10/3 實驗課一

KN language model

$$P_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c_{KN}(w_{i-n+1}^{i}) - d, 0)}{c_{KN}(w_{i-n+1}^{i-1})} + \lambda(w_{i-n+1}^{i-1}) P_{KN}(w_i \mid w_{i-n+2}^{i-1})$$

- LM = Bigram() + Pcontinuation()

- 用LM來分類句子

Dataset

- 12 category from reuters

```
➜  r_train ll
total 0
drwxr-xr-x  499 kelly  staff    16K Oct  3 00:23 acq
drwxr-xr-x  102 kelly  staff   3.2K Oct  3 00:48 corn
drwxr-xr-x  244 kelly  staff   7.6K Oct  3 00:38 crude
drwxr-xr-x  518 kelly  staff    16K Oct  3 00:35 earn
drwxr-xr-x  216 kelly  staff   6.8K Oct  3 00:56 grain
drwxr-xr-x  116 kelly  staff   3.6K Oct  3 01:12 interest
drwxr-xr-x  225 kelly  staff   7.0K Oct  3 00:46 money-fx
drwxr-xr-x   65 kelly  staff   2.0K Oct  3 01:02 oilseed
drwxr-xr-x  100 kelly  staff   3.1K Oct  3 00:41 ship
drwxr-xr-x   53 kelly  staff   1.7K Oct  3 01:06 soybean
drwxr-xr-x  258 kelly  staff   8.1K Oct  3 00:28 trade
drwxr-xr-x   99 kelly  staff   3.1K Oct  3 01:08 wheat
```

Dataset

```
  A Japanese businessman announced plans for   a new telecommunications firm in
which Britain's Cable and   Wireless Plc would be a core company.
However, the plan, unveiled by senior Federation of   Economic Organizations off
icial Fumio Watanabe, does not   specify what stake Cable and Wireless would hav
e.
"The share holdings of the core companies should be equal,"   Watanabe said in a
 statement.
"The actual percentage of   shareholdings should be agreed by the core companies
."
He said the eight core companies will provide directors for   the firm.
"The new company shall immediately set to work on the   feasibility study of con
structing a new cable for itself,"   Watanabe said.
Watanabe has acted as mediator between two rival groups,   one of which included
 C and W, seeking to compete against   &lt;Kokusai Denshin Denwa Co Ltd>, which
now monopolizes Japan's   overseas telephone business.
The Post and Telecommunications Ministry has said it wants   only one competitor
 to KDD and has backed Watanabe's efforts.
A British source, who declined to be identified further,   said the proposals co
uld open the door to further talks between   C and W &lt;CAWL.L> and the other f
irms involved.
C and W had earlier rejected a reported proposal which   would have given it a f
ive pct share in the new   telecommunications firm, compared to the less than th
ree pct   stake Watanabe originally proposed.
C and W has a 20 pct stake   in one of the two firms Watanabe has proposed shoul
training12409
```

Preprocessing

- Tokenize, Padding , word lower

raw: Today is a bad day.

—> [‘<s>’, ’today’, ’is’, ’a’, ‘bad’, ’day’, ’.’ , ’<\s>’]

KneserNeyLM

- KN($w_i \mid w_{i-1}$) = log(Bigram + $\lambda(w_{i-1})$* PCONTINUATION)
  - Bigram : max(C(w_{i-1}, w_i) -d)/$\sum C(w_{i-1}, w_*)$, 0)
  - PCONTINUATION: max($\sum$type(w_* ,w_i)- d)/ $\sum$type(w_{j-1} ,w_j) , 0)
  - $\lambda(w_{i-1})$ = (d / $\sum C(w_{i-1}, w_*)$)*$\sum$type(w_{i-1} ,w_*)
  - d = 0.75

Score

- Preprocessing
- Score
    - Bigram
    - Pcontinuation
    - Unseen words probability (use the min value from Pcontinuation)
- Article : max (12 scores) as label

Output example

```
testing
category_ 0   hit count:   22
category_ 1   hit count:   7
category_ 2   hit count:   45
category_ 3   hit count:   5
category_ 4   hit count:   84
category_ 5   hit count:   50
category_ 6   hit count:   22
category_ 7   hit count:   11
category_ 8   hit count:   5
category_ 9   hit count:   36
category_ 10   hit count:   92
category_ 11   hit count:   66
Accuracy:   0.6515373352855052
```

評分標準

- Baseline accuracy: 0.6
- 依照正確率計算分數
- 交 ipython notebook 檔, 檔名: lab01_學號
- Deadline: 10/09 23:59
- 每遲交一個禮拜扣十分

Python Useful Class

- from collections import Counter, defaultdict

- Counter([1, 2, 3, 4, 5, 6, 1, 2, 3])

```
In [1]: from collections import Counter, defaultdict

In [2]: sample_list = ['a','a','b','c','c','c']

In [3]: Counter(sample_list)
Out[3]: Counter({'a': 2, 'b': 1, 'c': 3})
```

- defaultdict (set default value for key not in dict.keys())

```
In [7]: sample_dict = defaultdict(int)

In [8]: sample_dict['A']
Out[8]: 0
```

Word tokenization

- Import nltk
- nltk.download(‘punkt’) #execute before calling word_tokenize

```
>>> from nltk import word_tokenize
>>> word_tokenize("Hello world.")
['Hello', 'world', '.']
>>>
```