



Министерство науки и высшего образования Российской Федерации  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Московский государственный технический университет  
имени Н.Э. Баумана  
(национальный исследовательский университет)»  
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ \_\_ Информатика и системы управления \_\_\_\_\_

КАФЕДРА \_\_ Системы обработки информации и управления \_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**  
**К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ**  
**ПО ОБРАБОТКЕ И АНАЛИЗУ ДАННЫХ**

**НА ТЕМУ:**

**Генерация твитов на основе LSTM сети**

Студент \_\_ ИУ5-34м \_\_  
(Группа)

\_\_\_\_\_  
(Подпись, дата)

**В.А. Макрушина**  
(И.О.Фамилия)

Руководитель

\_\_\_\_\_  
(Подпись, дата)

\_\_\_\_\_  
(И.О.Фамилия)

Консультант

\_\_\_\_\_  
(Подпись, дата)

\_\_\_\_\_  
(И.О.Фамилия)

Министерство науки и высшего образования Российской Федерации  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Московский государственный технический университет имени Н.Э. Баумана  
(национальный исследовательский университет)»  
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_  
(Индекс)

\_\_\_\_\_ В.М. Черненький \_\_\_\_\_  
(И.О.Фамилия)

« \_\_\_\_\_ » 20 \_\_\_\_ г.

## ЗАДАНИЕ

### на выполнение научно-исследовательской работы

по теме Генерация твитов на основе LSTM сети

Студент группы ИУ5-34м

Макрушина Виктория Александровна  
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

Учебная

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_ НИР \_\_\_\_\_

График выполнения НИР: 25% к \_\_\_\_ нед., 50% к \_\_\_\_ нед., 75% к \_\_\_\_ нед., 100% к \_\_\_\_ нед.

**Техническое задание** \_\_\_\_\_

#### **Оформление научно-исследовательской работы:**

Расчетно-пояснительная записка на \_\_\_\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

Дата выдачи задания « \_\_\_\_ » \_\_\_\_\_ 20\_\_ г.

**Руководитель НИР**

\_\_\_\_\_  
(Подпись, дата)

\_\_\_\_\_  
(И.О.Фамилия)

**Студент**

\_\_\_\_\_  
(Подпись, дата)

\_\_\_\_\_  
(И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

## Содержание

Цель работы .....	4
Задачи .....	4
Основная часть .....	4
1. Сбор данных .....	4
2. Анализ полученных данных .....	5
Длина твитов. ....	5
Количество слов в твите. ....	5
3. Очистка данных .....	6
4. Маркировка слов. ....	6
5. Создание тренировочного набора .....	7
Вариант 1: разделить по твиту. ....	7
Вариант 2: Без разделения на твиты. ....	7
6. Создание моделей для обучения. ....	8
Вариант 1 .....	8
Вариант 2 .....	8
7. Обучение модели .....	9
Модель № 1 .....	9
Модель № 2 .....	9
Модель № 3 .....	10
Модель № 4 .....	11
8. Создание новых твитов по обученной модели .....	11
Модель 1 .....	11
Модель 2 .....	12
Модель 3 .....	12
Модель 4 .....	12
Заключение .....	13
Список использованных источников. ....	14

## **Цель работы**

Создать нейронную сеть которая будет генерировать твиты определенного пользователя, на основе предыдущих твитов этого пользователя, полученных.

## **Задачи**

- Собрать данные из твитера с помощью twitter API.
- Проанализировать полученные данные.
- Очистить полученные данные от ссылок и некоторых символов.
- Токенизировать каждое слово из полученного набора.
- Создать обучающий набор.
- Обучить модель и получить результат

## **Основная часть**

### **1. Сбор данных**

Импорт библиотек для доступа к twitter API и составления файла .csv. Создание переменных с ключами доступа и авторизация для получения данных из twitter.

Набор данных с твитами ПАО Газпром был подобран опытным путем, первые несколько наборов данных были наборами твитов знаменитостей шоу бизнеса (певцы, актеры), но после обучения нейросети на этих данных был сделан вывод о том, что материал слишком разнороден мало повторяющихся слов, что приводит либо к полному переобучению модели либо наоборот вывод абсолютно не связанных слов или только знаков препинаний. После перечисленных неудач было принято решение взять более узкоспециализированную область, такого пользователя чтобы твиты имели много повторяющихся слов, однако не были абсолютно одинаковыми. Наилучшим кандидатом показался ПАО Газпром. У твиттера есть ограничение из-за которого мы не можем получить более 3200 твитов. Этот набор маловат для обучения сети поэтому были предприняты попытки объединения нескольких наборов похожих пользователей таких как ПАО Роснефть, однако при увеличении объема набора данных появилась новая проблема, время обучения сети сильно увеличилось, и объем занимаемой памяти тоже очень сильно вырос, что не позволило обучить сеть в Google Colab. После этого было принято решение обучать сеть на наборе данных ПАО Газпром.

Данные были приведены к csv формату для удобной работы с файлом.

355596055561256961  
 «Газовый „КАМАЗ“» успешно прошел все семь этапов «Шелкового пути»: Завершился между  
 в... <http://t.co/zUv8PUos7Q>  
 355702944370409474  
 000 «Газпром трансгаз Чайковский» открыл в Очере спорткомплекс «Газовик»: [http://t](http://t.co/n8RXAlc6I5)  
 355347278397050880  
 «Газпром» и E.ON провели IX заседание Координационного совета по научно-техническому  
 к... <http://t.co/h91k7T2Wj0>  
 355347278397050880  
 «Газпром» и Shell обсудили вопросы двустороннего сотрудничества в нефтегазовой сфер  
 ч... <http://t.co/n8RXAlc6I5>  
 355347278397050880  
 В Надыме прошло совещание по расширению использования газа в качестве моторного топ  
<http://t.co/LqgerJtXRu>  
 354940959345221632  
 «Газпром» и Правительство Оренбургской области будут развивать региональный рынок г  
 н... <http://t.co/XvcuQu2tgk>  
 354940958061760512  
 Виктор Зубков: «Газпром» готов проектировать дополнительные газовые заправки для се  
 а... <http://t.co/8XeztPhRfS>  
 Все ок: 3186

Рисунок 1. Результат (данные полученные с помощью твиттер API)

## 2. Анализ полученных данных

### Длина твитов.

Был проведен подсчет количества символов в каждом твите, после чего эти данные были записаны. Раньше у Твитера был лимит в 140 символов, теперь увеличился до 280. При рассмотрении вопроса о генерации твитов, необходимо генерировать определенное количество слов и последующих символов. Вместо того, чтобы генерировать произвольное количество символов между [140,280] или даже [0,280], мы будем генерировать длину при помощи функции генерации длинны твитов `get_tweet_length()`.

### Количество слов в твите.

Эта часть анализа и визуализации данных проводится для того чтобы определить параметр `seq_length`. Подсчитаем количество слов в каждом твите и сохраним эти значения. Визуализируем полученные данные. На графике видно соотношение количества твитов с определенным количеством слов.

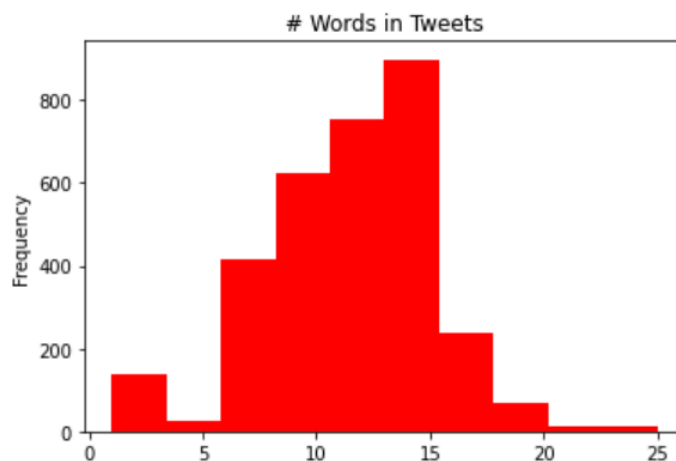


Рисунок 2. График количество твитов определенной длины.

После чего выведем описание этих данных.

```
count    3182.000000
mean      11.322124
std       3.699105
min       1.000000
25%       9.000000
50%      12.000000
75%      14.000000
max       25.000000
Name: num_words, dtype: float64
```

Рисунок 3. Описание длины твитов в наборе

По полученным данным видно что в среднем в твите содержится 11 слов, существуют твиты длиной всего 1 слово. Стоит заметить что любые твиты с количеством слов  $< seq\_length$  не будут участвовать в процессе тренировки вообще. То есть, если  $seq\_length < 9$  то любое значение больше 9 приведет к потере 25% твитов (при использовании версии 1 обучающего набора для сборки).

### 3. Очистка данных

Для того чтобы генерация проходила более успешно и в словаре не было слов, которые отличаются только кавычкой, скобкой и тому подобное, нам необходимо провести очистку данных.

- удаляем все ссылки
- заменяем упоминания пользователей твитера словом user
- удаляем знаки препинания
- удаляем скобки
- обозначения хэштега
- удаляем обозначение типа организации (ООО, ОАО)
- удаляем символы, которые могут появиться при декодировке полученного текста (`—\xa0, \xa0, amp\S*`)
- странные символы тоже удаляем (`©, ®, °, ™, №, '♪', ™`)
- стандартизируем
- приводим к нижнему регистру

Сохраним полученные очищенные данные в отдельный файл.

### 4. Маркировка слов.

Создадим переменную `all_text` которая будет хранить все твиты подряд одним текстом. Разделим каждый твит на набор слов.

Получим:

[ 'продолжается', 'реализация', 'проекта', 'создания', 'комплекса', 'по', 'переработке  
 [ 'на', 'первом', 'канале', 'состоялась', 'премьера', 'документального', 'фильма', 'эн  
 [ 'сегодня', 'на', 'телеканале', 'нтв', 'состоялась', 'премьера', 'фильма', 'вахта', '  
 [ 'газпром', 'назвал', 'новое', 'крупное', 'месторождение', 'на', 'шельфе', 'ямала', '  
 [ 'газпром', 'передал', '200', 'тысяч', 'защитных', 'костюмов', 'для', 'врачей', 'санк  
 [ 'члены', 'правления', 'газпрома', 'направили', 'личные', 'средства', 'на', 'борьбу',  
 [ 'отчетность', 'газпрома', 'по', 'международным', 'стандартам', 'финансовой', 'отчетн  
 [ 'в', 'ленэкспо', 'разворачивают', 'госпиталь', 'для', 'больных', 'с', 'легкой', 'фор  
 [ 'газпром', 'начинает', 'подготовку', 'к', 'работе', 'в', 'осенне', 'зимнем', 'период  
 [ 'правление', 'предлагает', 'дивиденды', 'по', 'итогам', '2019', 'года', 'в', 'размер  
 [ 'максимальная', 'суточная', 'производительность', 'подземных', 'хранилищ', 'газа', '  
 [ 'газпром', 'продолжает', 'реализацию', 'крупнейших', 'инвестпроектов']  
 [ 'правление', 'предлагает', 'провести', 'годовое', 'общее', 'собрание', 'акционеров',  
 [ 'газпром', 'продолжает', 'последовательную', 'работу', 'по', 'импортозамещению']  
 [ 'в', '2019', 'году', 'газпром', 'увеличил', 'продажи', 'природного', 'газа', 'в', 'к  
 [ 'газпрому', 'поручено', 'приступить', 'к', 'предынвестиционной', 'стадии', 'проекта'  
 [ 'газпром', 'с', 'комсомольской', 'правдой', 'открыли', 'совместный', 'проект', 'вечн  
 [ 'состоялась', 'рабочая', 'встреча', 'алексея', 'миллера', 'и', 'губернатора', 'санкт  
 [ 'фоторепортаж', 'суровые', 'правила', 'арктики']  
 [ 'газпром', 'и', 'user', 'вносят', 'изменения', 'в', 'условия', 'сделки', 'о', 'прода  
 [ 'милые', 'женщины', 'поздравляем', 'вас', 'с', 'наступающим', 'международным', 'женс  
 [ 'газпром', 'усиливает', 'меры', 'по', 'оптимизации', 'затрат', 'компаний', 'группы']

*Рисунок 4. Набор слов твите*

Сопоставим уникальные слова целым числам.

Токенизация (иногда – сегментация) по словам – это процесс разделения предложений на слова-компоненты.

Есть две основные разновидности кодировки для текстовых, символьных и словесных моделей. В моделях уровня символов мы токенизируем каждую букву в один горячий вектор из совокупности букв. В моделях на уровне слов мы токенизируем каждое слово в один горячий вектор, образующий совокупность слов.

В данной работе используется 2 вариант: токенизируем каждое слово.

В результате мы получили список уникальных слов, количество слов всего и количество уникальных слов.

## 5. Создание тренировочного набора

### Вариант 1: разделить по твиту.

- Создание обучающего набора из суммы длин всех твитов -  $\text{num\_tweets} * \text{seq\_length}$
- Это позволяет получить меньший набор данных, но набор данных имеет большую ценность из-за группировки слов (тематически незаметно).

### Вариант 2: Без разделения на твиты.

- Создает обучающий набор длины  $n\_words - \text{seq\_length}$
- Учебный набор для этой версии длиннее, но потенциально менее связный, так как твиты могут сталкиваться друг с другом.

Функция генерации слов по уже обученной модели. На вход принимает собственно саму модель. На выходе получается сгенерированный твит.

Структура функции:

- сгенерируем длину твита
- получим рандомное число и возьмем
- слово соответствующее этому числу из словаря слов
- пока длина нового твита меньше сгенерированной длины мы  
полняем цикл
- в цикле модель предсказывает слова на основании обучения
- выводим слова

## **6. Создание моделей для обучения.**

Обучение модели проводится на основе LSTM слоев. Долговременная краткосрочная память (Long Short-Term Memory, LSTM) - это название архитектуры, которая была вдохновлена устройством человеческого мозга, а точнее его краткосрочной памяти. Главная идея, лежащая в основе LSTM - это необходимость иметь некую ячейку памяти для хранения информации.

### **Вариант 1**

Создается модель для обучения нейросети.

Которая принимает параметры: размерность скрытого слоя и количество слоев.

Состоит из:

- LAYER\_NUM слоев LSTM.
- Далее идет слой, который считает ошибку на каждом этапе предсказания элемента последовательности.
- Слой с softmax функцией.
- Настройка модели функция потерь - кроссэнтропия, оптимизация - Методом адаптивного скользящего среднего градиентов.

Однако данная модель имеет недостаток, она быстро переобучается.

### **Вариант 2**

Немного изменим эту модель. После каждого уровня LSTM добавим слой Dropout, такой что он будет отключать 20% всех нейронов, для того чтобы не было переобучения модели.

В результате получаем следующую структуру:

- LAYER\_NUM слоев LSTM и Dropout.



- Далее идет слой, который считает ошибку на каждом этапе предсказания элемента последовательности.
- Слой с softmax функцией.
- Настройка модели функция потерь - кроссэнтропия, оптимизация - Методом адаптивного скользящего среднего градиентов.

## 7. Обучение модели

## Модель № 1

- Вариант генерации слов №1
- Функция генерации модели №1( без Dropout)

```
This is separate from the ipykernel package so we can avoid doing imports until
Epoch 1/5
7904/7904 [=====] - 37s 5ms/step - loss: 7.5543
Epoch 2/5
7904/7904 [=====] - 34s 4ms/step - loss: 7.1258
Epoch 3/5
7904/7904 [=====] - 36s 5ms/step - loss: 6.8249
Epoch 4/5
7904/7904 [=====] - 36s 5ms/step - loss: 6.2861
Epoch 5/5
7904/7904 [=====] - 36s 5ms/step - loss: 5.6035
total chars: 133
xii в газ в т млрд куб м м т июня м м м м м м м м м м м м м м м м м м м м м м м м м м м м м
```

Рисунок 5. Первый результат обучения (первые 5 эпох)

```
Epoch 1/5
7904/7904 [=====] - 37s 5ms/step - loss: 0.3873
Epoch 2/5
7904/7904 [=====] - 38s 5ms/step - loss: 0.3857
Epoch 3/5
7904/7904 [=====] - 36s 5ms/step - loss: 0.3858
Epoch 4/5
7904/7904 [=====] - 37s 5ms/step - loss: 0.3858
Epoch 5/5
7904/7904 [=====] - 35s 4ms/step - loss: 0.3852
total chars: 92
официальный канал газпрома на ютубе пополнился новым роликом видеодневник спартакиады 7 млрд
```

Рисунок 6. Последний результат обучения (95-100 эпохи)

## Модель № 2

- Вариант генерации слов №2
- Функция генерации модели №1 (без dropout)

```

This is separate from the ipykernel package so we can avoid doing imports until
Epoch 1/5
31066/31066 [=====] - 109s 4ms/step - loss: 7.1222
Epoch 2/5
31066/31066 [=====] - 107s 3ms/step - loss: 5.7355
Epoch 3/5
31066/31066 [=====] - 111s 4ms/step - loss: 4.4279
Epoch 4/5
31066/31066 [=====] - 109s 4ms/step - loss: 3.3965
Epoch 5/5
31066/31066 [=====] - 110s 4ms/step - loss: 2.5912
total chars: 108
надальнем в европе повышает надежность газоснабжения газоснабжения и газификации якутии до 2026 года газпром

```

*Рисунок 7. Первый результат обучения (первые 5 эпох)*

```

Epoch 1/5
31066/31066 [=====] - 119s 4ms/step - loss: 0.4758
Epoch 2/5
31066/31066 [=====] - 117s 4ms/step - loss: 0.4761
Epoch 3/5
31066/31066 [=====] - 120s 4ms/step - loss: 0.4757
Epoch 4/5
31066/31066 [=====] - 117s 4ms/step - loss: 0.4751
Epoch 5/5
31066/31066 [=====] - 117s 4ms/step - loss: 0.4749
total chars: 113
фрг геза андреас фон гайр обсудили актуальные направления сотрудничества с начала года газпром поставил в венгрию

```

*Рисунок 8. Последний результат обучения (115-120 эпохи)*

### Модель № 3

- Вариант генерации слов №1
- Функция генерации модели №2( с Dropout)

```

This is separate from the ipykernel package so we can avoid doing imports until
Epoch 1/5
7904/7904 [=====] - 33s 4ms/step - loss: 9.1067
Epoch 2/5
7904/7904 [=====] - 30s 4ms/step - loss: 8.7448
Epoch 3/5
7904/7904 [=====] - 31s 4ms/step - loss: 8.4855
Epoch 4/5
7904/7904 [=====] - 31s 4ms/step - loss: 8.0628
Epoch 5/5
7904/7904 [=====] - 32s 4ms/step - loss: 7.6004
total chars: 81
первые обсудили и сотрудничестве в сфере сфере и области и промышленности области

```

*Рисунок 9. Первый результат обучения (первые 5 эпох)*

```

Epoch 1/5
7904/7904 [=====] - 29s 4ms/step - loss: 3.4971
Epoch 2/5
7904/7904 [=====] - 28s 4ms/step - loss: 3.5073
Epoch 3/5
7904/7904 [=====] - 28s 4ms/step - loss: 3.5219
Epoch 4/5
7904/7904 [=====] - 28s 4ms/step - loss: 3.5064
Epoch 5/5
7904/7904 [=====] - 28s 4ms/step - loss: 3.4793
total chars: 143
рост спроса на российский газ в европе подтверждает важность строительства северного потока

```

*Рисунок 10. Последний результат обучения (95-100 эпохи)*

## Модель № 4

- Вариант генерации слов №2
- Функция генерации модели №2(с Dropout)

```
This is separate from the ipykernel package so we can avoid doing imports until
Epoch 1/5
31066/31066 [=====] - 139s 4ms/step - loss: 8.9565
Epoch 2/5
31066/31066 [=====] - 128s 4ms/step - loss: 8.3006
Epoch 3/5
31066/31066 [=====] - 130s 4ms/step - loss: 7.3407
Epoch 4/5
31066/31066 [=====] - 133s 4ms/step - loss: 6.5206
Epoch 5/5
31066/31066 [=====] - 135s 4ms/step - loss: 5.8814
total chars: 123
долгосрочном соглашении о взаимопонимании в сфере СПГ и Беларуси и янао в газовой сфере газпром
```

*Рисунок 11. Первый результат обучения (первые 5 эпох)*

```
Epoch 1/5
31066/31066 [=====] - 141s 5ms/step - loss: 3.5760
Epoch 2/5
31066/31066 [=====] - 141s 5ms/step - loss: 3.5713
Epoch 3/5
31066/31066 [=====] - 138s 4ms/step - loss: 3.5755
Epoch 4/5
31066/31066 [=====] - 139s 4ms/step - loss: 3.5833
Epoch 5/5
31066/31066 [=====] - 144s 5ms/step - loss: 3.5769
total chars: 114
медвежье 1972 год природа ямало-ненецкого автономного округа фото газпром добыча уренгой совет
```

*Рисунок 12. Последний результат обучения (115-120 эпохи)*

## 8. Создание новых твитов по обученной модели

### Модель 1

ПОСЛЕ ТРЕНИРОВКИ НА 20 ЭПОХАХ

Модель 1, Твит 1  
total chars: 96  
762 поддержке газпром трансгаз москва прошел концерт симфонической музыки соревнования инвалидов

ПОСЛЕ ТРЕНИРОВКИ НА 40 ЭПОХАХ

Модель 1, Твит 1  
total chars: 88  
компания газпром нефть вам лучше обратиться с этим вопросом в их твиттер user user пресс

ПОСЛЕ ТРЕНИРОВКИ НА 60 ЭПОХАХ

Модель 1, Твит 1  
total chars: 134  
выиграть на компрессорной станции гагацкая фото ольги филипповой газпром трансгаз ухта 2017 г фото газпром трансгаз ухта генеральным

ПОСЛЕ ТРЕНИРОВКИ НА 100 ЭПОХАХ

Модель 1, Твит 1  
total chars: 105  
всалехарде область накопила соглашение 8 активами 2017 года газпром gpr\_export заводе по территории поток

*Рисунок 13. Результат генерации твитов Модели 1*

## Модель 2

ПОСЛЕ ТРЕНИРОВКИ НА 20 ЭПОХАХ

Модель 2, Твит 1

total chars: 96

десять месяцев 2018 года объем поставок газпрома в германию достиг 476 млрд куб м что на 112 или

ПОСЛЕ ТРЕНИРОВКИ НА 120 ЭПОХАХ

Модель 2, Твит 1

total chars: 75

загрязнений подберезовики фото газпром добыча ямбург сургутская грэс 1 фото

*Рисунок 14. Результат генерации твитов Модели 2*

## Модель 3

Модель 3, Твит 5

total chars: 88

проблему работников нефтяной и газовой промышленности! фото газпром добыча ямбург ямбург

на 100 эпохах

Модель 3, Твит 1

total chars: 108

финишировал самый протяженный в мире автопробег на сжиженном природном газе будут перевозить user user мы мы

*Рисунок 15. Результат генерации твитов Модели 3*

## Модель 4

ПОСЛЕ ТРЕНИРОВКИ НА 60 ЭПОХАХ

Модель 4, Твит 1

total chars: 81

никита притыкин лучший молодой рационализатор газпрома газпром получил разрешение

ПОСЛЕ ТРЕНИРОВКИ НА 120 ЭПОХАХ

Модель 4, Твит 1

total chars: 92

клиентам как перевести автомобиль на природный газ в четыре шага? user газпром и wintershall

*Рисунок 16. Результат генерации твитов Модели 4*

## **Заключение**

Полученные результаты далеко не идеальны.

Это связано с тем, что

- используется довольно небольшой набор данных.
- В словаре встречаются одни и те же слова с разными окончаниями (эту проблему можно решить с помощью Byte Pair Encoding (BPE))
- В словаре встречаются цифры
- Используется всего 100 эпох (возможно если увеличить число эпох обучение покажет лучший результат, однако понадобится больше времени)

### **Список использованных источников.**

1. Лекции по предмету «Проектирование интеллектуальных систем», МГТУ им. Баумана, 2020г.
2. Генерация текста с помощью LSTM рекуррентных нейронных сетей [Электронный ресурс] URL: <https://www.machinelearningmastery.ru/text-generation-lstm-recurrent-neural-networks-python-keras/> (Дата обращения 12.05.2020)
3. Использование REST API для сбора исторических данных [Электронный ресурс] URL: <https://www.machinelearningmastery.ru/downloading-data-from-twitter-using-the-rest-api-24becf413875/> (Дата обращения 16.05.2020)
4. Николенко С.И., Кадури А.А., Архангельская Е.О. Глубокое обучение. – Издательский дом "Питер", 2017. — 480 с
5. Creating The Twitter Sentiment Analysis Program in Python with Naive Bayes Classification [Электронный ресурс] URL: <https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed> (Дата обращения 08.05.2020)