МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА

Факультет информатики и систем управления

Кафедра теоретической информатики и компьютерных технологий

Наброски дипломного проекта

«Автоматическое установление связей между сообщениями твиттера и новостными статьями»

Выполнил:
студент ИУ9-101
Выборнов А. И.

Руководитель:
Лукашевич Н.В.

Москва 2016

1. Руководство пользователя

Проект состоит из двух пакетов, каждый из которых устанавливается и используется в отдельности.

- twnews_consumer — консьюмер, который позволяет выкачивать твиты с твиттера и новости с rss каналов.

- twnews — пакет, позволяющий по твитам и новостям, произвести все необходимые преобразования данных и на основе полученных признаков произвести обучение и оценку модели.

Оба пакета ориентированы на работу в операционных системах из семейства linux. Для начала работы необходимо иметь установленный менеджер пакетов для языка Python — pip, а также установить setuptools:

```
$ pip install setuptools
```

Также необходимо выкачать git-репозиторий: https://github.com/art-vybor/twnews.git. Если установлен пакет git, то это можно сделать следующим образом:

```
$ git clone https://github.com/art-vybor/twnews.git
```

Для корректной работы пакетов, все указываемые в конфигурации директории должны быть заранее созданы.

1.1. Пакет twnews_consumer

Пакет twnews_consumer располагается в папке consumer в корне репозитория. Он позволяет выкачивать и сохранять в формате, удобном для дальнейшей работы пакета twnews, твиты и новости.

Конфигурирование пакета производится в файле twnews_consumer/defaults.py. Описание задаваемых параметров находится в таблице 1.

Результатом работы пакета является множество новостей и твитов, выкаченных за время работы программы. Для новостей сохраняются заголовок, краткое описание, ссылка на новость, время публикации и имя

ресурса, на котором новость была опубликована. Для твитов сохраняются текст, время публикации и информация о том, является ли он ретвитом (ретвит — твит, представляющий собой, ссылку на ранее созданный твит).

1.1.1. Установка

Для установки, необходимо зайти в папку consumer, находящуюся в корне репозитория и выполнить команду:

```
$ make install
```

Во время установки, нужно будет ввести пароль, для распаковки секретного ключа, который необходим для работы с API твиттера.

1.1.2. Использование

Для того, чтобы начать выкачивать новости, необходимо запустить команду:

```
$ twnews_consumer download --news
```

Для того, чтобы начать выкачивать сообщения твиттера, необходимо запустить команду:

```
$ twnews_consumer download --tweets
```

Узнать информацию о работе программы можно из файла лога. Пример:

```
$ tail -f  /var/log/twnews_consumer.log
2016-04-05 11:37:14: RSS> Start consume rss feeds
2016-04-05 11:37:17: TWITTER> Starting write to /mnt/yandex.
   disk/twnews_data/logs/tweets.shelve
2016-04-05 11:37:17: TWITTER> Starting to consume twitter
2016-04-05 12:33:32: TWITTER> ('Connection broken:
   IncompleteRead(0 bytes read, 512 more expected)',
   IncompleteRead(0 bytes read, 512 more expected))
```

1.2. Пакет twnews

Пакет twnews располагается в папке core в корне репозитория. Он позволяет обрабатывать данные полученные с помощью консьюмера с целью построения и оценки качества модели WTMF.

Конфигурирование пакета производится в файле twnews/defaults.py. Описание задаваемых параметров находится в таблице 2.

Результатом работы пакета является построенная модель WTMF, для которой измерено её качество. Перед построением модели, необходимо выполнить команду, которая разрешает ссылки (может занять длительное время).

1.2.1. Установка

Для установки, необходимо зайти в папку core, находящуюся в корне репозитория и выполнить команду:

```
$ make install
```

Для повышения производительности рекомендуется вручную собрать пакет numpy с использованием математической библиотеки OpenBLAS [9].

1.2.2. Использование

Для того, чтобы разрешить ссылки, необходимо запустить команду:

```
$ twnews_consumer --resolve
```

Для того, чтобы посмотреть статистику по упомянутым в коллекции твитов ссылкам нужно выполнить команду:

```
$ twnews_consumer download --analyse_urls
```

Для построения модели небходимо запустить команду:

```
$ twnews_consumer download --run_pipe
```

Узнать информацию о работе программы можно из файла лога. Пример:

```
$ tail -f /var/log/twnews.log
INFO:root:2016-04-08 10:20:08.256411: News successfully
    loaded
INFO:root:2016-04-08 10:20:23.006948: Function iteration
    started with time measure
INFO:root:2016-04-08 10:33:32.930520: Function iteration
    finished in 13m9.9234058857s
INFO:root:2016-04-08 10:33:32.940666: Function
    find_topk_sim_news_to_tweets started with time measure
INFO:root:2016-04-08 10:34:42.360326: Function
    find_topk_sim_news_to_tweets finished in 1m9.41950583458s
INFO:root:2016-04-08 10:34:42.587674: Function iteration
    started with time measure
INFO:root:2016-04-08 10:48:04.453983: Function iteration
    finished in 13m21.8661620617s
INFO:root:2016-04-08 10:48:04.466846: Function
    find_topk_sim_news_to_tweets started with time measure
INFO:root:2016-04-08 10:49:18.958096: Function
    find_topk_sim_news_to_tweets finished in 1m14.4910538197s
INFO:root:2016-04-08 10:49:19.171160: Function iteration
    started with time measure
```

Таблица 1: Описание конфигурации пакета twnews_consumer

Имя параметра	**Пример значения**	**Описание**
LOG_FILE	’/var/log/twnews_consumer.log’	Путь до файла с логом
LOG_LEVEL	logging.INFO	Уровень подробности лога
TWNEWS_DATA_PATH	’/home/avybornov/twnews_data/’	Путь до директории, в которую будут сохранены данные
RSS_FEEDS	{’ria’: {’rss_url’: ’http://ria.ru/export/rss2/index.xml’}, ’lifenews’: {’rss_url’: ’http://lifenews.ru/xml/feed.xml’}}	Новостные источники, которые требуется выкачать
TWEETS_LANGUAGES	[’ru’]	Список языков, твиты с использованием которых выкачиваются из твиттера

Таблица 2: Описание конфигурации пакета twnews

Имя параметра	**Пример значения**	**Описание**
LOG_FILE	’/var/log/twnews.log’	Путь до файла с логом
LOG_LEVEL	logging.INFO	Уровень подробности лога
TWNEWS_DATA_PATH	’/home/avybornov/twnews_data/’	Путь до рабочей директории в которой лежат выкаченные с помощью консьюмера данные
DATASET_FRACTION	1.0	Часть датасета, которая будет использована для обучения модели
TMP_FILE_DIRECTORY	’/tmp/twnews/’	Путь до директории в которую будут сохранены временные данные

Список литературы

[1] W. Guo, H. Li, H. Ji, and M. T. Diab. Linking tweets to news: A framework to enrich short text data in social media. - ACL, pages 239–249, 2013.

[2] Manos Tsagkias, Maarten de Rijke, Wouter Weerkamp. Linking Online News and Social Media. - ISLA, University of Amsterdam.

[3] T. Hoang-Vu, A. Bessa, L. Barbosa and J. Freire. Bridging Vocabularies to Link Tweets and News. - International Workshop on the Web and Databases (WebDB 2014), Snowbird, Utah, US, 2014.

[4] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, J. Sperling. TwitterStand: news in tweets. - 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2009, Seattle, Washington.

[5] Ou Jin, Nathan N. Liu, Kai Zhao, Yong Yu, and Qiang Yang. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In Proceedings of the 20th ACM international conference on Information and knowledge management.

[6] Weiwei Guo and Mona Diab. 2012a. Modeling sentences in the latent space. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.

[7] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[8] Nathan Srebro and Tommi Jaakkola. 2003. Weighted low-rank approximations. In Proceedings of the Twentieth International Conference on Machine Learning.

[9] Eric Huns. Hunseblog on Wordpress: URL: https://hunseblog.wordpress.com/2014/09/15/installing-numpy-and-openblas/.
