# Сколковский институт науки и технологий (Сколтех)

| | |
|---|---|
| Рабочая программа дисциплины | **Обработка естественного языка** |

| | |
|---|---|
| Преподаватель | Панченко Александр Иванович, PhD, доцент |

## Аннотация

Данный курс представляет собой введение в область обработки естественного языка (NLP). В рамках курса рассматриваются различные задачи и приложения в области NLP, такие как морфологический анализ, распознавание именованных сущностей и устранение неоднозначности смысла слов. Курс в значительной степени основан на классическом учебнике Jurafsky&Martin, но также содержит материалы по графовым моделям для NLP и аннотации данных/краудсорсингу лингвистических данных.
Курс «Deep Learning for Natural Language Processing» в Сколтехе дополняет материал из данного курса и в большей степени фокусируется на современных нейронных моделях и подходам, таких как Transformer и не покрывает различные приложения и смежные темы, такие как краудсорсинг, графовые методы и синтаксический парсинг.

## 1. Основная информация

| | |
|---|---|
| Академический уровень курса | Магистратура<br>Аспирантура |
| Количество кредитов | 3 |

Предварительные требования к курсу / рекомендации

Необходимые знания:

- Общие компьютерные науки на уровне бакалавра
- Язык программирования Python

Желательные знания:

- начальные знания в области машинного обучения
- начальные знания в области статистики

Тип оценки - дифференцированная

Отображение оценок в процентах

| | |
|---|---|
| **A:** | 86 |
| **B:** | 76 |
| **C:** | 66 |

| | |
|---|---|
| **D:** | 56 |
| **E:** | 46 |
| **F:** | 0 |

## 2. Содержание курса

| Тема | Краткое содержание | Лекции (час) | Семинары (час) | Лабораторные занятия (час) |
|---|---|---|---|---|
| Введение в естественный язык Обработка Приложения | Введение в область обработки естественного языка (NLP). Обзор задач NLP. Диалоговые системы как важное применение NLP. | 3 | 3 | 0 |
| Дистрибутивная семантика и смысл слова Устранение неоднозначности | Векторные пространства, основанные на разреженном подсчете. Встраивание слов, Word2Vec, GloVe и связанные с ними модели. WordNet, Устранение неоднозначности смысла слов, индукция смысла слов. | 3 | 3 | 0 |
| Последовательность Маркировка | Задача создания меток последовательности и ее приложения, такие как распознавание именованных объектов. Условное случайное поле и связанные с ним модели. | 3 | 3 | 0 |
| Языковые модели и машинный перевод | Языковые модели: униграмма/биграмма/n-грамм, цепочка Маркова, закон Ципфа, скрытая Марковская модели. Статистические и нейронные модели для машинного перевода. | 3 | 3 | 0 |
| Синтаксический разбор | Обзор различных типов синтаксического анализа, распространенных в NLP: разбиение на фрагменты, анализ зависимостей, анализ групп участников и другие. Методы анализа зависимостей | 3 | 3 | 0 |
| Графики для NLP: Лингвистические сети и кластеризация | Мотивация для представления графов в NLP. Типы графов в задачах NLP. Диаграмма Кластеризация, Китайский шепот, марковская кластеризация и связанные с ними алгоритмы. Приложения. | 3 | 3 | 0 |
| Аннотирование данных и краудсорсинг для NLP | Большинство успешных моделей NLP основаны на лингвистически аннотированных данных в той или иной форме. Часто в практических приложениях для данного языка и предметной области такой набор данных недоступен, что не позволяет применять контролируемые модели. В этой лекции вы узнаете, как настроить создание необходимых данных. | 3 | 3 | 0 |

| Вид задания | Краткое содержание задания |
|---|---|
| Домашнее задание | Определение смысла слова (WSI). Цель этой задачи - применить модели распространения, такие как word2vec, для различения различных |

| | |
|---|---|
| | значений одного слова, например, python как язык программирования и python как змея. Для достижения этой цели будет изучена кластеризация семантических векторов.<br>Решения будут представлены на платформе CodaLab с общедоступной таблицей лидеров. Будет предоставлена записная книжка Jupyter с кодом и результатами экспериментов. |
| Домашнее задание | Обогащение таксономии. Целью этого задания является автоматическое построение дерева заполнения / лингвистического дерева с использованием моделей распределения значений слов.<br>Решения будут представлены на платформе CodaLab с общедоступной таблицей лидеров. Необходимо предоставить записную книжку Jupyter с кодом и результатами экспериментов. |
| Домашнее задание | Семантическая ролевая маркировка (SRL). Цель этой задачи - обозначить текст семантическими ролями, указывающими на значение отдельных частей. Более конкретно, эта задача SRL будет представлена как проблема с маркировкой последовательности. Мы рассмотрим анализ сравнительных аргументов в предметной области: разбор текстов, в которых сравниваются два продукта по какому-либо аспекту, например, сравнение того, что лучше - Python или Matlab для NLP?<br>Решения будут представлены на платформе CodaLab с общедоступной таблицей лидеров. Необходимо предоставить записную книжку Jupyter с кодом и результатами экспериментов. |
| Тест/квиз | Отвечайте на вопросы по материалам каждой лекции. |

## 3. Результаты обучения

Результаты обучения в Сколтехе указаны в соответствии со структурой результатов обучения в Сколтехе

| Знания |
|---|
| Методы обработки естественного языка<br>Методы оценки систем обработки естественного языка<br>Библиотеки программного обеспечения и фреймворки для написания программ для обработки естественного языка<br>Ресурсы исследовательской литературы по обработке естественного языка |

| Навыки |
|---|
| Выбирать подходящие языковые модели и вычислительные методы для различных задач NLP.<br>Анализировать и оценивать статистические методы NLP в различных приложениях.<br>Разрабатывать статистические модели и методы для различных приложений NLP.<br>Проводить методологические исследования в области обработки естественного языка. |

| Опыт |
|---|
| Реализация ряда вычислительных методов и лингвистических моделей для решения различных NLP-задач. |

## 4. Задания и выставление оценок

| Тип назначения | % от итоговой оценки за курс |
|---|---|
| Тест/квиз | 25 |
| Домашние задания | 75 |

## 5. Критерии оценки

## 6. Учебники и интернет-ресурсы

| Необходимые учебники | ISBN-13 (or ISBN-10) |
|---|---|
| Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed.). Online. https://web.stanford.edu/~jurafsky/slp3/ | |
| Рекомендуемые учебники | |
| Manning, C. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. | |
| Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. | |

| Веб-ресурсы (ссылки) | Описание |
|---|---|
| https://web.stanford.edu/~jurafsky/slp3/ | Textbook online |

## 7. Оборудование

| Программное обеспечение |
|---|
| Python 3 |

| Оборудование |
|---|
| Доступ к вычислительному серверу с графическими процессорами (например, Nvidia 2080 Ti или аналогичному) может быть полезен, но обычно достаточно платформы Google CodaLab. |

## 8. Дополнительные примечания

Примеры заданий:

How many parameters would have a 3-gram language model with a vocabulary of 100 characters?
- 300
- 10 thousand
- 1 million
- 1 billion

Which architecture for language models will probably consume more memory while training with long sequences?
- n-gram language model
- convolutional neural network
- recurrent neural network
- transformer neural network

Which architecture for language models will probably be the slowest to train with long sequences (if one has a GPU)?
- n-gram language model
- convolutional neural network
- recurrent neural network
- transformer neural network

Which loss function for language models is NOT equivalent to all the others?
- cross-entropy
- mean squared error
- log likelihood
- perplexity

Which parameter was roughly the same for GPT, GPT-2 and GPT-3?
- number of layers
- size of the training corpus
- dimensionality of hidden state
- vocabulary size

With which task it would be difficult to create a prompt for GPT zero-shot application?
- Filling gaps in a text with appropriate words or phrases
- Text classification (e.g. by sentiment)
- Reading comprehension (answering a question for a text)
- Machine translation

Which parameter is meaningful with greedy (non-probabilistic) text generation?
- temperature
- top k
- repetition penalty
- top P

Which statement is NOT a reason why generation with positive lexical constaints is so difficult?
- Typically, a language model has not been trained to condition on anything except the already generated text.
- By simply forcing the model to generate certain words, we may prevent it from generating a fluent text.
- Because text generation is autoregressive, the model cannot naturally "plan" in advance to generate a text with certain words.
- During training, the model may have never seen a text when the given words have occured together.

Which approach to controllable text generation is the most resource-intensive during training?
- Gradient-based prompt tuning
- Model fine-tuning with reinforcement learning
- PPLM (steerable lanugage models)
- Model fine-tuning conditional on control codes

Which approach to controllable text generation is the most resource-intensive during inference?

- Gradient-based prompt tuning
- Model fine-tuning with reinforcement learning
- PPLM (steerable lanugage models)
- Model fine-tuning conditional on control codes

Which approach to controllable text generation does not require computing gradients of the model?
- Prompt tuning
- PPLM (steerable lanugage models)
- GeDi (generative discriminators)
- Model fine-tuning conditional on control codes