Chinese Word Segmentation

- Author: Otis B.C. Chung

Feature

- This project aims to segment Chinese words.
- The program will download dictionary automatically.
- The default sentence is "這裡是MI2S實驗室,位於資訊工程學系的65802室", you can use the command-line arguments to pass your own sentence.

Usage

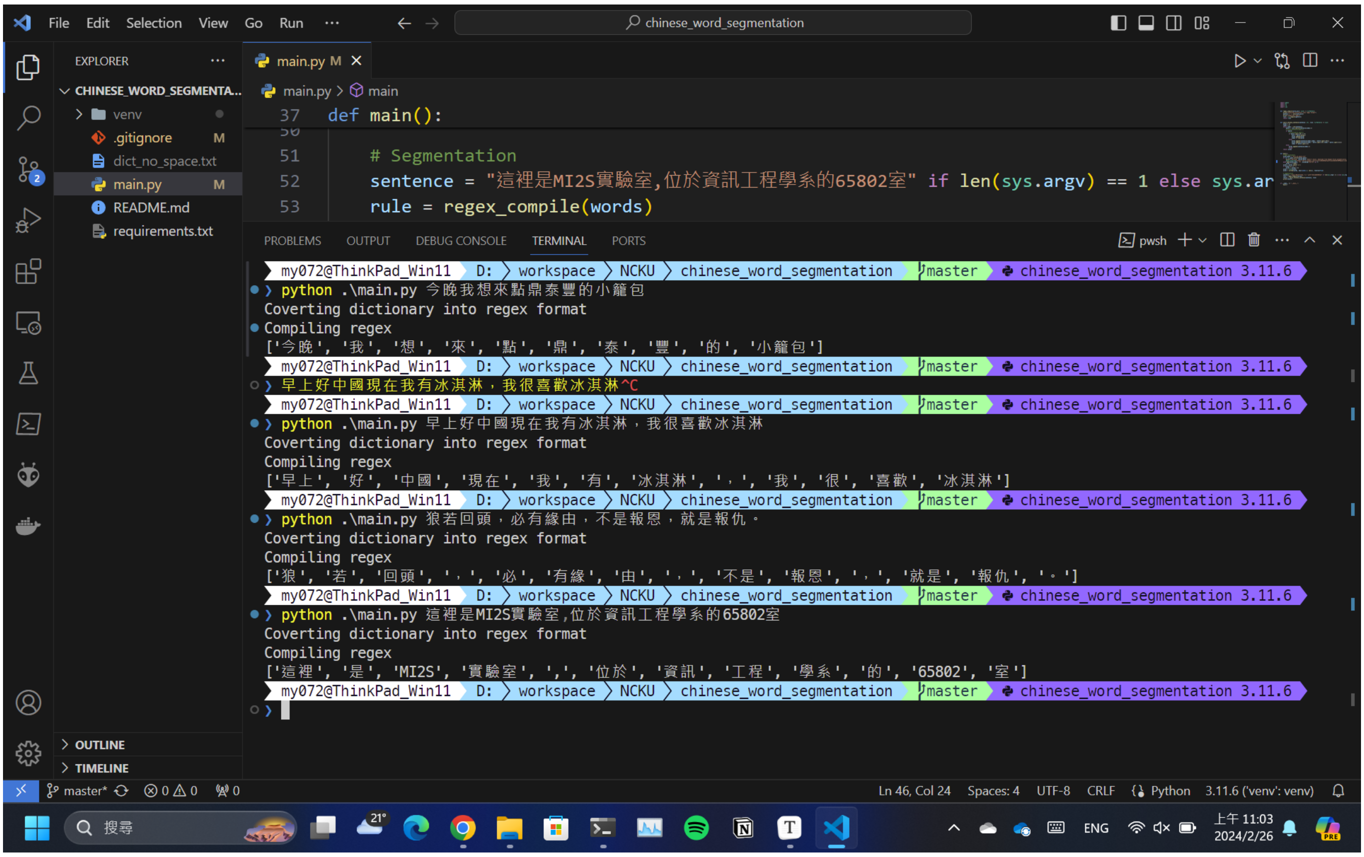
By default, you can use

```
python main.py
```

You can pass your sentence into program using command-line arguments

```
python main.py "狼若回頭，必有緣由，不是報恩，就是報仇。"
```

Result



```
def main():
        # Segmentation
        sentence = "這裡是MI2S實驗室,位於資訊工程學系的65802室" if len(sys.argv) == 1 else sys.ar
        rule = regex_compile(words)
```

```
 my072@ThinkPad_Win11  D:  workspace  NCKU  chinese_word_segmentation  master  chinese_word_segmentation 3.11.6
❯ python .\main.py 今晚我想來點鼎泰豐的小籠包
Coverting dictionary into regex format
Compiling regex
['今晚', '我', '想', '來', '點', '鼎', '泰', '豐', '的', '小籠包']
 my072@ThinkPad_Win11  D:  workspace  NCKU  chinese_word_segmentation  master  chinese_word_segmentation 3.11.6
❯ 早上好中國現在我有冰淇淋，我很喜歡冰淇淋^C
 my072@ThinkPad_Win11  D:  workspace  NCKU  chinese_word_segmentation  master  chinese_word_segmentation 3.11.6
❯ python .\main.py 早上好中國現在我有冰淇淋，我很喜歡冰淇淋
Coverting dictionary into regex format
Compiling regex
['早上', '好', '中國', '現在', '我', '有', '冰淇淋', '，', '我', '很', '喜歡', '冰淇淋']
 my072@ThinkPad_Win11  D:  workspace  NCKU  chinese_word_segmentation  master  chinese_word_segmentation 3.11.6
❯ python .\main.py 狼若回頭，必有緣由，不是報恩，就是報仇。
Coverting dictionary into regex format
Compiling regex
['狼', '若', '回頭', '，', '必', '有緣', '由', '，', '不是', '報恩', '，', '就是', '報仇', '。']
 my072@ThinkPad_Win11  D:  workspace  NCKU  chinese_word_segmentation  master  chinese_word_segmentation 3.11.6
❯ python .\main.py 這裡是MI2S實驗室,位於資訊工程學系的65802室
Coverting dictionary into regex format
Compiling regex
['這裡', '是', 'MI2S', '實驗室', ',', '位於', '資訊', '工程', '學系', '的', '65802', '室']
 my072@ThinkPad_Win11  D:  workspace  NCKU  chinese_word_segmentation  master  chinese_word_segmentation 3.11.6
❯
```