|  |  |  |
| --- | --- | --- |
| Казахский национальный университет им. аль-Фараби |  | Әл-Фараби атындағы Қазақ ұлттық университеті |

**Факультет информационных технологий**

|  |  |
| --- | --- |
| **“Согласовано”** | **“Утверждено”** |
| Зав. кафедрой «Информационные  Системы»  Есенгалиева Ж.С.\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | И.о. декана факультета информационных технологий  Тукеев У.А.\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  |  |
| «\_\_\_\_\_\_\_”\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г. | “\_\_\_\_\_\_\_”\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г. |
|  | (печать) |

ОТЧЕТ ПО НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ

РАБОТЕ МАГИСТРАНТА

ЗА ВЕСЕННИЙ СЕМЕСТР 2017-2018 УЧЕБНОГО ГОДА

|  |  |
| --- | --- |
| ФИО магистранта | Шалабаев Қазыбек Мадахметұлы |
| Специальность | 6M100200 – Системы информационной безопасности |
| Срок обучения | 2017-2019 года |
| Научный руководитель | к.ф-м.н. Мусиралиева Шынар Женисбековна |
| Тема диссертации | Контент анализ групп в социальных сетях |

**Алматы, 2018**

1. **Научно-исследовательская работа по теме диссертации.**

**Тема магистерской диссертации:** Контент анализ групп в социальных сетях.

За последние несколько лет появилось значительное количество социальных сетей (Facebook, Twitter, Linkedin, Google+, ВКонтакте, Одноклассники и проч.), Аудитория некоторых из них насчитывает более 1 миллиарда активных пользователей. В данных сетях люди имеют возможность находить своих одноклассников, родственников, знакомых, заводить новых друзей по интересам, объединяться в группы, просматривать видео, фото, слушать музыку, играть в игры и т.д.. Данный факт определяет важность создания систем которые будут быстро находить контент, интересует пользователей ориентируясь на их интересы, например для популяризации своих идей, рекламы продуктов, услуг или созданных в тех же социальных сетях сообществ, донесения объявлений.

В последнее время объединение в группы в социальных сетях приобретает все большую и большую популярность. Количество подписчиков часто измеряется сотнями тысяч и каждый из них может делать публикации в группе. Проверять тексты на популяризацию идей (например политических или террористических) становится все труднее, поэтому возникает необходимость в определенной степени автоматизировать этот процесс и оперативно находить такие публикации.

Исходя из этих предпосылок целью данной работы является анализ, исследования и разработка методов обработки данных небольшой части социальных сетей - групп по интересам (сообществ) чтобы иметь возможность определить самые интересные записи и их тему для аудитории, а поскольку функциональность системы будет направлена на анализ, исследования и разработка методов обработки данных небольшой части социальных сетей - групп по интересам чтобы иметь возможность определить тексты на популяризацию идей (например политических или террористических).

**Анализ статистических показателей**

Анализ статистических показателей социальной сети заключается в том, что выбираются записи (фотографии, видео, аудиозаписи, текстовые публикаций) и сортируются в порядке их популярности. Популярность определяется по количественным показателям меток "понравилось", "рассказать" и комментариях к записям. Наиболее популярные записи анализируются в первую очередь изза того что их увидит (так как каждый пользователь при нажатии на кнопку «рассказать» дает согласие на размещение текста публикации, видеозаписи, аудиозаписи или фото на своей странице и они автоматически там появляются, и при определенных настройках профиля отображаются в списке "новости" в друзей пользователя) больше людей и количество публикаций может измеряться десятками и даже сотнями тысяч, это будет нелегкая задача для компьютеров. Также анализ статистических (таких как количество подписчиков (основную аудиторию можно разделить по территории и возраста), фотографий, видео и аудио-записей и т.д.) показателей можно применять для обработки списка групп, будто в первую очередь исследоваться, опять же это обусловлено тем, что это все увидит большее количество людей.

**Алгоритм получения показателей по социальной сети**

Для того чтобы использовать в приложении который разрабатывается API функции социальной сети сначала нужно авторизоваться.

Тип программного продукта разрабатываемой - Standalone. Для доступа к API ВКонтакте с любого Standalone-приложения предусмотрен механизм клиентской авторизации на базе протокола OAuth 2.0. В качестве клиента может выступать любой Desktop / мобильное приложение, которое имеет возможность управления Web-браузером (например, компонент UIWebView для iOS).

Процесс авторизации приложения состоит из 3-х этапов:

1. Открытие окна браузера для аутентификации пользователя на сайте Вконтакте.

2. Разрешение пользователем доступа к своим данным.

3. Передача в программу ключа access\_token для доступа к API.

Для создания Standalone-приложение переходим на страницу [создания приложения](https://vk.com/editapp?act=create). Вбиваем название приложения, выбираем платформу «Standalone-приложение» и нажимаем кнопку «Подключить приложение». Окно создания Standalone-приложение можно увидеть на рисунке 1.

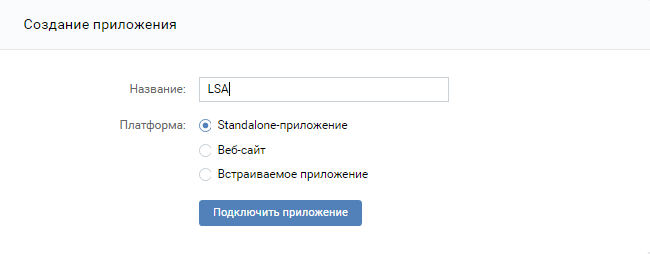


Рисунок – 1.Окно создания Standalone-приложение

При клиентской авторизации ключ доступа к API access\_token кажется приложении без необходимости раскрытия секретного ключа (выдается при регистрации приложения в социальной сети). Поэтому не рекомендуется указывать свой секретный ключ в середине кода программы, поскольку способом декомпиляции программы преступники могут получить его. Настройки приложение можно увидеть на рисунке 2.

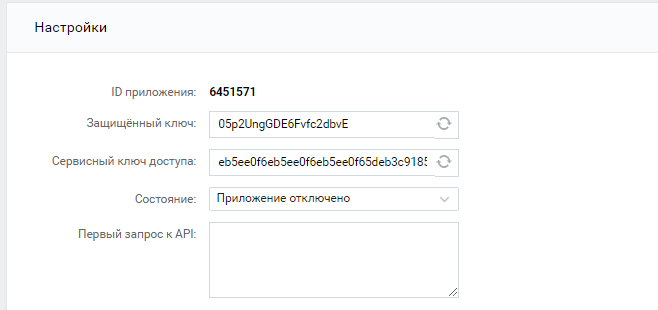


Рисунок – 2. Настройки приложение

Предоставление прав доступа

После успешного входа на сайт пользователю будет предложено авторизовать приложение, разрешив доступ к необходимым настройкам, востребованными вами с помощью параметра scope получение access\_token

После успешной авторизации приложении браузер пользователя будет перенаправлен по адресу REDIRECT\_URI, указанному при открытии диалога авторизации. При этом ключ доступа к API access\_token и другие параметры будут переданы в URL-фрагменте ссылки. Окно прав доступа можно увидеть на рисунке 3.

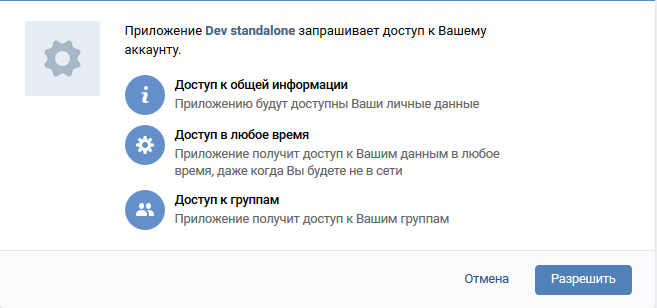


Рисунок – 3. Окно прав доступа

Вместе с ключом access\_token также будет указано время его жизни expires\_in, заданный в секундах. Если срок использования ключа закончился, то необходимо повторно провести все описанные выше шаги, но в этом случае пользователю уже не придется дважды разрешать доступ. Спрашивать access\_token также необходимо при изменении пользователем логина или пароля или удалением программы в настройках доступа. Кроме того, среди возвращенных параметров будет указан user\_id - идентификатор авторизованного пользователя в социальной сети.

В случае возникновения ошибки авторизации в качестве GET-параметров в REDIRECT\_URI будет передана информация об этой ошибке:

http: // REDIRECT\_URI? error = access\_denied & error\_description = The + user + or + authorization + server + denied + the + request.

После успешной авторизации можно осуществлять запросы к API.

Для того чтобы вызвать метод API ВКонтакте, необходимо осуществить POST или GET запрос по протоколу HTTPS на указанный URL:

https://api.vk.com/method/METHOD\_NAME?PARAMETERS& access\_token = ACCESS\_TOKEN

METHOD\_NAME - название метода из списка функций API,

PARAMETERS - параметры соответствующего метода API,

ACCESS\_TOKEN - ключ доступа, полученный в результате успешной авторизации приложения.

пример:

https://api.vk.com/method/getProfiles?uid=66748& access\_token = 533bacf01e11f55b536a565b57531ac114461ae8736d6506a3

В ответ на такой запрос будет получен ответ в формате JSON.

Для того, чтобы получить ответ в формате XML, необходимо добавить расширение формата после названия метода, METHOD\_NAME.xml

пример:

https://api.vk.com/method/getProfiles.xml?uid=66748&access\_token=533bacf01e11f55b536a565b57531ac114461ae8736d6506a

Передача параметров в API

Параметры могут передаваться как методом GET так и POST. Если передавать большие объемы данных и (более 2 килобайт) то следует использовать метод POST.

Каждый метод имеет собственный набор поддерживаемых параметров, однако существуют параметры, которые принимают все методы:

lang - определяет язык, на котором будут возвращаться различные данные, например названия стран и городов. Также если указанная не русский, украинский и белорусский языки, то имена будут автоматически транслируемые в латиницу. ru - Русский, ua - Украинский, be - Белорусский, en-Английский, es - Испанский, fi - Финская, de - Немецкий, it - Итальянский.

v - обозначает версию API, которая используется, например 4.95

Использование этого параметра применяет некоторые изменения в формате ответы различных методов.

https - 1 - позволяет получать https ссылки на фотографии и другие медиа. 0 - методы возвращают http ссылки (по умолчанию).

Для получения данных со стены публичных групп используется метод wall.get. Стоит заметить, что не все методы требуют access\_token, данный метод относится к таковым. Окно метода wall.get можно увидеть на рисунке 4.

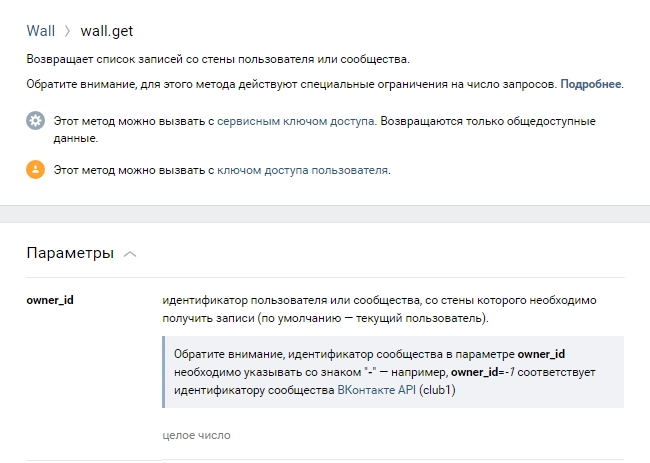


Рисунок – 4. Окно метода wall.get

Ниже на таблице 1указаны параметры метода wall.get.

Таблица -1.Параметры метода wall.get

|  |  |
| --- | --- |
| параметер | Описание |
| owner\_id | Идентификатор пользователя (по умолчанию - текущий пользователь). Чтобы получить записи со стены группы (публичной страницы, встречи), укажите ее идентификатор со знаком "минус": например, owner\_id = -1 соответствует группе с идентификатором 1. |
| domain | Краткая адрес пользователя или группы |
| offset | Смещение для выборки определенного множества сообщений |
| count | Количество сообщений, которое необходимо получить (но не более 100) |
| filter | Определяет, какие типы сообщений на стене необходимо получить. Возможны следующие значения параметра:  owner - сообщение на стене от ее владельца  others - сообщение на стенке не от ее владельца  all - все сообщения на стене  \* Если параметр не задан, то считается что он равен all |

*Результат*

После успешного выполнения возвращает общее количество записей и массив объектов post, каждый из которых содержит поля:

id - идентификатор записи на стене пользователя;

o\_id - идентификатор владельца записи;

from\_id - идентификатор пользователя, который создал запись;

date - время публикации записи в формате unixtime;

text - текст записи;

comments - информация о комментариях к записи, объект с полями:

- count - количество комментариев,

can\_post - информация о том, может ли текущий пользователь комментировать

likes - информация о брани дорожку объект с полями:

- count - число пользователей, которым понравилась запись,

user\_likes - наличие отметки «Мне нравится» от текущего пользователя (1 - есть, 0 - нет),

can\_like - информация о том, может ли текущий пользователь поставить отметку "Мне нравится" (1 - может, 0 - не может),

can\_publish - информация о том, может ли текущий пользователь сделать репост записи (1 - может, 0 - не может);

reposts - информация о репост записи ( "Рассказать друзьям"), объект с полями:

count - число пользователей, скопировали запись,

user\_reposted - наличие репоста от текущего пользователя (1 - есть, 0 - нет);

post\_source - информация о способе размещения записи (см. описание поля post\_source)

attachments - информация о вложении записи (фотографии, ссылки и т.п., см описание поля attachments)

geo - информация о местонахождении, содержит поля:

- place\_id - идентификатор места;

- title - название места;

- type - тип места;

- country\_id - идентификатор страны;

- city\_id - идентификатор города

- address - строка с указанием адреса места в городе;

signer\_id - идентификатор участника, если запись была опубликована от имени сообщества и подписано пользователем

copy\_owner\_id - идентификатор владельца стены, у которого была скопирована запись (если запись является копией записи с чужой стены)

copy\_post\_id - идентификатор скопированной записи на стене ее владельца (если запись является копией записи с чужой стены)

copy\_text - текст комментария, добавленного при копировании (если запись является копией записи с чужой стены).

Фрагмент ответа сервера в формате JSON:

"response": {

"count": 11572,

"items": [

{

"id": 28508,

"from\_id": -90689018,

"owner\_id": -90689018,

"date": 1525257849,

"marked\_as\_ads": 0,

"post\_type": "post",

"text": "Хазреті Әлидің дәрежесі \n \nБір күні Сахабалар хазреті Әлидің мәртебесін білу үшін Пайғамбарымыздан былай деп сұрады: \n- Ей Расулаллаһ! Хазреті Әлиге деген Біреуі Хазреті Әлиді шақыруға кетеді. Қалғандарынан пайғамбарымыз былай деп сұрайды:...",

"attachments": [

{

"type": "photo",

"photo": {

"id": 456247490,

"album\_id": -7,

"owner\_id": -90689018,

"user\_id": 100,

"photo\_75": "https://pp.userapi.com/c543105/v543105655/4c719/ddCJ-ZdjpfQ.jpg",

"photo\_130": "https://pp.userapi.com/c543105/v543105655/4c71a/vPnQ9WM9Nvw.jpg",

"photo\_604": "https://pp.userapi.com/c543105/v543105655/4c71b/I-VPP1P-jhg.jpg",

"width": 259,

"height": 194,

"date": 1525257849,

"post\_id": 28508,

"access\_key": "bc5f7b5415e09dae33"

}

}

],

"post\_source": {

"type": "api"

},

"comments": {

"count": 0,

"groups\_can\_post": true,

"can\_post": 0

},

"likes": {

"count": 13,

"user\_likes": 0,

"can\_like": 1,

"can\_publish": 1

},

"reposts": {

"count": 0,

"user\_reposted": 0

},

"views": {

"count": 274

}

},........}

Список литературы:

1. Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V., Varga, L.: The provenance of electronic data. Commun. ACM 51(4), 52–58 (2008)

2. Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S., Moreau, L.: An Architecture for Provenance Systems. Technical report, University of Southampton (2006)

3.Ukustov, S., S. Fomenkov and A. Polovinkin, 2013. of the Russian Fundamental Research Fund. Wrong metrics of success for open source software

4. Сайт Pearson Knowledge Technologies - http://www.k-a-t.com

5. Witten, L.H., Frank, E., Hall, M.A.: Data mining: practical machine learning tools and techniques, Amsterdam, p. 560 (2005)

6. Foltz, P. W. 1996. Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments and Computers. 28 197-202.

7. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering 13(2), 95–135 (2007)

8. The Python Website, [http://www.python.org](http://www.python.org/)

9. Дж. Голуб, Ч. Ван Лоун Матричные вычисления. - М., "Мир", 1999

10. Jackson, K.R.: PyGlobus: a Python interface to the Globus Toolkit. Concurrency and Computation: Practice and Experience 14(13-15), 1075–1083 (2002)

11. Qt: работа с Vkontakte API и Phonon - http://habrahabr.ru/post/115397/

12. Описание методов API - <https://vk.com/dev/methods>

13. Miles, S., Moreau, L., Groth, P., Tan, V., Munroe, S., Jiang, S.: Provenance Query Protocol. Technical report, University of Southampton (2006)

14. Groth, P., Tan, V., Munroe, S., Jiang, S., Miles, S., Moreau, L.: Process Documentation Recording Protocol. Technical report, University of Southampton (2006)

15. VKontakte: [electronic resource]-2013.-Mode of access: of people during the day depending on the day of the vk.com/dev/methods

16. VKontakte [electronic resource]-2013.-Mode of access: vk.com/dev/auth\_mobile

**II. Научные публикации за отчетный период.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Название**  **трудов** | **Рукопись или печатные** | **Наименование**  **издания и его выходные данные; наименование журнала (год, №),**  **№ авторского свидетельства** | **Кол-во печатных листов или стр.** | **Фамилии**  **соавторов**  **работы** | **Согласно индивидуальному учебному плану** |
|  |  |  |  |  |  |  |

**III. Участие в международных, республиканских конференциях, симпозиумах, «круглых столах», научных семинарах.**

Практическая конференция по актуальным проблемам информационной безопасности «Kaz’Hack’Stan», 27 октября 2017 года,Алматы.Семинар Масаловича А.И. «Конкурентная разведка в информационных технологиях»,28 октября 2017 года,Алматы.

**IV. Дополнительная отчетная информация** (участие в научных мероприятиях вне индивидуального плана магистранта; референтская научная работа или научно-переводческая работа в НИИ; выполнение смежных с тематикой диссертации научных исследований в НИИ и т.д.)**.**

Научно-исследовательская работа магистранта оценивается в \_\_\_\_баллов.

**Научный руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Мусиралиева Ш.Ж.**

**(подпись)**

**Эдвайзер \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Тукеев У.**

**(подпись)**

**Магистрант \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Шалабаев Қ.М.**

**(подпись)**