Доклад

Слайд 1:

Здравствуйте, уважаемые члены комиссии. Меня зовут Динь Вьет Ань из группы ИУ7И-84Б. Вашему вниманию представляется научно-исследовательская работа на тему: **«Метод классификации новостных текстов по тематике с помощью метода опорных векторов»**

Слайд 2:

Целью работы является разработка метода классификации новостных текстов по тематике с помощью метода опорных векторов. Для достижения поставленной цели необходимо выполнить следующие задачи:

+ проанализировать предметную область и основные подходы к классификации текстов;

+ разработать метод классификации новостных текстов по тематике с помощью метода опорных векторов;

+ разработать программное обеспечение, реализующее данный метод;

+ провести оценку качества классификации текстов.

Слайд 3:

Текст - зафиксированная на каком-либо материальном носителе человеческая мысль, в общем плане связная и полная последовательность символов. Классификация текста — это процесс присвоения предопределенной категории или метки предложениям, абзацам, текстовым отчетам или неструктурированного текста. Классификация новостей является одной из задач классификации текста. Это процесс определения категории или темы новости на основе ее содержания.

Слайд 4:

Для классификации текстов существует несколько основных методов. У каждого из них свои преимущества и недостатки, которые показаны на слайде.

Слайд 5:

На этом слайде показана функциональная схема обучения классификатора в виде IDEF0-диаграммы. Есть всего 4 этапа: Создание обучающей и тестовой выборки, очистка и предобработка текстов, извлечение признаков и многоклассовая классификация с помощью метода SVM.

Слайд 6:

Прежде всего, набор данных необходимо очистить и предварительно обработать. Очистка текстов включает в себя преобразование текста в нижний регистр, удаление лишних пробелов и не буквенно-цифровых символов. А предобработка текстов состоит из 3 процесса: токенизации, удаления стоп-слов и лемматизация.

Слайд 7:

Для извлечения численных значений признаков из текста используется мера TF-IDF. Это статистическая мера, используемая для оценки важности слова в контексте текста. TF – отношение числа вхождений некоторого слова к общему числу слов документа, а IDF – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Значения вычислены по формуле на слайде.

Слайд 8:

Для обучения классификатора было решено использовать набор данных «Dataset from Lenta.Ru». Этот набор данных содержит более 800 тысяч новостей на русском языке, соответствующих более чем 20 тематикам, с сентября 1999 года по декабрь 2019 года. Источник новостей – сайт lenta.ru , российское новостное интернет-издание, основанное в 1999 году. В наборе данных для каждой новости сохраняются текст, заголовок, тематик, дата публикации и ссылка на источник. Выборки, созданные из этого набора данных, состоит из 5 тематик: Наука и технологии, Спорт, Экономика, Культура и Мир. Количество текстов по каждой тематике показано в таблице на слайде.

Слайд 9:

На этом слайде показан интерфейс пользователя программы. Чтобы начинать работать с программой нужно выбрать файл для обучения классификатора. Эти файлы созданы из набора данных, упомянутого на предыдущем слайде, и могут быть обновлены с помощью кнопки «Обновить выбранный файл». После этого пользователь может ввести текст в поле ввода или выбрать текстовый файл. Программа поддерживает только txt файлы (текстовые файлы). При нажатии на кнопку «Выбрать файл» открывается файловая система компьютера, в которой необходимо выбрать интересующий текстовый файл. Когда пользователь выбирает файл, его содержимое отображается в текстовом поле. Когда классификатор завершит работать, на экран будет выведено название наиболее вероятного тематика, как показано на слайде.

Слайд 10:

Для оценки качества классификатора, обученного на обучающей выборках из набора данных, используются следующие две метрики: аккуратность и F1-мера. Эти метрики вычислены по формулам, показанным на слайде.

Слайд 11

Было произведено исследование зависимости времени обучения от количества текстов в выборке. На этом слайде показан результат исследования. Можно сделать вывод, что время обучения быстро увеличивается при увеличении количества текстов в обучающей выборке.

Слайд 12:

На этом слайде показан результат исследования зависимости качества классификатора от количества текстов в выборке. Точность быстро увеличивается при увеличении количества текстов в обучающей выборке, но при большем количестве текстов ускорение незначительно (при увеличении количества текстов с 1000 до 5000 текстов время увеличивается примерно в 27 раз, но точность увеличивается на 2%).

Слайд 13:

Следующее исследование - влияние ядер метода опорных векторов на качество классификатора. Из полученных результатов можно сделать вывод, что зависимости точности и F1-меры классификатора от различных ядер совпадают. Классификатор с полиномиальным ядром имеет наименьшее качество, а классификаторы с другими ядрами имеет примерно одинаковую точность.

Слайд 14:

В заключение поставленная цель была достигнута, все задачи были решены.

Слайд 15:

Направление дальнейшего развития можно рассмотреть ускорение работы метода и добавление возможности работы с различными языками.

Слайд 16:

Доклад окончен, спасибо большое за внимание!