УДК 811.581 + 81’33

АВТОМАТИЗАЦИЯ ВЫБОРКИ ТЕКСТОВ НА ЗАДАННУЮ ТЕМУ ИЗ КИТАЙСКОЙ СЕТЕВОЙ ЭНЦИКЛОПЕДИИ БАЙДУ МЕТОДАМИ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Б.Г. Фаткулин

Выборка текстов на заданную тематику является одной из основных подготовительных задач в исследовательских проектах в рамках сравнительно-исторического, типологического и сопоставительного языкознания. Ручная выборка текстов требует больших затрат по времени и не всегда отличается эффективностью. Работа с текстами на восточных языках отличается особой трудоемкостью. После того как появилась возможность оцифровки больших корпусов текстов, возникло множество прикладных утилит, использующих достижения современной информатики. Процедурное знание по применению этих утилит повышает эффективность работы исследователя. В статье содержатся процедурные знания по использованию утилиты baidubaike, основанной на возможностях языка Python и предназначенной для автоматизированной выборки текстовых коллекций по китайскому языку из сетевой энциклопедии Baidu.

Ключевые слова: выборка текстов, процедурное знание, прикладные утилиты, китайский язык, язык Python, исполняемый скрипт, Github, энциклопедия Байду, автоматизированная загрузка.

Электронные энциклопедии содержат большие структурированные массивы данных или текстовые коллекции. Развитие и поддержка электронных энциклопедий на национальных языках является приоритетным направлением в языковой политике любого государства. Китай в данном случае не является исключением. Данные энциклопедий используются лингвистами-текстологами для сбора информации по актуальной тематике. Поиск информации в иноязычной энциклопедии представляет собой довольно трудоемкое занятие. Зная заранее искомое слово или понятие, либо интертет-адреса страниц, мы можем автоматизировать этот процесс.

В основу характеристики предлагаемого нами решения мы ставим классификацию видов знания, данную Е.А. Буденковой. Согласно Е.А. Буденковой [1]:

«А. **Фактологические знание** включает в себя знание терминологии, а также специфических деталей и элементов информации, т.е. то, что учащемуся необходимо знать для введения в дисциплину или решения общих проблем в рамках данной дисциплины.

В. **Концептуальное знание** подразумевает под собой знание взаимосвязей, существующих между базовыми элементами структуры, которые

позволяют им совместно функционировать, те. владение знанием о классификациях и категориях; общих принципах и правилах (теориях, моделях и структурах.

С. **Процедурное знание** предполагает знание предметно-ориентированных навыков и алгоритмов [2], методов, техник и критериев, определяющих отбор соответствующих процедур для эффективного функционирования».

Умение пользоваться инструментарием прикладной лингвистики входит в категорию процедурного знания. Одним из методов прикладной лингвистики является автоматизированный поиск искомого слова в больших массивах структурированной информации. Для обработки этих массивов могут быть использованы различные процедуры и инструменты [3, 4].

В данной статье речь пойдет об утилите baidubaike, созданной в рамках языка Python. Все желающие легко могут найти эту утилиту на он-лайн репозитории Github. Ценным свойством утитилиты baidubaike является то, что она является программным обеспечением под лицензией GNU GPL. Лицензия GPL предоставляет получателям компьютерных программ следующие права:

- свободу изучения того, как программа работает, и ее модификации (предварительным условием для этого является доступ к исходному коду);
- свободу распространения копий как исходного, так и исполняемого кода;
- свободу улучшения программы, и выпуска улучшений в публичный доступ (предварительным условием для этого является доступ к исходному коду).

В общем случае распространитель программы, полученной на условиях GPL, либо программы, основанной на таковой, обязан предоставить получателю возможность получить соответствующий исходный код.

Порядок пользования утилитой baidubaike следующий:

- установить утилиту;
- создать файл, исполняемый для языка Python;
- включить в исполняемый файл программный скрипт, выставив необходимые настройки;
- в программный скрипт включить искомое слово либо адрес страницы Baidu;
- запустить файл в интерпретаторе языка Python (в нашем случае idle в ОС Linux Ubuntu);
- сохранить вывод команды в файл соответствующего формата.

Образцы скрипта содержат в себе следующие возможности:

1. Создать специальную страницу:

```
>>> from baidubaike import Page
>>> page = Page(' ')
```

2. Получить основную информацию о странице:
```
>>> info = page.get_info()
>>> print info['title'], info['url']
>>> print info.get('last_modify_time')
>>> print info.get('creator')
>>> print info.get('page_view')
```
3. Получить содержимое страницы и список ссылок на другие страницы:
```
>>> page.get_content()
>>> links = page.get_inurls()
>>> for word in links:
...     print word, links[word]
```
4. Получить список тэгов разметки на странице:
```
>>> page.get_tags()
```
5. Получить список ссылок на другие источники:
```
>>> ref = page.get_references()
>>> for r in ref:
...     print r['title']
...     print r['url']
```
6. Кроме того, вы можете загрузить страницу, зная ее адрес:
```
>>> page = Page('http://baike.baidu.com/view/105.htm')
```
7. Вы можете поменять кодировку страницы:
```
>>> page = Page('google', encoding='gbk')
```

После апробации работы утилиты мы получили следующую информацию (приводим только первые 25 строк):

- **俄**罗斯（欧亚大陆地跨欧、亚两洲的国家）
http://baike.baidu.com/subview/2403/14453555.htm;
- **俄**罗斯联邦http://baike.baidu.com/view/21945.htm;
- **俄**语http://baike.baidu.com/view/15862.htm;
- **欧**亚大陆http://baike.baidu.com/view/242367.htm;
- 库页岛http://baike.baidu.com/view/64559.htm;
- **南千**岛群岛http://baike.baidu.com/view/145948.htm;
- 东斯拉夫人http://baike.baidu.com/view/298349.htm;
- **伊凡三世**http://baike.baidu.com/view/201475.htm;
- **莫斯科大公国**http://baike.baidu.com/view/102186.htm;
- **伊凡四世**http://baike.baidu.com/view/59679.htm;
- **彼得大帝**http://baike.baidu.com/subview/14000/5110325.htm;
- **俄**罗斯帝国http://baike.baidu.com/view/97401.htm;
- **冷**战http://baike.baidu.com/subview/11198/11105329.htm.

Приведенный нами список представляет собой онтологию разделов по теме «Россия». Мы успешно использовали данную онтологию при проведении занятий по курсу «Страноведение» для студентов, изучающих китайский язык, а также составления глоссария по россиеведческой терминологии, используемой в КНР.

Библиографический список

1. Буденкова, Е.А. Управление результатами обучения в условиях реализации компетентностного подхода в системе ВПО / Е.А. Буденкова // Образовательные технологии. – 2014. – № 3. – С. 47–58.

2. Трифонов, А.А. Алгоритмы построения инвертированного индекса для коллекции текстовых данных / А.А. Трифонов // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2013. – Т. 3 (27). – С. 52–60.

3. Фаткулин, Б.Г. Извлечение имен собственных из текстов (named entity recognition) по тематике ШОС на китайском языке при помощи простейших скриптов UNIX / Б.Г. Фаткулин // Наука ЮУрГУ. Материалы 66-й научной конференции. – Челябинск: Издательский центр ЮУрГУ, 2014. – С. 1339–1342.

4. Фаткулин, Б.Г. Россия и Китай: Перспективы сотрудничества в области терминоведения и прикладной лингвистики (методы извлечения терминологии иранистики в китайском языке) / Б.Г. Фаткулин // Россия и Китай: история и перспективы сотрудничества. Материалы IV международной научно-практической конференции / под ред. Д.В. Кузнецов; отв. ред. Д.В. Буяров. – Благовещенск, 2014. – С. 421–424.