现代程序设计第十周作业

作业要求

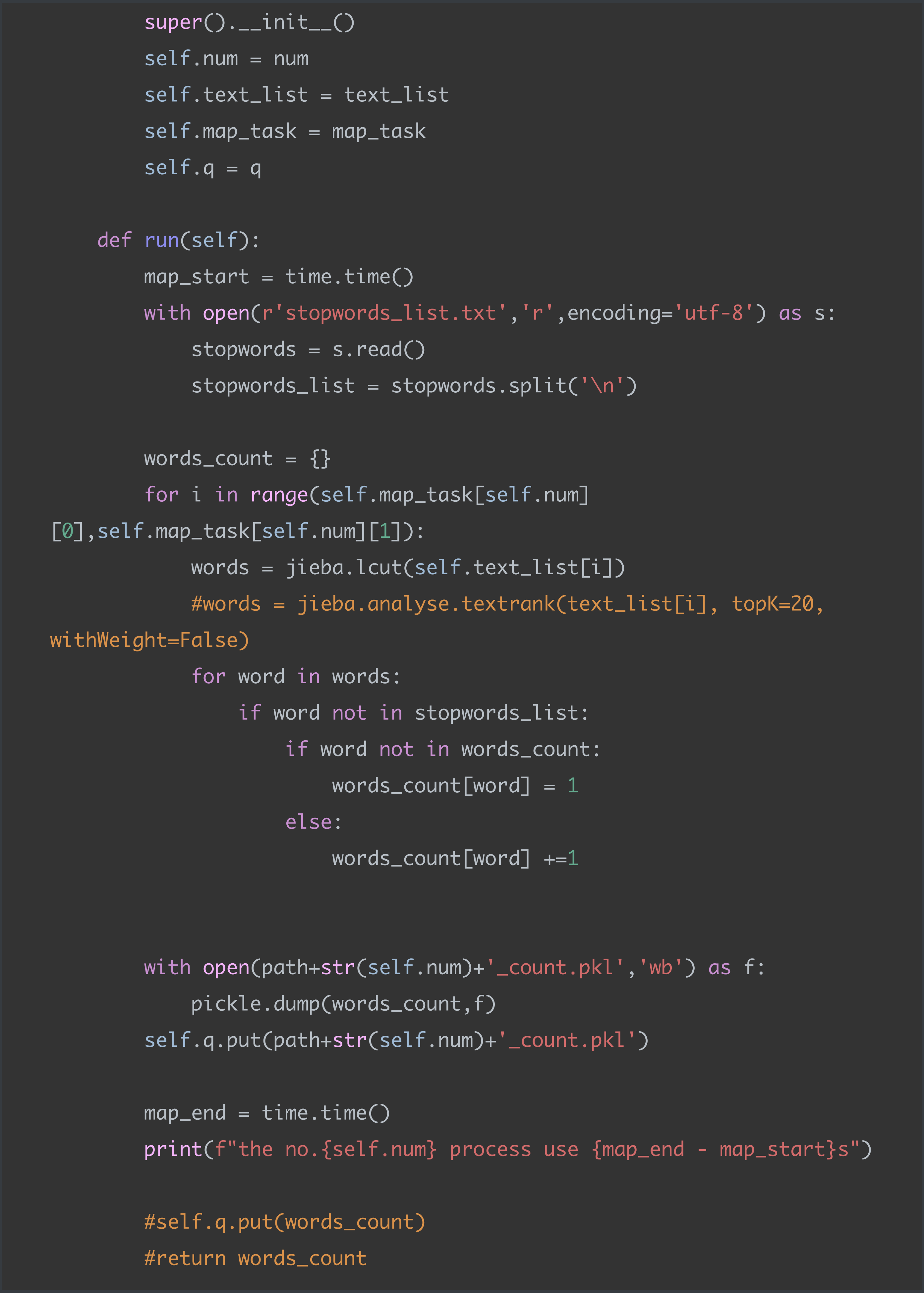
```
# MapReduce是利用多进程并行处理文件数据的典型场景。
# 作为一种编程模型，其甚至被称为Google的”三驾马车“之一(尽管目前由于内存计算等的普及已经被逐渐淘汰)。
# 在编程模型中，Map进行任务处理，Reduce进行结果归约。
# 本周作业要求利用Python多进程实现MapReduce模型下的文档库（搜狐新闻数据(SogouCS)，
# 下载地址：https://www.sogou.com/labs/resource/cs.php），注意仅使用页面内容，即新闻正文）词频统计功能。
```

库的导入

```python
from multiprocessing import Process,Queue
import time
import jieba
import jieba.analyse
import pickle
```

1. Map进程读取文档并进行词频统计，返回该文本的词频统计结果。

```python
class Map(Process):

    def __init__(self, num, text_list, map_task, q):
```



```
        super().__init__()
        self.num = num
        self.text_list = text_list
        self.map_task = map_task
        self.q = q

    def run(self):
        map_start = time.time()
        with open(r'stopwords_list.txt','r',encoding='utf-8') as s:
            stopwords = s.read()
            stopwords_list = stopwords.split('\n')

        words_count = {}
        for i in range(self.map_task[self.num]
[0],self.map_task[self.num][1]):
            words = jieba.lcut(self.text_list[i])
            #words = jieba.analyse.textrank(text_list[i], topK=20,
withWeight=False)
            for word in words:
                if word not in stopwords_list:
                    if word not in words_count:
                        words_count[word] = 1
                    else:
                        words_count[word] +=1



        with open(path+str(self.num)+'_count.pkl','wb') as f:
            pickle.dump(words_count,f)
        self.q.put(path+str(self.num)+'_count.pkl')

        map_end = time.time()
        print(f"the no.{self.num} process use {map_end - map_start}s")

        #self.q.put(words_count)
        #return words_count
```

2. Reduce进程收集所有Map进程提供的文档词频统计，更新总的文档库词频，并在所有map完成后保存总的词频到文件。

```
class Reduce(Process):

    def __init__(self,q):
        super().__init__()
        self.q = q

    def run(self):
        total_words_dict = {}
        for i in range(len(self.q)):
            with open(self.q[i]+'.pkl','rb') as f:
                word_dict = pickle.load(f)
            for word in word_dict:
                if word not in total_words_dict:
                    total_words_dict[word] = word_dict[word]
                else:
                    total_words_dict[word] += word_dict[word]
        with open(path+'total.pkl','wb') as f:
            pickle.dump(total_words_dict,f)
```

3. 主进程可提前读入所有的文档的路径列表，供多个Map进程竞争获取文档路径；或由主进程根据Map进程的数目进行分发；或者单独实现一个分发进程，与多个MAP进程通信。 \

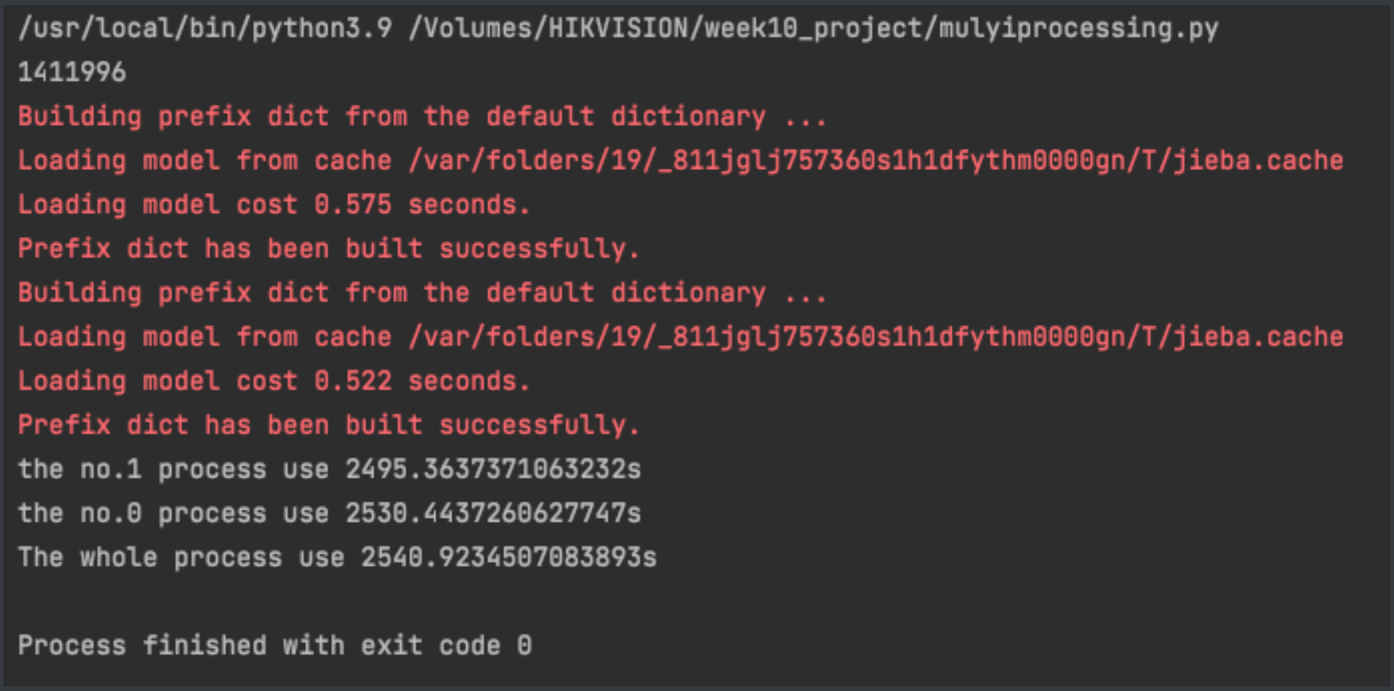
```
def read(path):
    with open(path,'rb') as f:
        sentences = f.readlines()
        #print(len(sentences))
        text_list = []
        for sentence in sentences:
            #print(sentence[2:11])
            if str(sentence[:9],encoding='gb18030') == '<content>':
                text_list.append(str(sentence[9:-9],encoding='gb18030'))
    return text_list
```

```
if __name__ == '__main__':

    text_list = read('news_sohusite_xml.dat')
    total_text = len(text_list)
    print(len(text_list))

    map_count = 8
    map_task=[[] for i in range(map_count)]
    for i in range(map_count-1):
        start = int((i/map_count)*total_text)
        end = int(((i+1)/map_count)*total_text)
        map_task[i].append(start)
        map_task[i].append(end)
    map_task[map_count-1].append(end)
    map_task[map_count-1].append(total_text)

    #print(map_task)
    # global word_count_queue
    word_count_queue = Queue()

    map_process = []
    reduce_process = []

    for i in range(map_count):
        p = Map(i,text_list,map_task,word_count_queue)
        map_process.append(p)

    main_start = time.time()

    '''for p in map_process:
        p.start()
    for p in map_process:
        p.join()'''

    q = [path+'0_count',path+'1_count',path+'2_count',path+'3_count',path+'4_count',path+'5_count',path+'6_count',path+'7_count']

    process = Reduce(q)
    process.start()
    process.join()

    main_end = time.time()
    print(f"The whole process use {main_end-main_start}s")
```

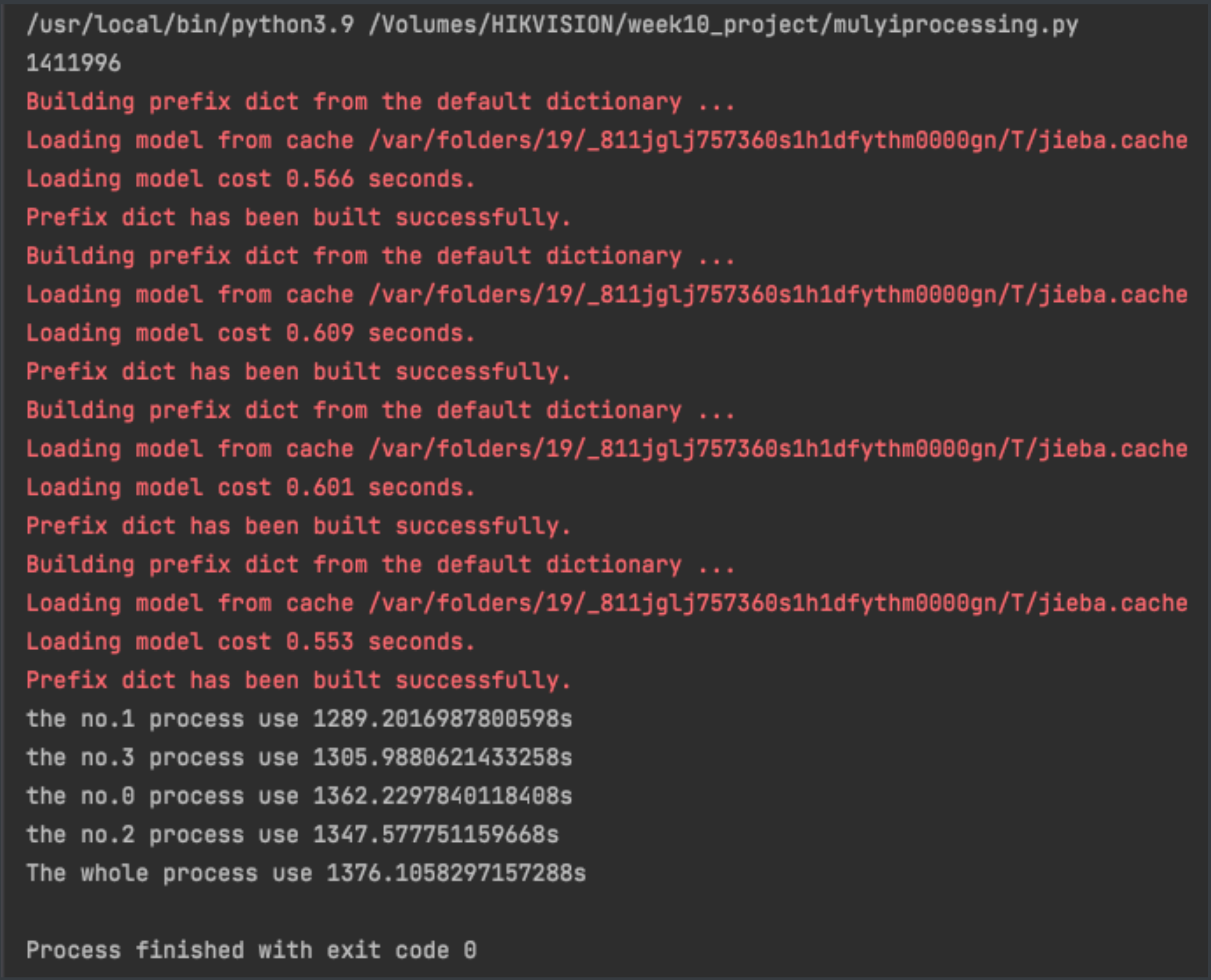
代码测试

分别从2，4，6，8，12，16进程对代码进行测试，并记录主进程和子进程的分别运行时长

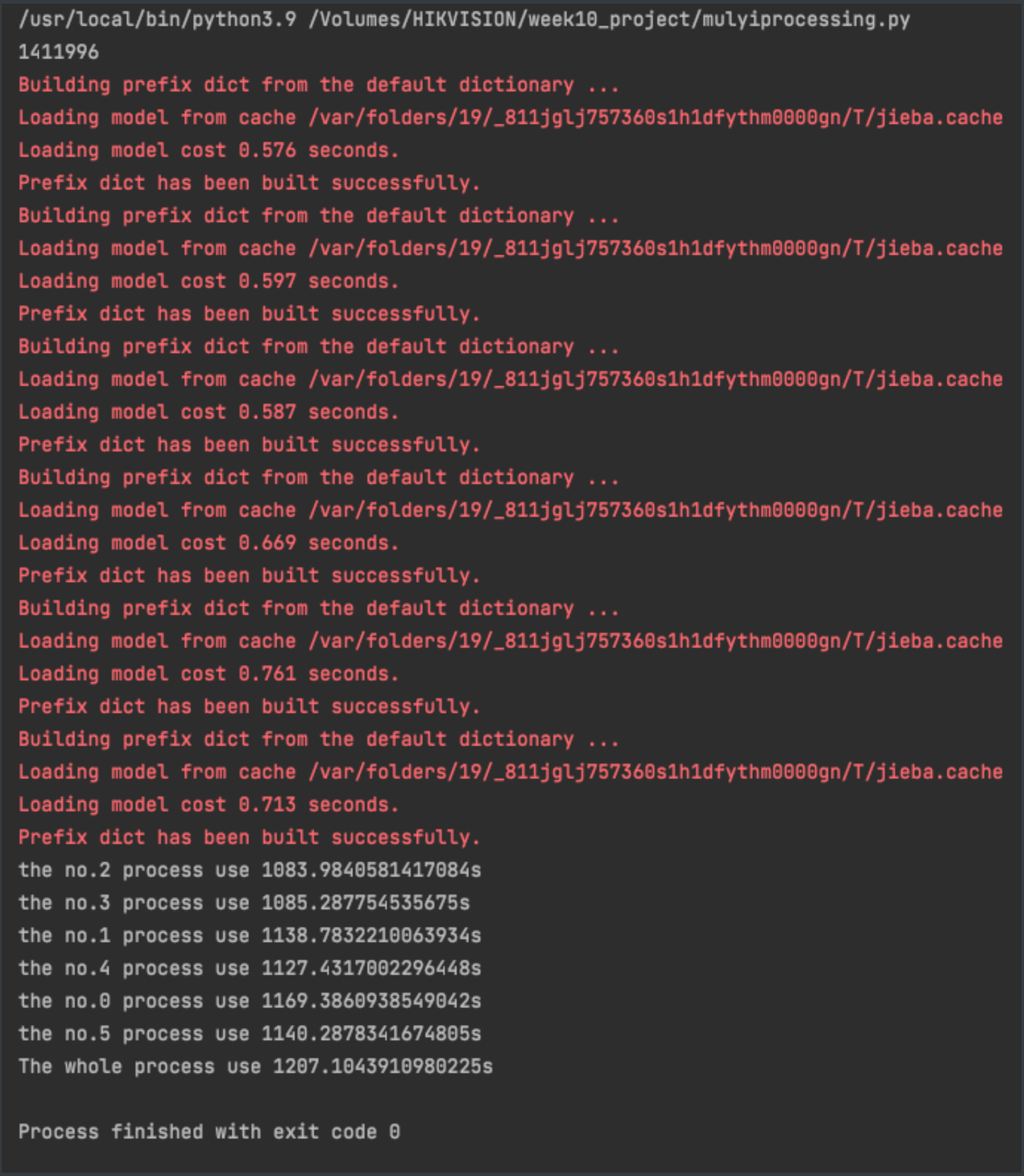
得到以下结果



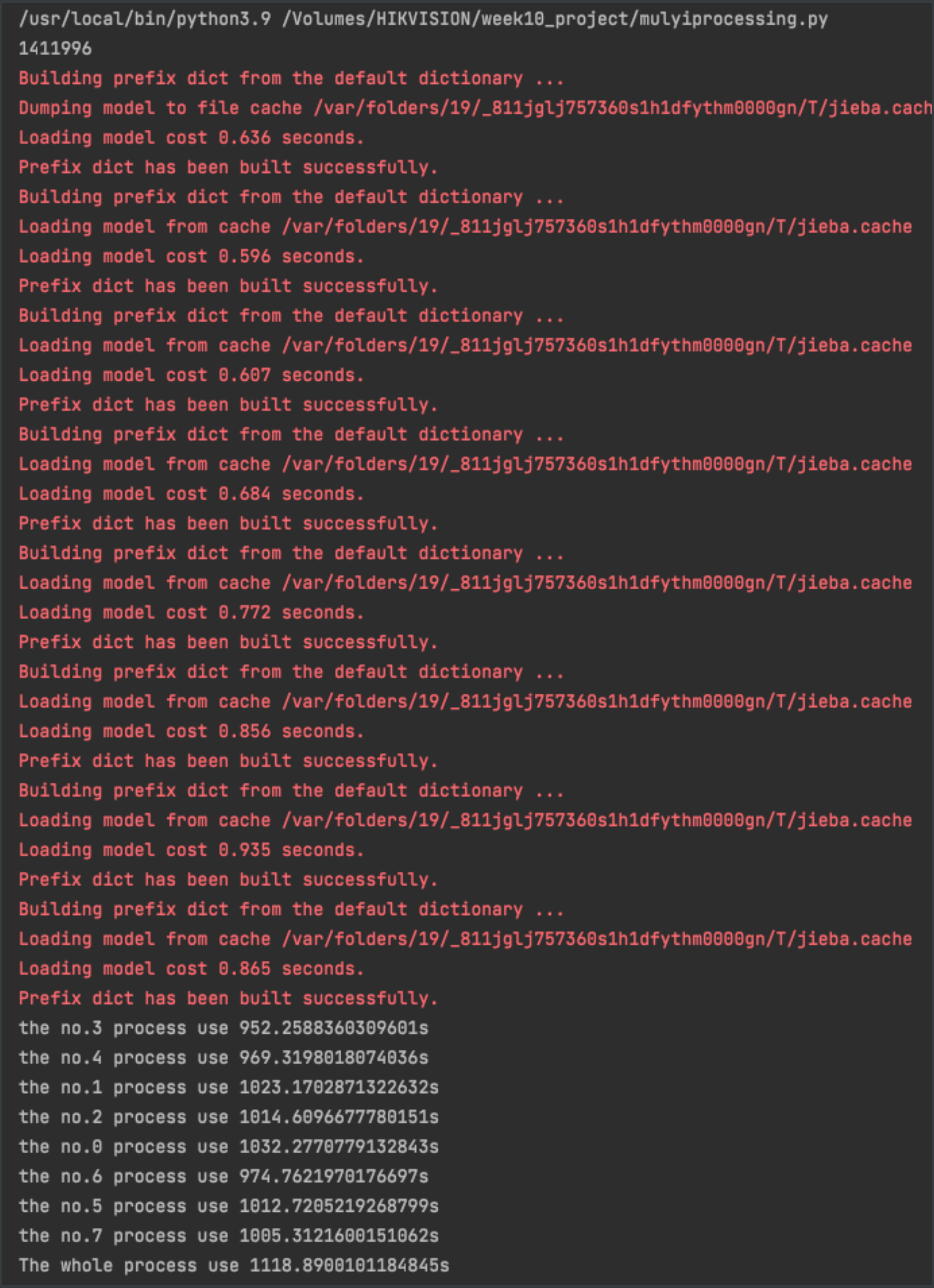
```
/usr/local/bin/python3.9 /Volumes/HIKVISION/week10_project/mulyiprocessing.py
1411996
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.575 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.522 seconds.
Prefix dict has been built successfully.
the no.1 process use 2495.3637371063232s
the no.0 process use 2530.4437260627747s
The whole process use 2540.9234507083893s

Process finished with exit code 0
```



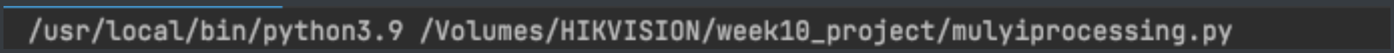
```
/usr/local/bin/python3.9 /Volumes/HIKVISION/week10_project/mulyiprocessing.py
1411996
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.566 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.609 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.601 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.553 seconds.
Prefix dict has been built successfully.
the no.1 process use 1289.2016987800598s
the no.3 process use 1305.9880621433258s
the no.0 process use 1362.2297840118408s
the no.2 process use 1347.577751159668s
The whole process use 1376.1058297157288s

Process finished with exit code 0
```



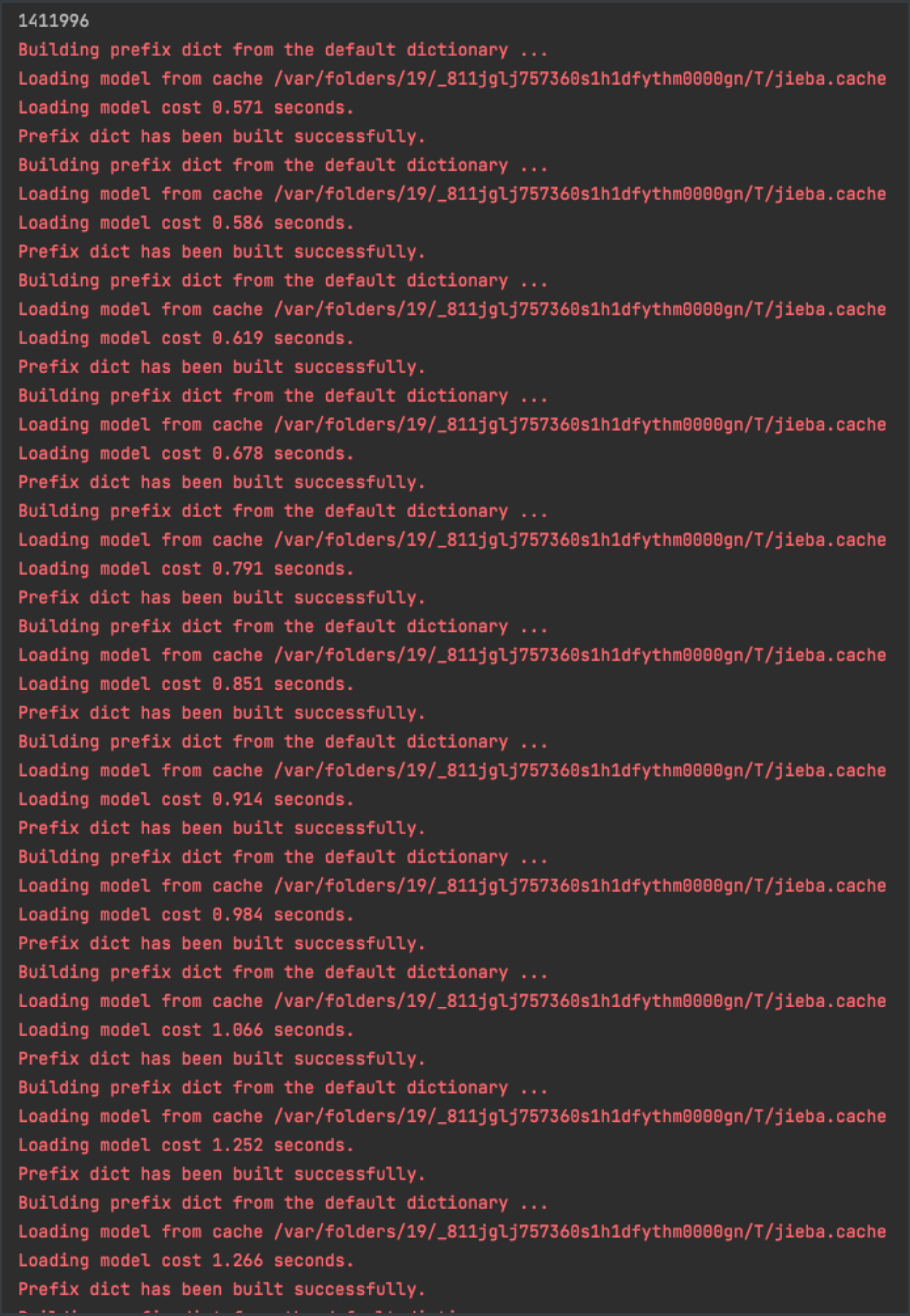
```
/usr/local/bin/python3.9 /Volumes/HIKVISION/week10_project/mulyiprocessing.py
1411996
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.576 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.597 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.587 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.669 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.761 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.713 seconds.
Prefix dict has been built successfully.
the no.2 process use 1083.9840581417084s
the no.3 process use 1085.287754535675s
the no.1 process use 1138.7832210063934s
the no.4 process use 1127.4317002296448s
the no.0 process use 1169.3860938549042s
the no.5 process use 1140.2878341674805s
The whole process use 1207.1043910980225s

Process finished with exit code 0
```



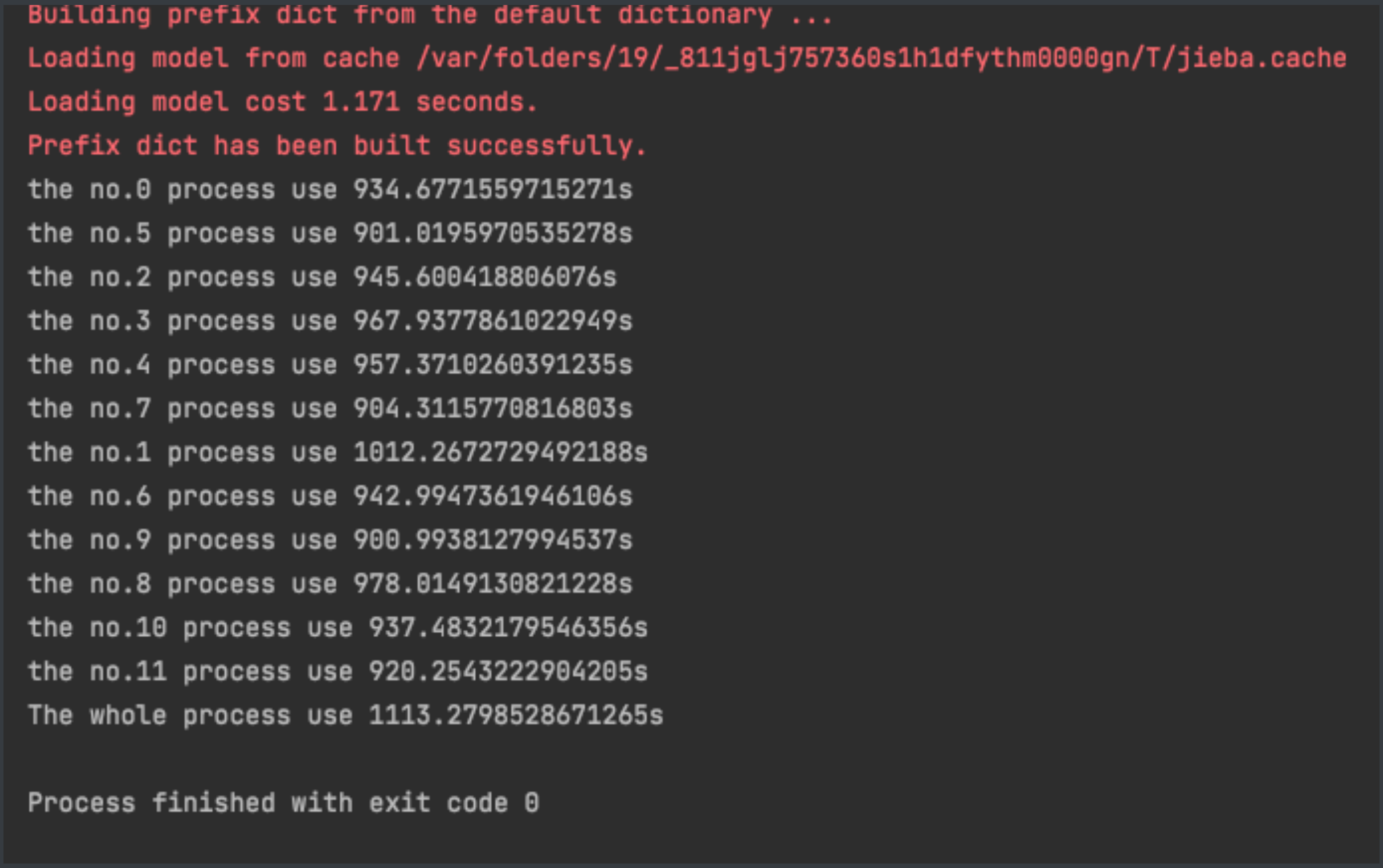
```
/usr/local/bin/python3.9 /Volumes/HIKVISION/week10_project/mulyiprocessing.py
1411996
Building prefix dict from the default dictionary ...
Dumping model to file cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cach
Loading model cost 0.636 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.596 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.607 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.684 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.772 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.856 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.935 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.865 seconds.
Prefix dict has been built successfully.
the no.3 process use 952.2588360309601s
the no.4 process use 969.3198018074036s
the no.1 process use 1023.1702871322632s
the no.2 process use 1014.6096677780151s
the no.0 process use 1032.2770779132843s
the no.6 process use 974.7621970176697s
the no.5 process use 1012.7205219268799s
the no.7 process use 1005.3121600151062s
The whole process use 1118.8900101184845s
```



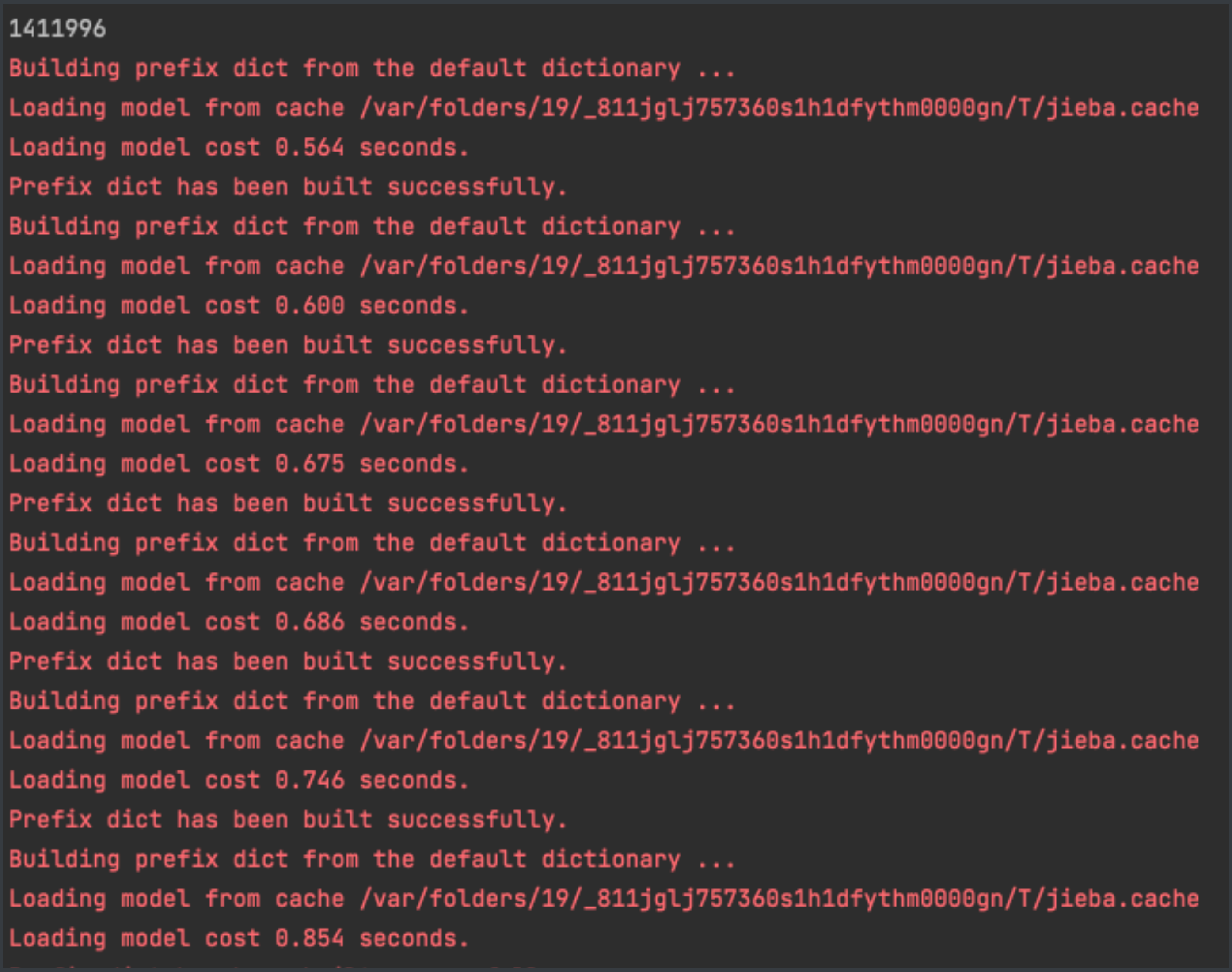
```
/usr/local/bin/python3.9 /Volumes/HIKVISION/week10_project/mulyiprocessing.py
```



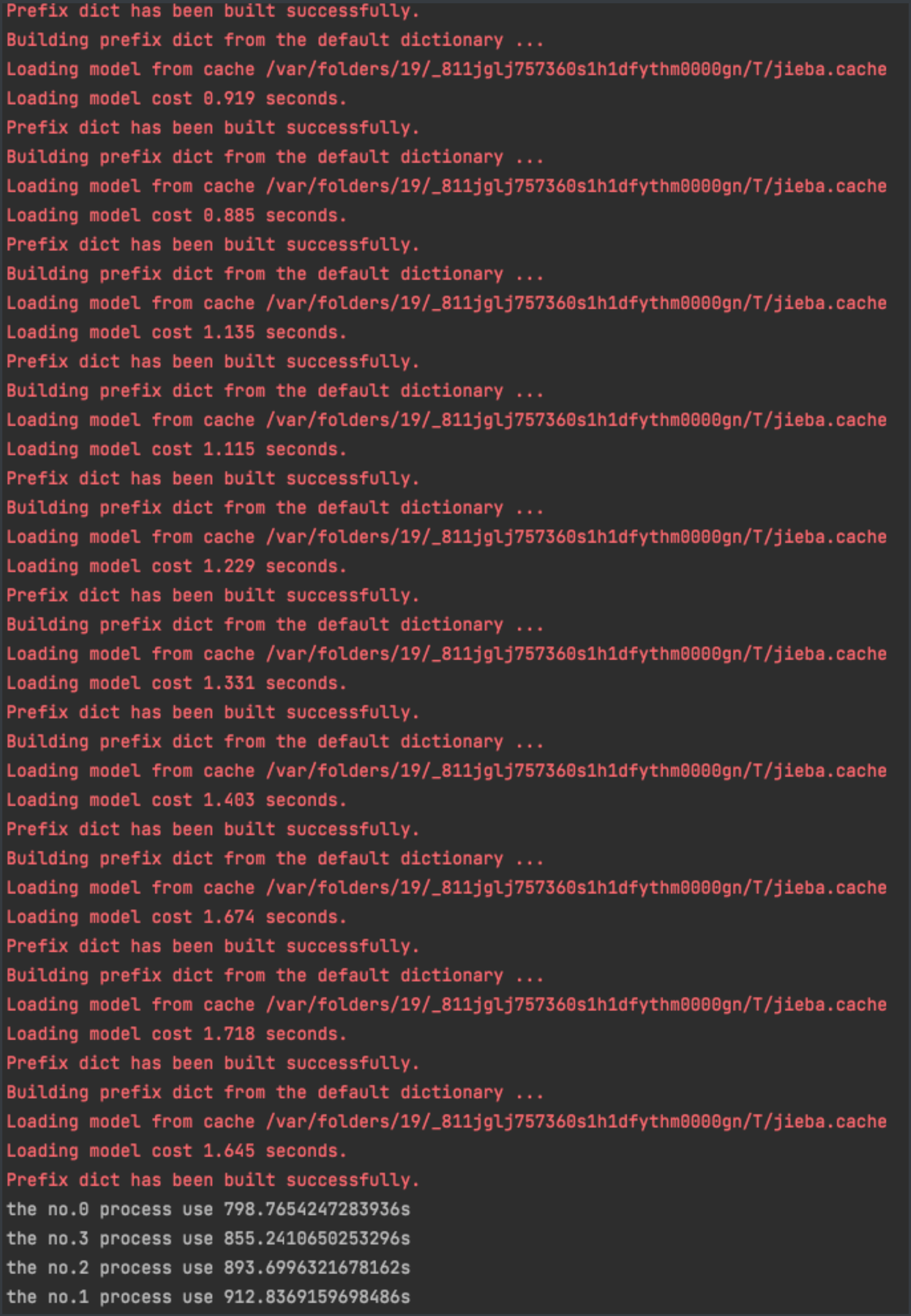
```
1411996
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.571 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.586 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.619 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.678 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.791 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.851 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.914 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.984 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.066 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.252 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.266 seconds.
Prefix dict has been built successfully.
```



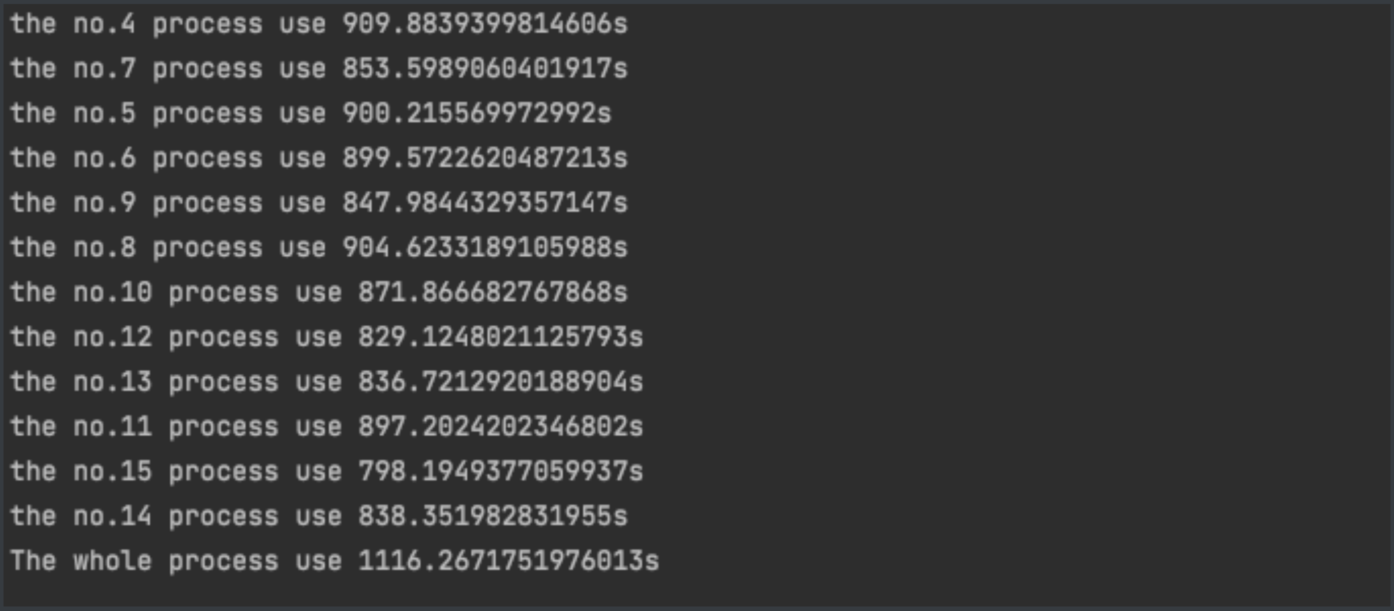
```
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.171 seconds.
Prefix dict has been built successfully.
the no.0 process use 934.6771559715271s
the no.5 process use 901.0195970535278s
the no.2 process use 945.600418806076s
the no.3 process use 967.9377861022949s
the no.4 process use 957.3710260391235s
the no.7 process use 904.3115770816803s
the no.1 process use 1012.2672729492188s
the no.6 process use 942.9947361946106s
the no.9 process use 900.9938127994537s
the no.8 process use 978.0149130821228s
the no.10 process use 937.4832179546356s
the no.11 process use 920.2543222904205s
The whole process use 1113.2798528671265s

Process finished with exit code 0
```



```
1411996
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.564 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.600 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.675 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.686 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.746 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.854 seconds.
```



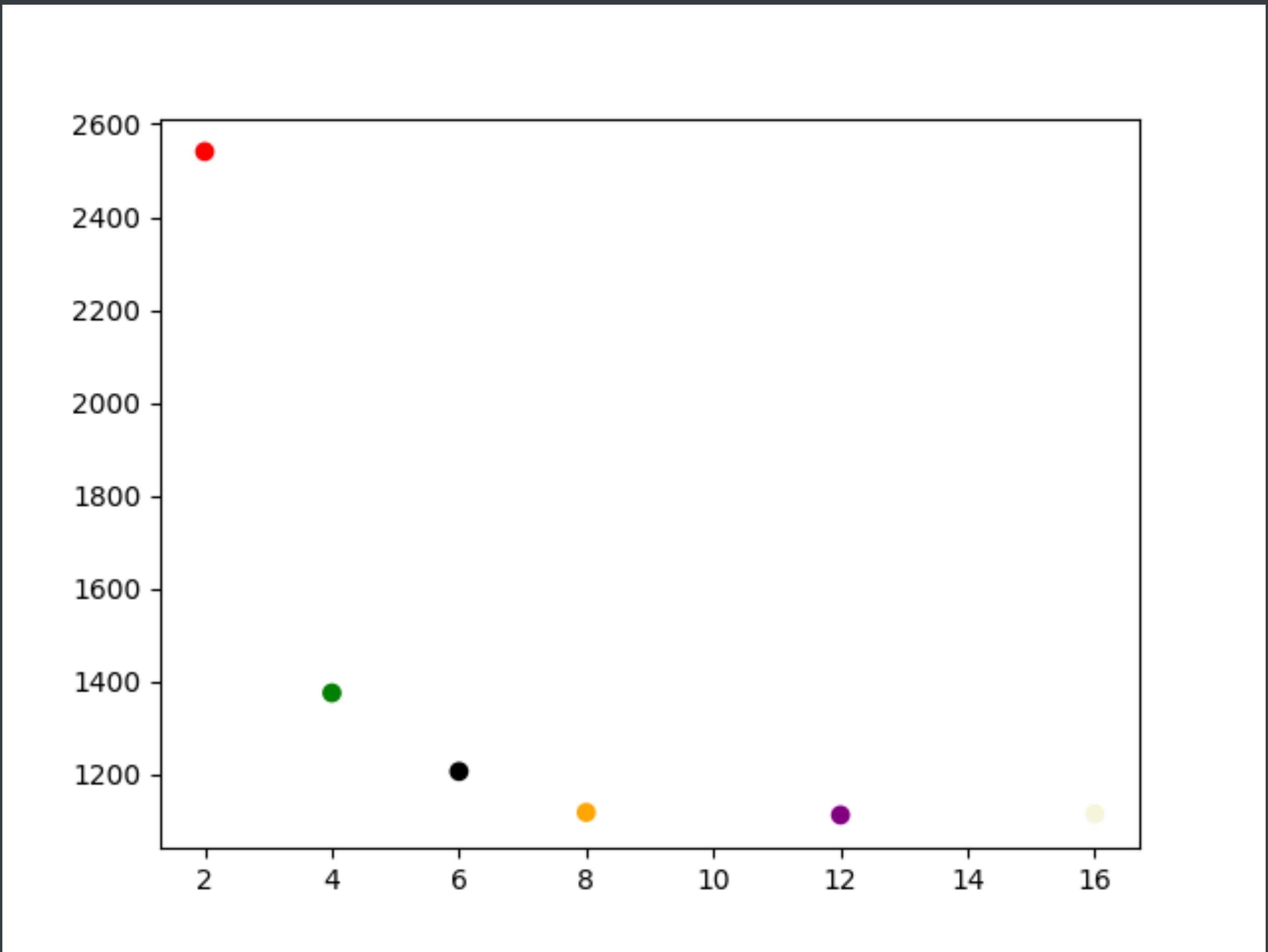
```
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.919 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 0.885 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.135 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.115 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.229 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.331 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.403 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.674 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.718 seconds.
Prefix dict has been built successfully.
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/19/_811jglj757360s1h1dfythm0000gn/T/jieba.cache
Loading model cost 1.645 seconds.
Prefix dict has been built successfully.
the no.0 process use 798.7654247283936s
the no.3 process use 855.2410650253296s
the no.2 process use 893.6996321678162s
the no.1 process use 912.8369159698486s
```



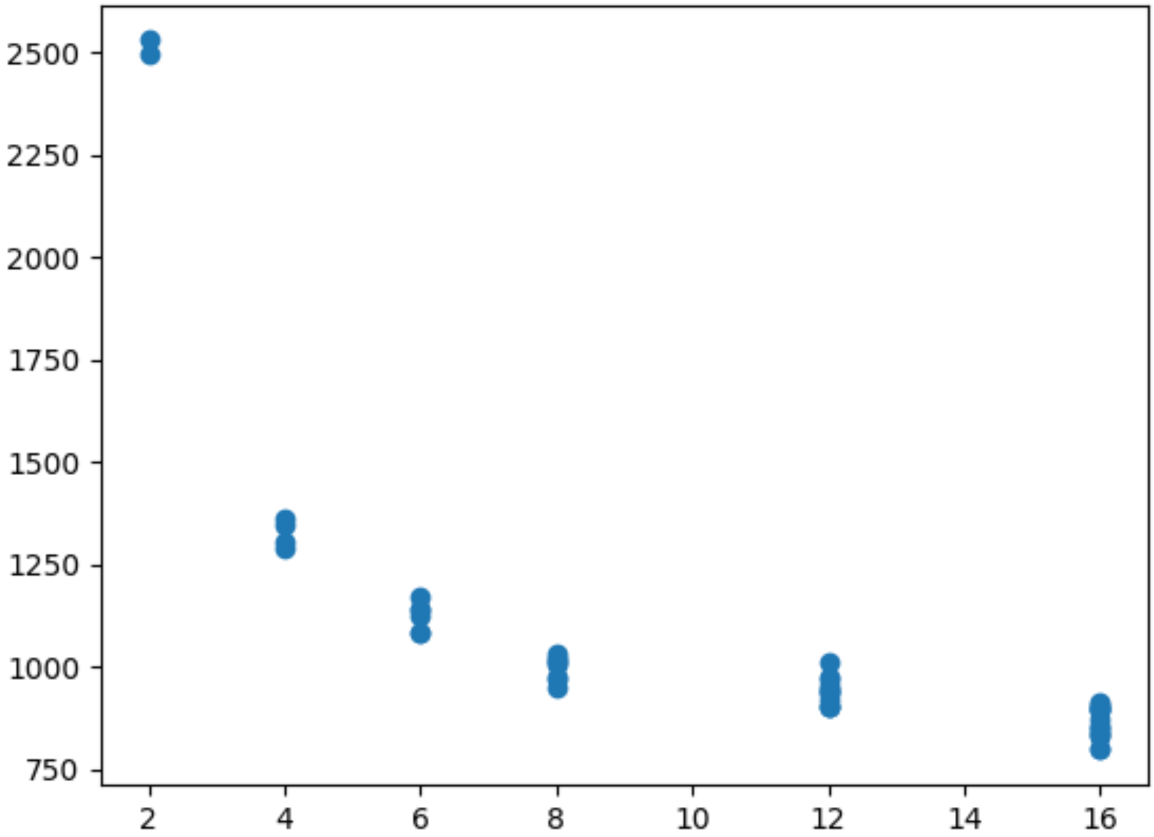
```
the no.4 process use 909.8839399814606s
the no.7 process use 853.5989060401917s
the no.5 process use 900.215569972992s
the no.6 process use 899.5722620487213s
the no.9 process use 847.9844329357147s
the no.8 process use 904.6233189105988s
the no.10 process use 871.866682767868s
the no.12 process use 829.1248021125793s
the no.13 process use 836.7212920188904s
the no.11 process use 897.2024202346802s
the no.15 process use 798.1949377059937s
the no.14 process use 838.351982831955s
The whole process use 1116.2671751976013s
```

统计后绘制图像

主进程



分支进程


2500
2250
2000
1750
1500
1250
1000
750
2
4
6
8
10
12
14
16