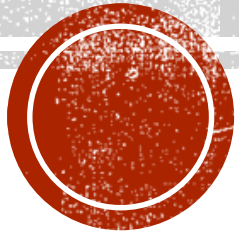
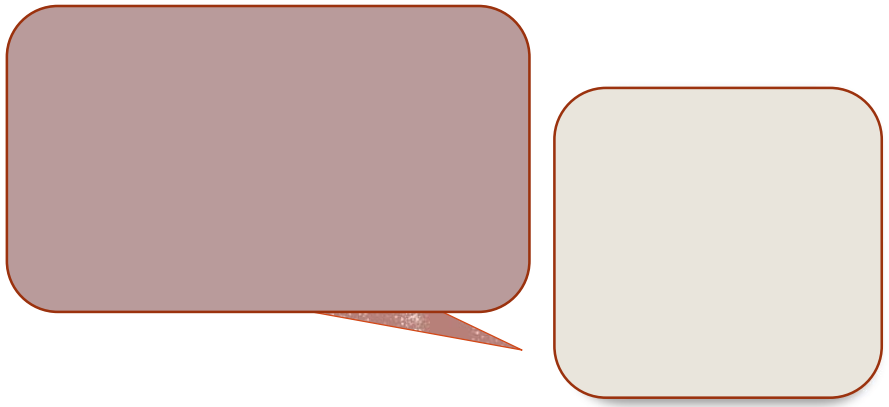
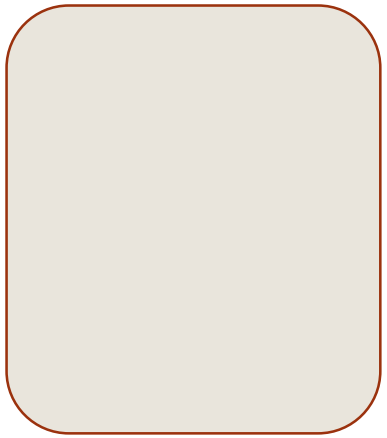
BEYOND KEYNESS(I): TF-IDF

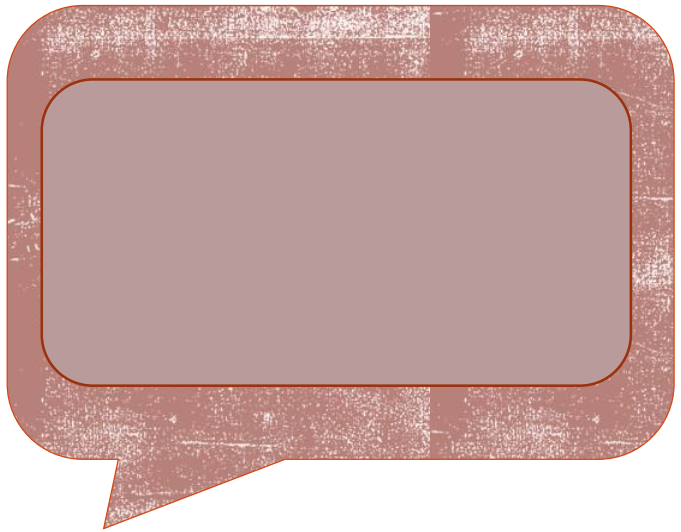
2020/12/13 Po-Ya Angela Wang 王伯雅 (Amber)

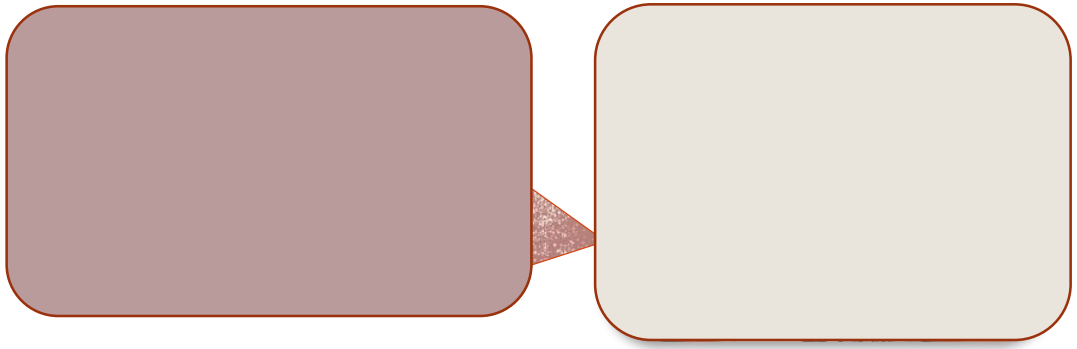


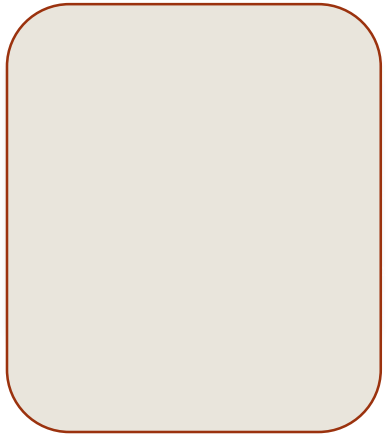
#打一句有聲音的句子

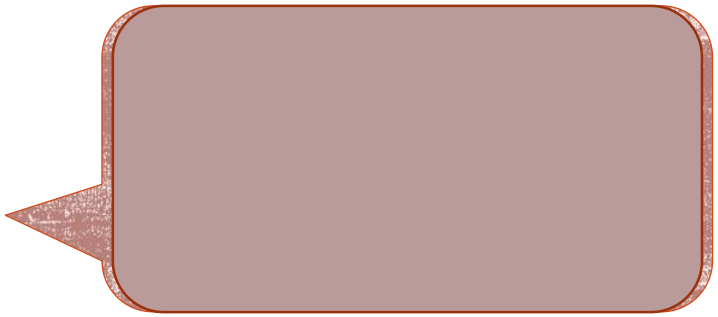












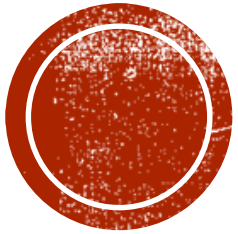
TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY(TF-IDF)

ROADMAP

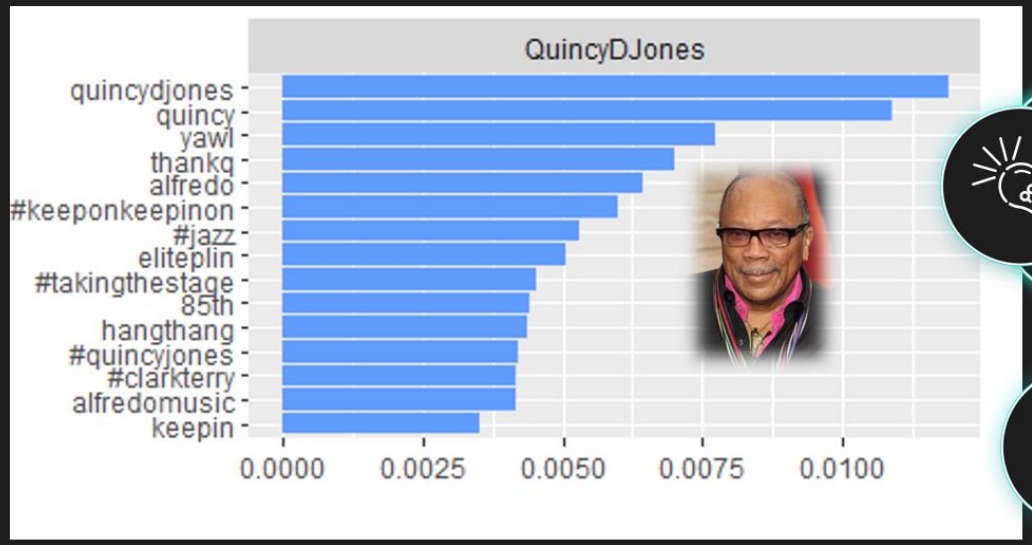
How to apply TF-IDF?

How to define TF-IDF?

How to code TF-IDF



HOW TO APPLY TF-IDF?

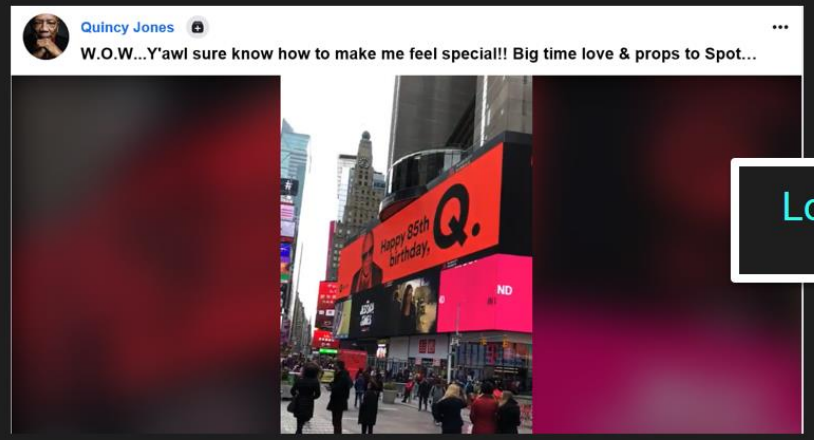


"yawl"

A southern dialect to say “you all. ”

"hang-thang"

“thange” is a southern dialect to say “thing. ”

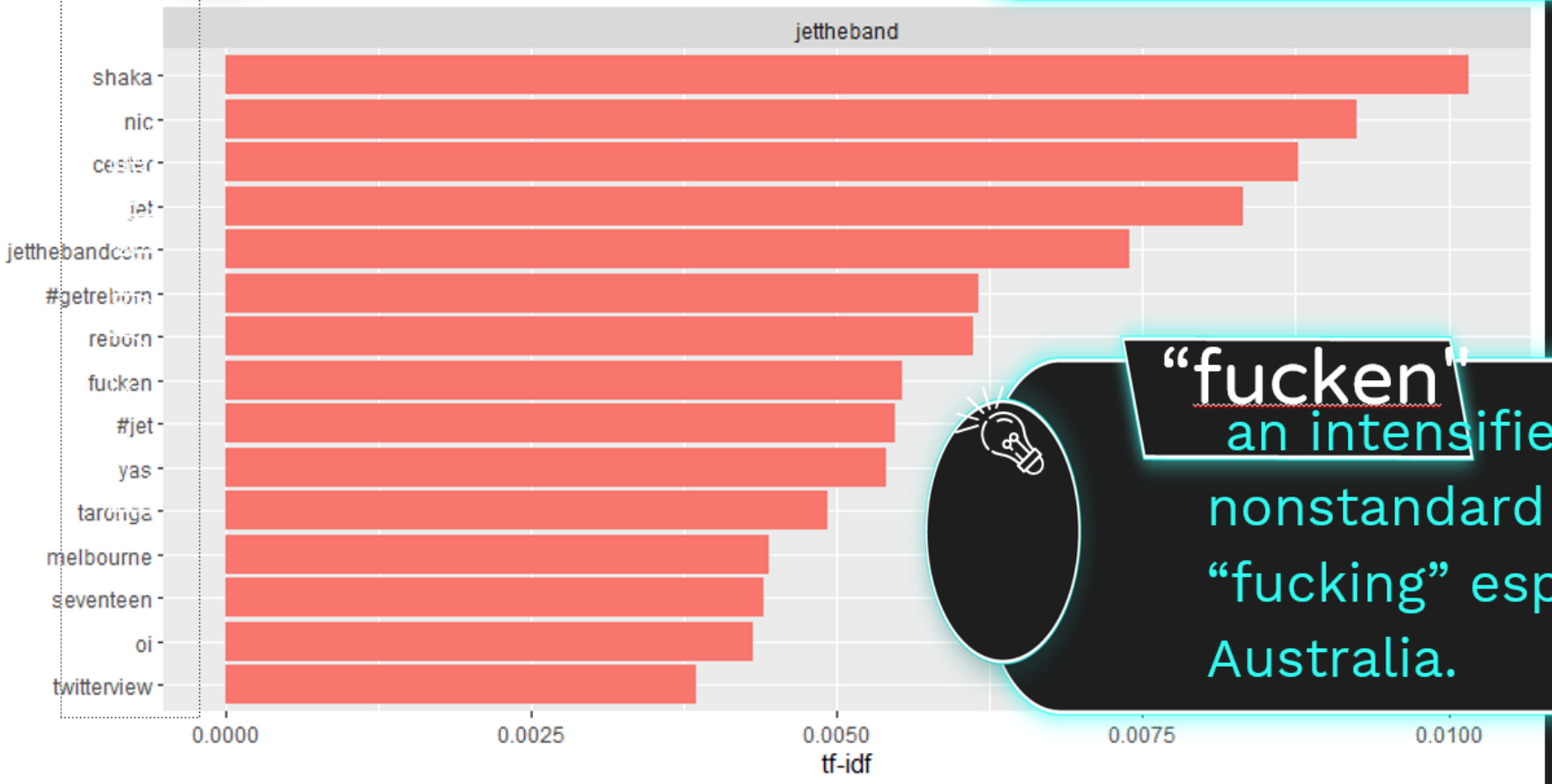


Local Angle

“ya’s”

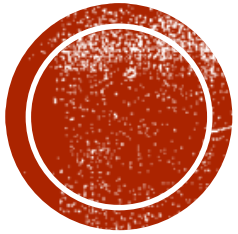
“you guys”

“NEW YORK CITY! See ya's next year....”



“fucken”

an intensifier and nonstandard form of “fucking” especially in Australia.

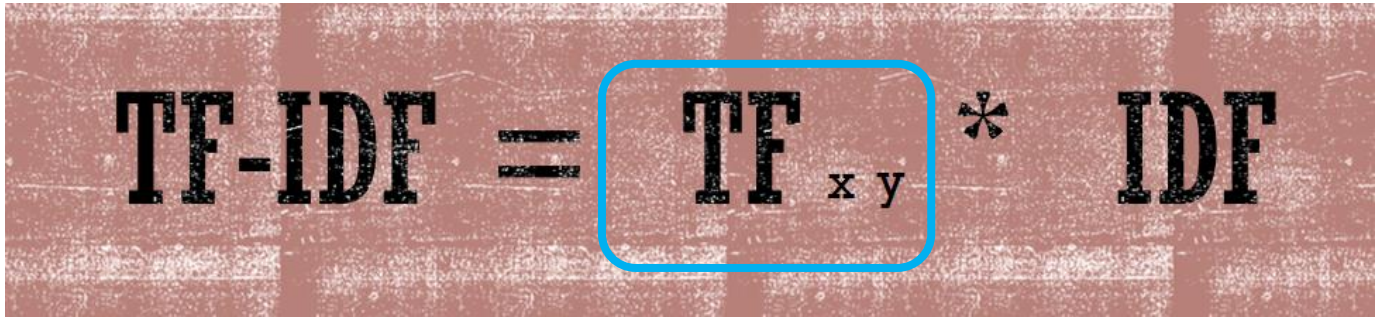


HOW TO DEFINE TF-IDF?

TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY(TF-IDF)

The importance of a word to a document in a collection.

$$TF\text{-}IDF = TF_{x\,y} * IDF$$



Term Frequency

Term **x** within
Document **y**

TF= Word Frequency/Total Word Count

$$TF\text{-}IDF = TF_{xy} * IDF$$

$$IDF = LOG(N/DF_x)$$

Inverse Document Frequency

N = total number of documents

DF= number of documents including x

N➔

DF of “的”➔

$$TF\text{-}IDF = TF_{xy} * IDF$$

$$LOG(N/DF_x)$$

Inverse Document Frequency

N = total number of documents

DF= number of documents including x

N➔ 3

DF of “的”➔ 2

DF of “人”➔ 1

t1=″下 次 有 人 在 色 瞇瞇 看 著 你 就 大方 的 秀 上述 照片 給 他 看″

t2=″當然 前提 是 你 自己 要 先 承受 的 了″

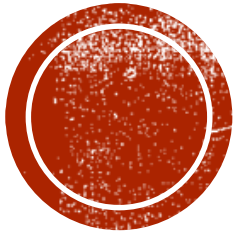
t3=″專門 秀 給 對方 看 可能 會 犯法 不如 把 照片″

$$TF\text{-}IDF = TF_{xy} * IDF$$

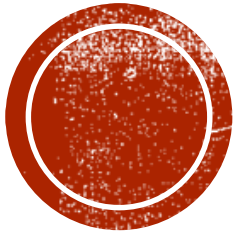
$$LOG(N/DF_x)$$

看	的	人
N= 3	N= 3	N= 3
DF = 1	DF = 2	DF = 1

t1=″下 次 有 人 在 色 瞇瞇 看 著 你 就 大方 的 秀 上述 照片 給 他 看″
t2=″當然 前提 是 你 自己 要 先 承受 的 了″
t3=″專門 秀 給 對方 看 可能 會 犯法 不如 把 照片″



HOW TO CODE TF-IDF?



TIME TO CODE

```
import re
import math
from collections import Counter
```

Workbook link: https://ppt.cc/fkDzBx

DATA OBSERVATION

```
t1="下 次 有 人 在 色 瞇瞇 看 著 你 就 大方 的 秀 上述 照片 給 他 看 小明 小明 小明 小明 小明 小明 小明 小明"
t2="當然 前提 是 你 自己 要 先 承受 的 了 小華 小華 小華 小華 小華 小華 小華 小華"
t3="專門 秀 給 對方 看 可能 會 犯法 不如 把 照片 小白 小白 小白 小白 小白 小白 小白"
```

t1="下 次 有 人 在 色 瞇瞇 看 著 你 就 大方 的 秀 上述 照片 給 他 看 小明 小明 小明 小明 小明 小明 小明 小明"
t2="當然 前提 是 你 自己 要 先 承受 的 了 小華 小華 小華 小華 小華 小華 小華 小華"
t3="專門 秀 給 對方 看 可能 會 犯法 不如 把 照片 小白 小白 小白 小白 小白 小白 小白"

Term Frequency

Term x
within
Document y

TF= Word Frequency/Total Word Count

- **1. Segment target document**
- **2. Count the frequency of target word**
- **3. Count the total word number**
- **4. Get Term Frequency**

KAHOOT

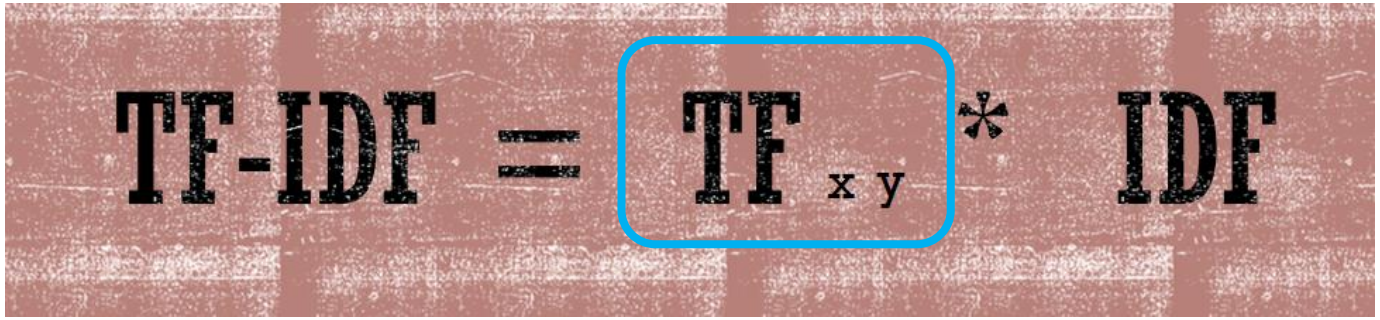
kahoot.it

Play Kahoot! - Enter game PIN here!

Join a game of kahoot here. Kahoot! is a free game-based learning platform that makes it fun to learn – any subject, in any language, on any device, for all ages!

9004235

19




```
def  tf(word, doc):
      words  =  re.split          doc)
      count  =  Counter(words)
      return
```

- **1. Segment target document**
- **2. Count the frequency of target word**
- **3. Count the total word number**
- **4. Get Term Frequency**

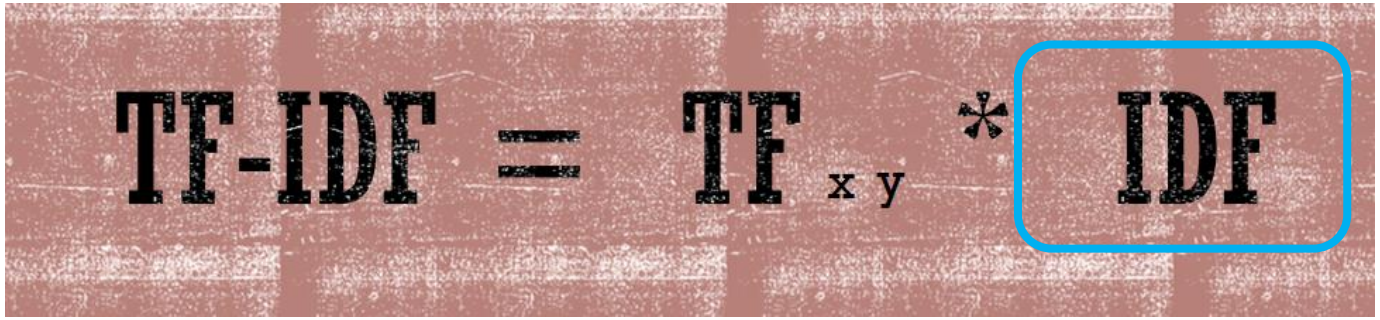
A. Count(words)/Counter(word)

B. count(word)/len(word)

C. count[word]/len(words)

TEST YOUR FUNCTION

```
#Test Your Function tf()
#Please calculate the tf() values of "的" in these 3 texts
t1="下 次 有 人 在 色 瞇瞇 看 著 你 就 大方 的 秀 上述 照片 給 他 看 小明 小明 小明 小明 小明 小明 小明 小明"
t2="當然 前提 是 你 自己 要 先 承受 的 了 小華 小華 小華 小華 小華 小華 小華 小華"
t3="專門 秀 給 對方 看 可能 會 犯法 不如 把 照片 小白 小白 小白 小白 小白 小白 小白"
```



Inverse Document Frequency

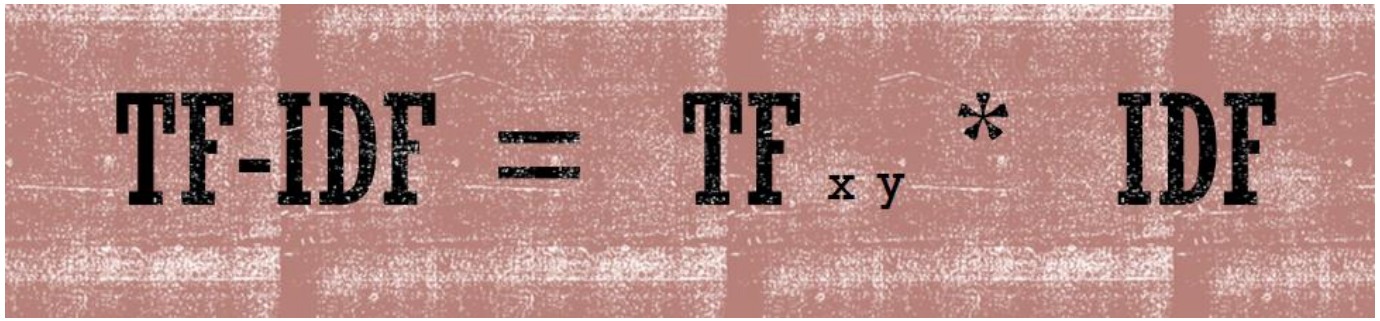
N = total number of documents
DF= number of documents including x

```
def  idf(word, docset):
    N  =                    #total  num  of  doc
    df  =
    for  doc  in  docset:



    idf=  math.log(N/df)
    return  idf
```

```
#Test Your Function idf()
#Please calculate the idf() values of "的"
t1="下 次 有 人 在 色 瞇瞇 看 著 你 就 大方 的 秀 上述 照片 給 他 看 小明 小明 小明 小明 小明 小明 小明 小明"
t2="當然 前提 是 你 自己 要 先 承受 的 了 小華 小華 小華 小華 小華 小華 小華 小華"
t3="專門 秀 給 對方 看 可能 會 犯法 不如 把 照片 小白 小白 小白 小白 小白 小白 小白"
```




```
def  tfidf(                    ):
    return  tf(word, doc)  *  idf(word, docset)
```

```
#Test  Your  Function    tf-idf()
#Please  calculate  the  tf-idf  values  of  "的"  in  different  documents
```

```
#Think about the meaning of the values
#The (larger) the tf-idf vlaue is , the more important the word is to the target document.
```

LET'S TRY ON SOME TEST DATA

Workbook Part 4