

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования «ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

09.03.04 - Программная инженерия

Профиль направления подготовки бакалавриата

«Системное и прикладное программное обеспечение»

ОТЧЁТ О ПРОХОЖДЕНИИ ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ

Выполнил:

студент 2 курса группы 22207

Богатырев Павел Валерьевич

Место прохождения практики:

Кафедра информатики и математического обеспечения

Сроки прохождения практики:

29.05-09.06

Руководитель практики:

к.т.н., доцент

Богоявленская Ольга Юрьевна

Оценка _____

Дата _____

Петрозаводск — 2023

Содержание

Введение	3
1 Теория	4
1.1 Генераторы текста	4
1.2 Цепи Маркова	5
2 Практическая часть	6
2.1 Задача	6
3 Заключение	7
Список литературы	8

Введение

В современном мире все больше внимания уделяется нейросетям. Их способности становятся все более развитыми, и на данный момент ИИ уже умеет вести диалоги с пользователем, генерировать картины или помогать в написании программ. Одной из главных задач в этой области является генерация текстов. Самые большие нейросети (например, ChatGPT) обучаются на огромных объемах данных, что позволяет им генерировать ответы на запросы, которые выглядят очень правдоподобно.

Однако не всегда пользователю требуется генерировать что-то сложное, а нужно лишь нечто похожее на осмысленный текст. Для этих целей существуют более простые алгоритмы, такие как цепи Маркова. Задачей данной практики заключается в том, чтобы ознакомиться с этим алгоритмом и реализовать одну из его частей.

1 Теория

1.1 Генераторы текста

Генератор текста — компьютерная программа, генерирующая тексты (сленг генерёнку), корректные с точки зрения большинства языковых норм, но, как правило, лишённые смысла.

Генерация текста путём составления из полностью случайных слов даёт мусорный результат: бессмысленный для человека и легко распознаётся анализаторами текстов, и поэтому не применяется. Обычно применяется генерация по вручную написанным фразам-шаблонам.

В случаях, когда не важен смысл генерируемого текста, он составляется из «мешанки» предложений из разных исходных текстов, или из частей предложений. Такой метод совместно с шаблонами и синонимизацией используется в чат-ботах и ботах-комментаторах в соцсетях и блогах. Такие боты копируют собеседнику фразы, записанные с других чатов или сайтов. Более продвинутые чат-боты сортируют фразы по ключевым словам, поэтому их ответ более приближен к теме диалога.

1.2 Цепи Маркова

Цепь Маркова — последовательность случайных событий с конечным или счётным числом исходов, где вероятность наступления каждого события зависит только от состояния, достигнутого в предыдущем событии. Характеризуется тем свойством, что, говоря нестрого, при фиксированном настоящем будущее независимо от прошлого. Названа в честь А. А. Маркова (старшего), который впервые ввёл это понятие в работе 1906 года.

Одной из первых научных дисциплин, в которой цепи Маркова нашли практическое применение, стала лингвистика (в частности текстология). Сам Марков для иллюстрации своих результатов исследовал зависимость в чередовании гласных и согласных в первых главах «Евгения Онегина» и «Детских годов Багрова-внука». В настоящий момент цепи Маркова получили широкое распространение в сфере генерации текстов, так как позволяют создавать правдоподобные тексты, хотя их осмысленность сильно уступает тем текстам, которые генерируют нейросети.

Для генерации текста с помощью цепей Маркова берется исходный текст, который затем разбивается на префиксы и суффиксы. Префикс — сочетание нескольких слов, которому сопоставляются один или несколько суффиксов (то есть слов, идущих в оригинале после него). Каждой паре префикс-суффикс также приписывается частота, с которой эта пара встречается в источнике. Затем на основе полученных данных алгоритм уже генерирует текст, выбирая каждое следующее слово на основе предыдущих и словаря префиксов.

2 Практическая часть

2.1 Задача

Задачей нашей команды было написание кода для поиска префиксов и суффиксов в уже разбитом на части тексте. Программа должна принимать на вход массив со словами из исходного текста и на основе него создавать словарь префиксов и суффиксов. Также для каждой пары должна быть определена частота использования в оригинальном тексте.

Для реализации был выбран язык программирования Python. Результат работы алгоритма записывается в файл на языке JSON.

Конкретно моей частью работы было написание функции заполнения словаря. В Python уже существует структура словарь, она и будет использоваться. В данной работе ключом является пара слов, а значением — массив слов. Сам алгоритм предельно прост, пройдя циклом по входному массиву до $n - 2$ элемента мы сможем обратиться к $i, i + 1, i + 2$ элементам, что и позволит создавать пары где ключом будет конкатенация $i, i + 1$ слов, а значением $i + 2$ слово. Далее стоит проверить существует ли очередной ключ, если его не существует то добавляем данную пару, иначе добавляем только слово-значение в массив по ключу. Код приведен ниже.

```
from IOPAS import *

def get_prefix(textAsArray: []):
    """
    :param textAsArray: Набор слов через пробел без знаков препинания
    :return: Словарь с префиксами и суффиксами
    """
    dct = {}
    for i in range(len(textAsArray) - 2):
        if textAsArray[i] + ' ' + textAsArray[i + 1] in dct:
            dct[textAsArray[i] + ' ' + textAsArray[i + 1]].append(
textAsArray[i + 2])
        else:
            dct.setdefault(textAsArray[i] + ' ' + textAsArray[i + 1], [
textAsArray[i + 2]])
    WritePrefixesAndSuffixes("test.json", dct)
```

3 Заключение

В ходе данной практики была выполнена задача по ознакомлению с алгоритмами генерации текста, а также реализована часть алгоритма цепей Маркова, отвечающая за поиск префиксов и суффиксов в исходном тексте.

Список литературы

1. Керниган Брайан У., Пайк Роб Практика программирования. - М.: Вильямс, 2015. - 288 с.