Федеральное государственное бюджетное образовательное учреждение
высшего образования
Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

УТВЕРЖДАЮ
декан факультета вычислительной
математики и кибернетики
_____ / И.А. Соколов /
«27» сентября 2023г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине

Основы обработки текстов

Уровень высшего образования:

бакалавриат

Направление подготовки / специальность:

02.03.02 "Фундаментальная информатика и информационные технологии" (3++)

Направленность (профиль) ОПОП:

Искусственный интеллект и анализ данных

Форма обучения:

очная

Рассмотрен и утвержден
на заседании Ученого совета факультета ВМК
(протокол №7, от 27 сентября 2023 года)

Москва 2023

1. ФОРМЫ И ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

В процессе и по завершении изучения дисциплины оценивается формирование у студентов следующих компетенций:

Планируемые результаты обучения по дисциплине (модулю)		
Содержание и код компетенции.	**Индикатор (показатель) достижения компетенции**	**Планируемые результаты обучения по дисциплине, сопряженные с индикаторами достижения компетенций**
ПК-9. Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	ПК-9.1. Участвует в реализации проектов в области сквозной цифровой субтехнологии «Компьютерное зрение» ПК-9.2. Участвует в реализации проектов в области сквозной цифровой субтехнологии «Обработка естественного языка»	ПК-9.2. З-1. Знает принципы построения систем обработки естественного языка, методы и технологии искусственного интеллекта для анализа естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка» ПК-9.2. У-1. Умеет применять методы и подходы к планированию и реализации проектов по созданию и поддержке системы искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка»

1.1. Текущий контроль успеваемости

Текущий контроль успеваемости осуществляется путем оценки результатов выполнения заданий практических (семинарских) занятий, самостоятельной работы, предусмотренных учебным планом и посещения занятий/активность на занятиях.

В качестве оценочных средств текущего контроля успеваемости предусмотрены:

решение индивидуальных заданий

Примеры индивидуальных заданий

Постановка задачи

Целью работы является создание системы, позволяющей оценивать эмоциональную окраску сообщений микроблога Twitter. Задача анализа эмоциональной окраски (или тональности) текста (англ. sentiment analysis или opinion mining) состоит в автоматическом выявлении в текстах эмоциональной оценки автора по отношению к некоторому объекту.

Примеры:

- "Начинается новый день" - нейтральная оценка
- "Какой прекрасный день" - позитивная
- "Ужасный день" - негативная

Предлагается разработать систему, которая на вход получает короткий текст (твит), а на выходе отдает одну из трех меток:

- neutral - нейтральная
- positive - позитивная
- negative - негативная

Решение задачи

Практические аспекты

Решения должны быть написаны на языке Python. Можно использовать все стандартные библиотеки, а также

- NLTK - инструменты для обработки текстов
- scikit-learn - алгоритмы машинного обучения
- numpy - работа с многомерными массивами

Доступ в Интернет на проверяющей машине закрыт.

Теоретические аспекты

Предполагается использование алгоритмов машинного обучения. Для обучения алгоритма требуется придумать признаки и дать ему на вход правильные примеры - обучающий корпус. Считается, что чем больше обучающий корпус, тем лучше работает алгоритм. Однако создание большого обучающего корпуса - довольно трудоемкая задача, непосильная одному человеку. Поэтому предлагается создать его с помощью коллективной работы. Чтобы облегчить эту работу, был сделан сайт: http://markup.at.ispras.ru.

Разметка обучающего корпуса

Для разметки корпуса необходимо зарегистрироваться на сайте http://markup.at.ispras.ru. Пожалуйста, вводите правильные данные, так как они будут использоваться при выставлении зачетов. Вне рамок практикума эти данные использоваться не будут.

После регистрации появится окно с тремя полями (Рис 1).

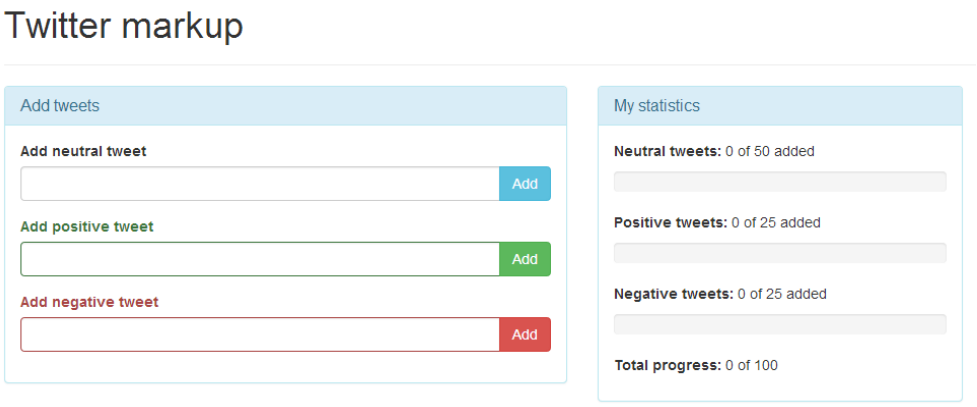

Рис 1. Форма для ввода твитов

Далее вы идете на сайт twitter.com выбираете любых пользователей, которые пишут **на русском языке**. Ищете у них сообщения, которые содержат эмоциональную окраску (либо однозначно позитивную, либо однозначно негативную) или не содержат никаких эмоций (например, констатация факта). Копируете эти сообщения в соответствующие поля формы и нажимаете add (enter тоже работает). Если есть сомнение в передаваемых эмоциях, лучше пропустить твит (см. следующий раздел). После того, как будет размечено не менее 25 позитивных, 25 негативных и 50 нейтральных сообщений, появится кнопка, позволяющая скачать размеченные твиты (см. раздел "тренировочный корпус").

Под формой ввода твитов будут появляться все твиты, размеченные и вашими коллегами. Эмоциональная окраска будет отмечена разными цветами (зеленый - позитивная, красный - негативная, серый - нейтральная). Вы можете оценить правильность разметки, нажав соответствующую кнопку рядом сообщением. Эта информация будет доступна в файле с тренировочным корпусом и ее можно использовать при обучении.

Какие твиты стоит размечать

Для разметки следует использовать твиты на русском языке. То есть твиты могут содержать иностранные названия, но основная часть текста должна быть представлена на русском языке кириллическими буквами.

ни дня без хорошей новости - recent studies suggest that sexual activity causes neurogenesis in the hippocampus.	Этот твит не считается твитом на русском языке. Такие твиты добавлять не стоит.
Мне понравилось видео "Anime | THE SIMPSONS | ANIMATION on FOX" (http://youtu.be/R94Q6NhuS3A?a) на @YouTube.	Этот твит нельзя считать твитом на русском языке, поскольку слов на английском языке значительно больше, чем на русском.
Жители Сан-Франциско атаковали автобус Google http://bit.ly/JZ9iZF	Этот твит можно считать твитом на русском языке, поскольку он содержит небольшое количество иностранных слов, которые являются именами собственными

При оценке эмоциональной окраски твита следует учитывать только субъективное мнение **автора текста** (твита) по отношению к описываемому объекту/явлению. Твиты могут содержать как явную эмоциональную окраску, так и не явную. Не следует путать печальные сообщения (по смыслу) с негативной эмоциональной окраской сообщения.

Фотосессии на ВМК всегда весёлые! http://instagram.com/p/jwm9GKx3Q5/	Это сообщение имеет явную положительную эмоциональную окраску. Такие твиты стоит добавлять.
Кочкин становится свидетелем того, как кортеж олигарха сбивает женщину, переходящую дорогу. #photo	Это эмоционально нейтральное сообщение. Несмотря на то, что описывает оно печальное событие.
Никогда ничего не покупайте в магазине Pleer.ru http://j.mp/KKiqSe	Это сообщение имеет неявную отрицательную эмоциональную окраску.

Не стоит добавлять твиты, содержащие одновременно и негативную и позитивную оценку каких-либо объектов

С одной стороны, я бы не хотела жить одна - скучно и одиноко, но с другой стороны - чистота в доме, свобода, не нужно готовить - рай.	Это сообщение содержит в себе две эмоционально окрашенные части: первая – отрицательная, вторая – положительная. Однозначно трактовать эмоциональную окраску всего предложения не возможно. Такие твиты добавлять не стоит.

Не стоит добавлять твиты, содержащие сарказм:

Члены партии единой России обладают великим искусством правильно подобрать варианты ответа к опросу http://er.ru/poll/5.html/	Сарказм. Без дополнительных знаний о контексте этого сообщения невозможно определить положительное оно или отрицательное. Такие сообщения добавлять не стоит.

Рекомендуется размечать максимально честно, так как от этого будет зависеть качество всех классификаторов. Если есть сомнения, к какому классу лучше отнести сообщение, то его стоит пропустить.

Тренировочный корпус

Тренировочный корпус будет доступен для скачивания в формате json. Для извлечение информации из этого файла рекомендуется использовать стандартную библиотеку Python с одноименным названием.

Для синхронизации обучения и тестирования в течении недели, корпус будет состоять из твитов, размеченных автором классификатора, плюс все твиты, размеченные в течении предшествующей недели.

Тестирование

Вместе с кнопкой скачивания тренировочного корпуса появится ссылка на форму для загрузки файла и личную страницу со статистикой. На личной странице находится статистика со всеми результатами в т.ч. результатами последнего тестирования (дата, описание, достоверность).

Загрузка решения. Загружаемый файл должен представлять собой zip архив с любым именем. Архив должен обязательно содержать:

- классификатор в файле SentimentAnalyzer.py. В файле должен содержаться класс SentimentAnalyzer. В классе должны присутствовать методы
    - train(self, training_corpus), где training_corpus - это список пар (text, class). Внимание: метод train будет вызываться отдельно, так что не стоит вызывать его в конструкторе класса.
    - getClasses (self,texts), который получает на вход список текстовых сообщений и возвращает список ответов классификатора. (Пример: [neutral, positive, positive, ...])
- описание применяемых алгоритмов в файле description.txt
- все используемые внешние библиотеки, кроме библиотек пакетов NLTK, scikit-learn и numpy (они доступны автоматически).

Результаты тестирования появятся на личной странице, как только закончится обучение и тестирование. При загрузке нового классификатора обучение будет производится на корпусе из твитов, размеченных автором классификатора, плюс все твиты, размеченные в течении предшествующей загрузке недели.

В течении недели студенты не видят прогресс своих коллег и могут посмотреть только свой результат. В конце каждой недели (каждый вторник в 23.59.59) будет производится переобучение последнего присланного решения от каждого студента на новом корпусе, а результаты тестирования будут показаны в сводной таблице.

Ограничения

1. каждую неделю можно послать только 10 версий программы (внимание! Итоговое тестирование будет проводится на последнем загруженном решении)
2. размер архива не может превышать 15Мб

В связи с первым ограничением, для тестирования на локальной машине рекомендуется использовать метод перекрестной проверки (http://en.wikipedia.org/wiki/Cross-validation_(statistics)). В библиотеке scikit-learn есть функции, которые могут помочь в использовании этого метода. Рекомендуется использовать метод StratifiedKFold().

Оценка качества

Для оценки качества используются метра достоверности (accuracy), которая равна отношению количества правильных ответов к общему количеству примеров в тестовой выборке.

$$\text{accuracy} = \frac{correct\ answers}{total\ questions}$$

Описание в документации к библиотеке scikit-learn: http://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

Baseline

Baseline 1. В качестве нижней границы используется классификатор, который дает всегда ответ "neutral". Достоверность этого метода равна 0.5.

Baseline 2. В качестве второй, более сложной нижней границы используется один из стандартных алгоритмов классификации с N-граммами в качестве признаков. Этот классификатор будет тренироваться на том же корпусе, что и присланные алгоритмы, и его достоверность будет меняться соответственно.

1.2. Промежуточная аттестация

Промежуточная аттестация осуществляется в форме экзамена
В качестве средств, используемых на промежуточной аттестации предусматривается:
Билеты

1.3. Типовые задания для проведения промежуточной аттестации

Вопросы к экзамену

Вопросы к экзамену.

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната
2. Регулярные выражения
3. Конечные автоматы, распознавание языка с помощью КА
4. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений
5. Методы поиска словосочетаний. Общая схема. Методы поиска кандидатов
6. Методы поиска словосочетаний. Проверка статистических гипотез
7. Модель N-грамм. Оценка вероятности высказывания
8. Модель N-грамм. Сглаживание (Лапласа и Откат)
9. Модель N-грамм. Оценка качества. Тренировочный и проверочный корпуса
10. Задача определения частей речи. Существующие подходы. Алгоритмы, основанные на правилах. Алгоритмы, основанные на трансформации.
11. Использование скрытой марковской модели для определения частей речи.
12. Скрытые марковские модели. Вероятность последовательности. Прямой алгоритм
13. Скрытые марковские модели. Наиболее правдоподобное объяснение. Алгоритм Витерби
14. Модели классификации. Наивный байесовский классификатор
15. Модели классификации. Логистическая регрессия
16. Модели классификации. Модель максимальной энтропии
17. Модели классификации. Марковская модель максимальной энтропии
18. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика
19. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев.
20. Синтаксический разбор. Разбор сверху вниз и снизу вверх
21. Синтаксический разбор. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик
22. Синтаксический разбор. Группировка (chunking)
23. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности
24. Моделирование языка. Обучение стохастических КС грамматик
25. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества
26. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества
27. Лексическая семантика. WordNet. Значения слов
28. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы оценки качества
29. Разрешение лексической многозначности. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества
30. Семантическая близость слов. Подходы на основе тезаурусов. Методы оценки качества
31. Семантическая близость слов. Подходы на основе статистик. Методы оценки качества
32. Информационный поиск. Ранжирование документов. Векторная модель. Взвешивание терминов. TF-IDF
33. Информационный поиск. Индексирование. Инвертированный индекс. Запросы с джокером. Исправление опечаток.

34. Вопросно-ответные системы. Общая архитектура. Обработка запроса
35. Вопросно-ответные системы. Общая архитектура. Извлечение фрагментов текста
36. Вопросно-ответные системы. Общая архитектура. Обработка ответа
37. Автоматическое реферирование. Общая архитектура
38. Машинный перевод. Классические подходы
39. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование
40. Статистический машинный перевод. Выравнивание слов. Модель IBM Model 1
41. Статистический машинный перевод. Выравнивание слов. Тренировка моделей выравнивания
42. Статистический машинный перевод. Методы оценки качества. BLUE

Пример экзаменационного билета

1. Методы поиска словосочетаний. Общая схема. Методы поиска кандидатов
2. Автоматическое реферирование. Общая архитектура

2. КРИТЕРИИ ОЦЕНКИ ПО ДИСЦИПЛИНЕ

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	**2 (не зачтено)**	**3 (зачтено)**	**4 (зачтено)**	**5 (зачтено)**
Знания **(виды оценочных средств: приведены в п. 1.2.)**	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения **(виды оценочных средств: приведены в п. 1.2.)**	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки **(владения, опыт деятельности)** **(виды оценочных средств: приведены в п. 1.2..)**	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач