NLTK: Natural Language Toolkit Overview and Application

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

2012 TAIWAN PYCON

Outline

1. An application based on NLP: 聚寶評
2. Introduction to Natural Language Processing
3. Brief History of NLTK
4. Overview of NLTK
5. Application of NLTK: Topic Classification on PTT

聚寶評 www.ezpao.com

美食搜尋引擎

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

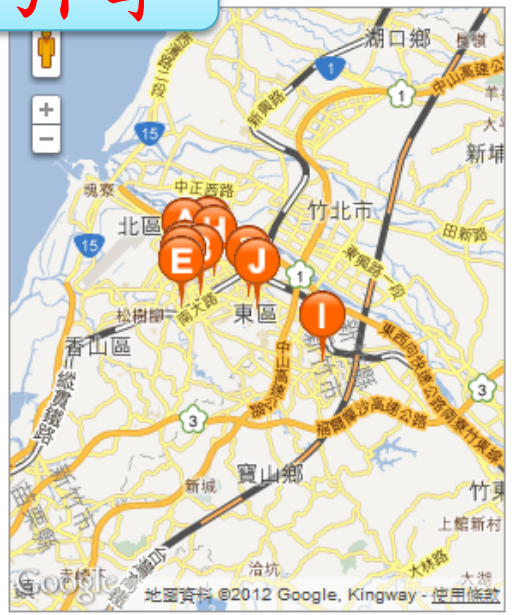
C 台北林季麻辣火鍋連鎖店 新竹市民生路266號

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

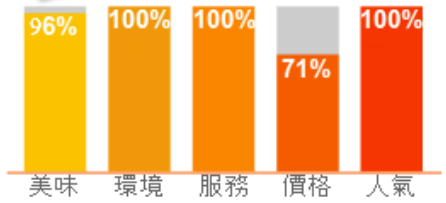
電話：03-5260688
營業時間：
消費價位：一般價位

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

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. **Overview of NLTK**
5. Application of NLTK: Topic Classification on PTT

Install NLTK

Python 2.6+

pip install numpy
pip install nltk

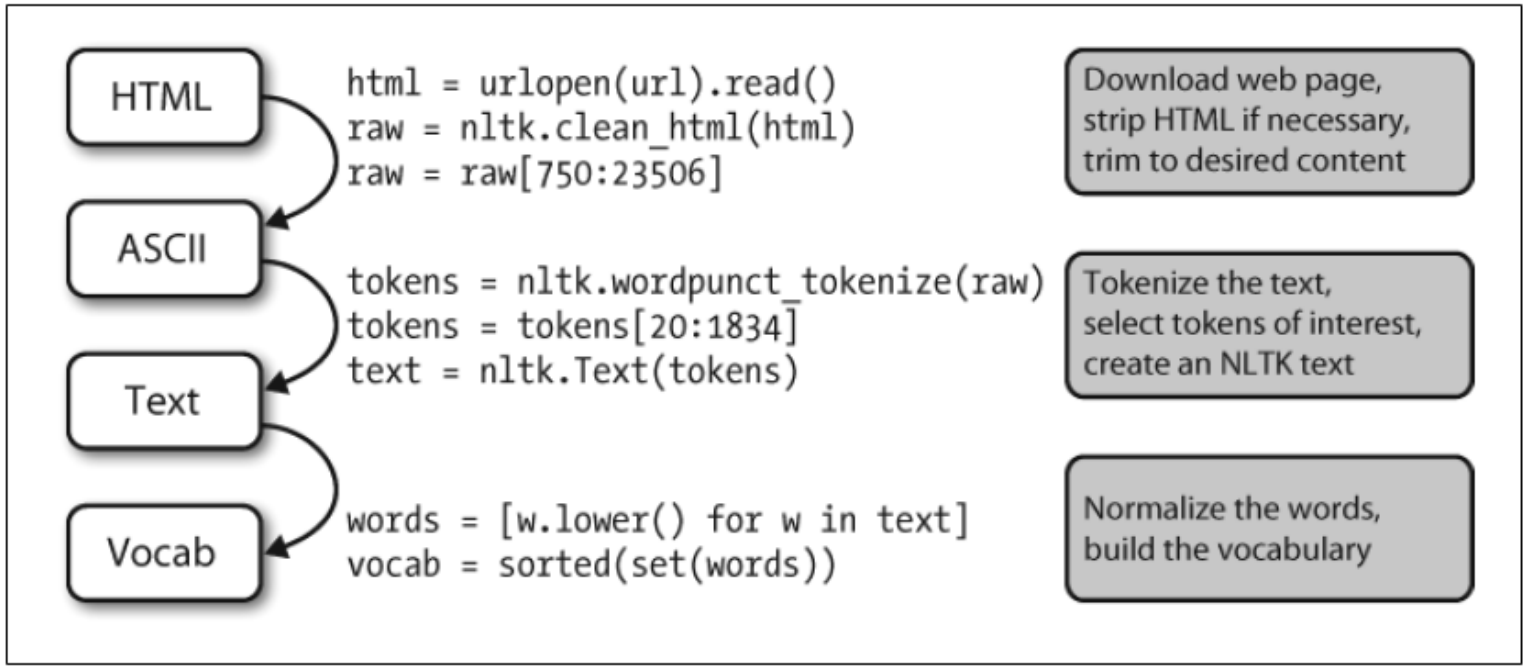
Annotated Text Corpora

```
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
```

```
#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)
```

NLP in NLTK –
Text Tokenization, Normalization

Text Processing Flow



Resources:
from nltk.tokenize import *

NLP in NLTK –
Part-of-speech Tagging

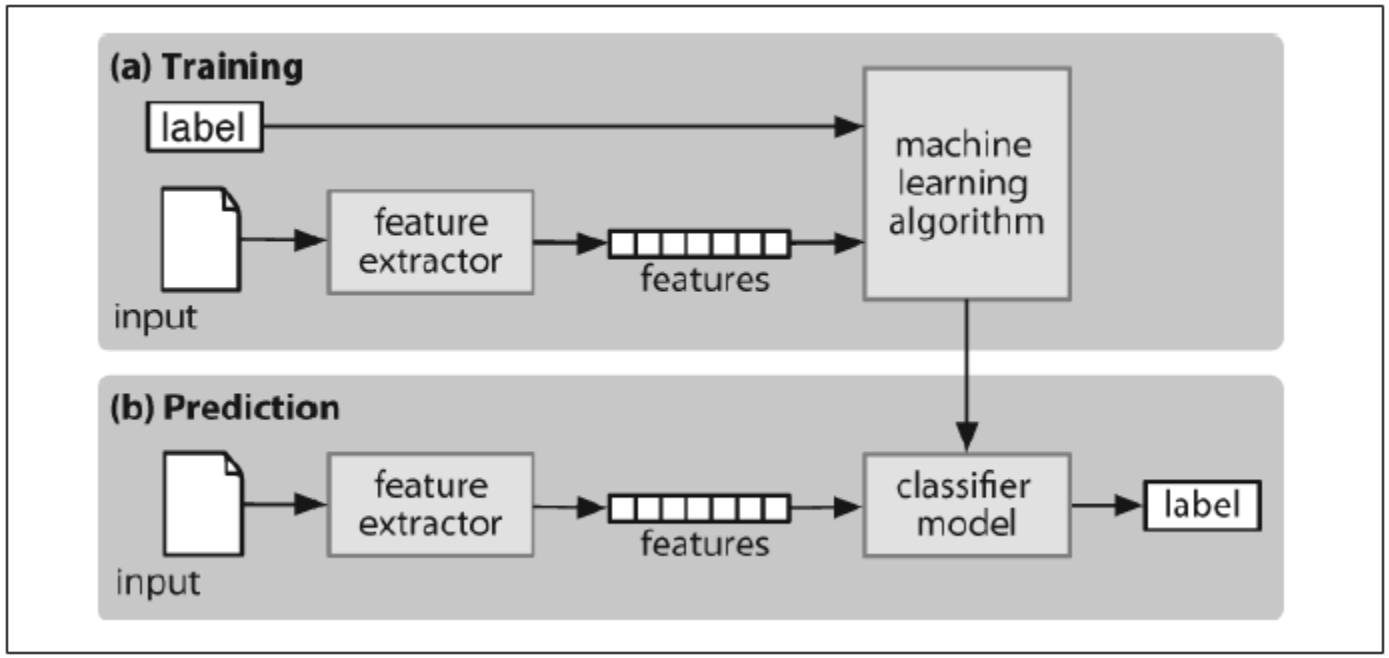
```
>>> text = nltk.word_tokenize("And now for something completely different")
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),
('completely', 'RB'), ('different', 'JJ')]
```

Tag	Meaning	Examples
ADJ	adjective	*new, good, high, special, big, local*
ADV	adverb	*really, already, still, early, now*
CNJ	conjunction	*and, or, but, if, while, although*
DET	determiner	*the, a, some, most, every, no*
EX	existential	*there, there's*
FW	foreign word	*dolce, ersatz, esprit, quo, maitre*
MOD	modal verb	*will, can, would, may, must, should*
N	noun	*year, home, costs, time, education*

Resources:
from nltk.tag import *

2012 TAIWAN PYCON

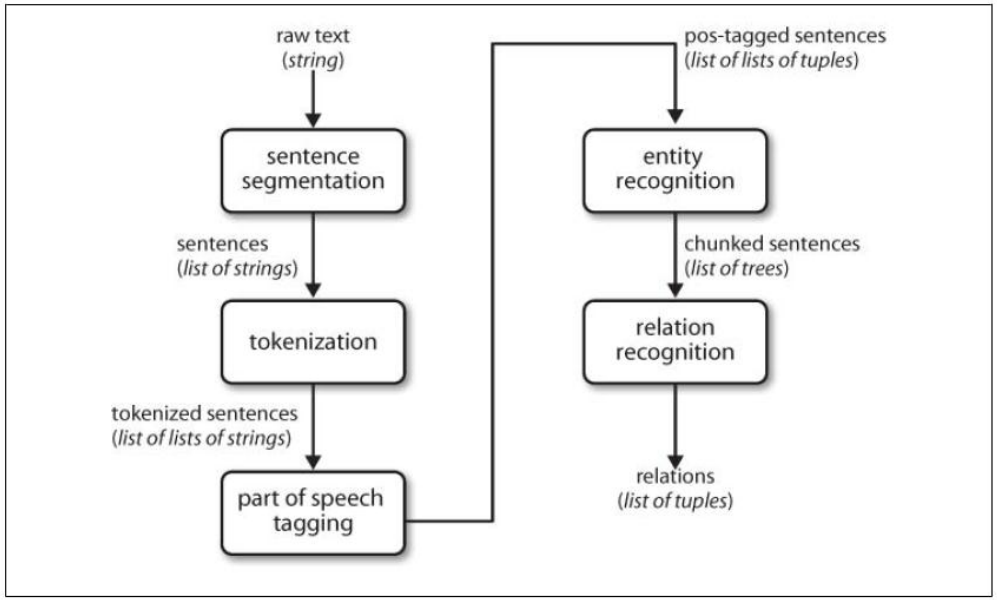
NLP in NLTK –
Text Classification



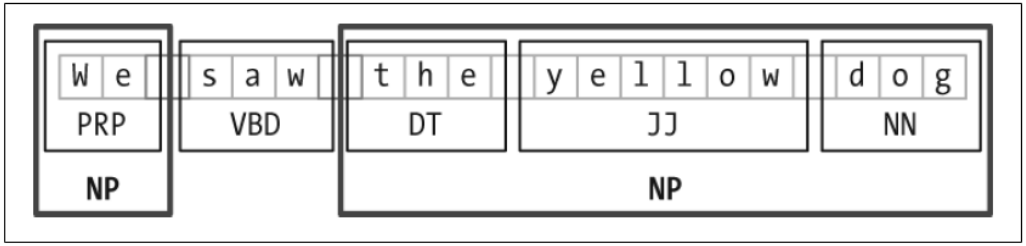

```
>>> classifier.classify(gender_features('Neo'))
'male'
>>> classifier.classify(gender_features('Trinity'))
'female'
```

Resources:
from nltk.classify import *

NLP in NLTK –
Entity Recognition




```
>>> grammar = "NP: {<DT>?<JJ>*<NN>}"

>>> cp = nltk.RegexpParser(grammar)
>>> result = cp.parse(sentence) ❹
```

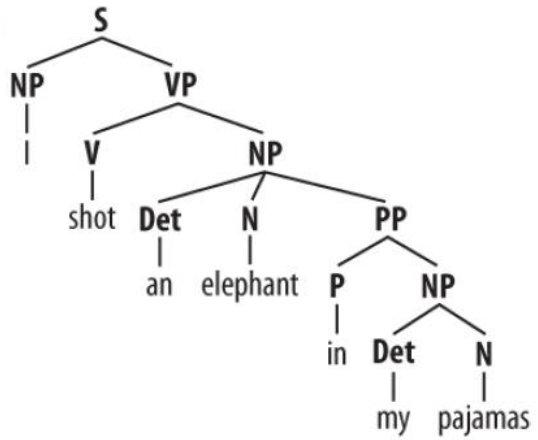


Resources:
from nltk.chunk import *

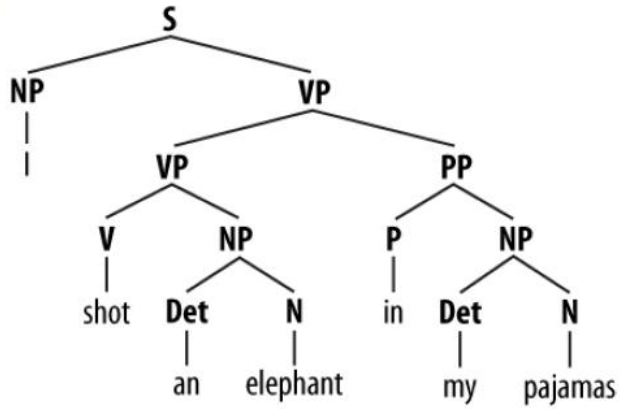
2012 TAIWAN PYCON

NLP in NLTK – Grammar Tree

a.



b.




```
grammar2 = nltk.parse_cfg("""
  S  -> NP VP
  NP -> Det Nom | PropN
  Nom -> Adj Nom | N
  VP -> V Adj | V NP | V S | V NP PP
  PP -> P NP
  PropN -> 'Buster' | 'Chatterer' | 'Joe'
  Det -> 'the' | 'a'
  N -> 'bear' | 'squirrel' | 'tree' | 'fish' | 'log'
  Adj  -> 'angry' | 'frightened' |  'little' | 'tall'
  V ->  'chased'  | 'saw' | 'said' | 'thought' | 'was' | 'put'
  P -> 'on'
  """)
```

Resources:
from nltk.parse import *

NLP in NLTK – Semantic of Sentence

- Propositional Logic
- First-Order Logic
- Disclosure Semantics

Boolean operator	Truth conditions		
negation (*it is not the case that ...*)	-φ is true in *s*	iff	φ is false in *s*
conjunction (*and*)	(φ & ψ) is true in *s*	iff	φ is true in *s* and ψ is true in *s*
Boolean operator	**Truth conditions**		
disjunction (*or*)	(φ | ψ) is true in *s*	iff	φ is true in *s* or ψ is true in *s*
implication (*if ..., then ...*)	(φ -> ψ) is true in *s*	iff	φ is false in *s* or ψ is true in *s*
equivalence (*if and only if*)	(φ <-> ψ) is true in *s*	iff	φ and ψ are both true in *s* or both false in *s*

Example	Description
=	Equality
!=	Inequality
exists	Existential quantifier
all	Universal quantifier

Resources:
from nltk.sem import *

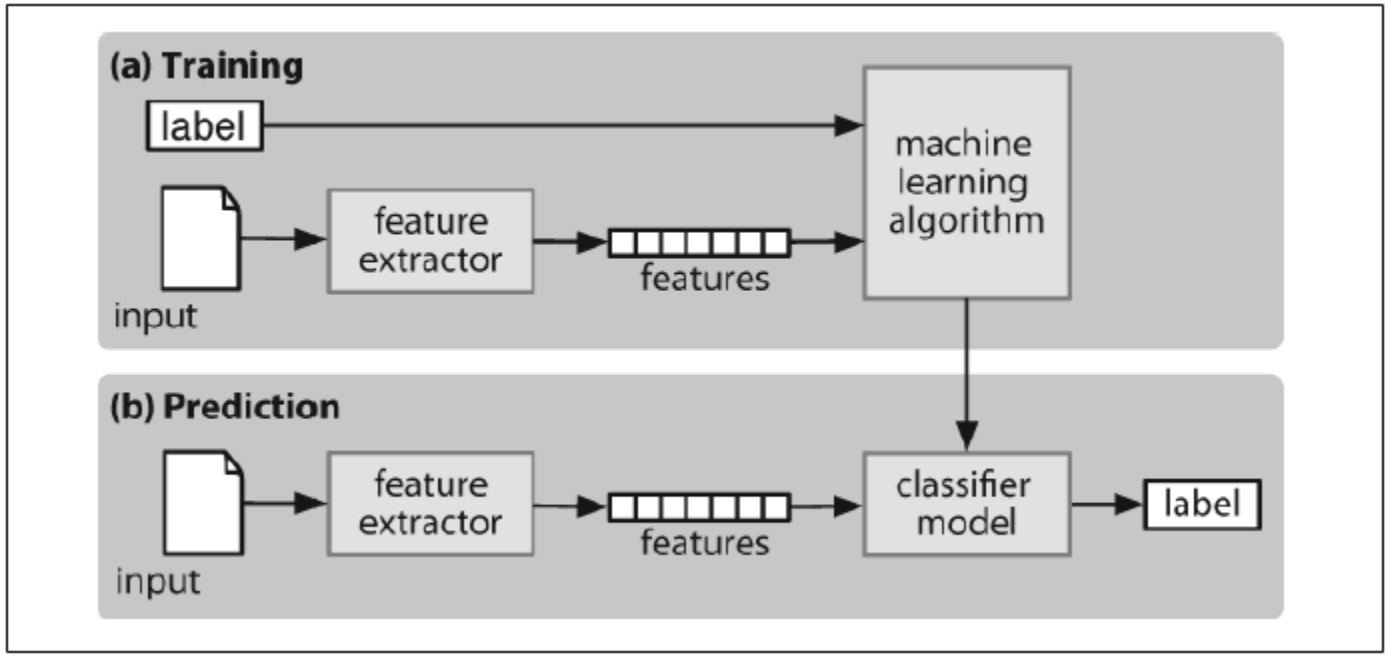
Outline

1. An application based on NLP: 聚寶評
2. Introduction to Natural Language Processing
3. Brief History of NLTK
4. Overview of NLTK
5. **Application of NLTK: Topic Classification on PTT**

Topic Classification on PTT

- 熱門看板:
  - 文章主題明確: Food(美食版), HatePolitics(政黑板), Baseball(棒球版), Stock(股票版), Boy-Girl(男女版)
  - 文章主題廣泛: Gossiping(八卦版)
- 目標: 將八卦版的文章依照主題分類，就可以只挑選有興趣的主題的文章來閱讀。

System Flow


(a) Training
label
machine learning algorithm
feature extractor
features
input
(b) Prediction
feature extractor
features
classifier model
label
input

Tokenization

```
print 'tokenization by unicode char'
words = article['content']
```

```
[ ] [ ] [看] [板] [ ] [ ] [B] [o] [y] [-]
[G] [i] [r] [l] [ ] [
] [ ] [標] [題] [ ]
[ ] [R] [e] [:] [ ] [[] [求] [助] []] [ ]
[（] [代] [p] [o] [）] [經] [濟] [能] [力] [不]
[對] [等] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[
] [ ] [時] [間] [ ] [ ] [S] [u] [n] [ ]
[F] [e] [b] [ ] [ ] [6] [ ] [1] [1] [:]
[3] [4] [:] [2] [8] [ ] [2] [0] [1] [1]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[ ] [ ] [ ] [ ] [ ] [ ] [ ] [
] [─] [─]
[─] [─] [─] [─] [─] [─] [─] [─] [─] [─]
[─] [─] [─] [─] [─] [─] [─] [─] [─] [─]
```

```
print 'tokenization by word_tokenize'
words = nltk.word_tokenize(article['content'])
```

```
[看] [板] [Boy-Girl] [標] [題] [Re] [:] [[] [求] [助]
[]] [（] [代] [po] [）] [經] [濟] [能] [力] [不]
[對] [等] [時] [間] [Sun] [Feb] [6] [11] [:] [34]
[:] [28] [2011] [─] [─] [─] [─] [─] [─] [─]
[─] [─] [─] [─] [─] [─] [─] [─] [─] [─]
[─] [─] [─] [─] [─] [─] [─] [─] [─] [─]
[─] [─] [─] [─] [─] [─] [─] [─] [─] [─]
[─] [─] [唉] [※] [引] [述] [《] [K5] [(] [Ｋ]
[我] [)] [》] [之] [銘] [言] [：] [:] [（] [代]
[po] [）] [:] [這] [個] [問] [題] [困] [擾] [我]
[好] [一] [陣] [子] [了] [.....] [.] [:] [我] [目]
[前] [二] [十] [八] [歲] [跟] [我] [男] [友] [交]
[往] [一] [年] [左] [右] [:] [我] [男] [友] [小]
[我] [三] [歲] [退] [伍] [沒] [多] [久] [正] [在]
[準] [備] [會] [計] [師] [考] [試] [:] [我] [目]
[前] [是] [有] [工] [作] [的] [一] [個] [月] [收]
[入] [約] [三] [萬] [五] [到] [四] [萬] [左] [右]
[:] [很] [多] [朋] [友] [知] [道] [我] [男] [友]
[在] [準] [備] [會] [計] [師] [考] [試] [都] [說]
[「] [啊] [～] [他] [考] [上] [、] [你] [就] [好]
```

```
[看板] [Boy] [-]
[Girl] [標題] [Re]
[:] [[] [求助]
[]] [（] [代po]
[）] [經濟能力不對等] [時間]
[Sun] [Feb] [6]
[11] [:] [34]
[:] [28] [2011]
[──────────────────────────────────────] [唉] [※]
[引述] [《] [K5]
[(] [Ｋ我] [)》]
[之銘言] [：] [:]
[（] [代po] [）]
[:] [這個問題] [困擾我好一陣子了]
[......] [:] [我目前二十八歲]
[跟我男友交往一年左右] [:] [我男友小我三歲]
[退伍沒多久] [正在準備會計師考試] [:]
```

```
print 'tokenization by wordpunct_tokenize'
words = nltk.wordpunct_tokenize(article['content'])
```

Text Classification

```
classes = [u'Food', u'HatePolitics', u'Baseball', u'Stock', u'Boy-Girl']

conn = pymongo.Connection()
coll = conn.noahsark.article
tagged_articles = []
for cls in classes:
    for article in coll.find({'board': cls}, limit=100):
        tagged_articles.append((article['content'], cls))

def get_feature_contain_word(article):
    feature = {}
    words = nltk.word_tokenize(article)
    for word in words:
        feature[u'contains(%s)' % word] = True
    return feature

feature_set = [(get_feature_contain_word(article), cls)
               for (article, cls) in tagged_articles]
random.shuffle(feature_set)
size = int(len(feature_set) * 0.9)
train_set = feature_set[:size]
test_set = feature_set[size:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print 'NaiveBayesClassifier accuracy=', nltk.classify.accuracy(classifier, test_set)
```

```
NaiveBayesClassifier accuracy= 0.932
```

Result – Boy-Girl

Boy-Girl
[H 作者 soulgel (習慣笑的像風鈴) 看板 Gossiping
標題 Re: [問卦] 有沒有很多女生想玩到30歲以後才想結婚 …
時間 Tue Feb 8 00:35:02 2011
- -

※ 引述《moebius2 (漫無目的)》之銘言：
: ※ 引述《littlest (讀冊人...)》之銘言：
: : 男人可以玩到3X歲才結婚，女人為什麼不行勒?
: : 反正都男女平等了，女人賺得錢也跟男人差不多了，
: 男人女人賺錢目地不同
: 假設兩個都同學歷
:　　　　男人　　女人
: --
:　18 ~ 24　　辛苦唸書　　普通唸書
:　　　　+ 運氣好的　　+容易找到家教
:　　　　兼些家教　　or 時薪高工作
:　　　　賺些小錢　　(show girl 酒促 或者..)
:　　　　照顧女生　　+很多殺俾使(宵夜 早餐)
:
: 25歲　　碩士畢業　　碩士畢業
: 26歲　　當時薪8元勞工　　開始工作賺錢
:　　　　賺的錢大該能買買泡麵 煙就沒了　　賺錢買包包
:　　　　存款=0　　存款=0
:
:　　　　男生要當兵
: --
: 27~28　　第一份工作　　工作賺錢
:　　　　努力存錢買車　　存錢出國遊學[7;3H----------

Result – HatePolitics

HatePolitics
[H 作者 cff9900ff (紫色) 看板 Gossiping
標題 [新聞] 台嫌遣陸事件 菲聲明表遺憾
時間 Mon Feb 7 18:04:24 2011

（中央社記者林行健馬尼拉7日專電）針對14名台籍跨國詐騙案嫌犯被遣送到中國，菲方7日發佈聲明表示「深切遺憾」，並正與台灣共同研議相關機制，以確保未來處理類似事件時，雙方能更密切協調。

聲明強調，台灣與菲律賓存在長久的友誼和共同價值，希望這次事件不會損害雙方人民間的緊密情誼，菲律賓仍持續歡迎台灣人來菲，並為守法的台灣人提供保護。

馬尼拉經濟文化辦事處（MECO）7日下午發表的聲明說，MECO對於這宗涉及台灣人的事件感到深切遺憾，也了解台灣當局和人民對於菲方動作的感受。

為表示慎重，MECO在發佈之前邀請駐菲代表李傳通到MECO辦公室過目，並由MECO理事主席培瑞斯（AmadeoPerez Jr.）親自把聲明交予李傳通。MECO駐台代表白熙禮、前駐台代表拉貝（Raul Rabe）也都在場。

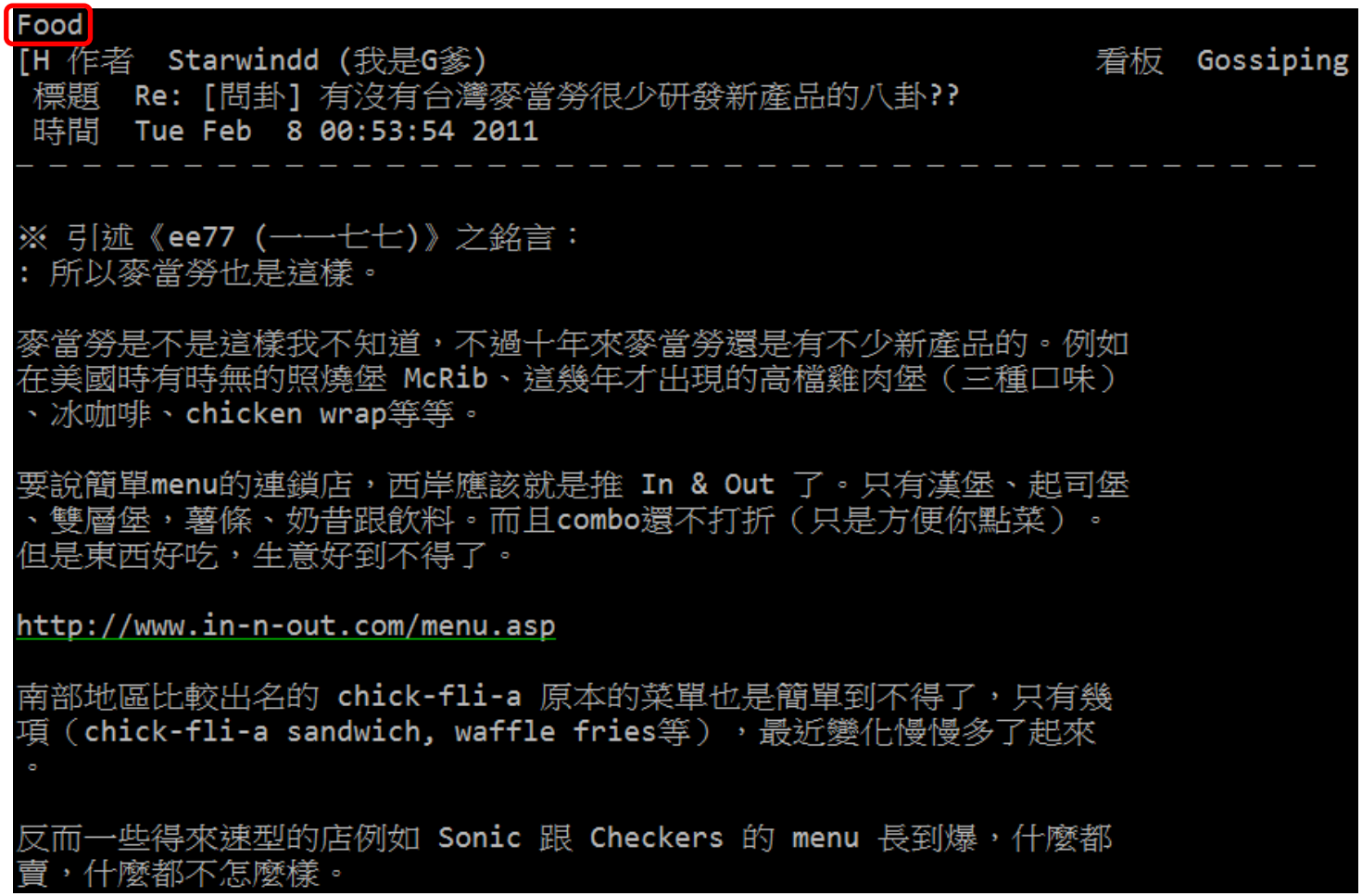
聲明說，菲方是依此案受害人為中國人、共犯也為中國人、以及本案在中國可以得到最妥適解決等考量，決定了遣送的動作。

聲明強調，跨國集團犯罪是重大的國際罪行，希望這次事件可以作為警示。

MECO也指出，中國和菲律賓簽有引渡條約，以及菲方了解兩岸簽訂共同打擊犯罪及司法互助協議，在此協議之下，兩岸可以共同處理本案。

聲明也提到，法院指示移民局制止遣送，司法部和移民局並未漠視任何相關程序，司法部相信菲方採取了適當的動作，維持領土不受國際犯罪所害，也感謝台灣及菲國警方協助菲律賓對付國際犯罪份子

Result – Food



```
Food
[H 作者  Starwindd (我是G爹)                                    看板  Gossiping
 標題  Re: [問卦] 有沒有台灣麥當勞很少研發新產品的八卦??
 時間  Tue Feb  8 00:53:54 2011
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

※ 引述《ee77 (一一七七)》之銘言：
: 所以麥當勞也是這樣。

麥當勞是不是這樣我不知道，不過十年來麥當勞還是有不少新產品的。例如
在美國時有時無的照燒堡 McRib、這幾年才出現的高檔雞肉堡（三種口味）
、冰咖啡、chicken wrap等等。

要說簡單menu的連鎖店，西岸應該就是推 In & Out 了。只有漢堡、起司堡
、雙層堡，薯條、奶昔跟飲料。而且combo還不打折（只是方便你點菜）。
但是東西好吃，生意好到不得了。

http://www.in-n-out.com/menu.asp

南部地區比較出名的 chick-fli-a 原本的菜單也是簡單到不得了，只有幾
項（chick-fli-a sandwich, waffle fries等），最近變化慢慢多了起來
。

反而一些得來速型的店例如 Sonic 跟 Checkers 的 menu 長到爆，什麼都
賣，什麼都不怎麼樣。
```

Result – Stock

```
Stock
[H 作者  CoffeeandTea (back to road!!!)                    看板  Gossiping
 標題  Re: [新聞] 推三環三線 發展火車頭
 時間  Sun Feb  6 18:15:04 2011
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

最近電視狂打廣告的  林口  世界首席  潛銷4x/p  坪數80~240
保守估計  80 x 35+車位=3000萬

拿得出3000萬的會買林口?  拿得出5000萬的會買新莊???
我一年前在home-sale板提出林口新成屋均價上看40/p  新莊新成屋均價上看60/p
為何房地產還好好的
因為低利
+游資(a.遺產稅由50%->10%,大批海外資金回流,b.歐美景氣不好,熱錢回亞洲)
+通膨(歐美狂灑錢,石油+原物料農產品都慢慢往上翻)

+金融風暴2年只繳利息優惠
+新北市預售屋而已,自備一兩百萬離交屋還有至少兩年

大家在玩死亡遊戲  等著看那些現在蓋的房子一旦成屋開始付房貸,一旦免繳本金截止
一旦美國開始升息(今年必開始升息,最快年中,最慢年底),而且利率到3%
很好玩

我算過  1500萬的房子  自備200萬  貸1300萬  分30年還  一個月本利和約還4萬
連台北市都很少有人會租到三萬以上的房子(特殊外商或營業用除外)
如果供幾量出來租不出去的話  會死人的
現在連台北市一堆地方房價上漲.房租卻不漲
新北市  科科  台北市現在已經是  大戶投資客+台商+陸資的天堂
一堆小投資客跟自住客被迫買新北市
```

Reference

- Steven Bird, Ewan Klein, and Edward Loper, "**Natural Language Processing with Python**", 2009. introduce: Python + NLP + NLTK
- Jacob Perkins, "Python Text Processing with NLTK 2.0 Cookbook", 2010
- 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.

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
- 市場研究及網路服務產品設計經理

- **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