Программа учебной дисциплины «Анализ текстов. Генеративные модели»

Утверждена

Академическим советом ООП

Протокол № 2.3-09/ 2706-01 от «27» июня 2018г.

Автор	Екатерина Леонидовна Черняк
Число кредитов	5
Контактная работа (час.)	64
Самостоятельная работа (час.)	126
Курс	1 курс
Формат изучения дисциплины	без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Данная дисциплина ставит своей целью изучение основных задач и методов обработки и анализа текстов, а также освоение программных систем и инструментов, в которых реализованы данные методы. Эти базовые знания и навыки необходимы в профессиональной деятельности специалистов по анализу данных и машинного обучения.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение (1 лекция и 1 семинар)

Основные задачи обработки и анализа текстов. Актуальность обработки и анализа текстов. Краткий исторический экскурс по обработке и анализу текстов. Обзор существующих систем обработки и анализа текстов. Классификация систем обработки и анализа текстов.

Тема 2. Методы сбора и хранения данных (1 лекция и 1 семинар)

Форматы данных, способы хранения, принципы работы интернета. Краулинг. Regexp. Unicode.

Тема 3. Частотный анализ текстов (1 лекция и 1 семинар)

Модель мешка слов. Закон Ципфа. Закон Хипса. Векторное представление текстов. Релевантность в векторной модели. Расширения модели мешка слов. Реализация модели мешка слов в библиотеках Gensim и NLTK.

Тема 4. Морфологический анализ и разрешение неоднозначности (1 лекция и 1 семинар)

Задача морфологического анализа. Типы языков. Алгоритмы морфологического разбора. Морфологическая разметка. Омонимия и неоднозначность. Алгоритм разрешения омонимии. Скрытые Марковские модели. Декодирование в скрытых Марковских моделях.

Тема 5. Синтаксический анализ. Универсальные зависимости (1 лекция и 1 семинар)

Задача синтаксического разбора предложений. Модель составляющих. Вероятностные контекстно-свободные грамматики. Модель зависимостей. Универсальные зависимости. Парсинг зависимостей. Архитектура SyntaxNet.

Тема 6. Выделение ключевых слов и словосочетаний (1 лекция и 1 семинар)

Лексический анализ. Словари и тезаурусы. Поиск синонимов. Частотные методы выделения ключевых слов и словосочетаний. Метрики совместной встречаемости. Выделение ключевых словосочетаний по морфологическим шаблонам. Выделение ключевых словосочетаний по синтаксическим шаблонам. Алгоритмы RAKE и TextRank.Программные средства для выделения ключевых слов: NLTK, Томита-парсер.

Тема 7. Векторная модель (1 лекция и 1 семинар)

Векторная модель документа, векторная модель слова. Поиск похожих текстов. Косинусная мера близости. Методы снижения размерности в векторной модели документа: сингулярное разложение, латентный семантический анализ. Связь с моделями скрытых тем. Латентное размещение Дирихле (LDA). Параметры модели. Выбор числа скрытых тем. Расширения модели LDA.

Дистрибутивная семантика, векторная модель слова. Построение матрицы PPMI. Поиск близких слов по значению. Снижение размерности и факторизация матрицы PPMI. Эмбеддинги: word2vec, GloVe, AdaGram. Обучение моделей word2vec. Отрицательное сэмплирование.

Тема 8. Классификация текстов (1 лекция и 1 семинар)

Задачи классификации текстов и предложений по теме, тональности и жанру. Метод наивного Байеса, метод максимальной энтропии. Сверточные нейронные сети. Архитектура FastText.

Тема 9. Языковые модели (1 лекция и 1 семинар)

Счетные языковые модели. Проблема нулевых вероятностей. Преобразование Лапласа, преобразование Гуд-Тьюринга. Вероятностные нейронные языковые модели. Генерация текстов. Рекуррентные нейронные сети.

Тема 10. Классификация последовательностей (1 лекция и 1 семинар)

Задача классификации последовательностей. Частеречная разметка, определение семантических ролей, извлечение именованных сущностей. IOB разметка, IOBES разметка. Условные случайные поля.

Тема 11. Суммаризация текстов, вопросно-ответные системы (1 лекция и 1 семинар)

Абстрактивная и генеративная суммаризация текстов. Алгоритм TextRank. Вопросно-ответные системы. Архитектура энкодера-декодоры для вопросно-ответных систем и чат-ботов.

Тема 12. Исправление опечаток (1 лекция и 1 семинар)

Модель зашумленного канала. Исправление опечаток по правила. Редакционное растояние.

Тема 13. Обработка речи, речевые технологии (1 лекция и 1 семинар)

Распознавание речи. Генерация речи.

Тема 14. Информационный поиск (1 лекция и 1 семинар)

Понятие релевантности. Использование векторной модели в задаче поиска. Косинусная мера релевантности. Использование языковой модели в задаче поиска. Обучение ранжированию. A|B - тестирование.

Тема 15. Мультимодальная обработка текстов (1 лекция и 1 семинар)

Связь обработки текстов с обработкой изображений. Генерация изображения по тексту. Поиск изображения по описанию.

III. ОЦЕНИВАНИЕ

Оценка по дисциплине формируется следующим образом:

Результирующая оценка рассчитывается по формуле:

$O_{итоговая} = 0.7 * O_{накопл} + 0.3 * O_{экз}$

Накопленная оценка рассчитывается по формуле:

$O_{накопл} = 0.3\ O_{сам.работа} + 0.7\ O_{дз}$

Условие получения автомата по курсу: накопленная оценка не менее 8 баллов из 10.

Все оценки округляются согласно правилам арифметического округления.

Итоговая оценка вычисляется по округленным оценкам $O_{накопл}$ и $O_{экз.}$.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примеры оценочных средств:

1. Домашнее задание содержит три задания на общую тему и на общем наборе данных разных уровней сложности. Первое задание предполагает реализацию базового алгоритма, второго – стандартного алгоритма, дающие лучшие показатели качества для рассматриваемой задачи, третье – реализацию небольшого исследовательского проекта на материале рассматриваемого набора данных.
2. Самостоятельная работа: содержит пять закрытых и открытых вопросов по материалу лекции, выдается студентам в виде электронной формы на фиксированное время после лекции.
3. Устный экзамен

VI. РЕСУРСЫ

1. Основная литература

1. Jurafsky, D., Martin J. H. Speech and Language Processing , 3 издание https://web.stanford.edu/~jurafsky/slp3/

2. Дополнительная литература

1. Болховитянов, А. В., Чеповский, А. М. Алгоритмы морфологического анализа компьютерной лингвистики. / – Москва : МГУП им. Ивана Федорова, 2013.
2. Ильвовский, Д. А., Черняк Е. Л. Системы автоматической обработки текстов // Открытые системы. – 2014. – № 1. – С. 51-53.
3. К.В.Воронцов. Лекции по вероятностным тематическим моделям [Электронный ресурс] / – Режим доступа: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf, свободный
4. Национальный корпус русского языка (НКРЯ) [Электронный ресурс] / – режим доступа: www.ruscorpora.org, свободный.
5. Чеповский, А. М. Неразрешимая проблема компьютерной лингвистики // Компьютерра. – 2002. – № 30. – С. 12-18.
6. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora // Journal of Language Resources and Evaluation. – № 43(3). – 2009. – С. 209-226.
7. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python / – O'Reilly Media, 2009.

8. Blei, D.M., Ng, D.M., Jordan, M.I. Latent Dirichlet allocation. // The Journal of Machine Learning Research. – № .3. – 2003. – С. 993-1022.
9. Chomsky, N. Syntactic structures / Walter de Gruyter, 2002.
10. Mitchell P. M., Marcinkiewicz, M. A., Santorini, B. Building a large annotated corpus of English: The Penn Treebank // Computational linguistics. – № 19(2). –1993. – С. 313-330.
11. Řehuřek, R., Sojka, P., Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – 2010. – С. 45-50.
12. Collobert, Ronan, et al. Natural language processing (almost) from scratch. // Journal of Machine Learning Research 12.Aug (2011): 2493-2537.
13. McCallum A. Efficiently inducing features of conditional random fields. InProceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence 2002 Aug 7 (pp. 403-410). Morgan Kaufmann Publishers Inc.
14. Большакова, Е.И., Ефремова, Н.Э., Шариков, Г.Ф., 2015. Инструментальные средства для разработки систем извлечения информации из русскоязычных текстов. Новые информационные технологии в автоматизированных системах, (18).
15. Ponte, Jay M., and W. Bruce Croft // A language modeling approach to information retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.
16. Bengio, Yoshua. Learning deep architectures for AI. // Foundations and trends® in Machine Learning 2.1 (2009): 1-127.
17. Mikolov, Tomas, et al. Efficient estimation of word representations in vector space. // arXiv preprint arXiv:1301.3781 (2013).

2. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Ubuntu 18	*Свободный*
2.	Python 3	*Свободный*

3. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Консультант Плюс	*Из внутренней сети университета (договор)*
2.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/

4. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для семинарских и самостоятельных занятий по дисциплине не требуют специального технического оснащения