А. В. Долгошей
(*ГрГУ им. Я. Купалы, Гродно*)

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБРАБОТКИ ТЕКСТА

Данная работа посвящена разработке системы для анализа и коррекции текстовой документации с использованием нейронных сетей. Решение проблемы сбора и накопления данных для обучения яв-

ляется задачей многомерной оптимизации в большинстве моделей нейронных сетей. Для классификации и выделении объектов необходимо иметь достаточно большой и хорошо сформированных набор исходных данных.

Актуальность темы обусловлена необходимостью разработки и внедрения модуля нейронной сети для анализа и коррекции специальной текстовой документации.

Благодаря таким программным средствам как Pandas, разрабатывается приложение для обработки текстовых документов. При написании данной программы используется такая библиотека, как Scikit-Learn, в которой реализовано большое количество алгоритмов машинного обучения.

Для начала обучения нам понадобилась выборка с большим количеством схожей документации, благодаря которой будет проходить обучение с помощью модели типа «обучение с учителем» (контролируемого обучения). Далее мы применим встроенный загрузчик наборов данных для выборки из scikit-learn. Чтобы мы смогли использовать машинное обучение на текстовой документации, необходимо перевести всё текстовое содержимое в числовой вектор признаков.

Такая статическая мера, как TF-IDF (Term Frequency, Inverse Document Frequency), используется для оценки важности слова в контексте текстового документа. Каждому слову будет присвоен определённый вес, который будет считаться по частоте употребления в данном тексте. Term frequency (частота слова) можно посчитать по формуле:

$$tf(t,d) = n_t / \sum_k n_k ,$$

где числитель – число вхождений слова t в текст, знаменатель – общее число слов в тексте. Inverse document frequency (обратная частота документа) можно рассчитать по следующей формуле:

$$idf(t, \mathrm{D}) = \log(\mathrm{D}/\mathrm{D}_t),$$

где числитель – число документов в коллекции, знаменатель – число документов из коллекции, в которых встречается слово t, если $\mathrm{D}_t \neq 0$.

Слова с наименьшими весами будут подвергаться сомнению и могут выделяться как ошибочные. Это позволит найти возможные ошибки в текстовых документах.