UNIT 12

語言搜尋引擎

December 9, 2014

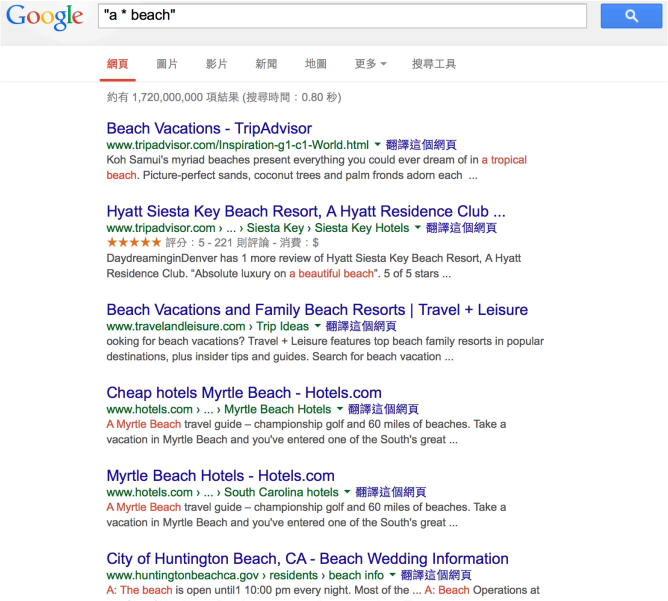
目標

- 建立一搜尋引擎用於搜尋英文詞語用法。
- 可輔助英語學習與文章寫作。

搜尋例子

- `adj. beach`: 即代表搜尋 beach 前面出現過的形容詞。
- `play * role`: 搜尋 play 與 role 中間最常出現的字詞組合。
- `go ?to home`: go 與 home 之間是否要放 to。
- `go * movie`: go 與 role 中間最常出現的字詞組合。
- `kill the _`: 最常被 kill 的東西是。

用 Google 查英文



"a * beach"

約有 1,720,000,000 項結果 (搜尋時間：0.80 秒)

Beach Vacations - TripAdvisor
www.tripadvisor.com/Inspiration-g1-c1-World.html 翻譯這個網頁
Koh Samui's myriad beaches present everything you could ever dream of in a tropical beach. Picture-perfect sands, coconut trees and palm fronds adorn each ...

Hyatt Siesta Key Beach Resort, A Hyatt Residence Club ...
www.tripadvisor.com › ... › Siesta Key › Siesta Key Hotels 翻譯這個網頁
★★★★★ 評分：5 - 221 則評論 - 消費：$
DaydreaminginDenver has 1 more review of Hyatt Siesta Key Beach Resort, A Hyatt Residence Club. “Absolute luxury on a beautiful beach”. 5 of 5 stars ...

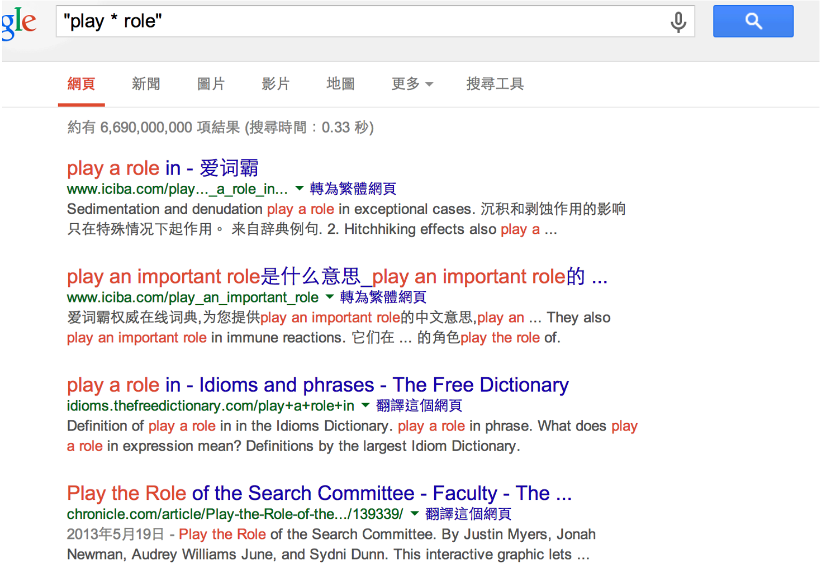
Beach Vacations and Family Beach Resorts | Travel + Leisure
www.travelandleisure.com › Trip Ideas 翻譯這個網頁
ooking for beach vacations? Travel + Leisure features top beach family resorts in popular destinations, plus insider tips and guides. Search for beach vacation ...

Cheap hotels Myrtle Beach - Hotels.com
www.hotels.com › ... › Myrtle Beach Hotels 翻譯這個網頁
A Myrtle Beach travel guide – championship golf and 60 miles of beaches. Take a vacation in Myrtle Beach and you've entered one of the South's great ...

Myrtle Beach Hotels - Hotels.com
www.hotels.com › ... › South Carolina hotels 翻譯這個網頁
A Myrtle Beach travel guide – championship golf and 60 miles of beaches. Take a vacation in Myrtle Beach and you've entered one of the South's great ...

City of Huntington Beach, CA - Beach Wedding Information
www.huntingtonbeachca.gov › residents › beach info 翻譯這個網頁
A: The beach is open until1 10:00 pm every night. Most of the ... A: Beach Operations at

用 Google 查英文



"play * role"

約有 6,690,000,000 項結果 (搜尋時間：0.33 秒)

play a role in - 爱词霸
www.iciba.com/play..._a_role_in... ▾ 轉為繁體網頁
Sedimentation and denudation play a role in exceptional cases. 沉积和剥蚀作用的影响只在特殊情况下起作用。 来自辞典例句. 2. Hitchhiking effects also play a ...

play an important role是什么意思_play an important role的 ...
www.iciba.com/play_an_important_role ▾ 轉為繁體網頁
爱词霸权威在线词典,为您提供play an important role的中文意思,play an ... They also play an important role in immune reactions. 它们在 ... 的角色play the role of.

play a role in - Idioms and phrases - The Free Dictionary
idioms.thefreedictionary.com/play+a+role+in ▾ 翻譯這個網頁
Definition of play a role in in the Idioms Dictionary. play a role in phrase. What does play a role in expression mean? Definitions by the largest Idiom Dictionary.

Play the Role of the Search Committee - Faculty - The ...
chronicle.com/article/Play-the-Role-of-the.../139339/ ▾ 翻譯這個網頁
2013年5月19日 - Play the Role of the Search Committee. By Justin Myers, Jonah Newman, Audrey Williams June, and Sydni Dunn. This interactive graphic lets ...

語法設計

語法	說明
_	單一任意字詞
*	零到多個任意字詞
?term	term 可有可無
term1 | term2	term1 或 term2
adj. det. n. v. prep	形容詞、冠詞、名詞、動詞、介繫詞

搜尋例子

- adj. beach: 即代表搜尋 beach 前面出現過的形容詞。
- play * role: 搜尋 play 與 role 中間最常出現的字詞組合。
- go ?to home: go 與 home 之間是否要放 to。
- go * movie: go 與 role 中間最常出現的字詞組合。
- kill the _: 最常被 kill 的東西是。

Lab 12

- 目標：完成語法第一項 _
  - 任意位置置入 _
  - 最長 4-gram

Query 範例

  - `play _ _ role`
  - `kill the _`
  - `a _ beach`

- 輸入資料：citeseerx 的許多句子
- 輸出結果：
  - key: 所有會有結果的 query
  - value: 符合 query 的前 100 名 ngram 與 count。

Lab 12 - 輸出

- key: 所有會有結果的 query
- value: 符合 query 的前 100 名 ngram 與 count。

輸出範例

Key	Ngrams	Counts
a _ beach	a sandy beach	486
	a private beach	416
	a beautiful beach	314
	a small beach	175
	...	
kill the _	kill the people	189
	kill the other	174
	kill the process	163
	kill the enemy	160
	...	

隨堂測驗

目標

- 依 MapReduce 架構，設計每階段 mapper, reduce 的輸入輸出來完成 Lab 12
- 在紙寫撰寫簡單輸入、輸出的 key-value 範例表達概念即可

小提示

- 可有 1 至多個 map, reduce 流程
- 考慮 mapper 的輸入資料切割影響
- mapper 輸入為 value 或 key-value，輸出為 key-value
- reducer 輸入為 grouped key-values，輸出為 key-value

Bi-gram Count

Bi-gram Count Mapper 範例

Input(value)	Output(key => value)
C D C D	C D => 2
	D C => 1
B C D A	B C => 1
	C D => 1
	D A => 1
C D A B	C D => 1
	D A => 1
	A B => 1

Reducer 範例

Input(key => value)	Output(key => value)
A B => 1	A B => 1
B C => 1	B C => 1
C D => 2	C D => 4
C D => 1	
C D => 1	
D A => 1	D A => 2
D A => 1	
D C => 2	C C => 2

Lab12

Lab12 Mapper 範例

Input(value)	Output(key => value)
A B C 200	A B C => A B C 200
	_ B C => A B C 200
	A _ C => A B C 200
	A B _ => A B C 200
	_ _ C => A B C 200
	_ B _ => A B C 200
	A _ _ => A B C 200
A D C 300	_ D C => A D C 300
	A _ C => A D C 300
	...
A E C 100	_ E C => A E C 100
	A _ C => A E C 100
	...

Lab12

Lab12 Reducer 範例

Input(value)	Output(key => value)
A _ C => A B C 200	A _ C => A D C 300,
A _ C => A D C 300	A B C 200,
A _ C => A E C 100	A E C 100
A B _ => A B C 200	A B _ => A B C 200
A D _ => A D C 300	A D _ => A D C 300
A E _ => A E C 100	A E _ => A E C 100
A _ _ => A B C 200	A _ _ => A D C 300,
A _ _ => A D C 300	A B C 200,
A _ _ => A E C 100	A E C 100
_ B C => A B C 200	_ B C => A B C 200
_ D C => A D C 300	_ D C => A D C 300
_ E C => A E C 100	_ E C => A E C 100
...	...

Lab12

需完成六支程式

- 產生 ngram count 的 mapper, reducer
- 產生 query result 的 mapper, reducer
- 將 query result 轉為 database
  (試試 python 內建的 shelve 或 sqlite3 套件)
- Database 介面程式，讓使用者輸入 query，即時取得 result

python shelve

```python
import shelve
d = shelve.open('data.shelve')
d['odds'] = [1, 3, 5, 7, 9]
print d['odds']
d['evens'] = [2, 4, 6, 8, 10]
d['hello'] = 'world'
del d['hello']
d['zipcodes'] = {'hsinchu': 300, 'zhongli': 320}
print d.keys()
d.close()
```

Google “python shelve” for official documents

Ngram Count I

```python
#!/usr/bin/env python
# -*- coding: utf-8 -*-


def ngrams(words):
    for length in range(1, 5 + 1):
        for ngram in zip(*(words[i:] for i in range(length))):
            yield ngram


def mapper(files):
    import fileinput
    from nltk.tokenize import word_tokenize
    from collections import Counter
    ngram_counter = Counter()
    for line in fileinput.input(files):
        line = line.decode('iso-8859-1')
```

Ngram Count II

```
        words = word_tokenize(line.lower())
        ngram_counter.update(ngrams(words))

    for ngram, count in ngram_counter.iteritems():
        print (u' '.join(ngram) + u'\t' + unicode(count)).encode('utf-8')


def line_to_ngram(line):
    line = line.decode('iso-8859-1')
    return line.split(u'\t', 1)[0]


def line_to_count(line):
    line = line.decode('iso-8859-1')
    return int(line.split(u'\t', 1)[1])


def reducer(files):
    import fileinput
```

Ngram Count III

```
    from itertools import groupby, imap

    for ngram, lines in groupby(fileinput.input(files), key=line_to_ngram):
        count = sum(imap(line_to_count, lines))
        print (ngram + u'\t' + unicode(count)).encode('utf-8')


if __name__ == '__main__':
    import argparse
    import sys
    parser = argparse.ArgumentParser(description='N-gram counter')
    parser.add_argument(
        '-r', '--reducer', action='store_true', help='reducer mode')
    parser.add_argument(
        '-m', '--mapper', action='store_true', help='mapper mode')
    parser.add_argument('files', metavar='FILE', type=str, nargs='*',
                        help='input files')

    args = parser.parse_args()
```

Ngram Count IV

```
    if (args.mapper and args.reducer
            or
            not args.mapper and not args.reducer):
        parser.print_help()
        sys.exit(1)

    if args.mapper:
        mapper(args.files)
    elif args.reducer:
        reducer(args.files)
```