Программа учебной дисциплины
«Интеллектуальный анализ текстов»

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

Разработчик	Шестаков Андрей Владимирович, Старший преподаватель, Департамент больших данных и информационного поиска
Число кредитов	4
Контактная работа (час.)	12
Самостоятельная работа (час.)	140
Курс, Образовательная программа	2 (М) курс, Прикладная политология
Формат изучения дисциплины	С использованием онлайн курса: https://ru.coursera.org/learn/language-processing

1. Цель, результаты освоения дисциплины и пререквизиты

Цели:

1. Формирование у студентов базовых теоретических знаний и практических навыков в области автоматической обработки естественного языка.

Планируемые результаты обучения (ПРО):

1. Понимание основного пайплайна обработки текстовой информации и умение применять его на практике
2. Понимание идеи языковых моделей и разработка языковой модели с помощью рекуррентной нейронной сети
3. Понимание принципов построения векторых представлений слов и текстов
4. Понимание принципов работы моделей машинного перевода
5. Понимание элементов архитектуры диалоговых систем
6. Разработка собственного чат-бота

Пререквизиты:

1. Базовый курс математической статистики и теории вероятности
2. Базовый курс по математическому анализу и линейной алгебре
3. Понимание работы методов машинного обучения
4. Программирование на Python
5. Английский язык на уровне Advanced

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	Объем в часах лк см onl/cp	Планируемые результаты обучения (ПРО), подлежащие контролю	Формы контроля
Введение в классфикикацию текстов	0 2 20	• Понимание основного пайплайна обработки текстовой информации и умение применять его на практике • Разработка собственного чат-бота	ОК, ПР, ЭК.
Языковые модели и разметка последовательностей	0 2 30	• Понимание идеи языковых моделей и разработка языковой модели с помощью рекуррентной нейронной сети • Разработка собственного чат-бота	ОК, ПР, ЭК.
Дистрибутивная семантика и тематические модели	0 4 40	• Понимание принципов построения векторых представлений слов и текстов • Разработка собственного чат-бота	ОК, ПР, ЭК.
Модели преобразования последовательности в последовательность	0 2 20	• Понимание принципов работы моделей машинного перевода • Разработка собственного чат-бота	ОК, ПР, ЭК.
Диалоговые системы	0 2 30	• Понимание элементов архитектуры диалоговых систем • Разработка собственного чат-бота	ОК, ПР, ЭК.
Часов по видам учебных занятий:	0 12 140		
Итого часов:	152		

Содержание разделов дисциплины:

1. **Введение в классфикикацию текстов**
   Изучение основных шагов в обработке текстовой информации; Разработка классификатора тэгов постов на ресурсе StackOverflow Применение методов грубокого обучения в NLP

2. **Языковые модели и разметка последовательностей**
   Применение LSTM для задачи распознавания именованных сущностей Моделирование языка с помощью n-gram и рекуррентных нейронных сетей Оценка качества моделей
3. **Дистрибутивная семантика и тематические модели**
   Модели word2vec, skipgram, CBOW, fastText и другие способы векторого представления слов Создание поисковой системы с помощью векторного представления предложений Обзор тематических моделей
4. **Модели преобразования последовательности в последовательность**
   Модели машинного перевода Обучение нейронной сети для решения задачи преобразования последовательностей Механизмы внимания
5. **Диалоговые системы**
   Архитектуры диалоговых систем Разработка чат-бота

3. Оценивание

- **ОК**, Не блокирующее, Оценка онлайн курса
  Оценка на платформе, приведенная к шкале 0-10
- **ПР**, Не блокирующее, Индивидуальный проект
  Дополнительный проект
- **ЭК**, Не блокирующее, Экзамен (письменный)
  Итоговый экзамен

Формула округления: Стандартное арифметическое округление

Вид формулы оценивания: Линейная
Формула оценивания:

Окончательная оценка = Округление(0.6 * ОК + 0.2 * ПР + 0.2 * ЭК)

Академическая этика
Преподаватель оставляет за собой право устроить устную защиту любой из форм контроля. Если при проверке работ (текущий и итоговый контроль) установлен факт нарушения академической этики, студент получает оценку «0» за данную работу. Работа студента, предоставившего свою работу для списывания, также аннулируется, к обоим студентам применяется дисциплинарное взыскание.
В случае нарушения правил проведения экзамена студент удаляется с экзамена с оценкой «0». К нарушениям правил проведения экзамена относятся: коммуникация с другими студентами во время выполнения работы, использование социальных сетей/телефона во время экзамена (с любой целью), списывание.

4. Примеры оценочных средств

https://www.coursera.org/learn/language-processing/home/welcome

1) Choose true statements about text tokens.
- Lemmatization is always better than stemming
- Stemming can be done with heuristic rules
- Lemmatization needs more storage than stemming to work
- A model without stemming/lemmatization can be the best

2) Let's consider the following texts:
good movie
not a good movie
did not like
i like it
good one
What is the sum of TF-IDF values for 1-grams in "good movie" text?

3) What models are usable on top of bag-of-words features (for 100000 words)?
What models are usable on top of bag-of-words features (for 100000 words)?
- Logistic Regression
- SVM
- Decision Tree
- Gradient Boosted Trees
- Naive Bayes

4) Consider the bigram language model trained on the sentence:
"This is the cow with the crumpled horn that tossed the dog that worried the cat that killed the rat that ate the malt that lay in the house that Jack built."
Find the probability of the sentence:
"This is the rat that worried the dog that Jack built."

5) Which of these models are generative i. e., which of them model the distribution p(x,y)?
- Conditional Random Fields
- Hidden Markov Models
- Maximum Entropy Markov Models

6) How are word embeddings usually evaluated (qualitatively or quantitively)?

7) How many parameters does PLSA topic model have?

8) Imagine you are analysing news flow for a company. You want to know what topics are being mentioned when people discuss the company, and how they change over time.
For each news article there are several modalities that you want to use: English text,

time, author and category. Your final goal is to track, how topics change over time. Which additive regularizers would you add to your topic model?

9) Which techniques would help if the data has rich morphology, informal spelling, and other sources of OOV tokens?
- Negative sampling
- Copy mechanism
- Byte-pair encoding
- Sub-word modeling
- Hierarchical softmax

10) What metrics do we use for NLU evaluation?

5. Ресурсы

5.1. Рекомендуемая основная литература

п/п	Наименование
1	*D. Jurafsky, J. H. Martin* Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition
2	*I. Goodfellow, Y. Bengio, A. Courville* Deep learning
3	*С. Паттанаяк* Глубокое обучение и TensorFlow для профессионалов: математический подход к построению систем искусственного интеллекта на Phyton

5.2. Рекомендуемая дополнительная литература

п/п	Наименование
1	*Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.*Автоматическая обработка текстов на естественном языке и анализ данных

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	https://www.anaconda.com/	Free

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.