Проект “новостные тренды”

С итальянского регионального новостного портала La Nazione

Войтович Екатерина

Цели и задачи проекта + используемые инструменты

- Разобраться с устройством html-страниц выбранного сайта
- Написать функции, которые извлекают из страницы сайта ***заголовок***, ***краткое описание*** и ***полный текст*** новости.
- Записать в базу данных эти данные как минимум с нескольких страниц сайта
- Провести морфологический анализ полученных текстов (найти самые частотные слова по отдельным статьям и общие)
- Создать и оформить сайт
- Внести на сайт результаты анализа базы данных
- Залить сайт на pythonanywhere

sqlite,
BeautifulSoup

nltk, sklearn,
tf-idf,
pymorphy

flask

Функции для работы с html-страницей сайта

```python
#функция для парсинга страницы со списокм новостей (блок одной новости)
def parse_news_page_block(one_block):
    block = {}
    a = one_block.find('a', {'class': 'title__link'})
    block['title'] = a.text
    block['href'] = a.attrs['href']
    block['short_text'] = one_block.find('div', {'class': 'abstract'}).text
    return block

page_number = 1
url = f'https://www.lanazione.it/cronaca?page={page_number}'
req = session.get(url, headers={'User-Agent': ua.random})
page = req.text
page = requests.get(url)
soup = BeautifulSoup(page.text, 'html.parser')
news = soup.find_all('div', {'class': 'card__text'})
for n in news:
    print(parse_news_page_block(n))
```

71] ✓ 0.7s Python

Output exceeds the size limit. Open the full output data in a text editor

{'title': 'Bugliani: "Giusto dare supporto" ', 'href': 'https://www.lanazione.it/massa-carrara/cronaca/bugliani-giusto-dare-supporto-1.7520181', 'short_text': None}

{'title': 'GiÃ\xa0 125 studenti ucraini in classe Tutti accolti bene dai compagni ', 'href': 'https://www.lanazione.it/lucca/cronaca/gia-125-studenti-ucraini-in-classe-tutti-accolti-bene-dai-compagni-1.7520171', 'short_text':
La dirigente Buonriposi fa il punto sullâ€™accoglienza dei ragazzi fuggiti dalle zone di conflitto. Pochi i mediatori linguistici, ma i giovani riescono lo stesso a comunicare fra di loro
}

{'title': 'Spaccio di hashish e coca Smantellata la â\x80\x9cbandaâ\x80\x9c ', 'href': 'https://www.lanazione.it/umbria/cronaca/spaccio-di-hashish-e-coca-smantellata-la-banda-1.7520166', 'short_text':
Gli agenti della Guardia di Finanza hanno eseguito cinque ordinanze di custodia cautelare. Quaranta chili di droga sequestrati nel corso dellâ€™indagine
}

{'title': 'Avis di Massa Marittima si rinnova 15 nuovi donatori tra i 18enni ', 'href': 'https://www.lanazione.it/grosseto/cronaca/avis-di-massa-marittima-si-rinnova-15-nuovi-donatori-tra-i-18enni-1.7520163', 'short_text': None}

{'title': 'Energia e materie prime "Intervenire sui costi" ', 'href': 'https://www.lanazione.it/umbria/cronaca/energia-e-materie-prime-intervenire-sui-costi-1.7520156', 'short_text':
Approfonditi in Commissione i temi delle ripercussioni sul comparto vitivinicolo e agroalimentare
}

```python
#функция для парсинга отдельной страницы новости
def parse_one_article(block):
    url_one = block['href']
    req = session.get(url_one, headers={'User-Agent': ua.random})
    page = req.text
    soup = BeautifulSoup(page, 'html.parser')
    block['full_text'] = soup.find('article', {'class': 'sc-1ig42x7-6 iSFXhF'}).text
    return block

page_number = 1
url = f'https://www.lanazione.it/cronaca?page={page_number}'
req = session.get(url, headers={'User-Agent': ua.random})
page = req.text
page = requests.get(url)
soup = BeautifulSoup(page.text, 'html.parser')
news = soup.find_all('div', {'class': 'card__text'})
print(parse_one_article(parse_news_page_block(news[1])))
```

[172] ✓ 1.7s Python

{'title': 'GiÃ\xa0 125 studenti ucraini in classe Tutti accolti bene dai compagni ', 'href': 'https://www.lanazione.it/lucca/cronaca/gia-125-studenti-ucraini-in-classe-tutti-accolti-bene-dai-compagni-1.7520171', 'short_text':

La dirigente Buonriposi fa il punto sullâ€™accoglienza dei ragazzi fuggiti dalle zone di conflitto. Pochi i mediatori linguistici, ma i giovani riescono lo stesso a comunicare fra di loro

, 'full_text': 'Sono centoventicinque i ragazzi ucraini che frequentano ad oggi le scuole del nostro territorio. Distribuiti tra Piana, Mediavalle, Gargagnana e Versilia, i giovani accolti dalle nostre istituzioni scolastiche si stanno, pur con la comprensibile fatica, inserendo nei diversi contesti scolastici. In alcuni di questi vi è la disponibilità del mediatore linguistico. Ma non in tutti. "Vi è un'oggettiva difficoltà a reperire queste figure – sostiene la dirigente dell'ufficio scolastico Donatella Buonriposi – una importante collaborazione arriva grazie alla disponibilità del sacerdote della chiesa cattolica, greco ucraina Volodymyr Lyupac che conosce molto bene la realtà di Lucca e non solo". \n\n\n\nL'istituto scolastico ha redatto, proprio in questi giorni, il monitoraggio da trasmettere alla prefettura: "Ad oggi – prosegue Buonriposi – i bambini e ragazzi inseriti nelle scuole sono 125, probabilmente dovremmo aggiungere qualche altra unità che si è inserita nelle ultime ore portando, probabilmente, la cifra a 130". L'arrivo a scuola dei ragazzi, è stata ben accolta da studenti e insegnanti; una bella gara di solidarietà scaturita grazie alla capacità dei ragazzi che, al di là della difficoltà linguistica, riescono comunque a comunicare attingendo all'universale linguaggio che è proprio dei giovani di oggi. Per quanto riguarda il Centro storico, sono cinque gli studenti al Comprensivo Lucca Centro, uno alla Ungaretti, uno a Lucca 3 e due a Lucca cinque. Nella Piana, i giovani studenti sono così suddivisi: Comprensivo di Altopascio due, Comprensivo di Lammari, cinque, Comprensivo di Camigliano, cinque. In Mediavalle incontriamo al Comprensivo di Borgo a Mozzano due studenti, a Coreglia Antelminelli cinque. In Grafagnana, una presenza al Comprensivo di Castelnuovo e quattro a Piazza al Serchio. \n\n\n\nImportante presenza al Cpia, il Centro provinciale di istruzione degli adulti che vede 39 persone inserite. Nei Comprensivi della Versilia, troviamo rispettivamente quattro studenti a Pietrasanta, due a Seravezza, nove a Migliarina, due al Darsena, 7 al Marco Polo-Viani e due a Torre del Lago. Passiamo infine agli istituti superiori. Al classici di Viareggio è presente uno studente, all'istituto superiore Chini-Michelangelo di Forte dei Marmi, sono tre, all'istituto superiore Piaggia di Viareggio uno, all'Artiglio due, a Pertini di Lucca uno, al liceo artistico Passaglia di Lucca, due. Conclude Buonriposi: "Sono soddisfatta di come ha funzionato fino a oggi la macchina scolastica dell'accoglienza; ancora una volta la scuola costituisce

Сайт, хоть он и неинформативный, выложен на pythonanywhere

http://ekaterinavoit.pythonanywhere.com/