# 商管程式設計（110-1）

# 作業六

## 作業設計：盧信銘

## 國立臺灣大學資訊管理學系

繳交作業時，請至PDOGS（http://pdogs.ntu.im/）為第一、二、三、四題各上傳一份 Python 3.9 原始碼（以複製貼上原始碼的方式上傳）。每位學生都要上傳自己寫的解答。不接受紙本繳交；不接受遲交。 這份作業的截止時間是2021年11月22日晚上九點。 為這份作業設計測試資料並且提供解答的助教是郭玟妍。

# 第一題

連續詞是數個連續字元組成的詞。比如說，考慮以下句子: 梁商成為輔政的大將軍 如果我們規定連續詞的字元長度為4，那這個句子有以下的連續詞:

梁商成為
商成為輔
成為輔政
為輔政的
輔政的大
政的大將
的大將軍

一般來說，連續詞不包含標點符號，而且不會跨越標點符號。這個題目使用的標點符號包含:

。，、；：「」『』（）—？！—…﹏《》〈〉·～ ,.; !"#$%&'()*+,-./:;<=>?@[\]^_`{¦}~

所以如果句子是: 永和元年（136年）成為河南尹。漢順帝時因梁妠的關係，梁商成為輔政的大將軍 ，那所有長度為4的連續詞為:

永和元年
136年
成為河南
為河南尹
漢順帝時
順帝時因
帝時因梁
時因梁妠
因梁妠的
梁妠的關
妠的關係
梁商成為
商成為輔
成為輔政
為輔政的
輔政的大
政的大將
的大將軍

寫一個程式，由標準輸入依序讀入連續詞長度與文字內容，並輸出所有指定長度的連續詞。

## 輸入輸出格式

系統將會提供連續詞長度與文字內容。輸入第一行為連續詞長度。數值為整數且至少為2。第二行為文字，長度為一萬個字元以內。如果連續詞長度小於2，則輸出ILLEGAL_N。

你的程式應依原始輸入順序輸出所有符合要求的連續詞，一行一個。

舉例而言，如果輸入是：

---

```
4
永和元年（136年）成為河南尹。漢順帝時因梁妠的關係，梁商成為輔政的大將軍
```

---

那輸出是:

---

```
永和元年
136年
成為河南
為河南尹
漢順帝時
順帝時因
帝時因梁
時因梁妠
因梁妠的
梁妠的關
妠的關係
梁商成為
商成為輔
成為輔政
為輔政的
輔政的大
政的大將
的大將軍
```

---

如果輸入是:

---

```
5
永和元年（136年）成為河南尹。漢順帝時因梁妠的關係，梁商成為輔政的大將軍
```

---

那輸出是:

---

```
成為河南尹
漢順帝時因
順帝時因梁
帝時因梁妠
時因梁妠的
因梁妠的關
梁妠的關係
梁商成為輔
商成為輔政
成為輔政的
為輔政的大
輔政的大將
政的大將軍
```

---

如果輸入是:

---

```
5
The windswept valleys surrounding the Hengill volcano in southwestern Iceland are dotted with h
```

```
ot springs and steam vents. Hikers from all over the world come here to witness its breath-taki
ng scenery. Even the sheep are photogenic in the soft Nordic light.
```

---

那輸出是:

---

```
winds
indsw
ndswe
dswep
swept
valle
alley
lleys
surro
urrou
rroun
round
oundi
undin
nding
Hengi
engil
ngill
volca
olcan
lcano
south
outhw
uthwe
thwes
hwest
weste
ester
stern
Icela
celan
eland
dotte
otted
sprin
pring
rings
steam
vents
Hiker
ikers
world
witne
itnes
tness
breat
reath
takin
aking
scene
cener
enery
sheep
photo
hotog
otoge
togen
ogeni
genic
Nordi
ordic
light
```

# 你上傳的原始碼裡應該包含什麼

你的.py 原始碼檔案裡面應該包含讀取測試資料、做運算，以及輸出答案的Python程式碼。當然，你應該寫適當的註解。針對這個題目，你**可以**使用上課沒有教過的方法。上傳至PDOGS的程式碼中**不能import任何套件**。

# 評分原則

- 這一題的24分會根據程式運算的正確性給分。PDOGS 會直譯並執行你的程式、輸入測試資料，並檢查輸出的答案的正確性。本題共有12 組測試資料，一筆測試資料佔2分。

# 第二題

(承接上題)抄襲偵測的原理是計算兩份文本的相似度。雖然文本相似不一定就是抄襲。但實務上由高相似文本作為調查抄襲的起點常是有效的作法。本題將練習實做簡單版本的抄襲偵測功能。

給定兩個字串，target與source，以及連續詞長度n，生成所有target中長度為n的連續詞。Target 中生成的連續詞數量稱為G。將每個連續詞與source比對，並計算有出現在source中的數量，稱之為m。target與source的相似程度為$m/G$。相似程度輸出到小數點後第四位。

舉例而言，令target與source為:

```
target: 最好哭的一幕 姜曉和智英一起聊生平、回到社會後的夢想智英發現自己完全沒有回去的理由，把存活的資格讓給姜曉。
source: 姜曉跟智英（李瑜美飾）決議兩人用一盤定生死，其餘時間用來聊天、交換祕密，反正這些傾訴最終只會被帶進墳墓。過程中，兩人產生深厚情誼，當智英知道姜曉為了家人，有必須獲勝的理由，而她自己沒有，最後選擇犧牲自己，讓姜曉獲勝。
```

如果令n=3，則source與target只有一個連續詞重疊: 的理由

而target的連續詞數量為41，因此相似度為 $1/41 \approx 0.0244$

再看另一個例子:

```
target: 販售蔥餅、豬肉蛋餅
source: 豬肉蛋餅
```

同樣令n=3，則source與target有兩個連續詞重疊:

```
豬肉蛋
肉蛋餅
```

由於target的連續詞數量為4，因此相似程度為2/4=0.5。

# 輸入輸出格式

系統將會提供測試的n, target, 與source。輸入第一行為n，數值為小於500的正整數。第二行為target，第三行為source。 你的程式在計算完成後應依序輸出

```
LEN=target的連續詞數量
MATCH_COUNT=連續詞重疊數量
SIMILARITY=相似度
MATCHED SEGMENTS:
重疊連續詞
重疊連續詞
...
```

(如果連續詞有重複則都需要印出)

舉例而言，如果輸入是：

---

```
3
販售蔥餅、豬肉蛋餅
豬肉蛋餅
```

---

那輸出是:

---

```
LEN=4
MATCH_COUNT=2
SIMILARITY=0.5000
MATCHED SEGMENTS:
豬肉蛋
肉蛋餅
```

如果輸入是:

---

```
4
販售蔥餅、豬肉蛋餅
豬肉蛋餅
```

---

那輸出是:

---

```
LEN=2
MATCH_COUNT=1
SIMILARITY=0.5000
MATCHED SEGMENTS:
豬肉蛋餅
```

---

如果輸入是:

```
3
今天東北季風增強，水氣增加，雨區擴大到整個北部及東半部地區，整天都有不定時短暫降雨的機會，特別是基隆北海岸雨勢會更明顯，並有局部較大雨勢發生的機率。
三立氣象專家吳德榮指出，今（21）日鋒面逐漸南下，北台灣及東半部轉為有雨，氣溫漸降，中南部日夜溫差大。明日起至週日鋒面影響及東北季風增強，平地最低溫下探18度，北台灣濕涼，降雨範圍更廣，苗栗以北、東部有較大雨勢。
```

---

那輸出是:

---

```
LEN=56
MATCH_COUNT=8
SIMILARITY=0.1429
MATCHED SEGMENTS:
東北季
北季風
季風增
風增強
及東半
東半部
較大雨
大雨勢
```

# 你上傳的原始碼裡應該包含什麼

你的.py 原始碼檔案裡面應該包含讀取測試資料、做運算，以及輸出答案的Python程式碼。當然，你應該寫適當的註解。針對這個題目，你**可以**使用上課沒有教過的方法。上傳至PDOGS的程式碼中**不能import其他套件**。

## 評分原則

- 這一題的26分會根據程式運算的正確性給分。PDOGS 會直譯並執行你的程式、輸入測試資料，並檢查輸出的答案的正確性。本題共有13 組測試資料，一筆測試資料佔2分。

# 第三題

在瞭解專業的文本時，我們可以透過建構縮寫與全名的對應掌握重要的名詞。本題將針對英文文本中的首字母縮寫(Acronym)建構縮寫與全文的對應。Acronym一般會以全大寫的形式出現，且前面通常會有這個縮寫的全名。舉例來說:

```
Nonparametric regressions provide a flexible approach for constructing nonlinear
maps from input features to real-valued outputs. Well-known examples include
kernel-based approaches, such as Gaussian process regression (GPR) [1] and
support vector regression (SVR) [2], ensemble-based approaches, such as random
forest [3], gradient boosting machines [4], and neural-network-based methods
[5].
```

在這段話中有兩個Acronym，分別是GRP與SVR。而GPR前面的三個單字，Gaussian process regression，為GPR的全名。SVR前面的三個單字，support vector regression，為SVR的全名。因此，我們可以透過Acronym出現的位置與其前面的單字，建構出Acronym與其全名的對照。

寫一個程式，由一段文字中建構出Acronym與其全名的對照。你可以假設Acronym會出現在刮號內，且為全大寫。如果一個Acronym有N個字母，則取其前面的N個單字作為全名的候選詞。如果這個候選詞的第一個字母組合起來(轉成大寫後)剛好跟這個Acronym一樣，則留下這個對應關係。否則就丟棄這個對照關係。留下的全名應保留原始的大小寫，且單字之間應只有一個空白。為了簡化問題。你可以假設刮號內不會再有刮號。也就是不會有像是 `(Hello (Hihi) World)` 這樣的情況。

## 輸入輸出格式

系統將會提供測試資料。輸入第一行為要分析的文本。在取的資料之後，進行縮寫與全文的對應的建構，並印出結果。印出結果的順序應與原始輸入文本出現的次序一致。並且只印出符合前述規則的對應。列印的格式為: `Acronym: Full Name` 。每行輸出一組對應。

舉例而言，如果輸入是：

---

```
Nonparametric regressions provide a flexible approach for constructing nonlinear maps from inpu
t features to real-valued outputs. Well-known examples include kernel-based approaches, such as
Gaussian process regression (GPR) [1] and support vector regression (SVR) [2], ensemble-based a
pproaches, such as random forest [3], gradient boosting machines [4], and neural-network-based
methods [5].
```

---

那輸出是:

---

```
GPR: Gaussian process regression
SVR: support vector regression
```

---

如果輸入是:

---

```
As interest has increased in the possible relationships between artificial intelligence (AI) an
d fashion, more and more approaches are being proposed for fashion recognition and understandin
g. Meanwhile, fashion retailers are using AI technologies in inventory management, clothing rec
ommendation, and virtual clothes fitting to improve their decision-making and competitive advan
tages [1]–[2][3].
```

---

那輸出是:

---

```
AI: artificial intelligence
```

---

如果輸入是:

---

```
Well-known examples include kernel-based approaches, such as Gaussian process regression model
(GPR) [1] and support vector regression (SVR) [2], ensemble-based approaches, such as random fo
rest [3], gradient boosting machines [4], and neural-network-based methods [5].
```

---

那輸出是:

---

```
SVR: support vector regression
```

---

# 你上傳的原始碼裡應該包含什麼

- 你的.py 原始碼檔案裡面應該包含讀取測試資料、做運算，以及輸出答案的Python程式碼。當然，你應該寫適當的註解。針對這個題目，你**可以**使用上課沒有教過的方法。上傳至PDOGS的程式碼中**不能import其他套件**。

# 評分原則

- 這一題的26分會根據程式運算的正確性給分。PDOGS 會直譯並執行你的程式、輸入測試資料，並檢查輸出的答案的正確性。本題共有13組測試資料，一筆測試資料佔2分。

# 第四題

小李的老闆希望他能將公司的資料檔用格式化的方式輸出。每個資料欄位寬度一樣，向右對齊。兩欄之間有一個空白字元。數值資料顯示到小數點第二位(四捨五入)。如果資料長度大於欄寬以致於無法完整顯示，則應在該欄最後一個字元顯示 ~ ，提示使用者顯示的資訊不完整。所有資料不應該超出設定的欄寬。

舉例而言，庫存資料的欄位依序為品名、數量、單價、狀態。品名與狀態是字串，而數量與單價為數值資料。如果欄位寬度為12，則資料列印的方式應為:

```
非常高級水果手機          12.00       2523.50          狀況良好
  真好吃大溪豆乾          55.00         45.00          已過期
```

須注意由於不同系統的顯示字型(有些字型空白的寬度跟其他字元不同)，這兩筆資料會看起來沒有完全對齊。我們在這裡只關心實際上字元的數量是否正確，暫不考慮字型所造成的問題。

如果我們將欄寬設為6，那第一欄會無法完整顯示品名，且第三欄也會有類似狀況。列印的結果會是:

```
非常高級水~  12.00 2523.~    狀況良好
真好吃大溪~  55.00  45.00     已過期
```

# 輸入輸出格式

系統將會提供測試資料。輸入第一行為欄寬，會是500以內的正整數。第二行是每個欄位的格式。 s 代表字串， f 代表數值資料。依照資料欄位次序排列，並以逗點分隔。之後每一行為一筆資料，欄位間以逗點分隔。輸入的最後一行為 RECORD_END ，代表所有輸入資料結束。

你的程式在讀取輸入之後，應該依照指定的欄寬與格式印出資料。

舉例而言，如果輸入是：

---

```
12
s,f,f,s
非常高級水果手機,12,2523.5,狀況良好
真好吃大溪豆乾,55,45,已過期
RECORD_END
```

---

那輸出是:

---

```
      非常高級水果手機          12.00       2523.50          狀況良好
       真好吃大溪豆乾          55.00         45.00          已過期
```

---

如果輸入是:

---

```
6
s,f,f,s
非常高級水果手機,12,2523.5,狀況良好
真好吃大溪豆乾,55,45,已過期
RECORD_END
```

---

那輸出是:

---

```
非常高級水~  12.00 2523.~   狀況良好
真好吃大溪~  55.00  45.00    已過期
```

---

如果輸入是:

---

```
6
s,f,f,s,s
非常高級水果手機,12,2523.5,狀況良好,還有庫存
真好吃大溪豆乾,55,45,已過期,已售完
RECORD_END
```

---

那輸出是:

---

```
非常高級水~  12.00 2523.~   狀況良好    還有庫存
真好吃大溪~  55.00  45.00    已過期     已售完
```

---

# 你上傳的原始碼裡應該包含什麼

- 你的.py 原始碼檔案裡面應該包含讀取測試資料、做運算，以及輸出答案的Python程式碼。當然，你應該寫適當的註解。針對這個題目，你**可以**使用上課沒有教過的方法。上傳至PDOGS的程式碼中**不能import其他套件**。

# 評分原則

- 這一題的24分會根據程式運算的正確性給分。PDOGS 會直譯並執行你的程式、輸入測試資料，並檢查輸出的答案的正確性。本題共有12組測試資料，一筆測試資料佔2分。