МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №16
по курсу «Информационный поиск»

«Автоматическое аннотирование»

Выполнил:
студент группы ИУ9-21М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2019

1 Цель работы

Аннотировать статьи 2 способами.

1.1 Ход работы

```
import re

import nltk
import pymorphy2

from collections import defaultdict


morph = pymorphy2.MorphAnalyzer()
nltk.download('punkt')  # required to split text into sentences

ARTICLES = ['cybersport.txt', 'fb.txt', 'telegram.txt']

SENTENCES_TO_OUTPUT = 4


class Sentence:
    def __init__(self, sentence: str):
        self.sent = sentence
        self.norm = normalize_sentence(sentence)
        self.words = self.norm.split(' ')

    def count_overall_frequency(self, wordcount: dict) -> int:
        freq = 0
        for w in self.words:
            freq += wordcount[w]
        return freq

    def count_avg_frequency(self, wordcount: dict) -> int:
        return self.count_overall_frequency(wordcount) //
         ↪  len(self.words)

    def __repr__(self):
        return self.__str__()

    def __str__(self):
        return self.sent


class Article:
```

```
40      def __init__(self, sentences: list):
41          self.sents = sentences
42          self.annt_overall = []
43          self.annt_average = []
44          self.wordcount = defaultdict(int)
45          self._count_word_freq()
46
47      def annotate(self):
48          self._annotate_average()
49          self._annotate_overall()
50
51      def _annotate_overall(self):
52          self.sents.sort(key=lambda sent:
            ↪ sent.count_overall_frequency(self.wordcount), reverse=True)
53          self.annt_overall = self.sents[:SENTENCES_TO_OUTPUT]
54
55      def _annotate_average(self):
56          self.sents.sort(key=lambda sent:
            ↪ sent.count_avg_frequency(self.wordcount), reverse=True)
57          self.annt_average = self.sents[:SENTENCES_TO_OUTPUT]
58
59      def _count_word_freq(self):
60          for s in self.sents:
61              words = s.norm.split(' ')
62              for w in words:
63                  self.wordcount[w] += 1
64
65
66  def normalize_sentence(sentence: str) -> str:
67      tags_to_remove = ['NPRO', 'PRED', 'PREP', 'CONJ', 'PRCL', 'INTJ']
68      normalized = []
69      for w in sentence.split(' '):
70          parsed = morph.parse(w)[0]
71          if (parsed.tag.POS not in tags_to_remove) and 3 <=
            ↪ len(parsed.normal_form):
72              normalized.append(parsed.normal_form)
73      return ' '.join(normalized)
74
75
76  def clean_up_sentence(s: str) -> str:
77      no_punct = re.sub(r'[^а-яё]', ' ', s.casefold())
78      no_duplicate_spaces = re.sub(r'\s+', ' ', no_punct)
79      return no_duplicate_spaces.strip()
80
81
```

```
def read_sentences(filename: str) -> list:
    text = open(filename, 'r', encoding='UTF-8').read()
    preprocessed = re.sub(r'[,:;-]', ' ', text)
    sentences = nltk.sent_tokenize(preprocessed)
    clean_sents = list(map(lambda s: clean_up_sentence(s), sentences))
    return list(filter(lambda s: len(s.split(' ')) >= 3, clean_sents))


def main():
    articles = []
    for article_filename in ARTICLES:
        sentences = read_sentences(article_filename)
        sentences = [Sentence(s) for s in sentences]
        articles.append(Article(sentences))

    [a.annotate() for a in articles]

    for a in articles:
        print('Overall')
        [print(f'\t{s}') for s in a.annt_overall]
        print('Average')
        [print(f'\t{s}') for s in a.annt_average]


if __name__ == '__main__':
    main()
```

2 Результаты

Статья 1:

Играть днями напролет и зарабатывать в разы больше, чем офисный клерк: этот сценарий стал реальным благодаря развитию стриминга — прямых трансляций онлайн-игр Доход 27-летнего американца Ninja достигает $500 тыс. в месяц, топовые российские стримеры в месяц пока зарабатывают 1–2 млн руб., но уже собирают тысячи зрителей в прямом эфире. Мир стримеров очень закрытый — они неохотно отвечают на предложения о сотрудничестве, да и рекламодатели часто побаиваются работать с 20-летними игроманами. Журнал РБК поговорил с тремя популярными стримерами и узнал, почему они не успевают тратить заработанное, как поднять-ся на стримах, если ты девушка, и почему Twitch победит YouTube. Для встречи с самым известным в России игроком в PUBG (PlayerUnknown's Battlegrounds) корре-спондент журнала РБК прилетел в город Ухту в республике Коми. В два часа дня здесь темно — полярная ночь. Сервис TripAdvisor здешним развлечением № 1 назы-вает вечный огонь, а ухтинцы гордятся тем, что в местном институте учился Роман Абрамович. «Я живу один, ходить особенно некуда, так что онлайн-игры для меня

— лучший досуг», — признается Дмитрий Пустоваров, больше известный под ником MakataO. 26-летний Пустоваров получил образование инженера оборудования нефтяного и газового промыслов, отслужил в ракетных войсках и устроился работать в котельную на шахте в соседнем с Ухтой поселке Ярега. По выходным подрабатывал диджеем на свадьбах. Компьютерными играми Дмитрий увлекся еще в школе, а в институте стал проводить за компьютером все свободное время. Играл в тактический шутер Battlefield и даже вошел в состав полупрофессиональной команды. Первые накопленные с зарплаты 100 тыс. руб. Пустоваров потратил на компьютер, который тянул онлайн-игры в высоком качестве. В 2016-м взял ник MakataO и решился постримить сам на платформе Twitch. Качество видео и звука было ужасным, веб-камеры не было, и за первые несколько дней на трансляцию не зашел ни один человек. Он не сдавался и каждый день выходил в эфир, разработав особую тактику: «Вечерние часы, когда зрители приходят с работы и учебы, заняты крупными стримерами, конкурировать с ними бесполезно. Поэтому я выбирал непопулярное время: утро, когда подключался Дальний Восток». После ночной смены в котельной ехал домой, завтракал и около девяти утра садился за компьютер на 8–10 часов.

Overall:

играть днями напролет и зарабатывать в разы больше чем офисный клерк этот сценарий стал реальным благодаря развитию стриминга прямых трансляций онлайн игр доход летнего американца достигает тыс

я живу один ходить особенно некуда так что онлайн игры для меня лучший досуг признается дмитрий пустоваров больше известный под ником

летний пустоваров получил образование инженера оборудования нефтяного и газового промыслов отслужил в ракетных войсках и устроился работать в котельную на шахте в соседнем с ухтой поселке ярега

качество видео и звука было ужасным веб камеры не было и за первые несколько дней на трансляцию не зашел ни один человек

Average:

я живу один ходить особенно некуда так что онлайн игры для меня лучший досуг признается дмитрий пустоваров больше известный под ником

первые накопленные с зарплаты тыс

пустоваров потратил на компьютер который тянул онлайн игры в высоком качестве

играть днями напролет и зарабатывать в разы больше чем офисный клерк этот сценарий стал реальным благодаря развитию стриминга прямых трансляций онлайн игр доход летнего американца достигает тыс

Статья 2:

Использование Facebook штабом Трампа, вирусное приложение GetContact — скандалов со скрытым сбором пользовательской информации все больше. Журнал РБК разобрался, как устроен бурно растущий рынок персональных данных в России политической рекламы. В конце февраля россияне увлеклись приложением GetContact — сервисом для проверки незнакомых телефонных номеров. Чтобы получить доступ к услуге, нужно разрешить доступ к своим контактам. Приложение быстро превратилось в сетевое развлечение — посмотреть, под каким именем ты записан в телефонах друзей и знакомых, потом выложить забавный скриншот со своими именами в

соцсетях. К середине марта турецкая Teknasyon, создатель GetContact и партнер сотового оператора Turkcell, собрала по всему миру более 3,5 млрд номеров с именами владельцев, указано на сайте приложения. Согласно пользовательскому соглашению эти данные разработчики могли использовать в маркетинговых целях или передавать третьим лицам (1 марта из документа исключили такую возможность).

Overall:

приложение быстро превратилось в сетевое развлечение посмотреть под каким именем ты записан в телефонах друзей и знакомых потом выложить забавный скриншот со своими именами в соцсетях

к середине марта турецкая создатель и партнер сотового оператора собрала по всему миру более млрд номеров с именами владельцев указано на сайте приложения

согласно пользовательскому соглашению эти данные разработчики могли использовать в маркетинговых целях или передавать третьим лицам марта из документа исключили такую возможность

использование штабом трампа вирусное приложение скандалов со скрытым сбором пользовательской информации все больше

Average:

использование штабом трампа вирусное приложение скандалов со скрытым сбором пользовательской информации все больше

журнал рбк разобрался как устроен бурно растущий рынок персональных данных в россии политической рекламы

в конце февраля россияне увлеклись приложением сервисом для проверки незнакомых телефонных номеров

чтобы получить доступ к услуге нужно разрешить доступ к своим контактам

Статья 3:

«Сталингулаг» стабильно занимает первое место в рейтинге самых популярных Telegram-каналов России: каждый пост в нем в среднем читают 300 тыс. человек. Журнал РБК разобрался, кто может быть автором этого анонимного проекта В конце 2008 года на старейшем российском форуме по интернет-маркетингу Searchengines.guru появилась запись от имени пользователя aelexandryo. Новичка интересовал вопрос, какую прибыль получит его блог, если вложить в продвижение 50 тыс. руб. Более опытные форумчане упражнялись в остроумии: «лучше пропить», «каникулы начались?», «потрать эти деньги на образование». Критика не остановила aelexandryo: через десять лет скрывавшийся за этим ником человек стал одним из самых популярных российских блогеров в мессенджере Telegram — площадке для сравнительно нового феномена под названием «анонимные каналы». Свое имя aelexandryo не раскрывает: 300 тыс. его подписчиков в Telegram знают его только как «Сталингулаг».

Сейчас проект «Сталингулаг» объединяет в себе аккаунты в трех социальных сетях. В Telegram публикуются небольшие посты публицистического характера, с едкой и подчас остроумной критикой российской действительности, в Twitter размещаются короткие комментарии той же направленности. Все публикации пишутся от первого лица, но личную информацию об авторе лучше искать в Instagram: здесь «Вождь и Учитель» — так, выдерживая стиль, называет себя автор «Сталингулаг», выкладывает фотографии из московских и лондонских ресторанов и театров и отчеты о полетах в бизнес-классе. Первый аккаунт «Сталингулаг» появился не в

Telegram, а в Twitter в июне 2013 года. Правда, это не был аккаунт нового пользователя: в @StalinGulag был переименован давно существовавший аккаунт @algorbunov, следует из твитов, написанных до весны 2013-го. Его создателем мог быть житель Махачкалы Александр Горбунов, которому сейчас 26 лет. Вести Twitter Горбунов стал с 2011 года, постоянно кидая фолловерам ссылки на свои посты в «Живом журнале» (ЖЖ) в блоге a1gorbunov. Блог не был анонимным: помимо имени и фамилии автор указал также дату рождения (данные WaybackMachine, сейчас страница удалена).

Twitter «Сталингулаг» первоначально был записан на ту же почту в сервисе Mail.Ru, на которую был зарегистрирован аккаунт матери Горбунова в «Одноклассниках» и его страница во «ВКонтакте», следует из сервисов восстановления пароля. Как блог в ЖЖ, страница Горбунова «ВКонтакте» сейчас удалена. Наконец, ту же почту использовали при регистрации сайта по личностному росту Moneymlm.ru: кроме нее в выходных данных были указаны полные имя и фамилия Горбунова, а также первая буква отчества и номер телефона с дагестанским кодом, указано в сервисе WhoIs. Сейчас этот номер недоступен. Горбунов не ответил на вопросы журнала РБК, отправленные ему на электронную почту.

Overall:

сталингулаг первоначально был записан на ту же почту в сервисе на которую был зарегистрирован аккаунт матери горбунова в одноклассниках и его страница во вконтакте следует из сервисов восстановления пароля

наконец ту же почту использовали при регистрации сайта по личностному росту кроме нее в выходных данных были указаны полные имя и фамилия горбунова а также первая буква отчества и номер телефона с дагестанским кодом указано в сервисе

журнал рбк разобрался кто может быть автором этого анонимного проекта в конце года на старейшем российском форуме по интернет маркетингу появилась запись от имени пользователя

блог не был анонимным помимо имени и фамилии автор указал также дату рождения данные сейчас страница удалена

Average:

первый аккаунт сталингулаг появился не в а в в июне года

его создателем мог быть житель махачкалы александр горбунов которому сейчас лет

блог не был анонимным помимо имени и фамилии автор указал также дату рождения данные сейчас страница удалена

сталингулаг первоначально был записан на ту же почту в сервисе на которую был зарегистрирован аккаунт матери горбунова в одноклассниках и его страница во вконтакте следует из сервисов восстановления пароля

3 Выводы

Статьи с усредненным алгоритмом аннотации выдали более качественные аннотации и «уловили суть».