|  |  |
| --- | --- |
|  | |
| 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](#_Toc23986741)

[ОСНОВНАЯ ЧАСТЬ 5](#_Toc23986742)

[1. ВВЕДЕНИЕ В МАРКОВСКИЕ ЦЕПИ 5](#_Toc23986743)

## 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.

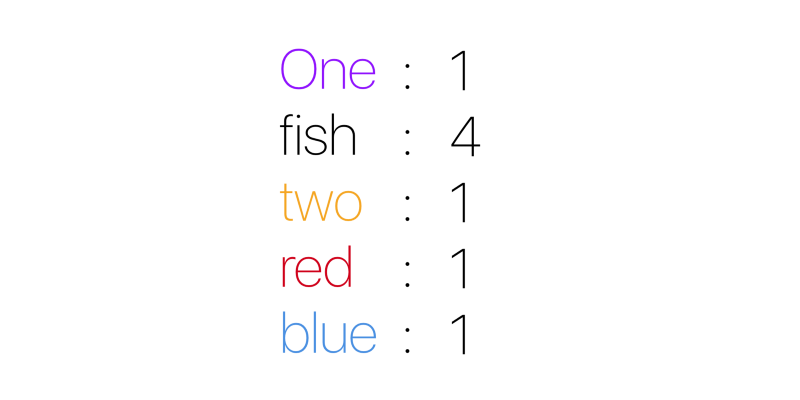


Рисунок 3 – «Количество вхождений слов в предложении»

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.

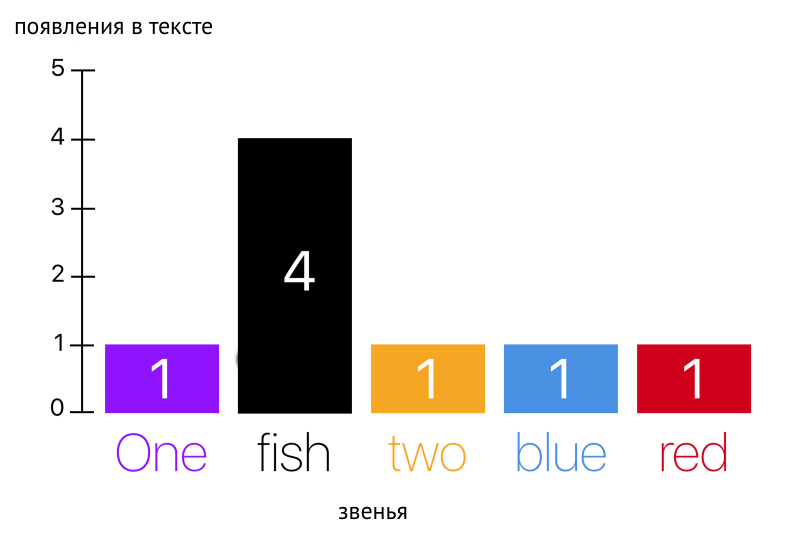


Рисунок 4 – «Гистограмма появления слов»

Сформируем пары по порядку в предложении.

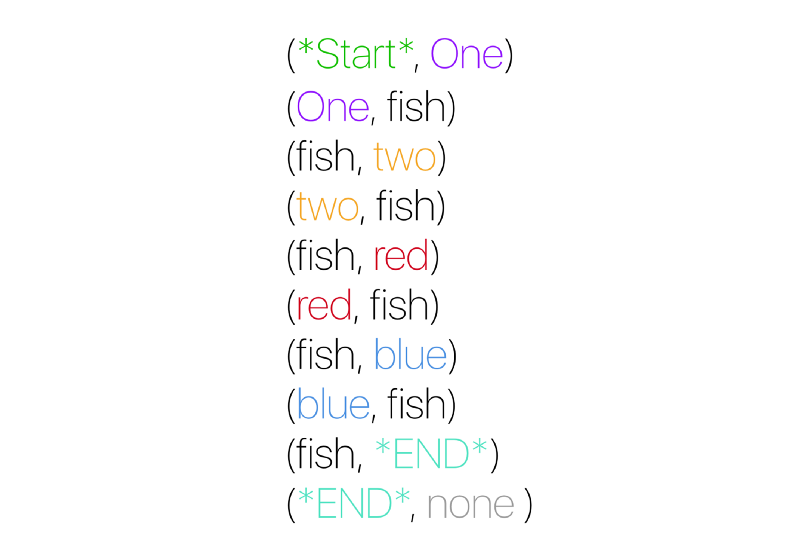


Рисунок 5 – «Пары слов по порядку»

Сгруппируем по первому слову.

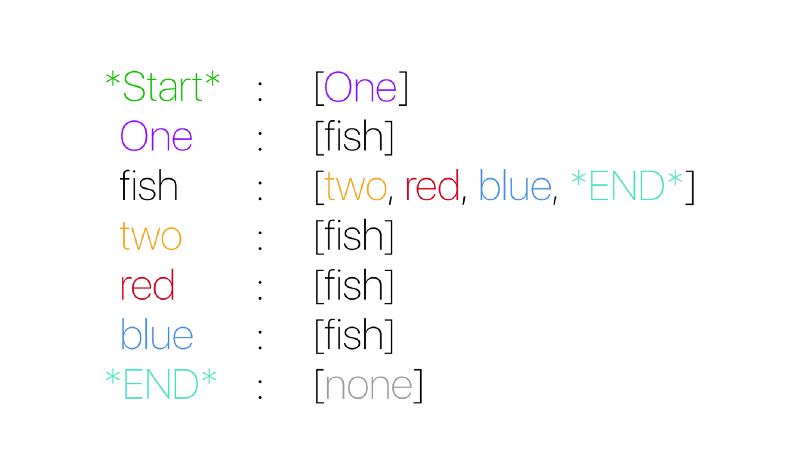


Рисунок 6 – «Сгруппированные слова»

1. **АЛГОРИТМ МАРКОВСКИХ ЦЕПЕЙ**
   1. **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.