

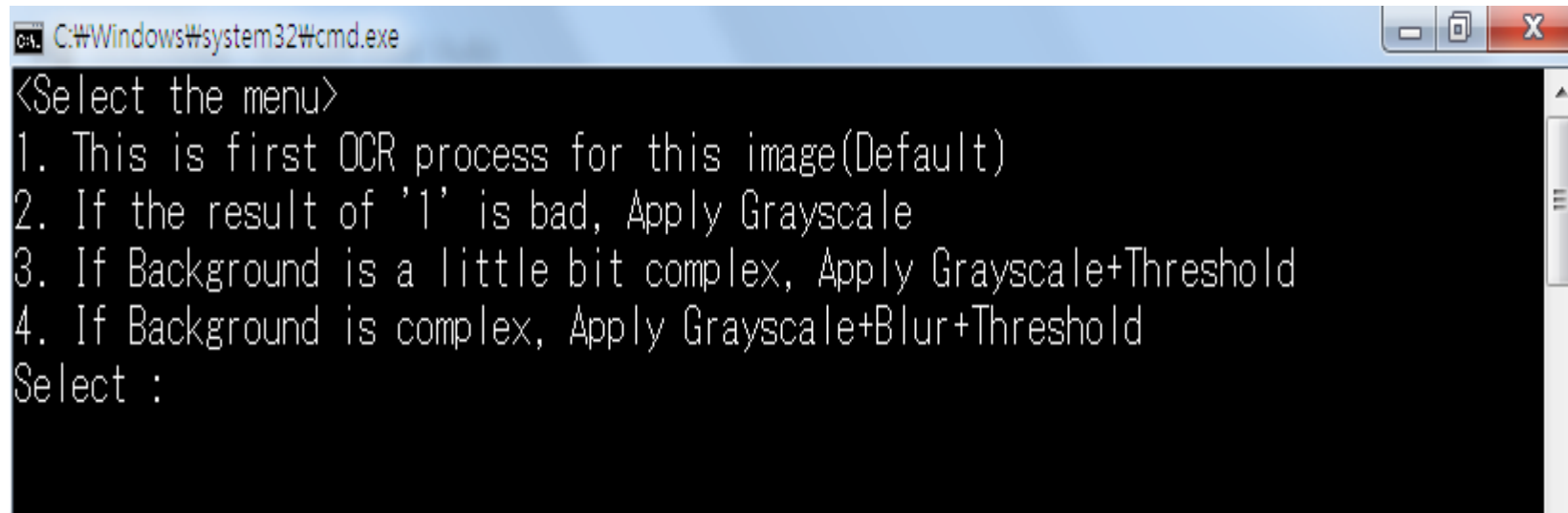
# Chinese Language OCR

Determine and Understand the Text in Image  
Using Tesseract and OpenCV

**PAK MINSEOK**  
**Hanyang University, Seoul**

# Overview

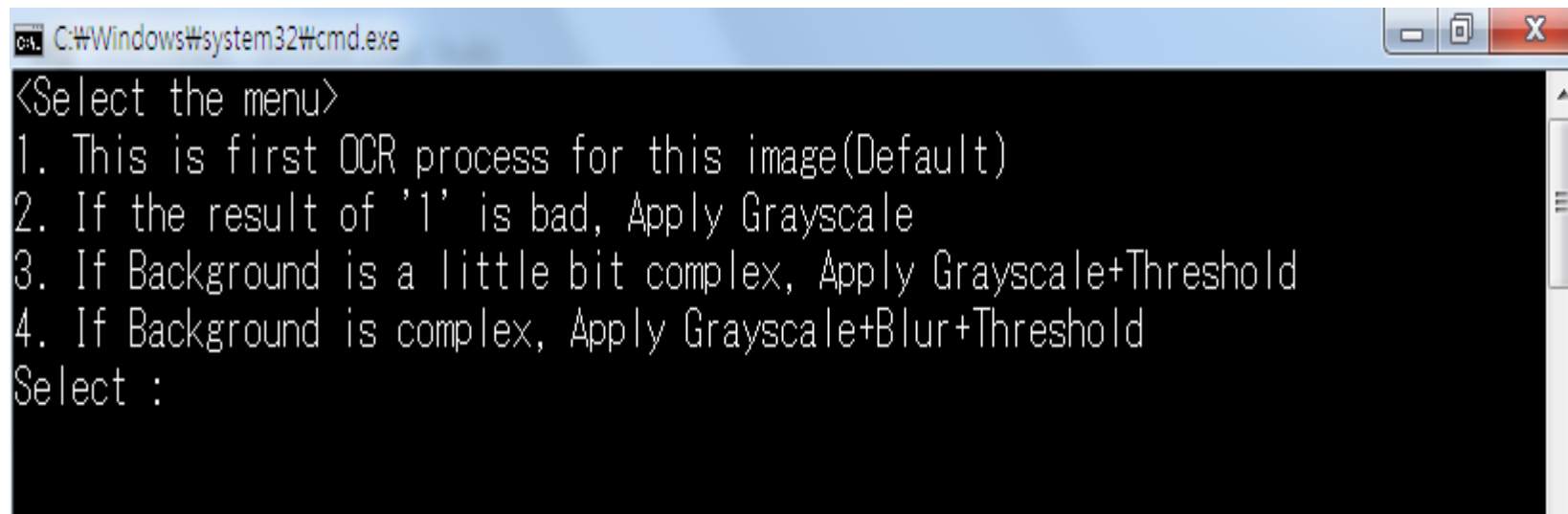
# Overview

A screenshot of a Windows command prompt window. The title bar shows the path 'C:\Windows\system32\cmd.exe'. The command prompt displays a menu with four numbered options for OCR processing, followed by a prompt 'Select :'.

```
C:\Windows\system32\cmd.exe
<Select the menu>
1. This is first OCR process for this image(Default)
2. If the result of '1' is bad, Apply Grayscale
3. If Background is a little bit complex, Apply Grayscale+Threshold
4. If Background is complex, Apply Grayscale+Blur+Threshold
Select :
```

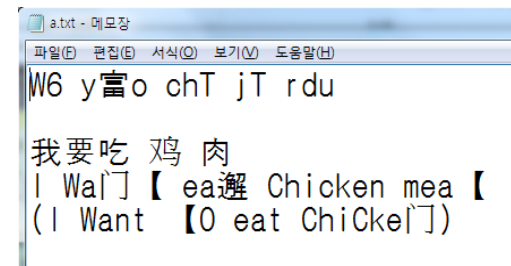
- When you execute this program, you can choose one of 4 options for OCR\_Chinese.
- First option is just using basic function of tesseract library. So, if the background of image is simple, the accuracy of OCR is great. However, if the background of image is complex, the accuracy of OCR is sometimes very poor.
- If background of image is a little bit complex(if distinguishing text and image is difficult) or using first option is bad, then use second and third option.
- Last option is good to apply when using OCR on image which has complex background.

# Overview



```
C:\Windows\system32\cmd.exe
<Select the menu>
1. This is first OCR process for this image(Default)
2. If the result of '1' is bad, Apply Grayscale
3. If Background is a little bit complex, Apply Grayscale+Threshold
4. If Background is complex, Apply Grayscale+Blur+Threshold
Select :
```

- After excuting the program, the text file is created.
- In this text file, there is a result of OCR(text in image) like this.

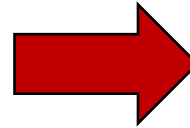


# Overview

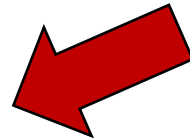
Image Preprocessing(Binarization) using OpenCV(Last Option).



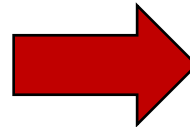
Original Image



Gray Scale



Blur



Threshold

# Demo

Compare Speed and Accuracy of each option to First Option.

(If First Option is not working, then compare each option to Second Option。 )

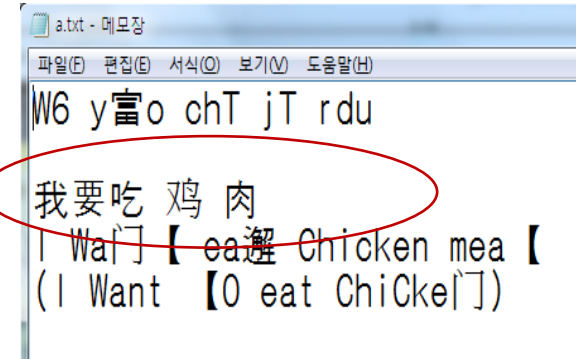
# DEMO for image1



Image\_1

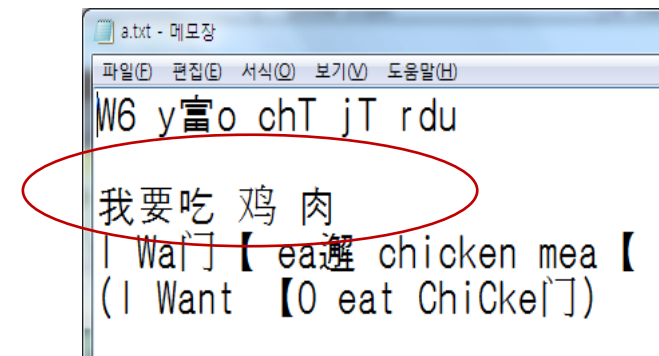


OCR



Using First Option

Elapsed time : 2400 milliseconds



Using Second Option

Elapsed time : 1940~1960milliseconds

Speed : up  
Accuracy : same

# DEMO for image1



Image\_1



OCR

a.txt - 메모장  
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)  
W6 yǎo chī jī ròu  
我要吃 鸡肉  
I Wa | eat chicken mea  
(I Want |0 eat ChiCke |)

Speed : up  
Accuracy : same

Using Third Option  
Elapsed time : about 2000 milliseconds

a.txt - 메모장  
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)  
W5 yǎo Chī jī ròu  
我腰吃 鸡肉  
I Wa | ea | Chicke | mea  
(I Wam !0 eat ChiCke |)

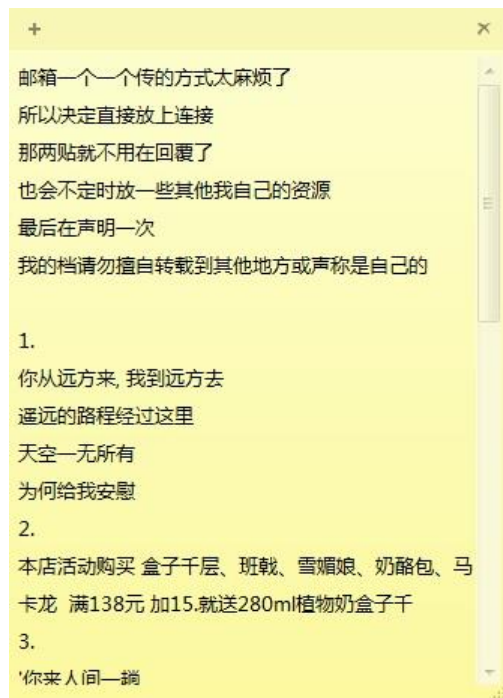
Speed : down  
Accuracy : down

Using Last Option  
Elapsed time : about 2600 milliseconds

Best is Second Option



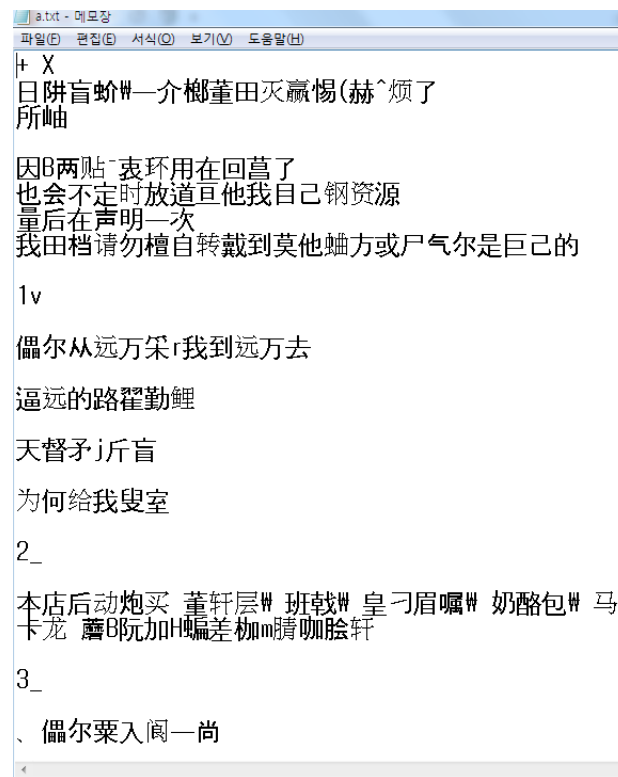
# DEMO for image2



Image\_2



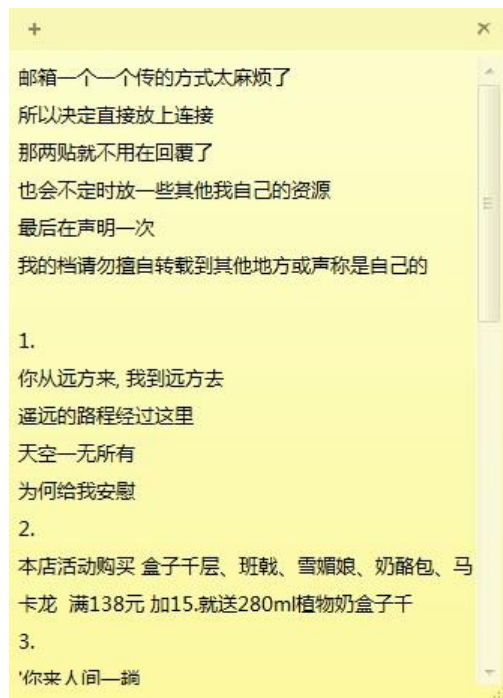
OCR



Using First Option

Elapsed time : 9600 milliseconds

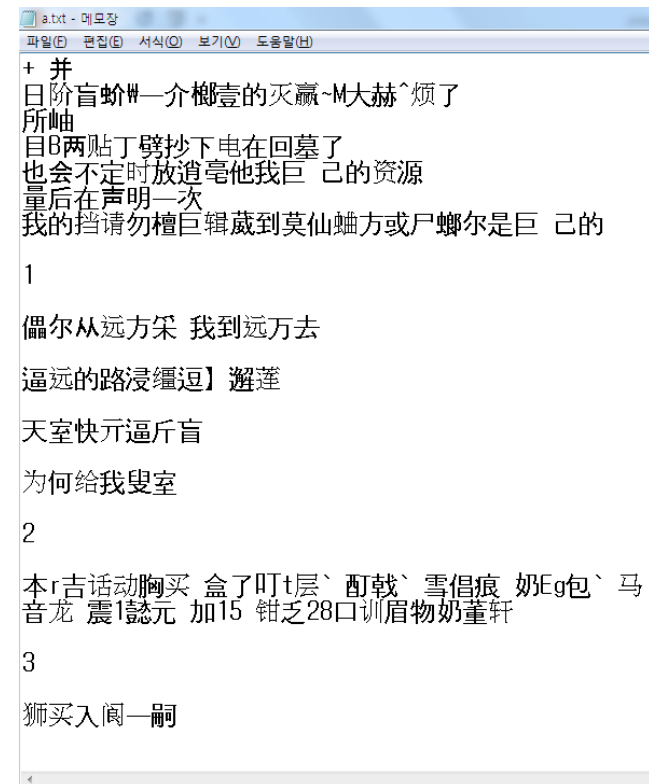
# DEMO for image2



Image\_2



OCR

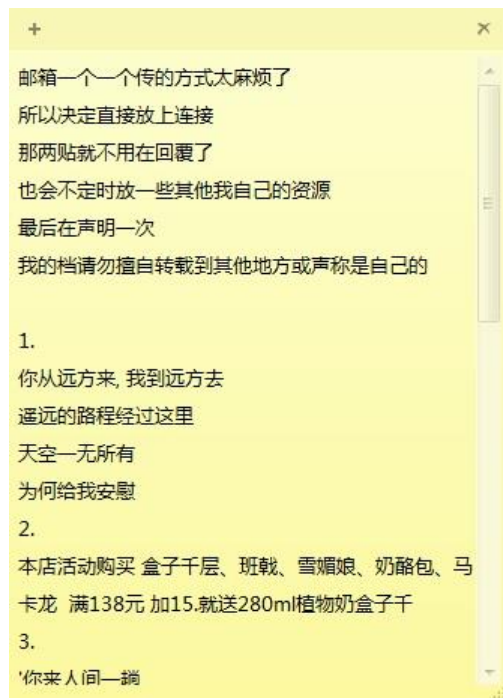


Speed : up  
Accuracy : up

Using Second Option  
Elapsed time : 9380 milliseconds

Best is Second Option

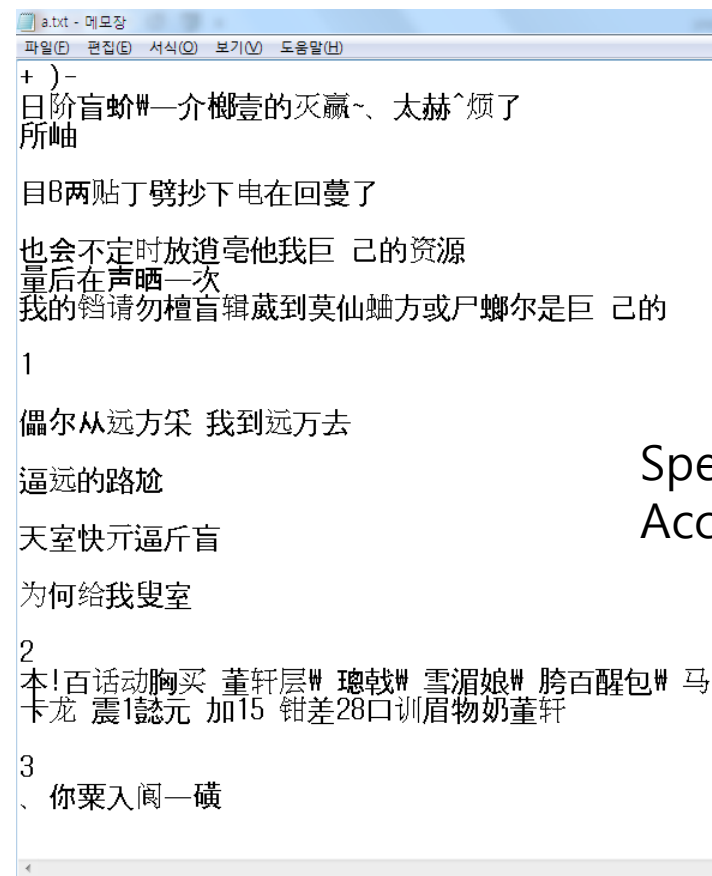
# DEMO for image2



Image\_2



OCR

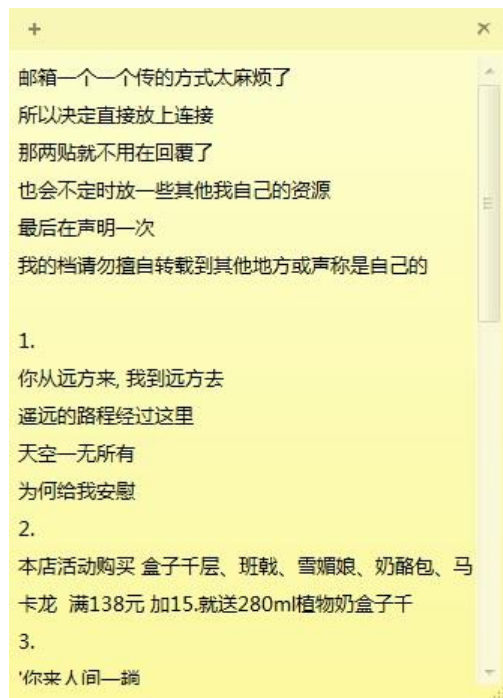


Using Third Option

Elapsed time : 8800 milliseconds

Speed : up  
Accuracy : down

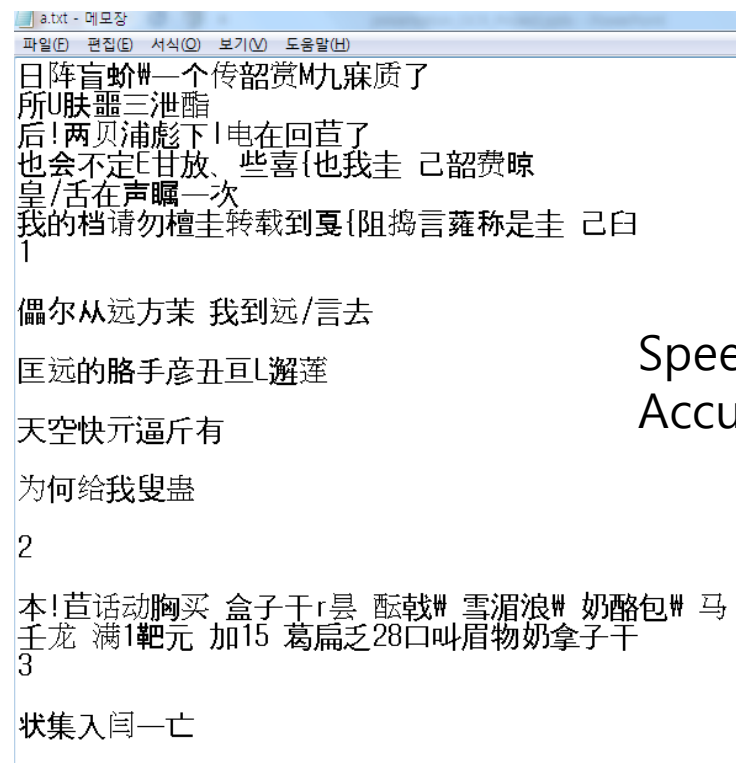
# DEMO for image2



Image\_2



OCR



Using Last Option

Elapsed time : 9260 milliseconds

Speed : up

Accuracy : similar

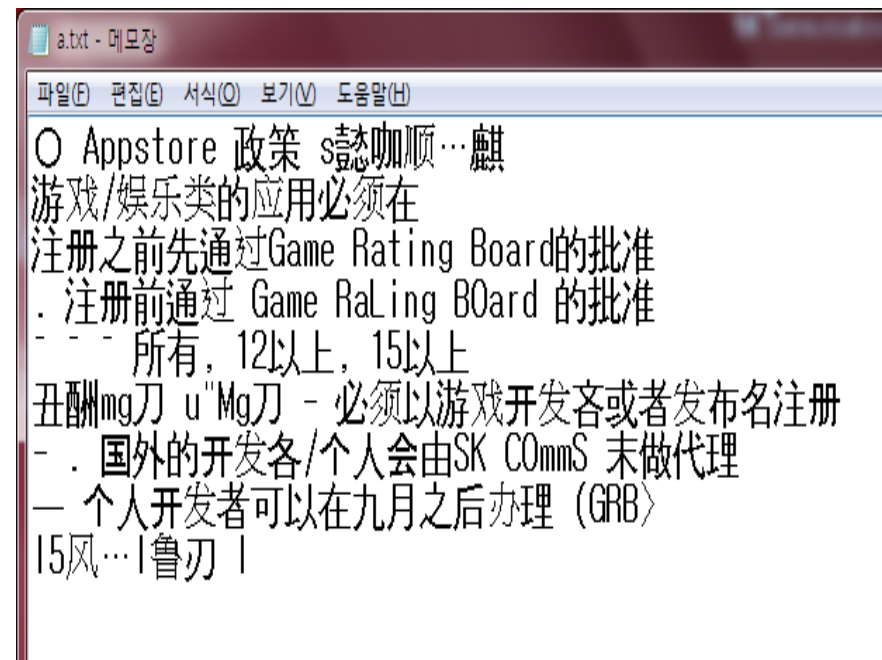
# DEMO for image3



Image\_3



OCR



Using First Option

Elapsed time : 6600 milliseconds

## Best is First Option

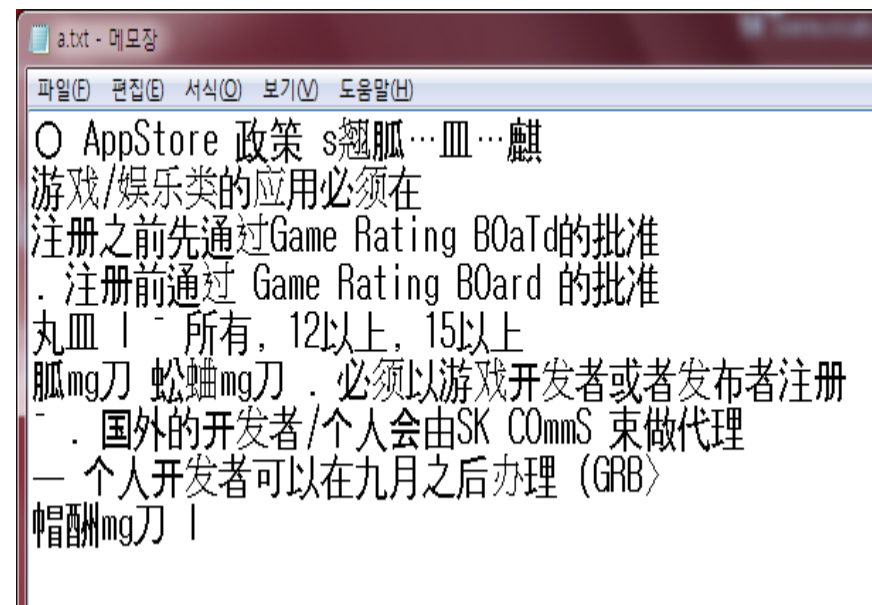
# DEMO for image3



Image\_3



OCR



Using Second Option  
Elapsed time : 6680 milliseconds  
Speed : down  
Accuracy : similar

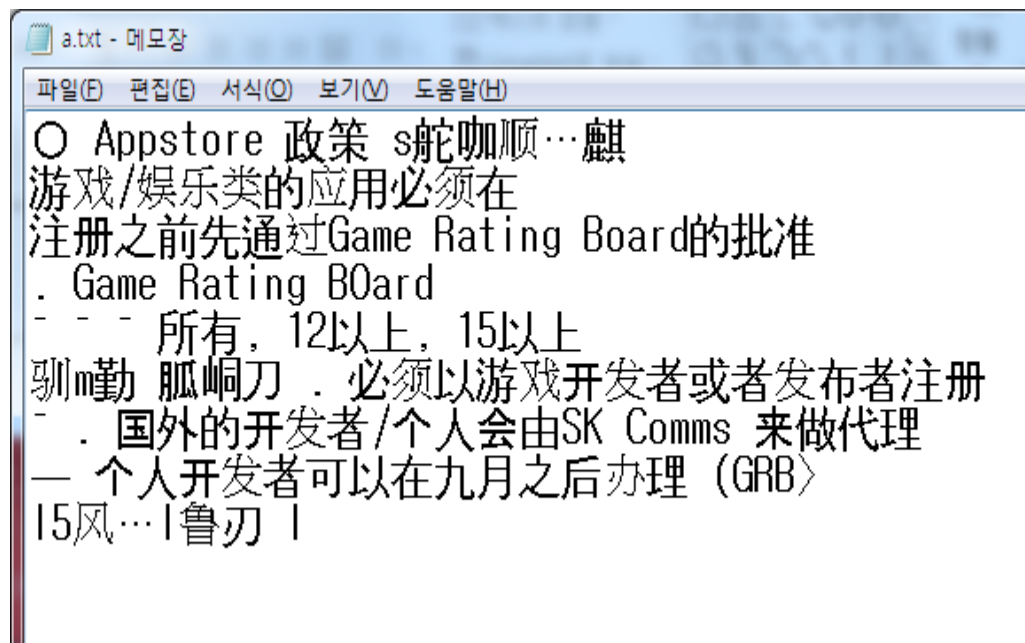
# DEMO for image3



Image\_3



OCR



Using Third Option

Elapsed time : 6440 milliseconds

Speed : up

Accuracy : down



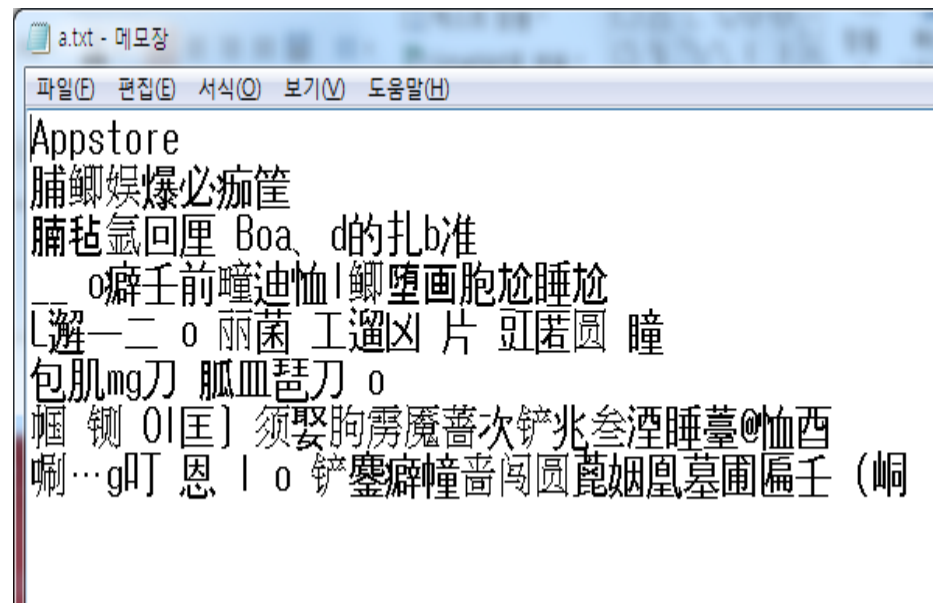
# DEMO for image3



Image\_3



OCR



Using Last Option

Elapsed time : 15650 milliseconds

Speed : down

Accuracy : down



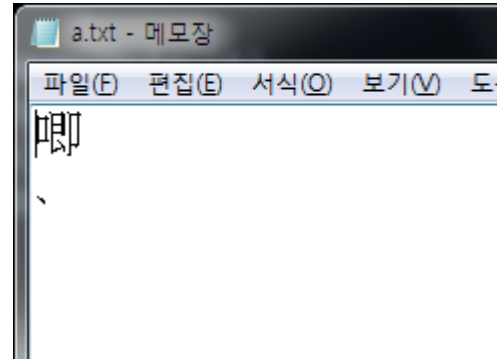
# DEMO for image4



Image\_4



OCR



Using First Option.  
Not working!

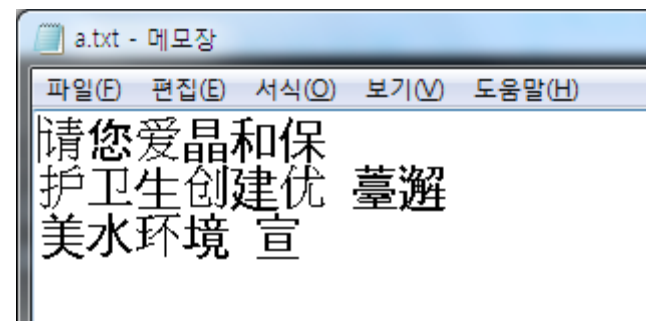
## DEMO for image4



Image\_4



OCR



Using Second Option  
Elapsed time : 1720 milliseconds

Best is Second Option

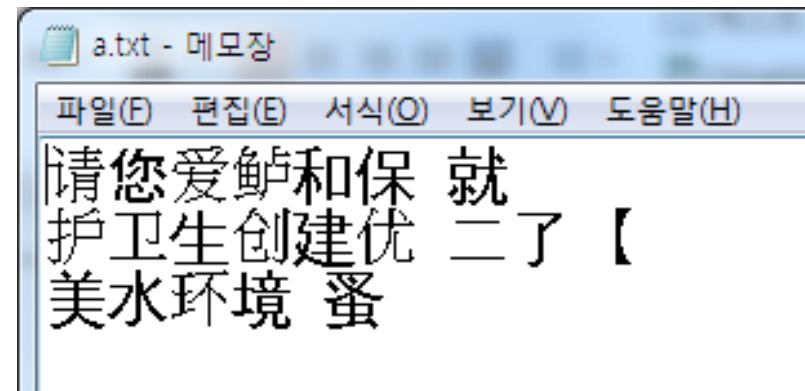
## DEMO for image4



Image\_4



OCR



Using Third Option

Elapsed time : 2100 milliseconds

Speed : down

Accuracy : similar

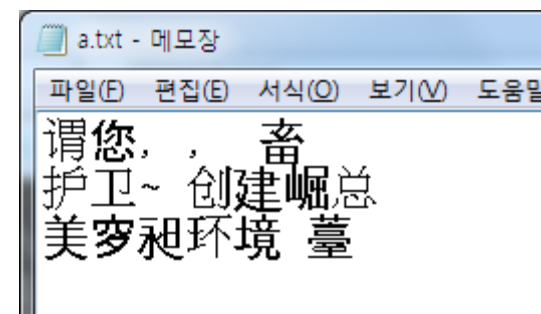
# DEMO for image4



Image\_4



OCR



Using Last Option

Elapsed time : 2350 milliseconds

Speed : down

Accuracy : down

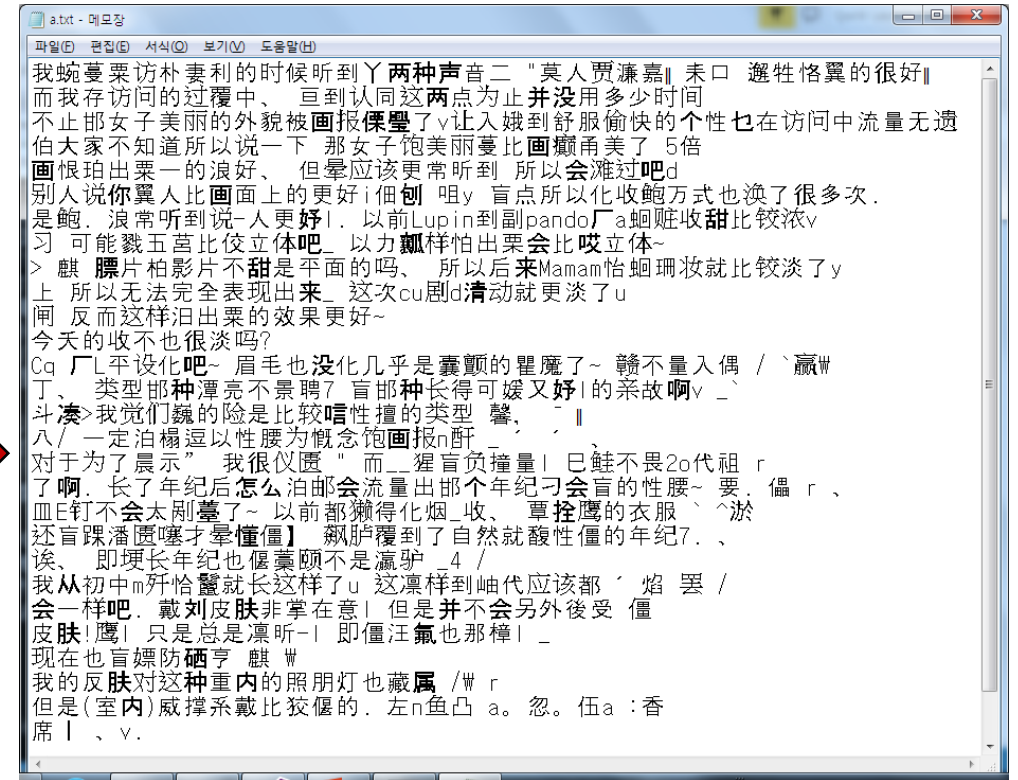
# DEMO for image5



Image\_5



OCR



Using First Option

Elapsed time : 33430 milliseconds

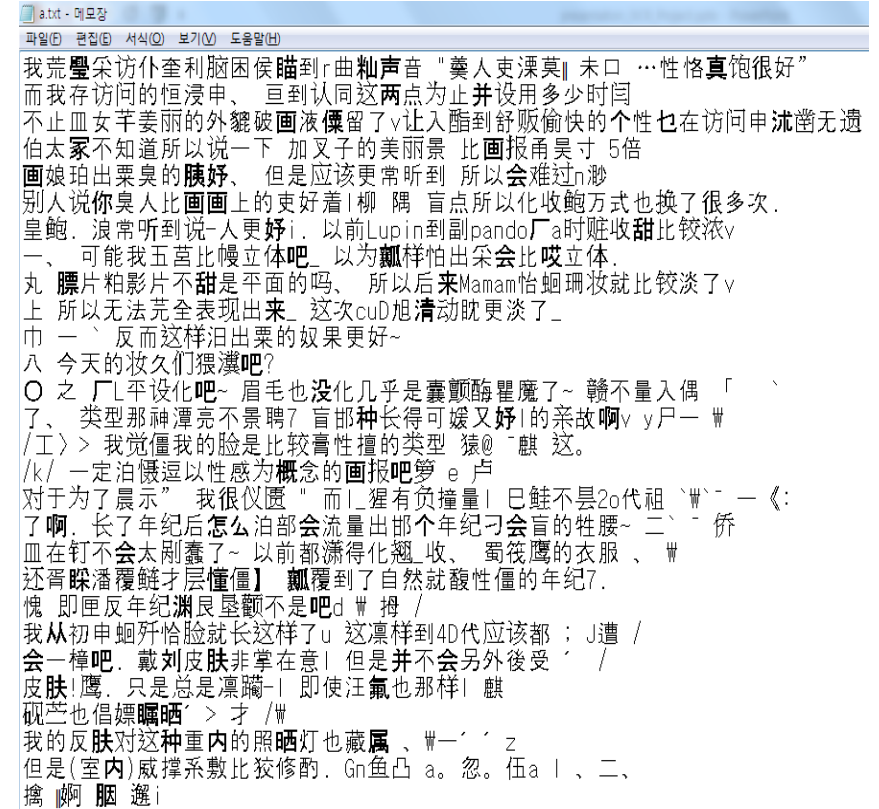


# DEMO for image5



Image\_5

OCR



Using Second Option  
Elapsed time : 32460 milliseconds  
Speed : up  
Accuracy : similar

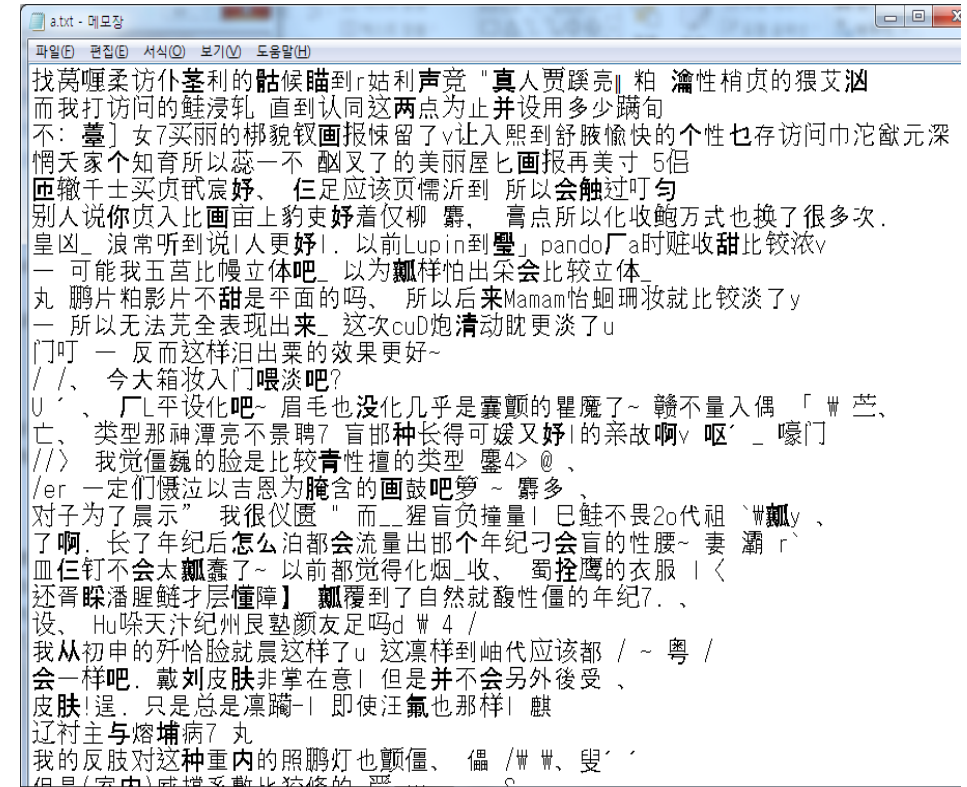
Best is Second Option

# DEMO for image5



Image\_5

OCR



Using Third Option

Elapsed time : 31490 milliseconds

Speed : up

Accuracy : down

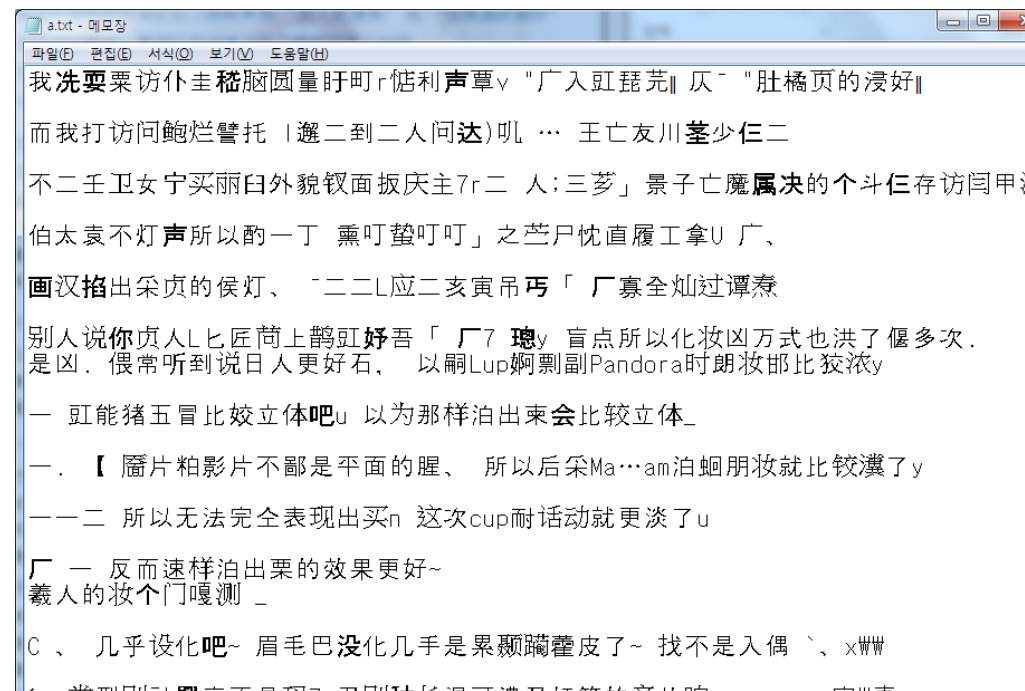
# DEMO for image5



Image\_5



OCR



Using Last Option

Elapsed time : 34060 milliseconds

Speed : down

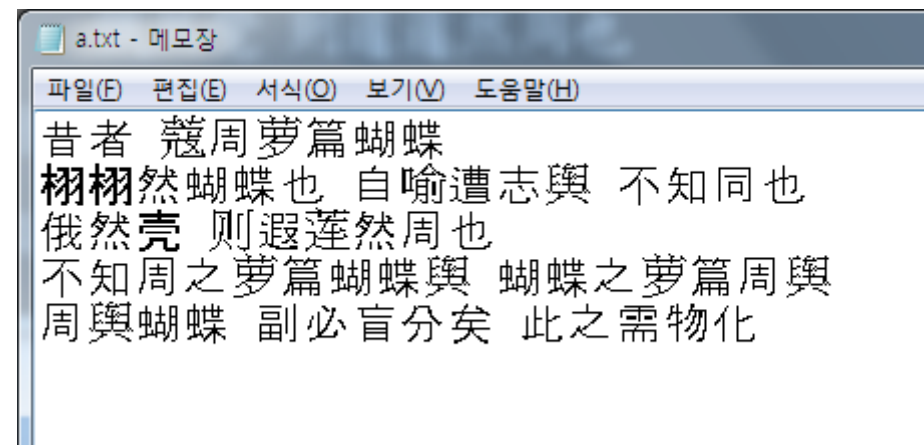
Accuracy : down



## DEMO for image6

昔者 莊周夢爲蝴蝶  
栩栩然蝴蝶也 自喻適志與 不知周也  
俄然覺 則蘧蘧然周也  
不知周之夢爲蝴蝶與 蝴蝶之夢爲周與  
周與蝴蝶 則必有分矣 此之謂物化

OCR



Image\_6

Using First Option

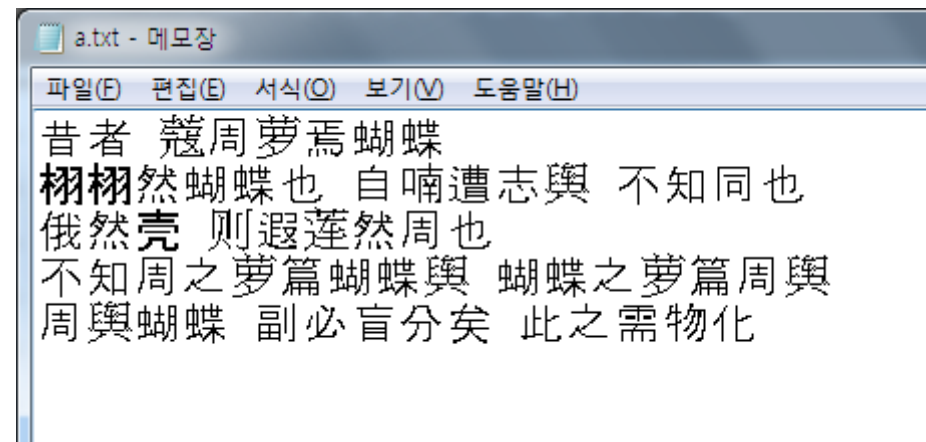
Elapsed time : 9290 milliseconds

# Best is First Option

# DEMO for image6

昔者 莊周夢爲蝴蝶  
栩栩然蝴蝶也 自喻適志與 不知周也  
俄然覺 則蘧蘧然周也  
不知周之夢爲蝴蝶與 蝴蝶之夢爲周與  
周與蝴蝶 則必有分矣 此之謂物化

OCR



Image\_6

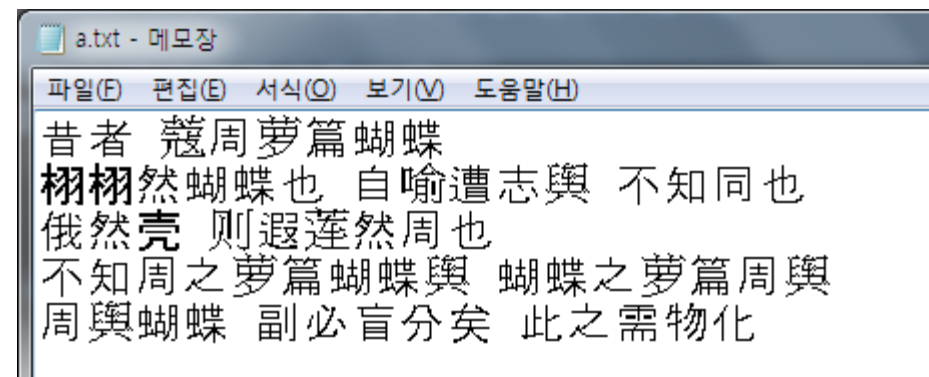
Using Second Option  
Elapsed time : 9820 milliseconds  
Speed : down  
Accuracy : similar

# DEMO for image6

昔者 莊周夢爲蝴蝶  
栩栩然蝴蝶也 自喻適志與 不知周也  
俄然覺 則蘧蘧然周也  
不知周之夢爲蝴蝶與 蝴蝶之夢爲周與  
周與蝴蝶 則必有分矣 此之謂物化



OCR



Image\_6

Using Third Option

Elapsed time : 9330 milliseconds

Speed : down

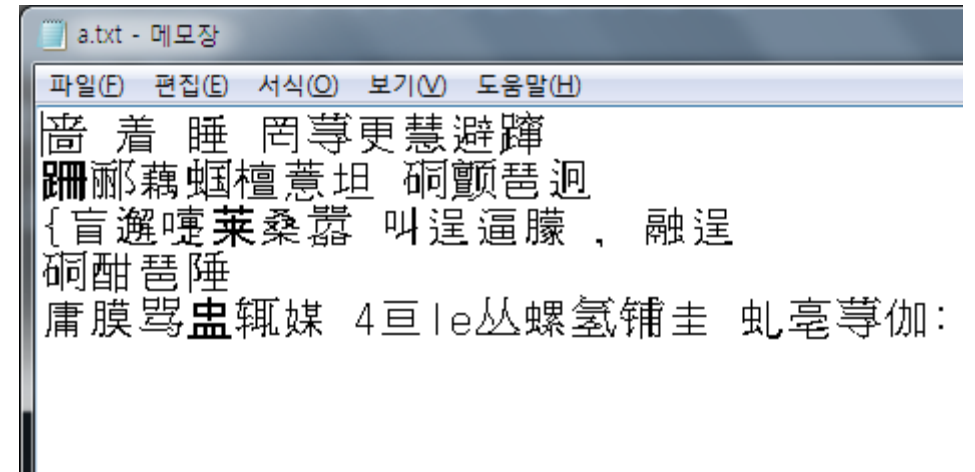
Accuracy : similar

# DEMO for image6

昔者 莊周夢爲蝴蝶  
栩栩然蝴蝶也 自喻適志與 不知周也  
俄然覺 則蘧蘧然周也  
不知周之夢爲蝴蝶與 蝴蝶之夢爲周與  
周與蝴蝶 則必有分矣 此之謂物化



OCR



Image\_6

Using Last Option

Elapsed time : 9930 milliseconds

Speed : down

Accuracy : down

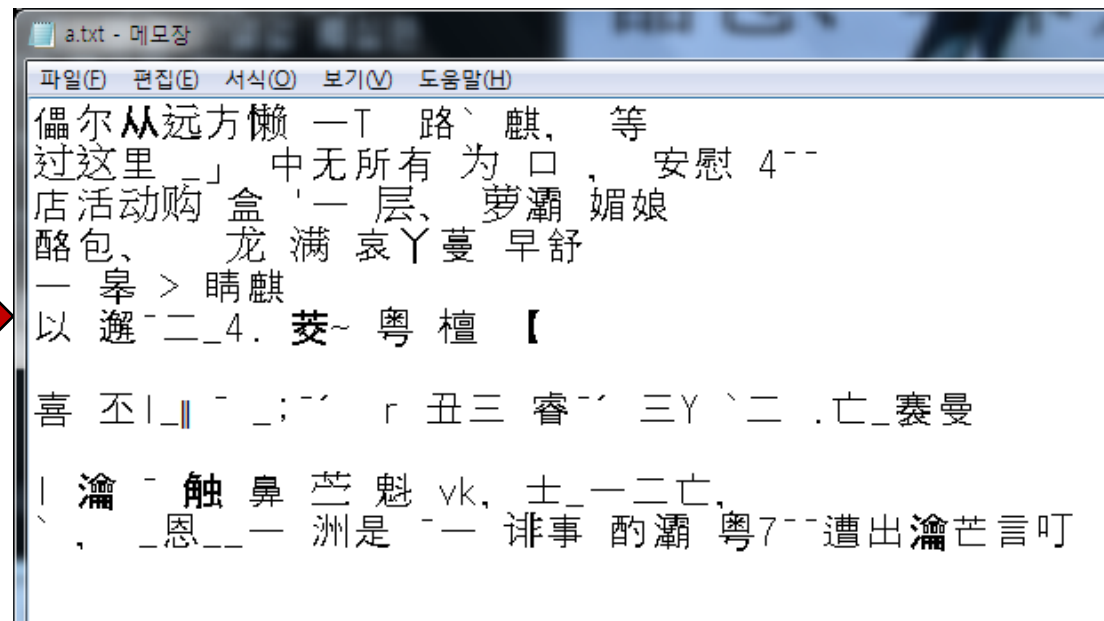
# DEMO for image7



Image\_7



OCR



Using First Option

Elapsed time : 6370 milliseconds

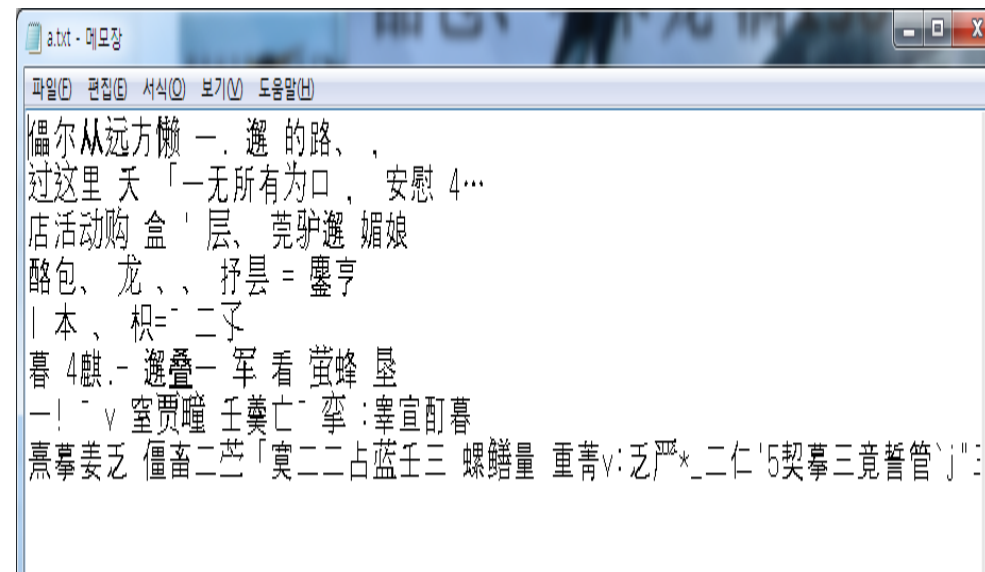
# DEMO for image7



Image\_7



OCR



Using Second Option  
Elapsed time : 11520 milliseconds  
Speed : down  
Accuracy : up



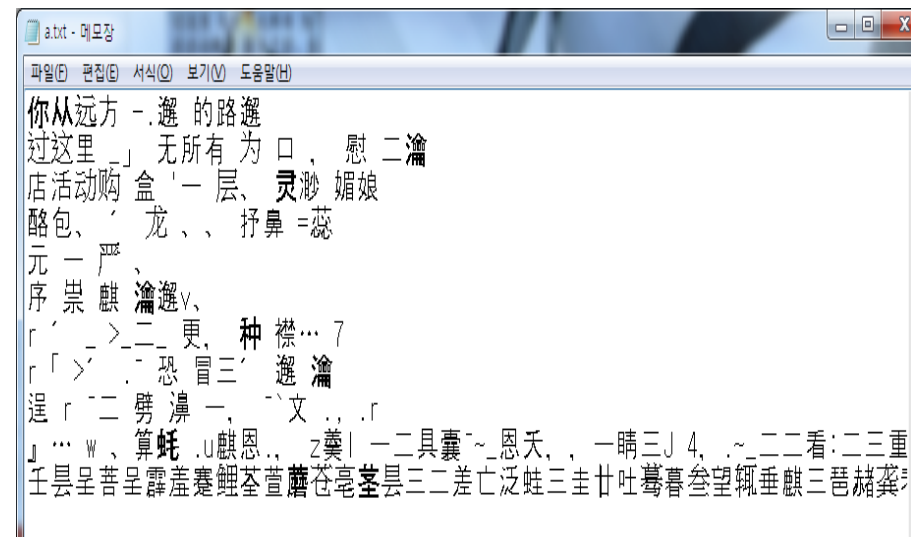
# DEMO for image7



Image\_7



OCR



Using Third Option

Elapsed time : 9220 milliseconds

Speed : down

Accuracy : up

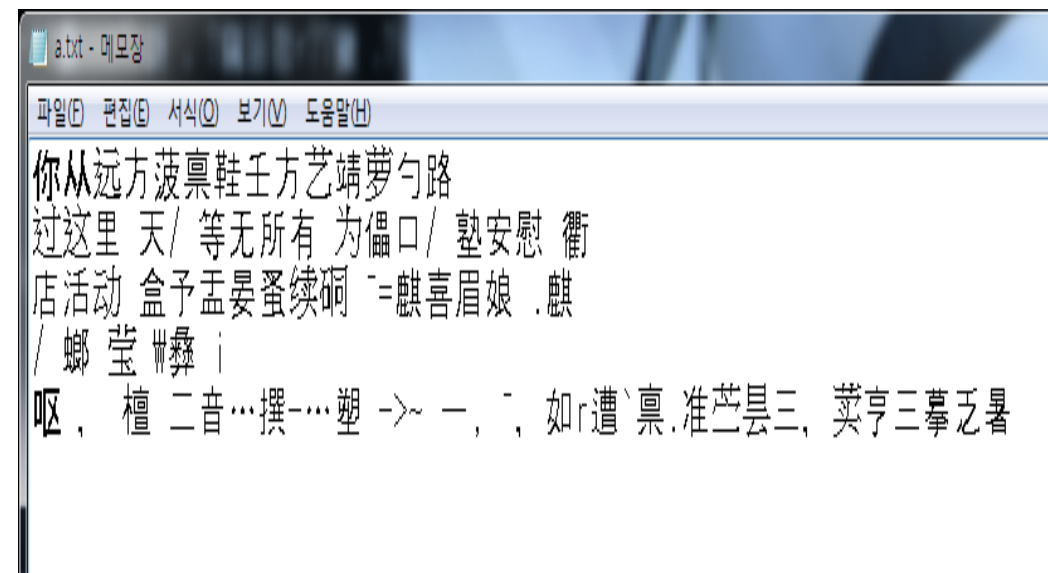
# DEMO for image7



Image\_7



OCR



Using Last Option

Elapsed time : 12200 milliseconds

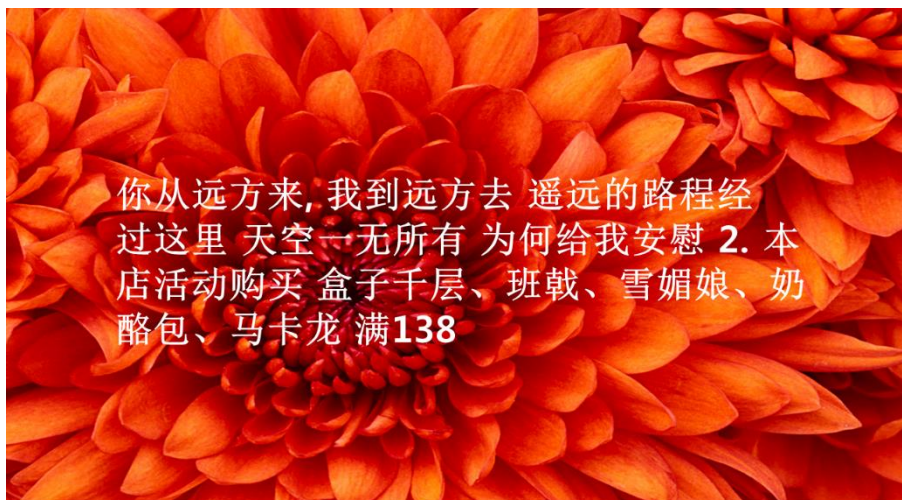
Speed : down

Accuracy : up(best of all)

## Best is Last Option



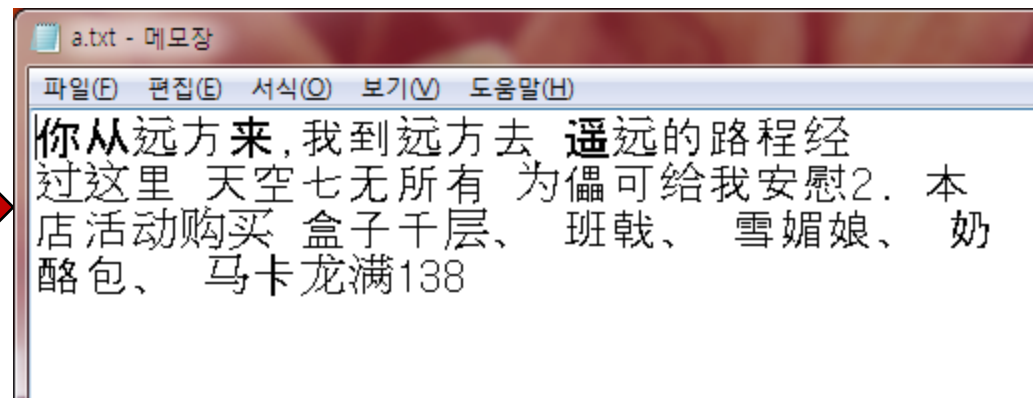
# DEMO for image8



Image\_8



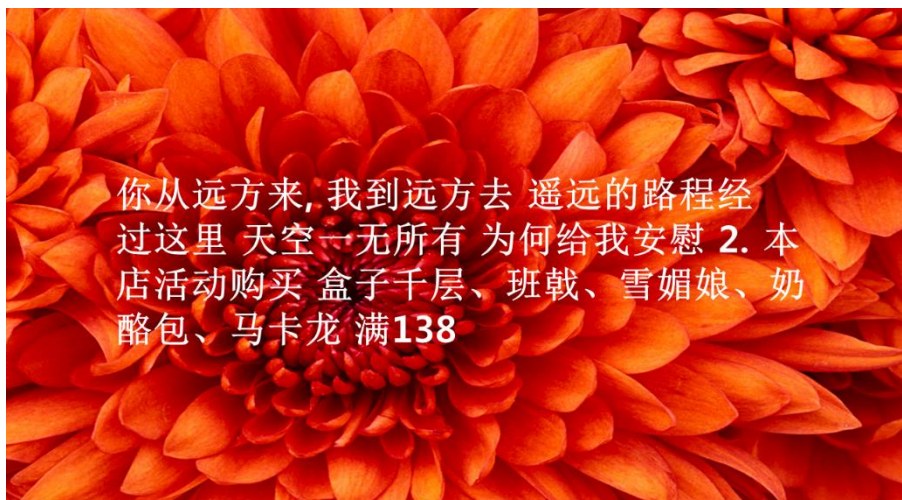
OCR



Using First Option

Elapsed time : 3490 milliseconds

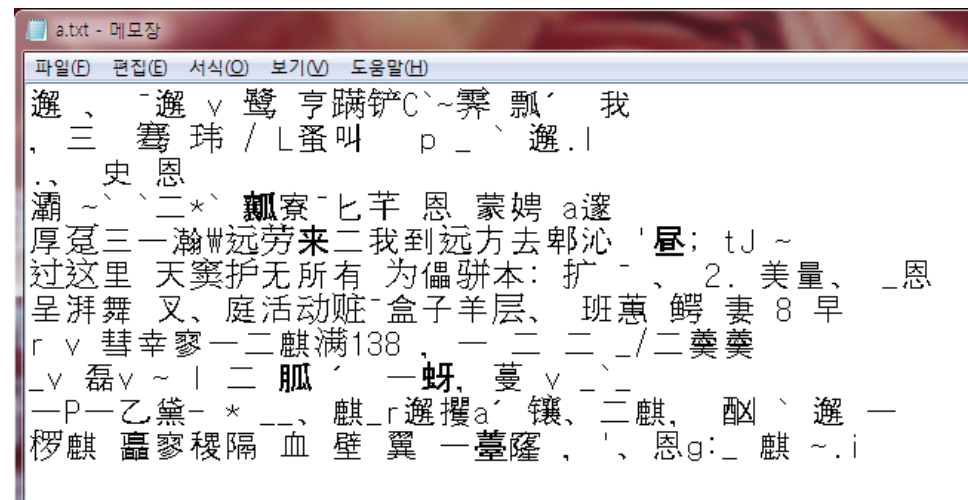
# DEMO for image8



Image\_8

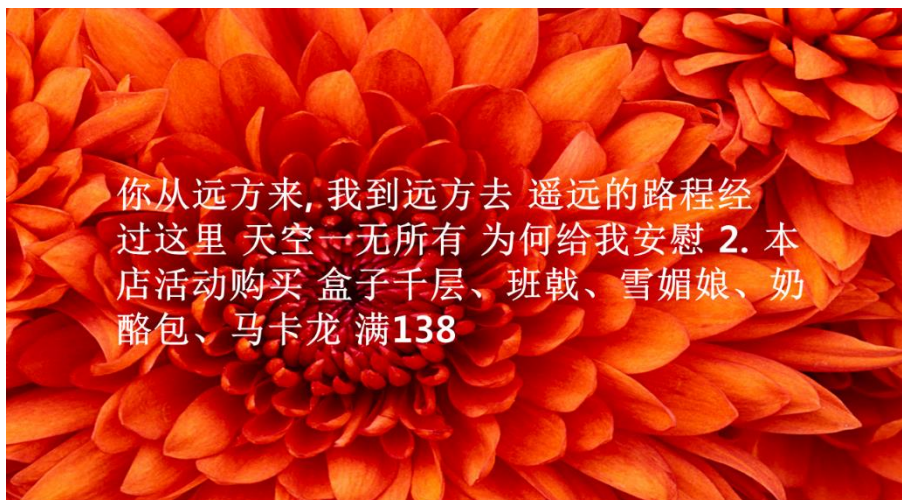


OCR



Using Second Option  
Elapsed time : 15400 milliseconds  
Speed : down  
Accuracy : down

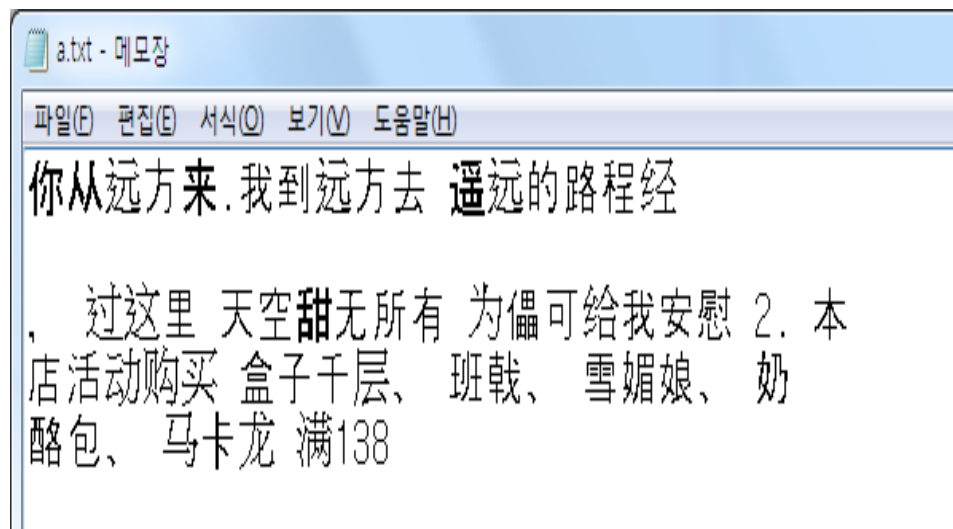
# DEMO for image8



Image\_8



OCR



Using Third Option

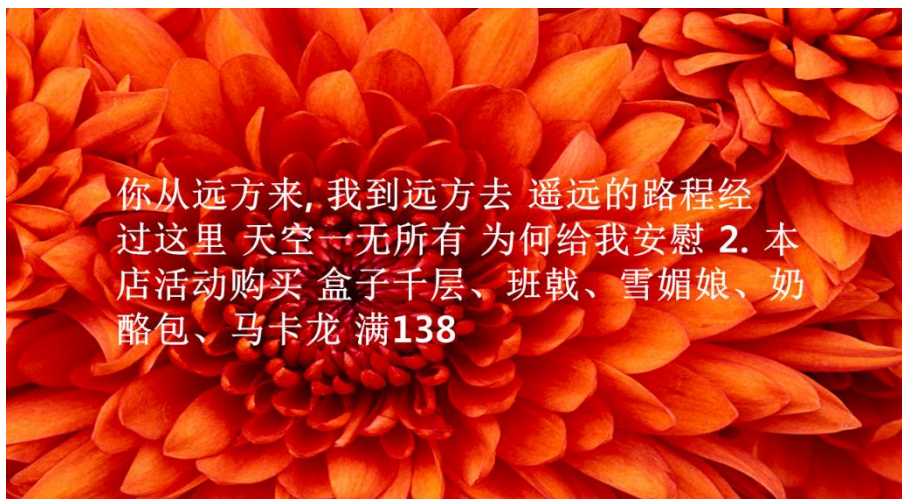
Elapsed time : 3240 milliseconds

Speed : up

Accuracy : similar

## Best is Third Option

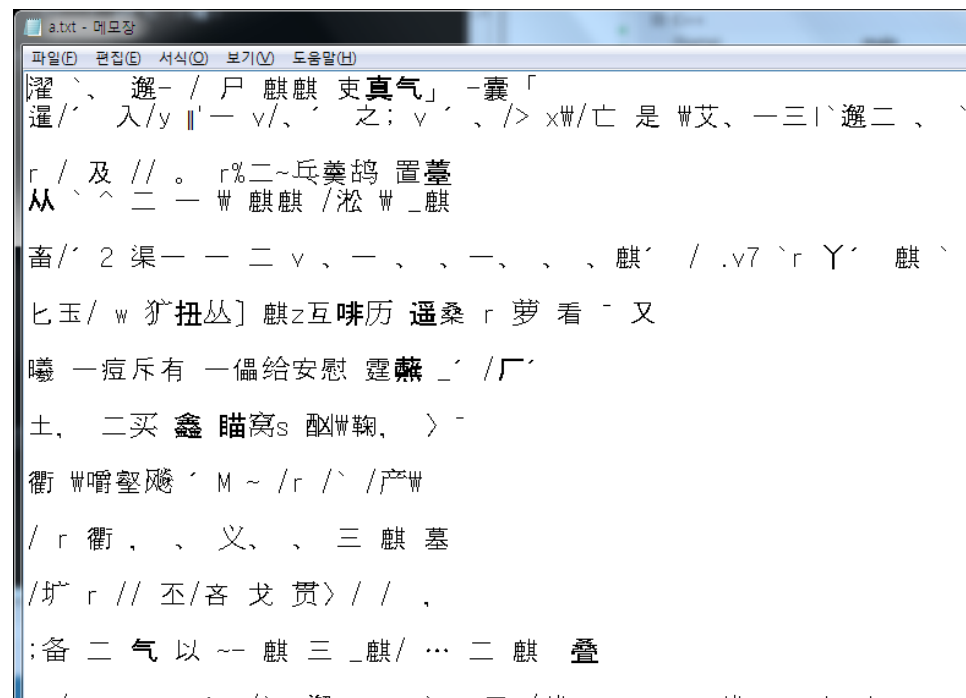
# DEMO for image8



Image\_8



OCR



Using Last Option

Elapsed time : 11340 milliseconds

Speed : down

Accuracy : down



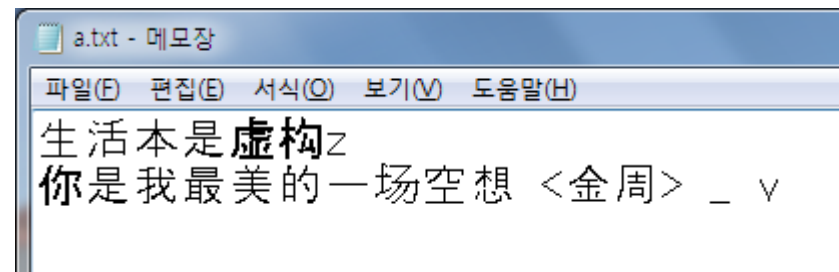
# DEMO for image9



Image\_9



OCR



Using First Option

Elapsed time : 1700 milliseconds

Speed : up

Accuracy : similar

## Best is First Option

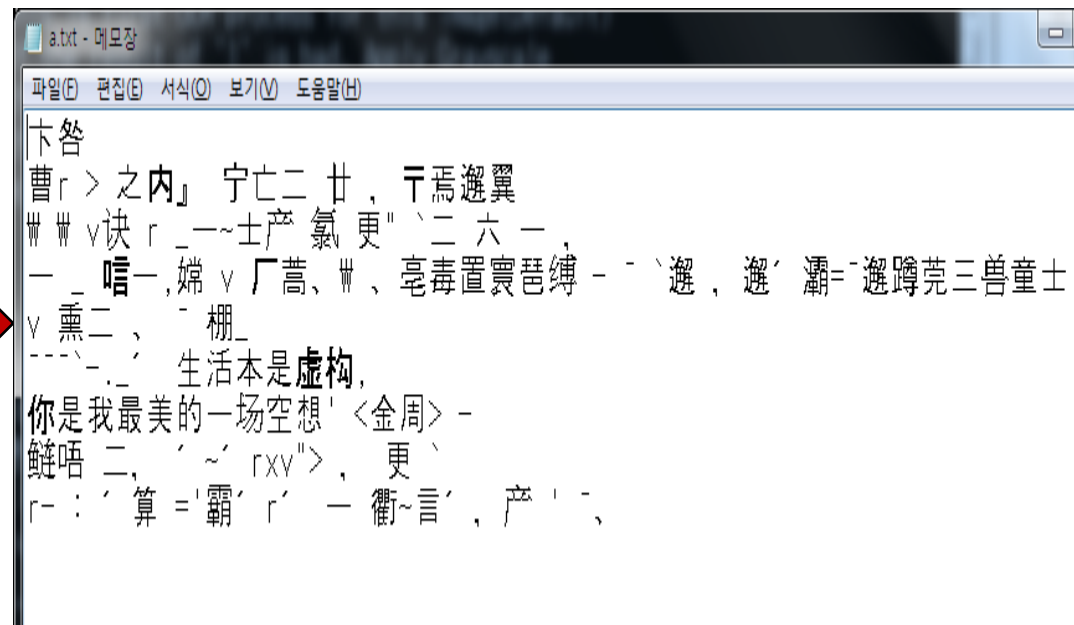
# DEMO for image9



Image\_9



OCR



Using SecondOption

Elapsed time : 5000 milliseconds

Speed : down

Accuracy : down

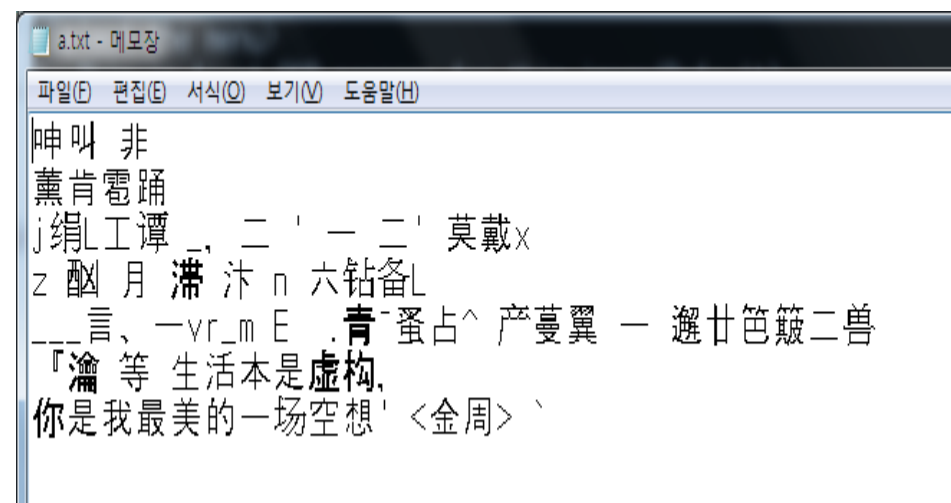
# DEMO for image9



Image\_9



OCR



Using Third Option

Elapsed time : 4600 milliseconds

Speed : down

Accuracy : down

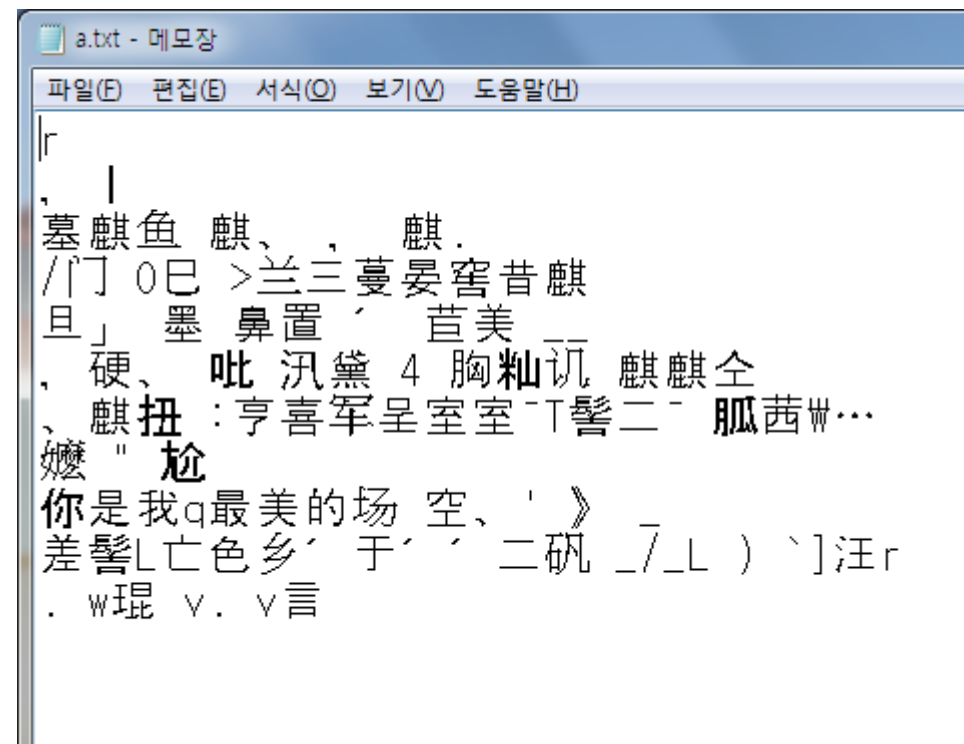
# DEMO for image9



Image\_9



OCR



Using Last Option

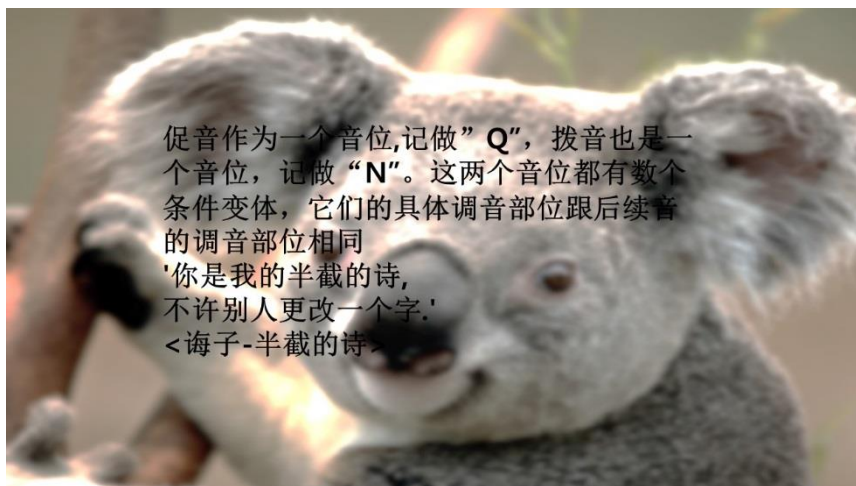
Elapsed time : 9220 milliseconds

Speed : down

Accuracy : down



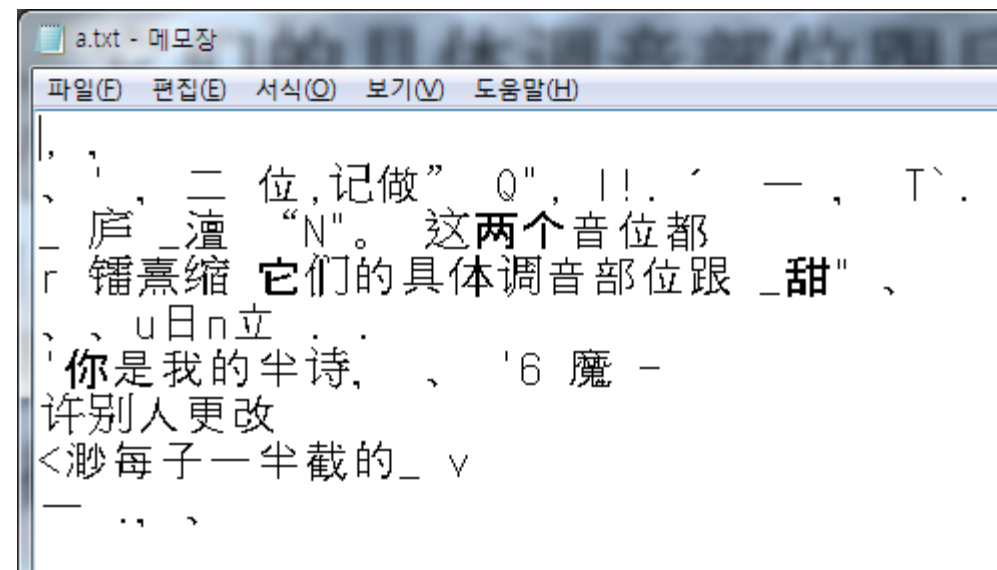
# DEMO for image10



Image\_10



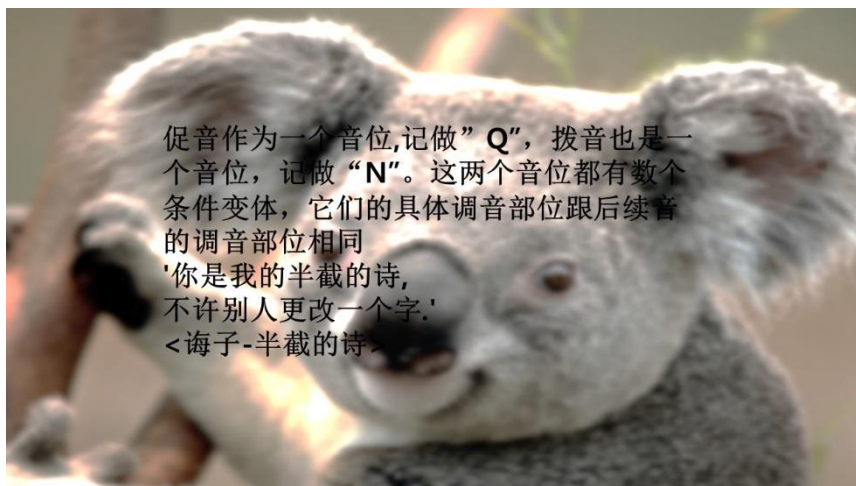
OCR



Using First Option

Elapsed time : 3610 milliseconds

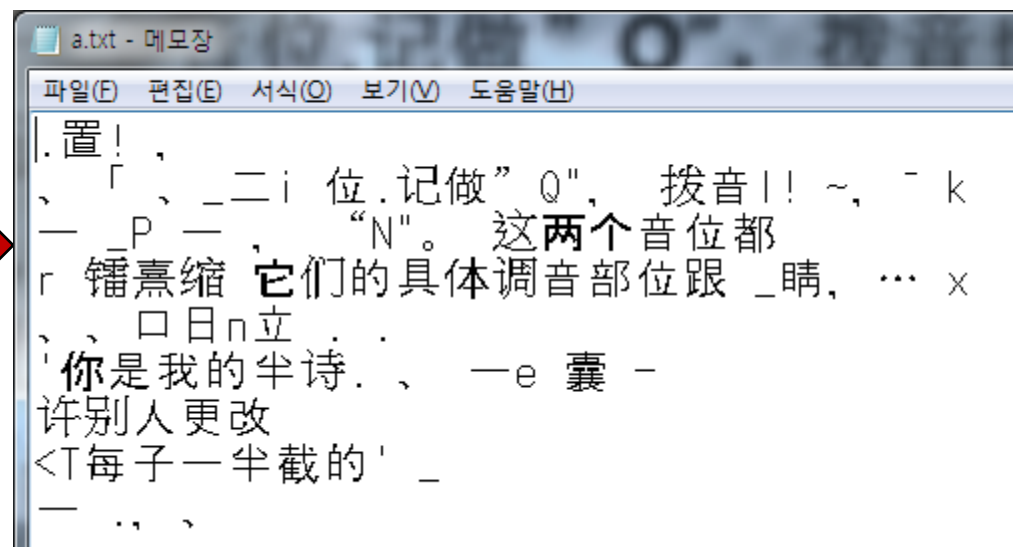
# DEMO for image10



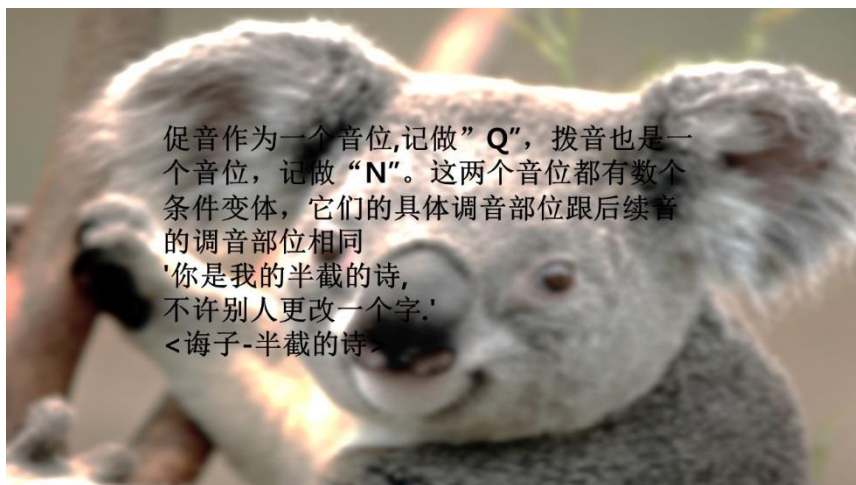
Image\_10



OCR



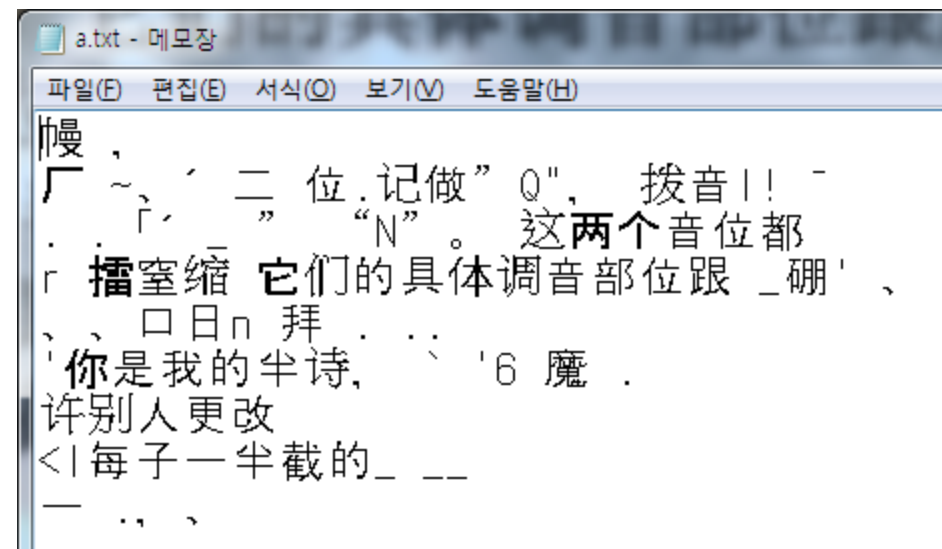
Using Second Option  
Elapsed time : 4060 milliseconds  
Speed : down  
Accuracy : similar



Image\_10



OCR



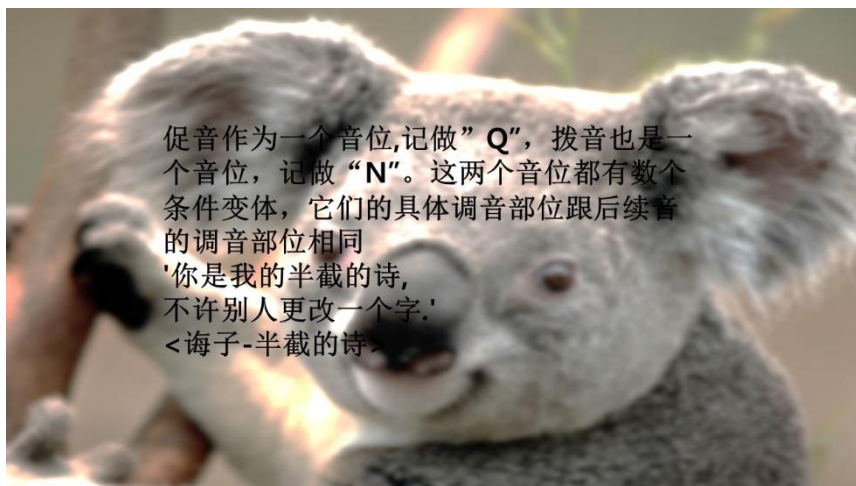
## Using Third Option

Elapsed time : 4340 milliseconds

Speed : down

Accuracy : similar

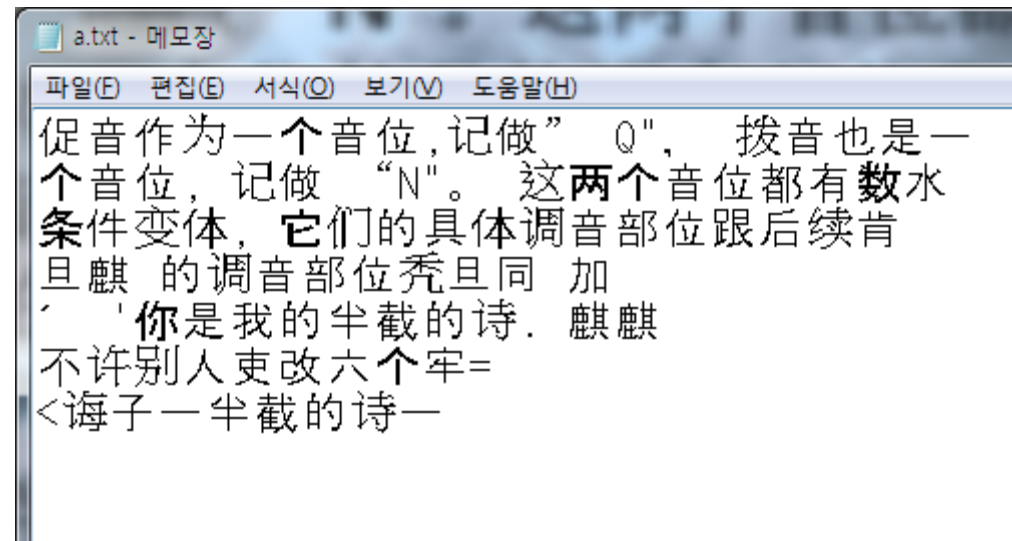
# DEMO for image10



Image\_10



OCR



Using Last Option

Elapsed time : 7730 milliseconds

Speed : down

Accuracy : up

## Best is Last Option

# Summary



# Summary

---

- More sentences in image, Longer elapsed time.
- Accuracy is more than 70%.
- If background of image is simple, option 1 or 2 is best.
- If background of image is simple, option 4 is worst.
- If background of image is complex, apply option 3 or 4 for more accuracy, but it takes long time.