Петрозаводский государственный университет
Институт математики и информационных технологий

Отчет о прохождении производственной практики

Выполнил студент группы 22207:
Соболев Евгений Владимирович

Место прохождения практики:
Кафедра информатики и математического обеспечения

Сроки прохождения практики:
29.05.23-08.06.23

Руководитель практики:
к.т.н., доцент
Богоявленская Ольга Юрьевна

Оценка _____

Дата _____

Петрозаводск — 2023

Содержание

Введение

В современном информационном обществе задача генерации текстов на естественном языке имеет большое значение. Она находит применение в различных областях, таких как машинный перевод, чат-боты, создание контента, реферирование, суммаризация и многое другое. С помощью алгоритмов генерации текста на основе корпусов можно создавать новые тексты, имитирующие стиль и содержание исходных данных. Например, такие алгоритмы могут создавать новости, стихи, резюме, отзывы и т.д.

Оптимальный выбор корпуса текста является ключевым фактором для генерации качественных и содержательных текстов. Корпус текста - это набор текстовых данных, который используется для обучения и тестирования алгоритмов генерации текста. Преимуществом этого подхода является его простота и эффективность, а недостатком - ограниченность и однородность сгенерированных текстов. Поэтому важно выбрать подходящий корпус текста, что позволит повысить разнообразие и релевантность генерируемых текстов.

Цель практики: выбор подходящего корпуса текста для дальнейшего применения в алгоритме генерации.

Задачи практики:

1. Проанализировать различные типы и характеристики корпусов текста, которые используются для обучения и тестирования алгоритмов генерации текста на основе корпусов.

2. Оценить качество и представительность разных корпусов текста по определенным критериям.

3. Выбрать подходящий корпус текста для дальнейшего применения в алгоритме генерации.

1 Теоретическая часть

1.1 Обзор существующих типов и характеристик корпусов текста

В этом разделе будут рассмотрены различные типы и характеристики корпусов текста, которые используются для обучения и тестирования алгоритмов генерации текста на основе марковских цепей.

- Типы корпусов текста: существует множество типов корпусов текста, которые отличаются по своей тематике, стилю, жанру и т.д. Например, можно выделить следующие типы корпусов текста:
  - Новостные корпусы: содержат новостные статьи из разных источников и средств массовой информации. Примеры: Reuters Corpus, Gigaword Corpus, News Commentary Corpus.
  - Литературные корпусы: содержат литературные произведения разных авторов и эпох. Примеры: Project Gutenberg, British National Corpus, Russian National Corpus.
  - Научные корпусы: содержат научные статьи и документы из разных областей знаний. Примеры: PubMed Central, arXiv, ACL Anthology.
  - И другие типы корпусов текста, такие как диалоговые корпусы, социальные корпусы, веб-корпусы и т.д.
- Характеристики корпусов текста: для оценки качества и представительности корпусов текста необходимо учитывать различные характеристики, которые характеризуют его качество и представительность. Например, можно выделить следующие характеристики:
  - Размер: это количество слов или других единиц, из которых состоит корпус текста. Размер корпуса текста влияет на его полноту и разнообразие. Чем больше размер корпуса текста, тем больше информации он содержит и тем больше вариантов генерации он предоставляет.
  - Тематика: это область или сфера, к которой относится корпус текста. Тематика корпуса текста влияет на его релевантность и интересность. Чем более специфична и актуальна тематика корпуса текста, тем более целевые и привлекательные будут генерируемые тексты.

– Стиль: это способ выражения мыслей и чувств в корпусе текста. Стиль корпуса текста влияет на его форму и тональность. Чем более определен и узнаваем стиль корпуса текста, тем более выразительные и убедительные будут генерируемые тексты.

– Энтропия: это мера неопределенности или хаоса в корпусе текста. Энтропия корпуса текста влияет на его сложность и непредсказуемость. Чем выше она, тем больше информации он несет и тем меньше шансов повторения генерируемых текстов.

– Коэффициент Жини: это мера неравномерности распределения частот слов в корпусе текста. Коэффициент Жини корпуса текста влияет на его равномерность и сбалансированность. Чем ниже коэффициент, тем более равномерно и сбалансированно он распределяет частоты слов и тем лучше он отражает естественный язык.

– Соотношение количества уникальных слов к общему количеству слов: это мера лексического разнообразия в корпусе текста. Соотношение количества уникальных слов к общему количеству слов влияет на широту и богатство языка в корпусе текста. Чем выше соотношение количества уникальных слов к общему количеству слов, тем больше различных слов и тематик используется в корпусе текста.

– Средняя длина слова: это мера сложности и специфичности лексики в корпусе текста. она влияет на уровень сложности и формальности языка в корпусе текста. Чем выше средняя длина слова, тем сложнее и специфичнее слова используются в корпусе текста.

– Средняя длина предложения: это мера сложности и развернутости синтаксиса в корпусе текста. Она влияет на уровень сложности и развернутости мысли в корпусе текста. Чем выше средняя длина предложения, тем сложнее и развернутее предложения используются в корпусе текста.

– Количество знаков препинания на 1000 слов влияет на структурную и стилистическую разнообразность в корпусе текста. Чем оно выше, тем больше структурных и стилистических средств используется в корпусе текста.

2 Практическая часть

2.1 Методика выбора корпуса текста

Для выбора подходящего корпуса текста для дальнейшего применения в алгоритме генерации на основе марковских цепей была разработана следующая методика:

1. Из общедоступных источников выбирается несколько корпусов текста различной тематики, такие как новостные, литературные, научные.

2. К каждому корпусу текста применяется модуль разбора текста, который позволит получить набор слов, составляющих корпус.

3. Строится статистика суффиксов и префиксов для каждого корпуса текста с помощью словаря. Вычисляется частота появления каждой комбинации префикса и суффикса в корпусе текста.

4. Оценивается качество и представительность каждого корпуса текста по ряду критериев, таких как общее количество слов, количество уникальных слов, соотношение количества уникальных слов к общему количеству слов, средняя длина слова, средняя длина предложения, количество знаков препинания на 1000 слов, энтропия и коэффициент Жини.

5. Сравниваются результаты оценки разных корпусов текста и выбирается наиболее подходящий.

2.2 Модуль statistics.py

Для оценки корпусов текста по этой методике был написан модуль statistics.py на языке Python, который позволяет обрабатывать корпусы текста и получать характеристики корпуса текста.

```
1 import nltk
2 import math
3 from collections import defaultdict, Counter
4 import re
5 import string
6 def parse_text(filename):
7     """
8     Функция для разбора текста на токены.
```

```
9
10      Args:
11          filename (str): Имя файла.
12
13      Returns:
14          list: Список токенов.
15      """
16      with open(filename, 'r') as file:
17          text = file.read()
18          tokens = nltk.word_tokenize(text)
19      return tokens
20
21  def build_suffix_prefix_stats(data):
22      """
23      Функция для построения статистики суффиксов и префиксов.
24
25      Args:
26          data (list): Список токенов.
27
28      Returns:
29          dict: Словарь статистики суффиксов и префиксов в виде {префикс: {суффикс: частота}}.
30      """
31      stats = defaultdict(Counter)
32      for i in range(len(data) - 1):
33          prefix = data[i]
34          suffix = data[i + 1]
35          stats[prefix][suffix] += 1
36
37      last_word = data[-1]
38      stats[last_word][""] += 1
39      return stats
40
41  def calculate_corpus_coefficients(stats, data):
42      """
43      Функция для расчета коэффициентов корпуса текста.
44
45      Args:
46          stats (dict): Словарь статистики суффиксов и префиксов.
47          data: Исходный текст
48
```

```
    Returns:
        dict: Словарь коэффициентов корпуса текста.
    """
    word_count = sum(sum(x.values()) for x in stats.values())
    unique_words = len(stats)
    entropy = calculate_probability(stats, calculate_entropy)
    gini_coefficient = calculate_probability(stats, calculate_gini_coefficient)
    unique_ratio = unique_words / word_count
    word_length = sum(len(x) for x in data) / word_count
    sentence_length = word_count / len(re.split(r'[.!?]+', ' '.join(data)))
    punctuation_count = sum(1 for x in data if x in string.punctuation) / word_count * 1000
    coefficients = {
        "Всего слов": word_count,
        "Уникальных слов": unique_words,
        "Соотношение уникальных слов": unique_ratio,
        "Средняя длина слова": word_length,
        "Средняя длина предложения": sentence_length,
        "Количество знаков препинания на 1000 слов": punctuation_count,
        "Энтропия": entropy,
        "Коэффициент Жини": gini_coefficient
    }
    return coefficients

def calculate_probability(stats, probability_function):
    """
    Функция для расчета вероятности.

    Args:
        stats (dict): Словарь статистики суффиксов и префиксов.
        probability_function (function): Функция для расчета вероятности по списку вероятностей.

    Returns:
        float: Результат расчета вероятности.
    """
    total_count = sum(sum(x.values()) for x in stats.values())
    probabilities = [count / total_count for x in stats.values() for count in x.values()]
    result = probability_function(probabilities)
    return result

def calculate_entropy(probabilities):
```

```
89      """
90      Функция для расчета энтропии.
91
92      Args:
93          probabilities (list): Список вероятностей.
94
95      Returns:
96          float: Энтропия.
97      """
98      entropy = -sum(p * math.log2(p) for p in probabilities)
99      return entropy
100
101 def calculate_gini_coefficient(probabilities):
102     """
103     Функция для расчета коэффициента Жини.
104
105     Args:
106         probabilities (list): Список вероятностей.
107
108     Returns:
109         float: Коэффициент Жини.
110     """
111     gini = 1 - sum(p ** 2 for p in probabilities)
112     return gini
113
114 def main():
115     filename = 'overheard.txt'
116     data = parse_text(filename)
117     stats = build_suffix_prefix_stats(data)
118     coefficients = calculate_corpus_coefficients(stats, data)
119
120     for key, value in coefficients.items():
121         print(f"{key}: {value}")
122
123 if __name__ == "__main__":
124     main()
125
```

Listing 1 – Модуль statistics.py

2.3 Результаты выбора корпуса текста

Для выбора оптимального корпуса текста была применена методика, описанная ранее. В результате были получены данные для корпусов текста:

Характеристика	**Новостной Корпус Русский**	**Литературный Корпус Английский**	**Научный Корпус Английский**
Количество слов	171343	210101	96137
Количество уникальных слов	38107	11619	9292
Коэффициент Жини	0.9998	0.9992	0.9998
Соотношение уникальных слов	0.2224	0.0553	0.0967
Средняя длина слова	5.2405	3.1963	4.8853
Средняя длина предложения	16.7327	10.7678	34.1517
Количество знаков препинания на 1000 слов	136.0779	203.7306	56.5131
Энтропия	16.0351	13.6274	14.9486

Таблица 1 – Результат анализа корпусов текста по разным характеристикам

Характеристика	**Новостной Корпус Русский**	**Литературный Корпус Английский**	**Научный Корпус Английский**
Количество слов	-	+	-
Количество уникальных слов	+	-	-
Коэффициент Жини	-	+	+
Соотношение уникальных слов	+	-	-
Средняя длина слова	+	-	-
Средняя длина предложения	-	-	+
Количество знаков препинания на 1000 слов	-	+	-
Энтропия	+	-	-
Итог	**4**	**3**	**2**

Таблица 2 – Результат анализа корпусов текста по разным характеристикам

По результатам анализа я сделал вывод, что **Новостной Корпус Русский** является наиболее подходящим корпусом текста для дальнейшего использования, так как он имеет:

- Большее количество уникальных слов, что свидетельствует о высокой лексической разнообразности и богатстве корпуса.
- Большее соотношение уникальных слов, что означает, что корпус содержит меньше повторяющихся слов и больше информации.
- Большую среднюю длину слова, что указывает на более сложную и развитую лексику.
- Большую энтропию, что означает, что корпус имеет более высокую степень неопределенности и непредсказуемости, что делает его более интересным и полезным для анализа.

Заключение

В рамках практики было проведено исследование разных типов и характеристик корпусов текста, применяемых для обучения и тестирования алгоритмов генерации текста на основе корпусов. Было оценено качество и представительность трех корпусов текста по определенным критериям, таким как количество слов, количество уникальных слов, коэффициент Жини, соотношение уникальных слов, средняя длина слова, средняя длина предложения, количество знаков препинания на 1000 слов и энтропия. Был выбран подходящий корпус текста для дальнейшего применения в алгоритме генерации.

По результатам анализа был сделан вывод, что **Новостной Корпус Русский** является наиболее подходящим корпусом текста для дальнейшего использования.

Список литературы

1. Corpora Russian News. [Электронный ресурс]. URL: https://wortschatz.uni-leipzig.de/en/download/Russian (дата обращения: 07.06.2023).

2. NLTK Corpora. [Электронный ресурс]. URL: https://www.nltk.org/nltk_data (дата обращения: 07.06.2023).