Московский Государственный Технический Университет имени Н. Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчёт по лабораторной работе № 6 по дисциплине «Проектирование интеллектуальных систем»

Исполнитель:
Саврасов П.А.
Группа ИУ5-24М
«__» _____ 2021 г.

Преподаватель:
Терехов В.И.
_____
«__» _____ 2021 г.

Москва, 2021 г.

1. Цель работы

Освоить методы предобработки текста для обучения нейросетей. Научиться использовать нейронные сети для классификации тектсов.

2. Предобработка текста

```
import numpy as np
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.datasets import fetch_20newsgroups
```

```
categories = ['comp.sys.ibm.pc.hardware', 'rec.sport.baseball', 'sci.med', 'alt.atheism']
ngAll = fetch_20newsgroups(subset='all', categories=categories)
print('text', ngAll.data[0])
print('category:', ngAll.target[0])
```

```
text From: sac@asdi.saic.com (Steve A. Conroy x6172)
Subject: Re: Darrrrrrrrryl
Organization: SAIC
Lines: 33

In article <mssC5KCru.5Ip@netcom.com>, mss@netcom.com (Mark Singer) writes:
|>
|>
|> The media is beating the incident at Dodger Stadium on Wednesday to
|> death, but I haven't seen anything in rsb yet.
|>
```

Для подготовки данных воспользуемся функцией векторизации текстов CountVectorizer из библиотеки scikit-learn модуля sklearn.feature_extraction.text

```
vectorizer = TfidfVectorizer()
data = vectorizer.fit_transform(ngAll.data).toarray()
inputDim = len(vectorizer.get_feature_names())

xTrain = data[:2500]
yTrain = ngAll.target[:2500]
xTest = data[2500:]
yTest = ngAll.target[2500:]
```

3. Обучение нейронной сети:

Параметры обучения

```
learning_rate = 0.01
training_epochs = 10
batch_size = 150
```

Параметры нейронной сети

```
n_hidden_1 = 500 # скрытый слой 1
n_hidden_2 = 300  # скрытый слой 2
n_hidden_3 = 100  # скрытый слой 3
n_classes = len(categories)
```

```
model = tf.keras.Sequential(
    [
        layers.InputLayer(input_shape=(inputDim)),
        layers.Dense(n_hidden_1, activation='relu'),
        layers.Dense(n_hidden_2, activation='relu'),
        layers.Dense(n_hidden_3, activation='relu'),
        layers.Dense(n_classes, activation='relu')
    ]
)
```

```
model.compile (
    loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
    optimizer=keras.optimizers.Adam(learning_rate=learning_rate),
    metrics=['accuracy']
)
```

```
model.fit(xTrain, yTrain, batch_size=batch_size, epochs=training_epochs, verbose=2)
model.evaluate(xTest, yTest, verbose=2)
```

```
Epoch 1/10
17/17 - 2s - loss: 0.8732 - accuracy: 0.6064
Epoch 2/10
17/17 - 2s - loss: 0.6783 - accuracy: 0.7212
Epoch 3/10
17/17 - 1s - loss: 0.6695 - accuracy: 0.7232
Epoch 4/10
17/17 - 2s - loss: 0.6664 - accuracy: 0.7240
Epoch 5/10
17/17 - 1s - loss: 0.6662 - accuracy: 0.7248
Epoch 6/10
17/17 - 2s - loss: 0.6651 - accuracy: 0.7240
Epoch 7/10
17/17 - 2s - loss: 0.6650 - accuracy: 0.7256
Epoch 8/10
17/17 - 2s - loss: 0.6638 - accuracy: 0.7260
Epoch 9/10
17/17 - 2s - loss: 0.6642 - accuracy: 0.7260
Epoch 10/10
17/17 - 1s - loss: 0.6638 - accuracy: 0.7260
40/40 - 1s - loss: 0.7055 - accuracy: 0.7233

[0.7054914832115173, 0.7233201861381531]
```

4. Выводы по работе:

Освоил методы предобработки текста для обучения нейросетей. Научился использовать нейронные сети для классификации тектсов.

5. Контрольные вопросы

5.1. Какие вы знаете задачи обработки текстов, в чем они заключаются?

Классификация текстов и предложений
Машинный перевод
Интеллектуальная обработка текстов
Генерация описаний по входному объекту
Автоматическая обработка текстов

5.2. Зачем нужна предобработка текста для машинного обучения?

Это нужно для использования текстовых данных в нейронных сетях их нужно представить в виде матрицы, каждый элемент будет соответствовать словам.

5.3. Какие виды предобработки текста вы знаете?

Все тексты на естественном языке имеют большое количество слов, которые не несут информации о данном тексте. К примеру, в английском языке такими словами являются артикли, в русском к ним можно отнести предлоги, союзы, частицы. Данные слова называют шумовыми или стоп-словами. Для достижения лучшего качества классификации на первом этапе предобработки текстов обычно необходимо удалять такие слова. Второй этап предобработки текстов — приведение каждого слова к основе, одинаковой для всех его грамматических форм. Это необходимо, так как слова несущие один и тот же смысл могут быть записаны в разной форме. Например, одно и то же слово может встретиться в разных склонениях, иметь различные приставки и окончания.

5.4. Что такое стемминг?

Стемминг — это процесс нахождения основы слова для заданного исходного слова. Основа слова не обязательно совпадает с морфологическим корнем слова. Задача нахождения основы слова представляет собой давнюю проблему в области компьютерных наук. Первая публикация по данному вопросу датируется 1968 годом. Стемминг применяется в поисковых системах для расширения поискового запроса пользователя, является частью процесса нормализации текста.

5.5. Что такое 20 Newsgroups?

20 Newsgroups – это набор, состоящий из примерно 20 тысяч постов.

5.6. Чему должно равняться число входных и выходных нейронов в задаче классификации текстов?

Размер входного слоя должен быть равен количеству уникальных слов в текстах, а размер выходного слоя — количеству классов.