Министерство образования и науки Российской Федерации

Московский авиационный институт (национальный исследовательский университет)

ЖУРНАЛ

ПО ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ

Наименование практики: *исследовательская*
Студент: И. Т. Батяновский
Факультет №8, курс 3, группа 7

Практика с 29.06.20 по 12.07.20

Москва, 2020

ИНСТРУКЦИЯ

о заполнении журнала по производственной практике

Журнал по производственной практике студентов имеет единую форму для всех видов практик.

Задание в журнал вписывается руководителем практики от института в первые три-пять дней пребывания студентов на практике в соответствии с тематикой, утверждённой на кафедре до начала практики. Журнал по производственной практике является основным документом для текущего и итогового контроля выполнения заданий, требований инструкции и программы практики.

Табель прохождения практики, задание, а также технический отчёт выполняются каждым студентом самостоятельно.

Журнал заполняется студентом непрерывно в процессе прохождения всей практики и регулярно представляется для просмотра руководителям практики. Все их замечания подлежат немедленному выполнению.

В разделе «Табель прохождения практики» ежедневно должно быть указано, на каких рабочих местах и в качестве кого работал студент. Эти записи проверяются и заверяются цеховыми руководителями практики, в том числе мастерами и бригадирами. График прохождения практики заполняется в соответствии с графиком распределения студентов по рабочим местам практики, утверждённым руководителем предприятия. В разделе «Рационализаторские предложения» должно быть приведено содержание поданных в цехе рационализаторских предложений со всеми необходимыми расчётами и эскизами. Рационализаторские предложения подаются индивидуально и коллективно.

Выполнение студентом задания по общественно-политической практике заносятся в раздел «Общественно-политическая практика». Выполнение работы по оказанию практической помощи предприятию (участие в выполнении спецзаданий, работа сверхурочно и т.п.) заносятся в раздел журнала «Работа в помощь предприятию» с последующим письменным подтверждением записанной работы соответствующими цеховыми руководителями. Раздел «Технический отчёт по практике» должен быть заполнен особо тщательно. Записи необходимо делать чернилами в сжатой, но вместе с тем чёткой и ясной форме и технически грамотно. Студент обязан ежедневно подробно излагать содержание работы, выполняемой за каждый день. Содержание этого раздела должно отвечать тем конкретным требованиям, которые предъявляются к техническому отчёту заданием и программой практики. Технический отчёт должен показать умение студента критически оценивать работу данного производственного участка и отразить, в какой степени студент способен применить теоретические знания для решения конкретных производственных задач.

Иллюстративный и другие материалы, использованные студентом в других разделах журнала, в техническом отчёте не должны повторяться, следует ограничиваться лишь ссылкой на него. Участие студентов в производственно-технической конференции, выступление с докладами, рационализаторские предложения и т.п. должны заноситься на свободные страницы журнала.

Примечание. Синьки, кальки и другие дополнения к журналу могут быть сделаны только с разрешения администрации предприятия и должны подшиваться в конце журнала.

Руководители практики от института обязаны следить за тем, чтобы каждый цеховой руководитель практики перед уходом студентов из данного цеха в другой цех вписывал в журнал студента отзывы об их работе в цехе.

Текущий контроль работы студентов осуществляется руководители практики от института и цеховыми руководителями практики заводов. Все замечания студентам руководители делают в письменном виде на страницах журнала, ставя при этом свою подпись и дату проверки.

Результаты защиты технического отчёта заносятся в протокол и одновременно заносятся в ведомость и зачётную книжку студента.

Примечание. Нумерация чистых страниц журнала проставляется каждым студентом в своём журнале до начала практики.

С инструкцией о заполнении журнала ознакомились:

« » _____ 2020 г. Студент Батяновский И. Т. _____

(дата) (подпись)

ЗАДАНИЕ

кафедры 806 по вычислительной/исследовательской практике:
парсить новостные ленты, новости, названия и модель, которая их классифицирует.

Руководитель практики от института:

« » _____ 2020 г. Кухтичев А. А. _____
(дата) (подпись)

ТАБЕЛЬ ПРОХОЖДЕНИЯ ПРАКТИКИ

Дата	**Содержание или наименование проделанной работы**	**Место работы**	**Время работы**		**Подпись цехового руководителя**
			Начало	Конец	
29.06.2019	Получение задания	МАИ	9:00	18:00	
01.07.2019	Чтение литературы по Web-scraping	МАИ	9:00	18:00	
02.07.2019	Установка необходимого ПО	МАИ	9:00	18:00	
03.07.2019	Изучение html страниц https://meduza.io/	МАИ	9:00	18:00	
04.07.2019	Написание парсера	МАИ	9:00	18:00	
05.07.2019	Написание кролера(crawler)	МАИ	9:00	18:00	
06.07.2019	Изучение html страниц https://lenta.ru/	МАИ	9:00	18:00	
07.07.2019	Написание парсера и кролера(с загрузкой данных на MySQL)	МАИ	9:00	18:00	
09.07.2019	Поиск информации по написанию модели классификации текста	МАИ	9:00	18:00	
10.07.2019	Написание программы перевода данных с MySQL в дата фрейм pandas	МАИ	9:00	18:00	
11.07.2019	Написание модели	МАИ	9:00	18:00	
12.07.2018	Сдача журнала	МАИ	9:00	18:00	

Отзывы цеховых руководителей практики

Студент Батяновский И. Т. разработал веб кролер(внутри него парсер и загрузка данных на MySQL) и модель классификации текста.

Презентация защищена на комиссии кафедры 806. Работа выполнена в полном объёме. Рекомендую на оценку « ». Все материалы сданы на кафедру.

ПРОТОКОЛ

ЗАЩИТЫ ТЕХНИЧЕСКОГО ОТЧЁТА

по *производственной практике*

студентами: Батяновский Иван Тарасович

Слушали:	**Постановили:**
Отчёт практиканта	считать практику выполненной и защищённой на Общая оценка: _____

Руководители: Зайцев В. Е. _____

Кухтичев А. А. _____

Дата: 12 июля 2020 г.

МАТЕРИАЛЫ ПО РАЦИОНАЛИЗАТОРСКИМ ПРЕДЛОЖЕНИЯМ

В коде кролера стоит проверить необходимость конструкций try-except-finally: есть потенциально опасные места, где кролер может сломаться, и наоборот, где эта проверка замедляет работу кролера. Возможно, стоит поискать другие решения по отсеиванию уже пройденных страниц(есть вероятность, что в архиве страницы не повторяются и тогда роль pages = set() становиться сомнительной).

База данных состоит из новостей за последние 30 дней. Для более точных показателей модели, стоит запустить кролера на большее число дней. И стоит доработать новости с видеозаписями(проблема заключается в том, что tag с текстом новости другой в подобных страницах и это приводит к вылету программы с ошибкой).

Хотя вероятность попадания на страницы, которые запрещены в https://lenta.ru/robots.txt мала, лучше напрямую запретить эти страницы парсить(User-agent: GoogleBot, Disallow: /search, Disallow: /check_ed и тд).

Так как на написание модели оставалось мало времени не удалось должным образом продумать ее параметры. Из-за того же недостатка было получено мало данных для обучения. Стоит проверить результаты модели на разных метриках и использование других оптимизаторов. Увеличение или уменьшение слоев и изменение их параметров(количество нейронов, функции активациии тд). Проверить результаты на разных алгоритмах.

ТЕХНИЧЕСКИЙ ОТЧЁТ ПО ПРАКТИКЕ

Архитектура

```
from keras.models import Sequential
from keras import layers

embedding_dim = 50

model2 = Sequential()
model2.add(layers.Embedding(input_dim=vocab_size,
                            output_dim=embedding_dim,
                            input_length=maxlen))
model2.add(layers.GlobalMaxPool1D())
model2.add(layers.Dense(256, activation='relu'))
model2.add(layers.Dense(11, activation='softmax'))
model2.compile(optimizer='adam',
               loss='categorical_crossentropy',
               metrics=['accuracy'])
model2.summary()
```

Описание

Последовательность запускаемых файлов и их смысл:

1. lentaScrapper.py

   Эта программа запускает кролера, который просматривает определенное количество страниц в архиве lenta.ru с разных рубрик. В целом никаких параметров изменять не надо, кроме одного - limit. limit - количество дней, которые нужно просмотреть в ленте начиная с текущего.

   Что насчет описания работы самого кролера, то страница архива ленты представляет собой(имеется ввиду где нужная информация находится) 3 блока span4, в которой назодятся гиперссылки на статья.

   Для получения ссылок на архив разных рубрик программа парсит эти ссылки с главной страницы. Конкатенируя ссылки на рубрики и текущую дату, получаем ссылку на страницу архива, через которую сканируем ссылки на статьи и проходим по архиву через кнопку <, которая хранит ссылку на предыдущий день. Поэтому в теории можно пройтись по всем статьям сайта, но моей главной задачей стояла прежде всего получить опыт в написании кролера, а не полный сбор данных с сайта(кроме того это замет слишком много времени).

   Для работы Selenium нужно для своего браузера гугл установить соот версию chromedriver.exe

2. Save-.py

   Эта программа переводит данные с MySQL сервера в файл формата .csv, с которым придется в дальнейшем работать при обучении модели.

3. output_with_sport.csv

   Итоговая таблица данных(id, title, genre, content)

   - genre - это название рубрики. Это название стоило бы поменять на rubrics(genre - это унаследованное название с кролера, который я написал сначала для Meduza, однако я

решил переделать его для Lenta, т.к. многие статьи в Meduza выложены без конкретного описания тематики)

4. classText.ipynb

   Ноутбук где последовательно обрабатывается текст, пишется модель, обучается и записаны результаты.

Реализация

```
def crawler(url="", rubric="NULL", date=date.today(),
depth=0, limit=3, main=False):
    global pages
    if main is True:
        newUrl = "https://lenta.ru" + url + str(date)[:4] + "/" + str(date)[5:7]
    else:
        newUrl = "https://lenta.ru" + url
    driver.get(newUrl)
    print(newUrl, rubric)
    try:
        # Waiting for block of news to appear
        element = WebDriverWait(driver, 10) \
            .until(EC.presence_of_element_located
            ((By.CLASS_NAME, "b-layout.js-layout.b-layout_archive")))
    except:
        print("Can't locate news block")
        return
    finally:
        bsObj = BeautifulSoup(driver.page_source, "html.parser")

    span4s = bsObj.find("section",
    {"class": "b-layout js-layout b-layout_archive"})
    .findAll("div", {"class": "span4"})
    for span4 in span4s:
        items = span4.findAll("div",
        {"class": "item news b-tabloid__topic_news"})
        for item in items:
            if items is not None:
                newLink = item.find("a", {"href": re.compile("^(\/.*\/).*")})
                .attrs["href"]
                if newLink not in pages:
                    pages.add(newLink)
                    parser(newLink, rubric)
    if depth < limit:
        print("Push some buttons sometimes(dep, lim):", depth + 1, limit)
        crawler(url=bsObj.find("a", {"class": "control_mini"})
        .attrs["href"], rubric=rubric, depth=depth + 1, limit=limit, date=date)
```

Тестирование

text = "В наступившем году должны состояться первые с июля 2011о пилотируемые полеты США к МКС на собственных космических кораблях (до этого США отправляли своих астронавтов на околоземную орбиту при помощи многоразовых космических кораблей Space Shuttle). Скорее всего, первым из них в первом полугодии стартует Crew Dragon компании SpaceX, в декабре 2019-го успешно завершивший испытания парашютной системы, к которой ранее у НАСА были претензии, а до этого, в марте того же года, выполнивший первый (в беспилотном режиме) полет к МКС."

```
re = tokenizer.texts_to_sequences([text])
resu = pad_sequences(re, padding='post', maxlen=maxlen)
otvet = model2.predict(resu)

print(np.where(otvet == np.amax(otvet))[1])
print(maper)
```

```
[6]
{"'russia'": 1, "'world'": 2, "'ussr'": 3, "'economics'": 4,
"'forces'": 5, "'science'": 6, "'culture'": 7, "'sport'": 8,
"'media'": 9, "'style'": 10}
```

В данном примере я взял новость из категории наука https://lenta.ru/articles/2020/01/08/2020/. Ответ [6] означает, что данная новость относится к категории наука, что верно.

Ссылка на GitHub

https://github.com/Ivan-Batyanovsky/Practice2020