Программа учебной дисциплины
«Анализ неструктурированных данных»

Утверждена
Академическим советом ОП
Протокол №_____ от _____._____.20_____

Разработчик	Артемова Екатерина Леонидовна, Доцент, Департамент больших данных и информационного поиска
Число кредитов	5
Контактная работа (час.)	60
Самостоятельная работа (час.)	130
Курс, Образовательная программа	4 (Б) курс, Прикладная математика и информатика
Формат изучения дисциплины	Без использования онлайн курса

1. Цель, результаты освоения дисциплины и пререквизиты

Цели:

1. Изучение базовых задач и методов обработки и анализа текстов
2. Изучение современных нейросетевых моделей для обработки и анализа текстов
3. Освоение программных систем и инструментов для обработки и анализа текстов

Пререквизиты:

1. Математический анализ
2. Линейная алгебра
3. Теория вероятностей и математическая статистика
4. Машинное обучение

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах**	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
	лк		
	см		
	onl/cp		
Введение. Статистический анализ текстов	2		
	2		
	8		
Векторные модели представления слов	2		
	2		
	8		
Классификация текстов	2		
	2		
	18		

Классификация последовательностей	2 2 8		
Предобученные языковые модели	2 2 8		
Синтаксический анализ	2 2 8		
Машинный перевод	2 2 8		
Генерация текстов	2 2 8		
Разметка данных, активное обучение.	2 2 8		
Вопросное-ответные системы	2 2 8		
Мультимодальные методы	2 2 8		
Мультиязычные методы	2 2 8		
Обработка текстов в медицине	2 2 8		
Информационный поиск	2 2 8		
Этические вопросы в обработке текстов	2 2 8		
Часов по видам учебных занятий:	30 30 130		
Итого часов:	190		

Содержание разделов дисциплины:

1. **Введение. Статистический анализ текстов**

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. Описательные статистики, оцениваемые по тексту. Методы излечения ключевых слов и словосочетаний. Закон Хипса, Закон Ципфа. Токенизация на основе регулярных выражений. Обучаемая сегментация предложений.

2. **Векторные модели представления слов**
   Векторная модель документа, векторная модель слова. Поиск похожих текстов. Косинусная мера близости. Методы снижения размерности в век-торной модели документа: сингулярное разложение, латентный семантический анализ. Связь с моделями скрытых тем. Латентное размещение Дири-хле (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 факто-ризация матрицы PPMI. Эмбеддинги: word2vec, GloVe, AdaGram. Обучение моделей word2vec. Отрицательное сэмплирование.
3. **Классификация текстов**
   Задачи классификации текстов и предложений по теме, тональности и жанру. Метод наивного Байеса, метод максимальной энтропии. Сверточные нейронные сети. Архитектура FastText. Аугментация данных. Классификация при небольших объемах размеченных данных.
4. **Классификация последовательностей**
   Задача классификации последовательностей. Частеречная разметка, определение семантических ролей, извлечение именованных сущностей. IOB раз-метка, IOBES разметка. Условные случайные поля. Рекуррентные нейронные сети. Модели последовательностей на основе сверточных сетей и трансформеров. Переход от токенизации к BPE кодированию.
5. **Предобученные языковые модели**
   Предобученные языковые модели на основе рекуррентных нейронных сетей и трансформеров. Архитектуры ELMo, BERT, ULMFit, XLNET, GPT2 и др. GLUE оценка.
6. **Синтаксический анализ**
   Задача синтаксического разбора предложений. Модель составляющих. Вероятностные контекстно-свободные грамматики. Модель зависимостей. Универсальные зависимости. Корпус. Universal Dependencies. Парсинг зависимостей. Архитектура SyntaxNet и архитектура UDPipe.
7. **Машинный перевод**
   Статистический машинный перевод. Нейросетевой машинный перевод и модели класса энкодер-декодер. Механизм внимания.
8. **Генерация текстов**
   Контролируемая генерация тектов. Диалоговые системы общего назначения.
9. **Разметка данных, активное обучение.**
   Системы разметки данных. Краудсорсинговые платформы. Коэффициенты согласия аннотаторов. Стратегии активного обучения.
10. **Вопросное-ответные системы**
    Типология вопросно-ответных системы. Архитектуры BiDAF, QANet, DRQ&A. Машинное чтение. Задача SQUAD.
11. **Мультимодальные методы**
    Задачи, связывающие анализ изображений и анализ текстов. Распознование текстов [optical character recognition].
12. **Мультиязычные методы**
    Перенос обучения между различными предметными областями. Перенос обучения с одного языка на другой.
13. **Обработка текстов в медицине**
    Анонимизация и подготовка медицинских текстов к анализу. Обзор задач, возникабщих при анализе медицинских текстов. Источники данных, онтологии, таксономии и графы знаний в медицине.
14. **Информационный поиск**
    Современный поисковые системы: индексация, поиск по векторному представлению. Связь с вопросно-ответными системами и рекомендательными системами.

15. **Этические вопросы в обработке текстов**
Предвзятость в предобученных моделях и способы ее компенсации. Детектирование ложных нвоостей и пропаганды.

3. Оценивание

- **1**, Не блокирующее, Домашнее задание
Четыре домашних задания
- **2**, Не блокирующее, Компьютерное тестирование
Квизы по итогам каждой лекции
- **3**, Не блокирующее, Домашнее задание
Проект на основе SemEval
- **4**, Не блокирующее, Экзамен (устный)
Экзамен по итогам курса

Формула округления: Стандартное арифметическое округление
Шкала оценки: Десятибалльная
Вид формулы оценивания: Линейная
Формула оценивания:

Пром1 = Округление(0.2 квиз + 0.25 ДЗ1 + 0.25 ДЗ2 + 0.3 Проект1)
Пром2 = Округление(0.2 квиз + 0.25 ДЗ3 + 0.25 ДЗ4 + 0.3 Проект2)
Окончательная оценка = Округление(0.4 Экзамен + 0.3 Пром1 + 0.3 Пром2)

Автомат: при 1/2 Округление (Пром1+Пром2) >= 8, автоматически выставляется оценка за Экзамен = 1/2 Округление (Пром1+Пром2)

Проект1 и Проект2 – две части проекта, предполагающего участие в соревновании SemEval

Предполагается 10-балльная шкала оценивания.

4. Примеры оценочных средств

Примеры заданий и вопросов могут быть найдены по ссылке https://github.com/MelLain/hse-nlp

5. Ресурсы

5.1. Рекомендуемая основная литература

п/п	Наименование
1	*Dan Jurafsky, James H. Martin* Speech and Language Processing
2	*Jacob Eisenstein* Natural Language Processing

5.2. Рекомендуемая дополнительная литература

п/п	Наименование
1	*Yoav Goldberg* Neural Network Methods for Natural Language Processing
2	*Shay Cohen* Bayesian Analysis in Natural Language Processing, 2 издание

5.3. Программное обеспечение

п/п	Наименование	Условия доступа/скачивания
1	Microsoft Windows 7 Professional RUS Microsoft Windows 8.1 Professional RUS Microsoft Windows 10	*Из внутренней сети университета (договор)*
2	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*

3	Библиотека torch	свободный
4	Библиотека nltk	свободный
5	Библиотека gensim	свободный
6	Библиотека sklearn	свободный

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

п/п	**Наименование**	**Условия доступа/скачивания**
	Профессиональные базы данных, информационно-справочные системы	
1	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1	Открытое образование	URL: https://openedu.ru/

5.5. Материально-техническое обеспечение дисциплины
Учебные аудитории для лекционных 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 оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:
6.1.1. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.
6.1.2. *для лиц с нарушениями слуха:* в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.
6.1.3. *для лиц с нарушениями опорно-двигательного аппарата:* в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.