МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

«Утверждаю»
Заведующий кафедрой ТО и ЗИ


05.07.2018 г.

А.А. Сирота

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.ДВ.3.1. Компьютерная лингвистика

1. Шифр и наименование направления подготовки/специальности:
45.03.03 Фундаментальная и прикладная лингвистика

2. Профиль подготовки/специализации:-

3. Квалификация (степень) выпускника:бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: кафедра Технологий обработки и защиты информации

6. Составители программы: Гаршина Вероника Викторовна, канд.тех.наук, доцент кафедры Технологий обработки и защиты информации

7. **Рекомендована:** Научно-методическим советом ФКН, протокол № 6 от 25.06.2018 г.

8. Учебный год: 2020/2021 **Семестр(-ы):**5

9. Цели и задачи учебной дисциплины:Изучить формальные, программно-реализуемые подходы к изучению структур и закономерностей естественных языков, ознакомится с основными прикладными практическими задачами компьютерной лингвистики

Основные задачи дисциплины:

- Изучить основные принципы иметодов обработки естественного языка. Получение навыков разработки и интеграции программных систем обработки естественного языка в программные продукты.
- Знакомство с принципами проектирования программных систем, ориентированных на обработку естественно-языковых текстов.

10. Место учебной дисциплины в структуре ООП:дисциплинаБ1.Б.28 Технологии обработки текста и звучащей речи входит вбазовую часть ООП. Для изучения дисциплины необходимы знания, умения и компетенции, сформированные дисциплин: Б1.Б.15 Введение в теорию языка, Б1.Б.27 Общая и компьютерная лексикография,Б1.Б.26Технологии корпусной лингвистики, Б1.Б.14 Информатика и основы программированияБ1.В.ОД.2 Проектирование баз данных,

11. Планируемые результаты обучения по дисциплине (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями выпускников):

Компетенция		Планируемые результаты обучения
Код	Название	
ПК-2	Владение основными методами инструментального анализа звучащей речи	знать: цели и задачи теоретической и практической фонетики уметь: сопоставлять фонетические факты английского и родного языков; делать фонетический анализ, объяснять фонетические явления, ис- пользовать теоретические знания о фонетической системе английского языка на практике для решения конкретных лингвистических задач владеть (иметь навык(и)): навыками фонологического анализа; фонетической терминологией
ПК-5	Владение основными способами описания и формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации,	Знать: методы описания денотативной, концептуальной, коммуникативной и прагматической информации Уметь: использовать лингвистически-ориентированные программные системы. Владеть:

	содержащейся в тексте на естественном языке	основами дисциплин, необходимых для формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур.

12. Объем дисциплины в зачетных единицах/часах в соответствии с учебным планом — 3ЗЕТ / 108 час.

Форма промежуточнойаттестациизачет.

13. Виды учебной работы:

Вид учебной работы	Трудоемкость (часы)			
	Всего	По семестрам		
		№ сем.5	№ сем.	…..
Аудиторные занятия	38	38		
в том числе: лекции				
практические	0	0		
лабораторные	38	38		
Самостоятельная работа	70	70		
Итого:	108	108		
Форма промежуточной аттестации (зачет)				

13.1Содержание дисциплины:

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины
1. Лекции		
Не предусмотрены учебным планом		
2. Практические занятия		
Не предусмотрены учебным планом		
3. **Лабораторные занятия**		
3.1	Задачи компьютерной лингвистики в изучении естественного языка (ЕЯ).	История научного направления. Классификация языков Хомского. Особенности ЕЯ. Знакомство с направлениями современных исследований.
3.2	Лингвистический процессор. Алгоритмы лингвистического разбора и анализа текста. ПарсерыЕЯ-предложений.	Лингвистический процессор - функциональная структура. Методы морфологического анализа, используемые в лингвистических процессорах. Морфологические словари. Алгоритмы синтаксического и семантического анализа для автоматических систем обработки текстов. Парсеры ЕЯ. Прикладные системы-спэлчекеры, текстовые редакторы, системы профессионального редактирования.
3.3	Формальные методы исследования структуры ЕЯ текста	Статистические методы анализа структур ЕЯ текста на морфологическом, синтаксическом, семантическом уровнях. Метод позиционных статистик. Приложение методов для задач дешифровки ЕЯ текстов на

		неизвестных языках. Марковские цепи.
3.4	Формальные методы классификации полнотекстовых документов	Математическая постановка задачи распознавания образов и классификации. Формальные методы определения сходства ЕЯ документов на различных уровнях лингвистического анализа (морфологическом, синтаксическом, семантическом): кластерный анализ, деревья принятия решений, векторные методы, Байесовский классификатор. Применение методов классификации для задач определения авторства текстов.
3.5	Проблемы построения систем семантического анализа текстов (TextMining)	Автоматическое извлечение знаний изЕЯ текстов. Формирование онтологии предметной области по тексту. Построение семантической модели текста. Семантическая классификация и кластеризация текстов.

13.2 Темы (разделы) дисциплины и виды занятий:

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Задачи компьютерной лингвистики в изучении естественного языка (ЕЯ).			4	10	14
2	Лингвистический процессор. Алгоритмы лингвистического разбора и анализа текста. ПарсерыЕЯ-предложений.			10	16	26
3	Формальные методы исследования структуры ЕЯ текста			6	14	20
4	Формальные методы классификации полнотекстовых документов			8	18	26
5	Проблемы построения систем семантического анализа текстов (TextMining)			10	12	22
	Итого:			38	70	108

14. Методические указания для обучающихся по освоению дисциплины:

1) При изучении дисциплины рекомендуется использовать следующие средства**:**

- рекомендуемую основную и дополнительную литературу;
- методические указания и пособия;
- контрольные задания для закрепления теоретического материала;
- электронные версии учебников и методических указаний для выполнения лабораторно - практических работ (при необходимости материалы рассылаются по электронной почте).

2) Для максимального усвоения дисциплины рекомендуется проведение письменного опроса (тестирование, решение задач) студентов по материалам лекций и лабораторных работ. Подборка вопросов для тестирования осуществляется на основе изученного теоретического материала. Такой подход позволяет повысить мотивацию студентов при конспектировании лекционного материала.

3) При проведении лабораторных занятий обеспечивается максимальная степень соответствия с материалом лекционных занятий и осуществляется экспериментальная проверка методов, алгоритмов и технологий обработки информации, излагаемых в рамках лекций.

13. Перечень основной и дополнительной литературы, ресурсов Интернет, необходимых для освоения дисциплины:

а) основная литература:

№ п/п	Источник
1	Кипяткова И.С., Ронжин А.Л., Крапов А.А. Автоматическая обработка разговорной русской речи. - Санкт-Петербург: ГУАП, 2013.
2	Фролов А.В. Фролов Г.В. Синтез и распознавание речи. Современные решения.- М.: Связь, 2003.
3	Лукашевич Н. В. Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011
4	Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
5	Леонтьева Н.Н. Автоматическое понимание текстов. М., 2006

б) дополнительная литература:

№ п/п	Источник
1	Белоногов Г.Г.Компьютерная лингвистика и перспективные информационные технологии.- М.:Русский мир, 2004.
2	Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. - М.: Академия, 2004.
3	Всеволодова А.В. Компьютерная обработка лингвистических данных. М.:Наука, 2007.
4	Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Академия, 2006.
5	Добров Б.В. Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие / Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, В.Д. Соловьев. / - М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2009.
1	Автоматическая обработка текстов на естественном языке и анализ данных : учеб.пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с
2	Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб.пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
3	Карпов А. А., Кипяткова И. С., Ронжин А. Л. Проектирование речевых интерфейсов для информационно-управляющих систем : учебное пособие.- С.-Петерб. гос. ун-т аэрокосм. приборостроения. - Санкт-Петербург : ГУАП, 2012.
4	Сегаран. Т. «Программируем коллективный разум» Символ-Плюс, 2008 год, 368 с.
5	Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Академия, 2006.
6	Кипяткова И.С., Ронжин А.Л., Крапов А.А. Автоматическая обработка разговорной русской речи. - Санкт-Петербург: ГУАП, 2013.
7	Всеволодова А.В. Компьютерная обработка лингвистических данных. М.:Наука, 2007.
8	Белоногов Г.Г.Компьютерная лингвистика и перспективные информационные технологии.- М.:Русский мир, 2004.
9	Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. - М.: Академия, 2004.
10	Марчук Ю. Н. Компьютерная лингвистика.Учебник для Вузов. АСТ-2007, 320 с.
11	Марчук Ю. Н. Компьютерная лингвистика. Учебник для Вузов. АСТ-2007, 320 с.

12	Люггер Дж. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем / Дж. Ф. Люггер. – М. : Вильямс , 2003.
13	Рассел С., Норвиг П. Искусственный интеллект: современный подход. – М.: Вильямс , 2006.
14	Б. В. Костров, В. Н. Ручкин, В. А. ФулинИскусственный интеллект и робототехника Издательство: Диалог-МИФИ, 2008 г.
15	И. М. Макаров, В. М. Лохин, С. В. Манько, М. П. Романов. Искусственный интеллект и интеллектуальные системы управления.- М.: Наука, 2006 г.

в)базы данных, информационно-справочные и поисковые системы**:**

№ п/п	Источник
1	Электронный каталог Научной библиотеки Воронежского государственного университета. – (http // www.lib.vsu.ru/).
2	Образовательный портал «Электронный университет ВГУ».– (https://edu.vsu.ru/)
3	ЭБС «Издательства «Лань», Договор №3010-06/71-14 от 25.11.2014, ЭБС «Университетская библиотека online», Договор №3010-06/70-14 от 25.11.14, Национальный цифровой ресурс «РУКОНТ», Договор №ДС-208 от 01.02.2012
4	Международная конференция по компьютерной лингвистике. http://www.dialog-21.ru/
5	Лаборатория компьютерной лингвистики Института проблем передачи информации РАН. http://proling.iitp.ru/
6	Лаборатория общей компьютерной лексикологии и лексикографии МГУ. http://www.philol.msu.ru/~lex/library.htm
7	Научно-практический журнал РЕЧЕВЫЕ ТЕХНОЛОГИИ http://speechtechnology.ru/

16. Перечень учебно-методического обеспечения для самостоятельной работы:

№ п/п	Источник
1.	Боярский К. К. Введение в компьютерную лингвистику. Учебное по-собие. – СПб: НИУ ИТМО, 2013. – 72 с.

17. Информационные технологии, используемые для реализации учебной дисциплины, включая программное обеспечение и информационно-справочные системы (при необходимости):

Для реализации учебного процесса используются:

1) ПО Microsoft в рамках подписок «Imagine»,ежегодные сублицензионные договоры № 56035/BPH3739 и № 56036/BPH3739 от 07.10.2016.

3) Персер русского языка ТОМИТА (Свободно-распространяемое ПО)

4) Язык программирования Python,IDEPysharm.

4) ПОРедактор онтологий и фреймворк для построения баз знаний Protege. Свободно-распространяемое ПО.

18. Материально-техническое обеспечение дисциплины

Мультимедийная лекционная аудитория, персональный компьютер (ПК) рабочее место преподавателя: проектор, видеокоммутатор, персональные компьютеры (ПК), наушники и микрофоны по числу студентов, специализированная мебель: доска меловая., столы,стулья; выход в Интернет, доступ к фондам учебно-методической документации и электронным изданиям

19. Фонд оценочных средств:

19.1 Перечень компетенций с указанием этапов формирования и планируемых результатов обучения

Код и содержание компетенции	Планируемые результаты обучения (показатели	Этапы формирования	ФОС* (средства

(или ее части)	достижения заданного уровня освоения компетенции посредством формирования знаний, умений, навыков)	компетенции (разделы (темы) дисциплины или модуля и их наименование)	оценивания)
ПК-2 Владение основными методами инструментального анализа звучащей речи	**знать:** математические и алгоритмические методы обработки текстовой информации, принципы организации и этапы разработки лингвистически ориентированных программных продуктов.	Разделы 1-5	Контрольная работа по соответствующим разделам.
	уметь: прилагать полученные теоретические и практические знания к задачам обработки текста и речи	Разделы 1-5	Лабораторные работы 1,2
	владеть: программными библиотеками для разработки систем, ориентированных на обработку звучащей речи и текста.	Разделы 1-5	Лабораторные работы 1,2
ПК-5 Владение основными способами описания и формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации, содержащейся в тексте на естественном языке	**знать:** способы формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации	Разделы 4,5	Контрольная работа по соответствующим разделам.
	уметь:использовать лингвистически-ориентированные программные системы, программные среды разработки, библиотеки программ.	Разделы 4,5	Лабораторные работы 3-5
	владеть:навыками использования, интеграции и разработки программных систем обработки лингвистической информации.	Разделы 4,5	Лабораторные работы 3-5
Промежуточная аттестация			Комплект КИМ

* В графе «ФОС» в обязательном порядке перечисляются оценочные средства текущей и промежуточной аттестаций.

19.2. Описание критериев и шкалы оценивания компетенций (результатов обучения) при промежуточной аттестации

Для оценивания результатов обучения на зачете используются следующие содержательные показатели (формулируется с учетом конкретных требований

дисциплины):

знать: математические и алгоритмические методы обработки текстовой информации, принципы организации и этапы разработки лингвистически ориентированных программных продуктов.

уметь: прилагать полученные теоретические и практические знания к задачам обработки текста и речивладеть: программными библиотеками для разработки систем, ориентированных на обработку звучащей речи и текста.

знать: способы формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации

уметь:использовать лингвистически-ориентированные программные системы, программные среды разработки, библиотеки программ.

владеть:навыками использования, интеграции и разработки программных систем обработки лингвистической информации.

Различные комбинации перечисленных показателей определяют критерииоценивания результатов обучения (сформированности компетенций) на зачете:

- высокий (углубленный) уровень сформированности компетенций;
- повышенный (продвинутый) уровень сформированности компетенций;
- пороговый (базовый) уровень сформированности компетенций.

Для оценивания результатов обучения на зачете используется – зачтено (выше порогового уровня), не зачтено (ниже порогового уровня) по результатам тестирования.

Соотношение показателей, критериев и шкалы оценивания результатов обучения на государственном экзамене представлено в следующей таблице.

Критерии оценивания компетенций и шкала оценок

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся демонстрирует полное соответствие знаний, умений, навыков по приведенным критериям свободно оперирует понятийным аппаратом и приобретенными знаниями, умениями, применяет их при решении практических задач.	Повышенный уровень	Отлично
Ответ на контрольно-измерительный материал не полностью соответствует одному из перечисленных выше показателей, но обучающийся дает правильные ответы на дополнительные вопросы. При этом обучающийся демонстрирует соответствие знаний, умений, навыков приведенным в таблицах показателям, но допускает незначительные ошибки, неточности, испытывает затруднения при решении практических задач.	Базовый уровень	Хорошо
Обучающийся демонстрирует неполное соответствие знаний, умений, навыков приведенным в таблицах показателям, допускает значительные ошибки при решении практических задач. При этом ответ на контрольно-измерительный материал не соответствует любым двум из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы.	Пороговый уровень	Удовлетворительно
Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки	–	Неудовлетворительно

19.3 Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности,

характеризующие этапы формирования компетенций в процессе освоения образовательной программы

19.3.1 Примерный перечень применяемых оценочных средств

№ п/п	Наименование оценочного средства	Представление оценочного средства в фонде	Критерии оценки
1	2	3	4
1	Устный опрос	Вопросы по темам/разделам дисциплины	Правильный ответ – зачтено, неправильный или принципиально неточный ответ - не зачтено
2	Контрольная работа по разделам дисциплины	Теоретические вопросы по темам/разделам дисциплины	Шкала оценивания соответствует приведенной в разделе 19.2
3	Лабораторная работа	Содержит 9 лабораторных заданий, предусматривающихосвоение программных системобработки текста и речи.	При успешно выполнении работы ставится оценка зачтено и осуществляется допуск к зачету, в противном случае ставится оценка не зачтено и обучающийся не допускается к зачету.
4	КИМ промежуточной аттестации	Каждый контрольно-измерительный материал для проведения промежуточной аттестации включает 2 заданий вопросов для контроля знаний, умений и владений в рамках оценки уровня сформированности компетенции.	Шкалы оценивания приведены в разделе 19.2

19.3.2. Примерный перечень вопросов к зачету

1. Компьютерная лингвистика как междисциплинарная область.
2. Основные задачи, решаемые компьютерной лингвистикой. Направления исследований.
3. Проблемы моделирования естественного языка в компьютерной лингвистике: виды и особенности моделей.
4. Лингвистические ресурсы, используемые для обработки текста и речи.
5. Прикладные задачи компьютерной лингвистики.
6. Компьютерная лингвистика как междисциплинарная область.
7. Основные задачи компьютерной лингвистики, направления исследований.
8. Проблемы моделирования естественного языка в компьютерной лингвистике: виды и особенности моделей.
9. Лингвистические ресурсы, используемые для обработки текста и речи.
10. Классификация прикладных систем в области компьютерной лингвистики.
11. История математической лингвистики, основные этапы развития: докомпьютерная эпоха.
12. Появление компьютерной лингвистики. Этапы становления.
13. Компьютерная лингвистика в России.
14. Направления перспективных исследований в области Компьютерной лингвистики.

15. Проблемы автоматизации синтеза текста. Алгоритмы синтеза текста для вербализации заданного содержания. Семантические, морфологические, синтаксические проблемы синтеза.
16. Автоматическое аннотирование и индексирование научно-технической документации. Автоматическое реферирование.
17. Проблемы автоматической обработки ошибок в печатных текстах. Автоматические корректоры.
18. Вероятностно-статистические характеристики текста, его элементов. Их применение в задачах лингвистики

19.3.3. Пример задания для выполнения лабораторной работы

Лабораторная работа № 4 datamining и классификация документов.

Цель: усвоить основы интеллектуального анализа данных и применение наивного байесовского классификатора для задачи категоризации текстовых документов.

Основные теоретические сведения.

Интеллектуальный анализ данных (*DataMining*) — это процесс обнаружения в множестве данных ранее неизвестных, полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Английский термин «DataMining» не имеет однозначного перевода на русский язык (интеллектуальный анализ данных, добыча данных, вскрытие данных, информационная проходка, извлечение данных/информации) поэтому в большинстве случаев используется в оригинале.

Методы DataMining разделяются на две группы:

1. **статистические** (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов)
2. **кибернетические** (искусственные нейронные сети, байесовские сети, эволюционное программирование, генетические алгоритмы, ассоциативная память, нечеткая логика, деревья решений, системы обработки экспертных знаний)

Визуальные инструменты DataMining позволяют проводить анализ данных предметными специалистами (аналитиками), не владеющими соответствующими математическими знаниями.

Задачи, решаемые иад

1. **Классификация** — отнесение входного вектора (объекта, события, наблюдения) к одному из заранее известных классов.
2. **Кластеризация** — разделение множества входных векторов на группы (кластеры) по степени «похожести» друг на друга.
3. **Сокращение описания** — для визуализации данных, лаконизма моделей, упрощения счета и интерпретации, сжатия объемов собираемой и хранимой информации.
4. **Ассоциация** — поиск повторяющихся образцов. Например, поиск «устойчивых связей в корзине покупателя» (*marketbasketanalysis*) — вместе с пивом часто покупают орешки.
5. **Прогнозирование** – предсказание следующего состояния системы по наблюдаемым.
6. **Анализ отклонений** — выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы.
7. …

В литературе можно встретить еще ряд классов задач. Базовыми задачами являются первые три. Остальные задачи сводятся к ним тем или иным способом.

Алгоритмы решения задач иад

Для задач классификации характерно «обучение с учителем», при котором построение (обучение) модели производится по выборке содержащей входные и выходные векторы.

Для задач кластеризации и ассоциации применяется «обучение без учителя», при котором построение модели производится по выборке, в которой нет выходного параметра. Значение выходного параметра («относится к кластеру …», «похож на вектор …») подбирается автоматически в процессе обучения.

Для задач сокращения описания характерно *отсутствие разделения на входные и выходные векторы*. Начиная с классических работ К. Пирсона по методу главных компонент, основное внимание здесь уделяется аппроксимации данных.

Этапы обучения

Можно выделить типичный ряд этапов решения задач методами ИАД:

1. Формирование гипотезы;
2. Сбор данных;
3. Подготовка данных (фильтрация);
4. Выбор модели;
5. Подбор параметров модели и алгоритма обучения;
6. Обучение модели (автоматический поиск остальных параметров модели);
7. Анализ качества обучения, если неудовлетворительный переход на п. 5 или п. 4;
8. Анализ выявленных закономерностей, если неудовлетворительный переход на п. 1, 4 или 5.

Ниже приведено решение задачи классификации документов на основе наивной байесовской модели (naiveBayesianmodel).

Классификация документов — одна из задач информационного поиска, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Классификация может осуществляться полностью вручную, либо автоматически с помощью созданного вручную набора правил, либо автоматически с применением методов машинного обучения. Следует отличать классификацию текстов от кластеризации, в последнем случае тексты также группируются по некоторым критериям, но заранее заданные категории отсутствуют.

Области применения задачи классификации текстов:

1. фильтрация спама
2. составление интернет-каталогов
3. подбор контекстной рекламы
4. в системах документооборота
5. автоматическое реферирование (составление аннотаций)
6. снятие неоднозначности при автоматическом переводе текстов
7. ограничение области поиска в поисковых системах
8. определение кодировки и языка текста
9. …

Подходы к классификации текстов

Выделяют три подхода к задаче классификации текстов:

1. Во-первых, классификация не всегда осуществляется с помощью компьютера. Например, в обычной библиотеке тематические рубрики присваиваются книгам вручную библиотекарем. Подобная *ручная классификация* дорога и неприменима в случаях, когда необходимо классифицировать большое количество документов с высокой скоростью.

2. Другой подход заключается в *написании правил*, по которым можно отнести текст к той или иной категории. Например, одно из таких правил может выглядеть следующим образом: "если текст содержит слова производная и уравнение, то отнести его к категории математика". Специалист, знакомый с предметной областью и обладающий навыком написания регулярных выражений, может составить ряд правил, которые затем автоматически применяются к поступающим документам для их классификации. Этот подход лучше предыдущего, поскольку процесс классификации автоматизируется и, следовательно, количество обрабатываемых документов практически не ограничено. Более того, построение правил вручную может дать лучшую точность классификации, чем при машинном. Однако создание и поддержание правил в актуальном состоянии требует постоянных усилий специалиста.
3. Наконец, третий подход основывается на *машинном обучении*. В этом подходе набор правил или, более общо, критерий принятия решения текстового классификатора, вычисляется автоматически из обучающих данных (другими словами, производится обучение классификатора). Обучающие данные — это некоторое количество хороших образцов документов из каждого класса. В машинном обучении сохраняется необходимость ручной разметки (термин *разметка* означает процесс приписывания класса документу). Но разметка является более простой задачей, чем написание правил. Кроме того, разметка может быть произведена в обычном режиме использования системы. Например, в программе электронной почты может существовать возможность помечать письма как спам, тем самым формируя обучающее множество для классификатора — фильтра нежелательных сообщений. Таким образом, классификация текстов, основанная на машинном обучении, является примером обучения с учителем, где в роли учителя выступает человек, задающий набор классов и размечающий обучающее множество.

Постановка задачи классификации текстов

1. Имеется множество категорий (классов, меток) $C = \{c_1, \ldots, c_{|C|}\}$.
2. Имеется множество документов $D = \{d_1, \ldots, d_{|D|}\}$.
3. Неизвестная целевая функция $F : C \times D \to \{0,1\}$.
4. Необходимо построить классификатор $F\square$, максимально близкий к F.
5. Имеется некоторая начальная коллекция размеченных документов $R \subset C \times D$, для которых известны значения F. Обычно её делят на «обучающую» и «проверочную» части. Первая используется для обучения классификатора, вторая — для независимой проверки качества его работы.
6. Классификатор может выдавать точный ответ $F\square: C \times D \to \{0,1\}$ или степень подобия $F\square: C \times D \to [0,1]$.

Этапы решения задачи классификации текстов

1. **Индексация документов.** Построение некоторой числовой модели текста, например, в виде многомерного вектора слов и их веса в документе. Уменьшение размерности модели.
2. **Построение и обучение классификатора.** Могут использоваться различные методы машинного обучения: решающие деревья, наивный байесовский классификатор, нейронные сети, метод опорных векторов и др.
3. **Оценка качества классификации.** Можно оценивать по критериям полноты, точности, сравнивать классификаторы по специальным тестовым наборам.

Применение наивной байесовской модели к задаче классификации текстов

Наивная байесовская модель является вероятностным методом обучения. Вероятность того, что документ *d* попадёт в класс *c* записывается как $P(c|d)$. Поскольку цель классификации - найти самый подходящий класс для данного документа, то в наивной байесовской классификации задача состоит в нахождении наиболее вероятного класса c_m:

$$c_m = \arg\max_{c \in C} P(c \mid d)$$.

Вычислить значение этой вероятности напрямую невозможно, поскольку для этого нужно, чтобы обучающее множество содержало все (или почти все) возможные комбинации классов и документов. Однако, используя формулу Байеса, можно переписать выражение для $P(c|d)$:

$$c_m = \arg\max_{c \in C} \frac{P(c)P(d \mid c)}{P(d)} = \arg\max_{c \in C} P(c)P(d \mid c)$$.

где знаменатель $P(d)$ опущен, так как не зависит от c и, следовательно, не влияет на нахождение максимума; $P(c)$ - вероятность того, что встретится класс c, независимо от рассматриваемого документа; $P(d|c)$ - вероятность встретить документ d среди документов класса c.

Используя обучающее множество, вероятность $P(c)$ можно оценить как

$$\hat{P}(c) = \frac{N_c}{N}$$

где N_c - количество документов в классе c, N - общее количество документов в обучающем множестве. Здесь использован другой знак для вероятности, $\hat{P}$, поскольку с помощью обучающего множества можно лишь оценить вероятность, но не найти её точное значение.

Чтобы оценить вероятность $P(d \mid c) = P(t_1, t_2, ..., t_{n_d} \mid c)$, где t_k - терм из документа d, n_d - общее количество значимых термов в документе, необходимо ввести упрощающие предположения (1) о условной независимости термов и (2) о независимости позиций термов. Другими словами, мы пренебрегаем, во-первых, тем фактом, что в тексте на естественном языке появление одного слова часто тесно связано с появлением других слов (например, вероятнее, что слово интеграл встретится в одном тексте со словом уравнение, чем со словом бактерия), и, во-вторых, что вероятность встретить одно и то же слово различна для разных позиций в тексте. Именно из-за этих грубых упрощений рассматриваемая модель естественного языка называется наивной (хотя она является достаточно эффективной в задаче классификации). Итак, в свете сделанных предположений, используя правило умножения вероятностей независимых событий, можно записать

$$P(d \mid c) = P(t_1, t_2, ..., t_{n_d} \mid c) = P(t_1 \mid c)P(t_2 \mid c) \cdots P(t_{n_d} \mid c) = \prod_{k=1}^{n_d} P(t_k \mid c)$$

Оценка вероятностей $P(t|c)$ с помощью обучающего множества будет

$$\hat{P}(t \mid c) = \frac{T_{ct}}{T_c}$$

где T_{ct} - количество вхождений терма t во всех документах класса c (и на любых позициях - здесь существенно используется второе упрощающее предположение, иначе пришлось бы вычислить эти вероятности для каждой позиции в документе, что невозможно сделать достаточно точно из-за разреженности обучающих данных - трудно ожидать, чтобы каждый терм встретился в каждой позиции достаточное количество раз); T_c - общее количество термов в документах класса c. При подсчёте учитываются все повторные вхождения.

После того, как классификатор "обучен", то есть, найдены величины $\hat{P}(c)$ и $\hat{P}(t \mid c)$, можно отыскать класс документа

$$c_m = \arg\max_{c \in C} \hat{P}(c)\hat{P}(d \mid c) = \arg\max_{c \in C} \hat{P}(c) \prod_{k=1}^{n_d} \hat{P}(t_k \mid c)$$

Чтобы избежать в последней формуле переполнения снизу, когда из-за большого числа

$$\hat{P}(c) \prod_{k=1}^{n_d} \hat{P}(t_k \mid c) = 0$$

сомножителей выражение , на практике вместо произведения обычно используют сумму логарифмов. Логарифмирование не влияет на нахождение максимума, так как логарифм является монотонно возрастающей функцией. Поэтому в большинстве реализаций вместо последней формулы используется

$$c_m = \arg\max_{c \in C} [\log(\hat{P}(c)\hat{P}(d \mid c))] = \arg\max_{c \in C} [\log \hat{P}(c) + \sum_{k=1}^{n_d} \log \hat{P}(t_k \mid c)]$$

Эта формула имеет простую интерпретацию. Шансы классифицировать документ часто встречающимся классом выше, и слагаемое $\log \hat{P}(c)$ вносит в общую сумму соответствующий вклад. Величины же $\log \hat{P}(t_k \mid c)$ тем больше, чем важнее терм *t* для идентификации класса *c*, и, соответственно, тем весомее их вклад в общую сумму.

Необходимо также отметить возможность использования различных терминов, как и их количества, для описания различных классов документов. В этом случае некоторые значения вероятностей $\hat{P}(t_k \mid c)$ будут равны нулю и не будут вовлечены в вычислениеc_m, в связи с сингулярностью выражения $\log \hat{P}(t_k \mid c)$. При возникновении подобной ситуации, вероятности $\hat{P}(t_k \mid c)$ будут игнорироваться и финальное значениеc_m должно быть нормировано на фактическое значение числа слагаемых $\log \hat{P}(t_k \mid c)$.

Задание.

Разработать систему классификации текстов на основе наивного байесовского классификатора. Например, создание почтового фильтра, отмечающего к какой категории отнести входящее сообщение (спам, личное или рабочее письмо). Для определения категоризирующих терминов необходимо произвести частотный анализ текста как всех сообщений из обучающего множества так и предклассифицированных.

19.3.4. Пример контрольно-измерительного материала

УТВЕРЖДАЮ
Заведующий кафедрой технологий обработки и защиты информации

_____ А.А. Сирота
__.__.2017

Направление подготовки / специальность

45.03.03 Фундаментальная и прикладная лингвистика

Дисциплина Б1.В.ДВ.3.1. Компьютерная лингвистика

Форма обучения Очное

Вид контроля Зачет

Вид аттестации Промежуточная

Контрольно-измерительный материал № 1

1. Проблемы автоматизации синтеза текста. Алгоритмы синтеза текста для вербализации заданного содержания. Семантические, морфологические, синтаксические проблемы синтеза.
2. Автоматическое аннотирование и индексирование научно-технической документации. Автоматическое реферирование.

Преподаватель _____ В.В.Гаршина

19.4. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Оценка знаний, умений и навыков, характеризующая этапы формирования компетенций в рамках изучения дисциплины осуществляется в ходе текущей и промежуточной аттестаций.

Текущая аттестация проводится в соответствии с Положением о текущей аттестации обучающихся по программам высшего образования Воронежского государственного университета.Текущая аттестация проводится в формах устного опроса (индивидуальный опрос, фронтальная беседа) и письменных работ (контрольные, лабораторные работы). При оценивании могут использоваться количественные или качественные шкалы оценок.

Промежуточная аттестация может включать в себя теоретические вопросы, позволяющие оценить уровень полученных знаний и/или практическое (ие) задание(я), позволяющее (ие) оценить степень сформированности умений и навыков.

При оценивании используется количественная шкала. Критерии оценивания приведены выше в таблице раздела 19.2.