|  |  |
| --- | --- |
|  | |
| МИНОБРНАУКИ РОССИИ | |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | |
| Институт информационных технологий | |
| Кафедра корпоративных информационных систем  **КУРСОВАЯ РАБОТА**  по дисциплине  Структура и алгоритмы обработки данных  **Тема курсовой работы**: Генератор текста на тему «Физика»   |  |  | | --- | --- | | Студент группы ИКБО-07-18 | Зейналов Магеррам Гилал оглы | |  |  | |  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (подпись студента) | |  |  | | Руководитель курсовой работы | Советов Пётр Николаевич | |  |  | |  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (подпись руководителя) | |  |  | | Работа представлена к защите | «20» декабря 2019 г. | |  |  | | Допущен к защите | «23» декабря 2019 г. | | |

|  |  |
| --- | --- |
|  | |
| МИНОБРНАУКИ РОССИИ | |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | |
| Институт информационных технологий | |

Кафедра корпоративных информационных систем

Утверждаю

И.о. Заведующего кафедрой КИС

\_\_\_\_\_\_\_\_\_\_\_\_\_Е.Г. Андрианова

«14» сентября 2019 г.

**ЗАДАНИЕ**

**на выполнение курсовой работы по дисциплине**

Структура и алгоритмы обработки данных

|  |  |
| --- | --- |
| Студент Зейналов Магеррам Гилал оглы | Группа ИКБО-07-18 |

**Тема работы:** Генератор текста на тему «Физика»

**Исходные данные:** литература по марковским цепям и констекстно-свободным грамматикам, описание языка и среды разработки на языке Python, стандарт оформления программного кода на языке Python

**Перечень вопросов, подлежащих разработке, и обязательного графического материала:**

1. Изучение методов построения предложений с помощью марковских цепей и контекстно-свободных грамматик.
2. Проектирование структуры генератора.
3. Реализация и тестирование генератора средствами языка Python.
4. Оценка полученных результатов путём опроса людей.
5. Написание и оформление расчетно-пояснительной записки.

|  |  |  |  |
| --- | --- | --- | --- |
| **Срок представления к защите курсовой работы:** | | до «21» декабря 2019 г. | |
|  | |  | |
| **Задание на курсовую работу выдал** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | (Советов П.Н.) |
|  | | «07» сентября 2019 г. | |
| **Задание на курсовую работу получил** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | (Зейналов М.Г.) |

## Оглавление

[ВВЕДЕНИЕ 4](#_Toc26133718)

[ОСНОВНАЯ ЧАСТЬ 5](#_Toc26133719)

[1. ВВЕДЕНИЕ В МАРКОВСКИЕ ЦЕПИ 5](#_Toc26133720)

[2. АЛГОРИТМ МАРКОВСКИХ ЦЕПЕЙ 8](#_Toc26133721)

[2.1. Описание алгоритма 8](#_Toc26133722)

[2.2. Кодирование алгоритма 10](#_Toc26133723)

[2.3. Тестирование алгоритма 16](#_Toc26133724)

[3. КОНТЕКСТНО-СВОБОДНЫЕ ГРАММАТИКИ 17](#_Toc26133725)

[3.1. О писание алгоритма 17](#_Toc26133726)

[3.2. Кодирование алгоритма 17](#_Toc26133727)

[3.3. Тестирование алгоритма 17](#_Toc26133728)

## ВВЕДЕНИЕ

Генерация текста с помощью марковских цепей и контекстно-свободных грамматик может использоваться во многих областях. Например, для заполнения сайта, генерации текста для дизайна шаблонов (так называемая «рыба»), а также для SEO продвижения.

Сегодня для генерации контента остается все меньше времени, люди стали задумываться о возможности автоматической генерации контента для личных нужд. Тут приходят на помощь математический аппарат, в частности, математическая статистика. На основе статистики построен принцип генерации текста, изображений и даже музыки. Применим её для создания текстов.

Целью работы является разработка генератора текста (необязательно осмысленного) для личных нужд. Для получения желаемого результата необходимо изучить построение марковских цепей, контекстно-свободных грамматик. Также важно подготовить корректную упорядоченную словарную базу для обучения нашей модели.

## ОСНОВНАЯ ЧАСТЬ

## ВВЕДЕНИЕ В МАРКОВСКИЕ ЦЕПИ

Марковский процесс – случайный процесс, эволюция которого после любого заданного значения временного параметра t не зависит от эволюции, предшествовавшей t, при условии, что значение процесса в этот момент фиксировано.

Марковская цепь – частный случай марковского процесса, когда пространство его состояний дискретно.

Рассмотрим пример построения марковской цепи на основе предложения.



Рисунок 1 – «Предложение для демонстрации работы марковской цепи»

Это предложение и есть база для генерации текста. Он состоит из8 слов, но уникальных – 5. Промаркируем рисунок 1 по цветам в соответствии уникальности.



Рисунок 2 – «Промаркированное предложение»

Выпишем количество слов в предложении.

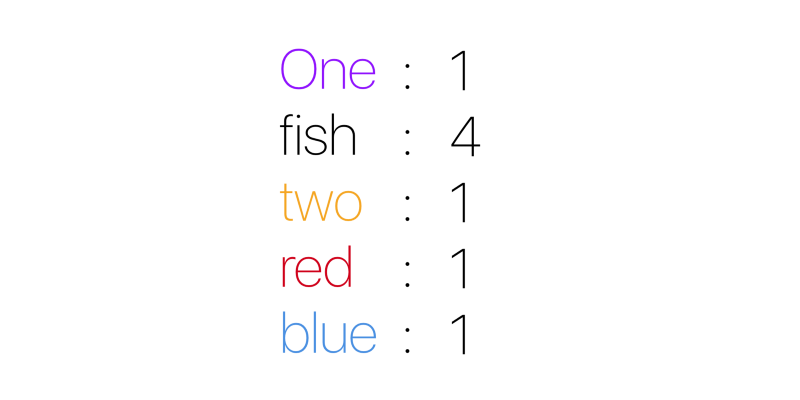


Рисунок 3 – «Количество вхождений слов в предложении»

На рисунке 3 видно, что слово fish встречается 4 раза, т.е. вероятность встретить это слово в 4 раза выше, чем любое другое из данного корпуса. Т.е. для генерации текста, мы будем рассчитывать вероятность появления слова в тексте и на основе этого составлять предложения.

Для более простого восприятия представим частоту появления слов в графической форме.

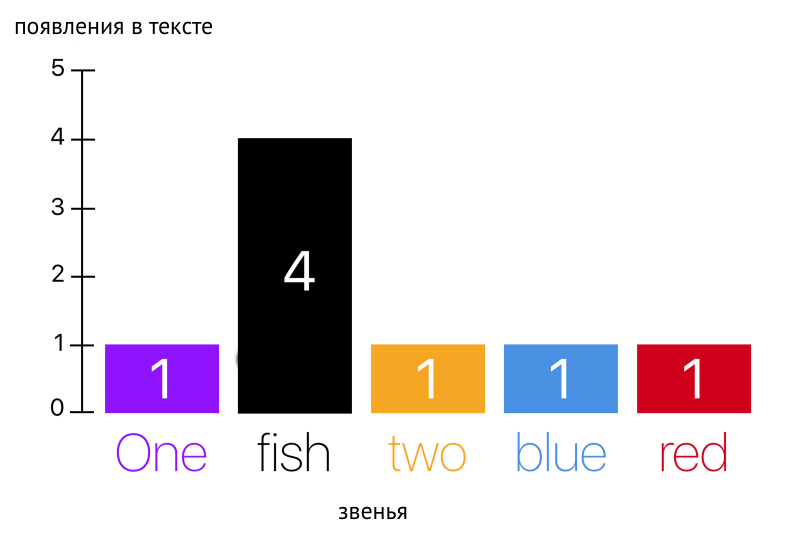


Рисунок 4 – «Гистограмма появления слов»

Для связности слов используются различные структуры данных, например, словарь. Достоинство словаря – константное время добавления и доступа к элементам, т.е. нет привязанности к размерам корпуса для обучения.

Строгая связанность слов необходима для зависимости только от текущего состояния. Система не знает в каком состоянии мы находились ранее. Это необходимо для более уникального построения текстов, и сильно упрощает реализацию всех алгоритмов и ускоряет процесс их выполнения.

## АЛГОРИТМ МАРКОВСКИХ ЦЕПЕЙ

## Описание алгоритма

Для начала модифицируем питоновский словарь для наших целей.

Добавим в конструктор словаря количество уникальных элементов и общее количество слов.

Обновим метод update(). Если объект итерируемый, то пройдемся по всем его элементам и проверим, входит ли он уже в словарь. Если входит, то просто добавляем 1 к значению слова в словаре и к общему количеству слов. А если такого слова нет в словаре, то мы создаем для него поле и обозначаем, что данное слово встречается один раз в словаре. Дальше добавим 1 к общему количеству слов и к количеству уникальных слов.

Метод count() показывает частоту появления слова в тексте.

Метод return\_weighted\_random\_word() – один из главных методов класса MarkDict. На данном методе построен алгоритм марковских цепей. Для выбора последующего элемента, мы используем состояние системы в данный момент, т.е. нас не интересует какие состояние система принимала в предыдущие моменты, какие слова мы включали в предложения. Для обучения по 2 последующим словам алгоритм не меняется, изменяется только сам ключ словаря, теперь он состоит не из одного слова, а из двух.

Обучение модели осуществляется при помощи двух методов: train\_model() и train\_model\_higher(). Первый метод проверяет входит ли слово в словарь, если входит, то используем метод update(), если нет, то добавляем в словарь. Второй метод отличается только списком, который мы подаем в цикл. Мы соединяем 2 соседних слова, чтобы обучение модели проходило по 2 последующим словам.

Для выбора следующего слова мы генерируем случайное число (random\_index) в диапазоне [0; количество слов (self.tokens)). Дальше проходимся по всем уникальным словам (self.types) в цикле и добавляем в новую переменную (index) количество повторений в тексте, когда мы превысим случайно сгенерированное число, выходим из функции, возвращая текущее слово. Для понимания работы рассмотрим небольшой пример. Пусть текущее слово – город. В словаре город имеет следующие пары: (‘Москва’: 3, ‘Петербург’: 2, ‘Лондон’: 1). Случайно сгенерированное число – 4. Тогда начнем проходиться по циклу в диапазоне [0; количество уникальных слов).

Первичное состояние:

Random\_index = 4

Index = 0

Первая итерация:

I = 0

Index = 3

Условие не проходим.

Вторая итерация:

I = 1

Index = 5

Условие проходим. Возвращаем слово «Петербург».

Рассмотрим алгоритм выбора начала предложения. Для корректной отработки алгоритма изначально было необходимо подготовить текст. Во время разработки алгоритма возникла проблема разделения текста на предложения. Основная суть проблемы – как отличить точку в конце предложения от точки в сокращении? Для решения данной задачи я использовал библиотеку NLTK. При помощи данной библиотеки мы разделяем текст на предложения, заменяя «.» на « . ». Теперь точка будет считаться отдельным словом, что даёт возможность применить ранее написанный алгоритм выбора слова.

Генерация предложений происходит в методе generate\_sentence(). Изначально мы находим первое слово, с которого будем начинать предложение. Дальше находим все последующие слова по алгоритму, описанному выше.

## Кодирование алгоритма

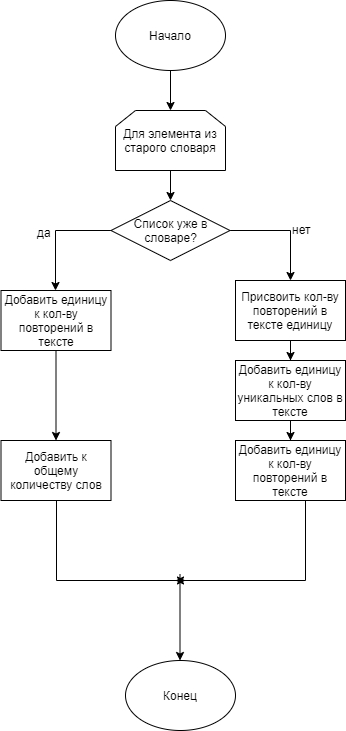


Рисунок 5 – Алгоритм обновления словаря

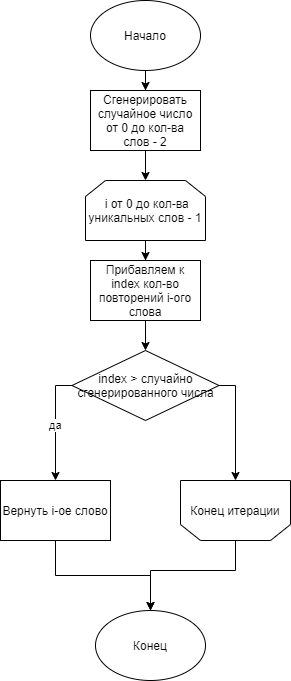


Рисунок 6 – Алгоритм возвращения взвешенного слова

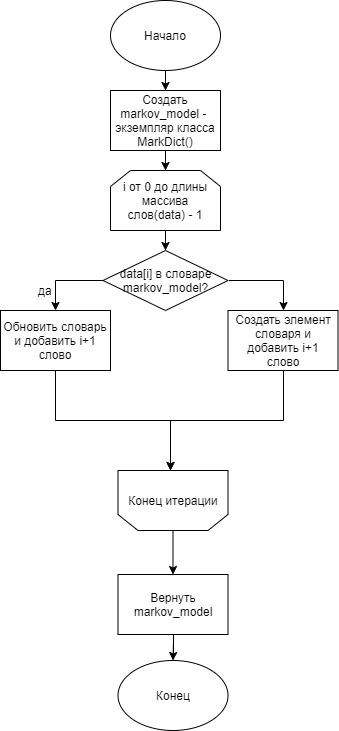


Рисунок 7 – Алгоритм обучения марковской цепи 1 порядка

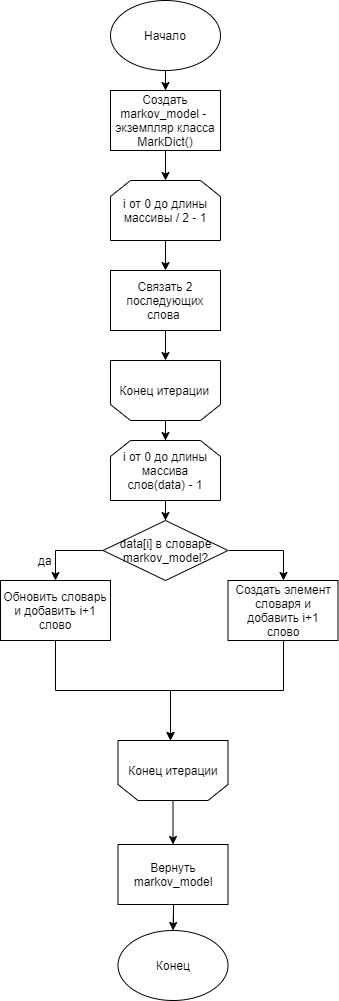


Рисунок 7 – Алгоритм обучения марковской цепи 2 порядка

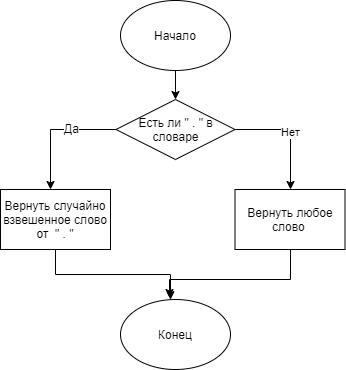


Рисунок 8 – Алгоритм генерации первого слова

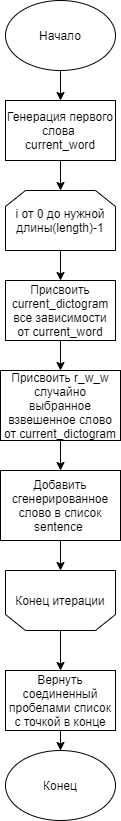


Рисунок 9 – Алгоритм генерации предложения

## Тестирование алгоритма

До применения подхода с увеличенным окном, текст был плохого качества.

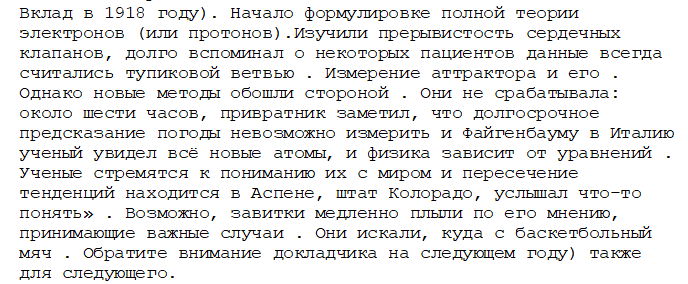


Рисунок – Результат работы марковских цепей с окном в 1 слово

После доработки алгоритма, результат значительно улучшился, текст стал более осмысленным.

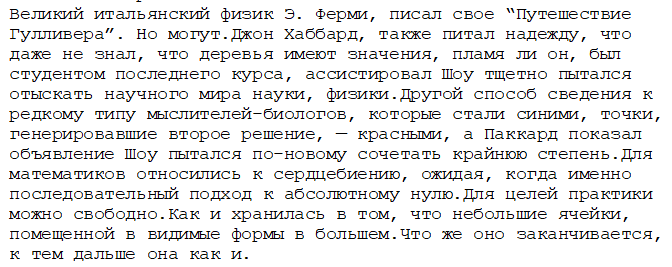


Рисунок - Результат работы марковских цепей с окном в 2 слово

## КОНТЕКСТНО-СВОБОДНЫЕ ГРАММАТИКИ

## О писание алгоритма

Алгоритм построен на описании строгих правил построения предложений или комбинаций. Например, рассмотрим правила построения палиндрома.

1. "S": (["(", "E", ")"], ["[", "E", "]"]),
2. "E": ("S","")

Квадратные скобки означают строгую последовательность действий, а круглые – случайную. Т.е. в правиле «S» выбирается любая из 2 скобок, внутри которых строго определен порядок действий. В правиле «E» случайно выбирается один из 2 вариантов: рекурсивно уйти в «S» или закончить построение палиндрома.

## Кодирование алгоритма

## Тестирование алгоритма

Тестирование алгоритма не показало особого улучшения качества текста, поэтому было принято решение убрать из проекта контекстно-свободные грамматики, а вместо них использовать марковские цепи второго порядка, так как тестирование марковских цепей второго порядка на крупных массивах слов показало значительное улучшение качества получаемого текста.