 也_她 他_一個 就_大 就_我們
就_這 的_他 說_中國 都_樹

#collocation
In [3]: text.collocations()
Building collocations list
全 國; 國家 公園; 就 是; 身 上; 都 是; 才 能; 站 在;
還 有; 最 大; 河 邊; 的 時候; 變 得; 心 裡; 這 是; 各
地; 帶 著; 院子 裡; 一個 人; 是 一個; 就 會

#concordance
In [4]: text.concordance('台灣', lines=10)
Displaying 10 of 161 matches:
??  ＩＢＭＣｌｕｂ 的 組織 台灣 ＩＢＭ 為 配合 國內 勞工
?? 男子 獨 專 成功 立名 的 台灣 為 女性 揚眉吐氣 她們 是
?? 智慧 的 她 女性 形象 在 台灣 和 中國 大陸 小說 是 解放
腐  的 同時 在 雙十節 宣稱 台灣 要 以 統一 大業 為 己任 ?
資源  的 條件 因為 今天 的 台灣 並 沒有 發展出 一個 具備
有  一位 廣播 電視 廳長 是 台灣 基隆 人 一個 人 如果 已 ?
?? 粗俗 的 水泥 步道 時 在 台灣 經濟 發展 的 歷史 赤裸裸
?? 的 自然 的 力量 為 避免 台灣 原住民族 文化 特質 與 總?
?? 註 明 清 時期 漢人 嫁給 台灣 平埔族 女子 大部份 摻雜 ?
國  報紙 上 看到 一則 介紹 台灣 近況 的 報導 尋找 自我 認
```

Text Distribution Analysis

nltk.probability

```
In [1]: import nltk
In [2]: dist = nltk.probability.FreqDist(nltk.corpus.shakespeare.words('hamlet.xml'))
In [3]: dist.tabulate(20)
   ,    .  the    '  and   to   of    I    ;    :  you    a   my    ?   in HAMLET   it   is  not  his
3211 1289  996  909  705  640  631  606  580  519  497  467  435  417  413  389  362  324  300  281
In [4]: dist.plot(20)
In [5]: dist['the']
Out[5]: 996

In [6]: cdist = nltk.probability.ConditionalFreqDist(
            (len(word), word) for word in nltk.corpus.shakespeare.words('hamlet.xml'))
In [7]: float(cdist[3]['the']) / float(cdist[3].N())
Out[7]: 0.14619110523998238
```



Part-of-speech Tagging (1/2)
nltk.tag

- Part of Speech Tagging: 詞性標記，標記每個單字的詞性。
- 同一單字的不同詞性其語義不同，如Book名詞是書，動詞是預定。
- 透過POS Tagging，可以賦予文字更多語義資訊。

Tag	Meaning
ADJ	adjective
ADV	adverb
CNJ	conjunction
DET	determiner
EX	existential
FW	foreign word
MOD	modal verb
N	noun
NP	proper noun
NUM	number
PRO	pronoun
P	preposition

Part-of-speech Tagging (2/2)

nltk.tag

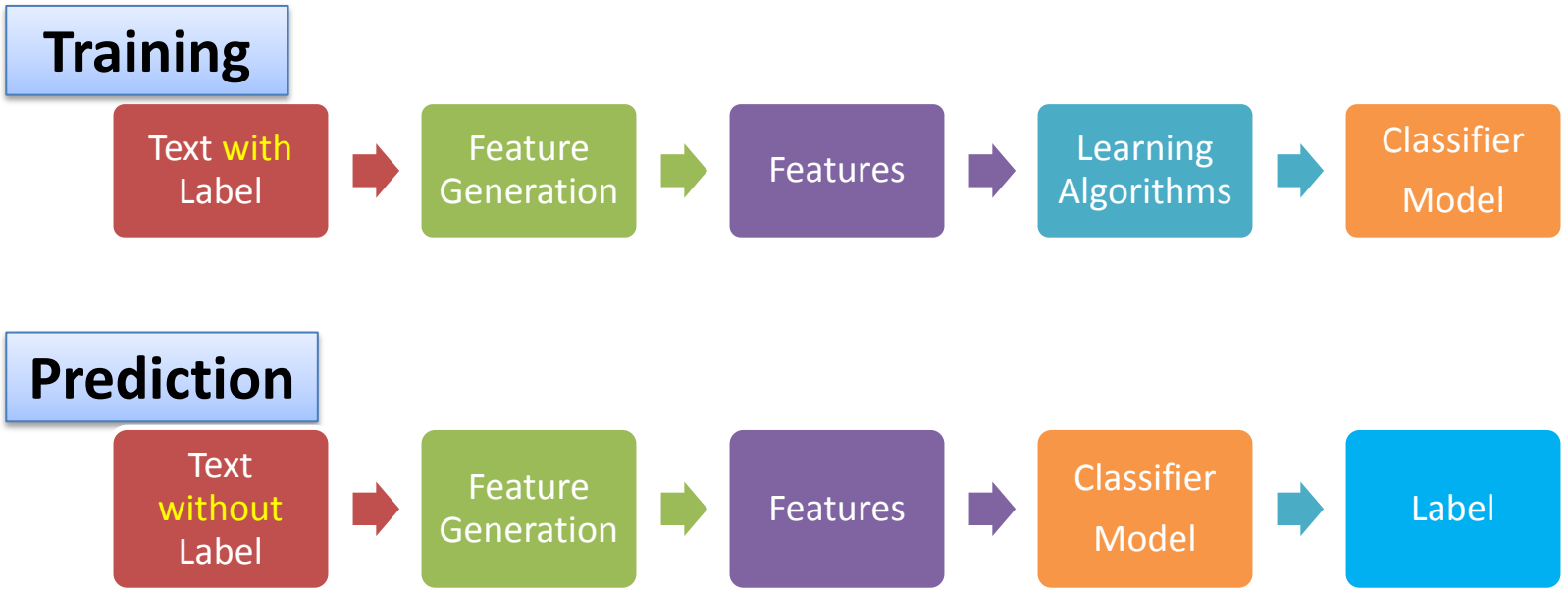
```
In [1]: print nltk.corpus.brown.tagged_words()[:15]
[('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'),
('Jury', 'NN-TL'), ('said', 'VBD'), ('Friday', 'NR'), ('an', 'AT'),
('investigation', 'NN'), ('of', 'IN'), ("Atlanta's", 'NP$'), ('recent', 'JJ'),
('primary', 'NN'), ('election', 'NN'), ('produced', 'VBD')]

#an off-the-shelf tagger
In [2]: print nltk.tag.pos_tag(nltk.corpus.brown.words()[:15])
[('The', 'DT'), ('Fulton', 'NNP'), ('County', 'NNP'), ('Grand', 'NNP'), ('Jury', 'NNP'),
('said', 'VBD'), ('Friday', 'NNP'), ('an', 'DT'), ('investigation', 'NN'), ('of', 'IN'),
("Atlanta's", 'JJ'), ('recent', 'JJ'), ('primary', 'JJ'), ('election', 'NN'),
('produced', 'VBN')]

#train our tagger
In [3]: unigram_tagger = nltk.tag.UnigramTagger(
            nltk.corpus.brown.tagged_sents(categories='news')[:1000])
In [4]: unigram_tagger.evaluate(nltk.corpus.brown.tagged_sents(categories='news')[1000:2000])
Out[4]: 0.6709000287549123

In [5]: affix_tagger = nltk.tag.AffixTagger(
            nltk.corpus.brown.tagged_sents(categories='news')[:1000], backoff=unigram_tagger)
In [6]: affix_tagger.evaluate(nltk.corpus.brown.tagged_sents(categories='news')[1000:2000])
Out[6]: 0.6848461612192083

In [7]: bigram_tagger = nltk.tag.BigramTagger(
            nltk.corpus.brown.tagged_sents(categories='news')[:1000], backoff=affix_tagger)
In [8]: bigram_tagger.evaluate(nltk.corpus.brown.tagged_sents(categories='news')[1000:2000])
Out[8]: 0.7011406115211348
```

Text Classification (1/3)
nltk.classify

- Text Classification: 文字分類，分析文字後將文字分到預先定義的類別裡。
- 基於統計的機器學習演算法，著名的演算法為：
  - Naïve Bayes Classifier
  - Decision Tree
  - Support Vector Machine

Text Classification (2/3)
nltk.classify

Machine Learning Approach Work Flow



Text Classification (3/3)

nltk.classify

```
In [1]: name_label_list = []
In [2]: name_label_list.extend([ (name, 'female') for name in nltk.corpus.names.words('female.txt')])
In [3]: name_label_list.extend([ (name, 'male') for name in nltk.corpus.names.words('male.txt')])
In [4]: random.shuffle(name_label_list)
In [5]: name_label_list[:5]
Out[6]:
[('Giffard', 'male'),
 ('Melosa', 'female'),
 ('Minerva', 'female'),
 ('Debora', 'female'),
 ('Adrien', 'male')]

In [7]: def get_feature(name):
   ...:     features = {'first_letter':name[0], 'last_letter':name[-1],
   ...:                 'first_2_letter': name[:2], 'last_2_letter': name[-2:],
   ...:                 'first_3_letter': name[:3], 'last_3_letter': name[-3:]}
   ...:      return features
   ...:

In [8]: feature_set = [(get_feature(name), label) for (name, label) in name_label_list ]
In [9]: size = int(len(feature_set) * 0.9)
In [10]: train_set, test_set = feature_set[:size], feature_set[size:]
In [11]: classifier = nltk.classify.NaiveBayesClassifier.train(train_set)
In [12]: nltk.classify.accuracy(classifier, test_set)
Out[13]: 0.8377358490566038
```

Named Entity Recognition (NER) (1/2)

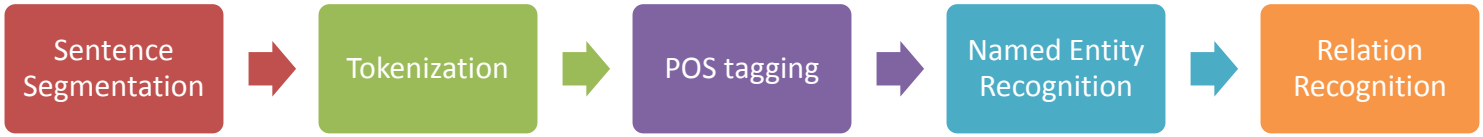
nltk.tag, nltk.chunk

- Named Entity Recognition: 從文字中擷取出命名實體，命名實體是具有完整語義的複合單字。例如：人名、地名、事件。

Named Entity Recognition (NER) (2/2)

nltk.tag, nltk.chunk

NER General Work Flow




```
In [1]: import nltk
In [2]: tagger = nltk.tag.stanford.NERTagger(
                 'stanford-ner-2012-05-22/classifiers/english.conll.4class.distsim.crf.ser.gz',
                 'stanford-ner-2012-05-22/stanford-ner-2012-05-22.jar')
In [3]: tagger.tag('Rami Eid is studying at Stony Brook University in NY'.split())
Out[3]:
[('Rami', 'PERSON'),
 ('Eid', 'PERSON'),
 ('is', 'O'),
 ('studying', 'O'),
 ('at', 'O'),
 ('Stony', 'ORGANIZATION'),
 ('Brook', 'ORGANIZATION'),
 ('University', 'ORGANIZATION'),
 ('in', 'O'),
 ('NY', 'O')]
```

Reference

- Steven Bird, Ewan Klein, and Edward Loper, "**Natural Language Processing with Python**", 2009. #includes: Python + NLP + NLTK
- Jacob Perkins, "Python Text Processing with NLTK 2.0 Cookbook", 2010.
- Matthew A. Russell, "Mining the Social Web", 2011.
- Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. Proceedings of the ACL02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics.

Thank you for your attention.
Q & A

We are hiring!

- 核心引擎演算法研發工程師
- 系統研發工程師
- 網路應用研發工程師

本公司提供企業客戶語義分析服務，
獲日本創投"軟體銀行"投資與肯定。
歡迎對創業有熱情的你加入我們！

Oxygen-Intelligence Taiwan Limited
引京聚點 知識結構搜索股份有限公司

- 公司簡介、職缺簡介：http://goo.gl/18vvQ
- 請將履歷寄到 jimmy.lai@oi-sys.com

2012 TAIWAN PYCON