Web スクレイピング

問題

金沢工業大学の研究室ガイド (https://kitnet.jp/laboratories/) のページをスクレイピングした結果を JSON 形式でファイル `labs.json` に書き出す Python の関数 `scrape_labs()`と、`labs.json` を読み込み、与えられた文字列 `keyword` を研究室のキーワードとして含んでいる研究室のリストを返す関数 `search_labs(keyword)`を実装せよ。[10 点]

仕様

`labs.json` のフォーマット（読みやすいように改行・インデントを入れているが、これらは無くて良い）：

```
[{'dept': '機械工学科',
  'labs': [{'keywords': ['キャビテーション', 'マイクロバブル', 'ウォータジェット', '流体力学'],
            'lab': '佐藤・杉本研究室'},
           {'keywords': ['形状記憶合金', '応力誘起変態', '新材料'], 'lab': '矢島・岸研究室'},
           {'keywords': ['コンピュータシミュレーション', '自動車', '事前予測技術'], 'lab': '山部昌研究室'},
           {'keywords': ['カスタムメイド', '医療・福祉', '人工関節', '骨再生'], 'lab': '新谷一博研究室'},...]},
...
{'dept': '情報工学科',
  'labs': [{'keywords': ['並列処理技術', 'PC クラスタ', '分散型並列計算機'], 'lab': '津田伸生研究室'},
           {'keywords': ['セキュリティ', '教育アプリ', '不正アクセス', '不正侵入', 'サイバー攻撃検知'],
            'lab': '五十嵐寛研究室'},
           {'keywords': ['自然計算', '量子コンピュータ', '分子コンピュータ', '粘菌', 'アメーバ'],
            'lab': '蜷川繁研究室'},...]},
...
{'dept': '応用化学科',
  'labs': [...,
           {'keywords': ['植物工場', 'アグロメディカルフーズ', '機能性野菜', '農業 ICT'],
            'lab': '松本恵子研究室'}]}]
```

まず、`labs.json` は UTF-8 でエンコードすること。トップレベルでは学科ごとに学科名（キーは `'dept'`）および研究室のリスト（キーは `'labs'`）を保存した辞書を要素として持つリストである。各学科の研究室リストは研究室名（キーは `'lab'`）およびキーワードのリスト（キーは `'keywords'`）を保存した辞書を要素として持つリストである（上記の JSON では keywords、lab の順で表示されているが）。

search_labs の出力例:

```
>>> from 作成したファイル名から.py の拡張子を除いた部分 import search_labs
>>> search_labs('画像認識')
[]
>>> search_labs('ビッグデータ')
[('情報工学科', '中野淳研究室')]
>>> search_labs('データマイニング')
[('情報工学科', '元木光雄研究室'), ('情報工学科', '林亮子研究室'), ('応用バイオ学科', '相良純一研究室')]
```

上記のように(学科名, 研究室名)のタプルからなるリストを返すこと。

注意

- この課題に取り組むにあたっては、非標準ライブラリではあるが requests と Beautiful Soup を使うこと。
- プログラム中に研究室に関する情報や URL をハードコーディングしてはならない。
- 研究室ごとのキーワードは研究室ガイドのトップページには記載されておらず、トップページから各研究室へのリンクをたどった先のページにある。その URL を引数 url として、研究室のキーワードを取り出すには次の関数を用いること（研究室ごとの URL はトップページから自分で見つけ出さなくてはならない）。

```
import time
import requests
from bs4 import BeautifulSoup

def get_keywords(url):
    time.sleep(0.5)
    r = requests.get(url)
    soup = BeautifulSoup(r.content, 'lxml')
    keywords_csv = soup.find('meta', attrs={'name': 'keywords'}).get('content')
    return keywords_csv.split(',')
```

提出方法

e シラバスにて 2 つの関数 scrape_labs と search_labs を含んだ Python スクリプト（1 つのファイル）をアップロード。
提出期限は 12 月 4 日（火）深夜。

できる限り早く取りかかること！