

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

| | |
|----------------------------|--|
| | Рабочая программа дисциплины (модуля) |
| по дисциплине: | Основные задачи и модели NLP |
| по направлению: | Информатика и вычислительная техника |
| профиль подготовки: | Физтех-школа Прикладной Математики и Информатики кафедра компьютерной лингвистики |
| курс: | 4 |
| квалификация: | бакалавр |

Семестры, формы промежуточной аттестации:

7 (осенний) - Дифференцированный зачет

8 (весенний) - Дифференцированный зачет

Аудиторных часов: 90 всего, в том числе:

лекции: 60 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составил: Д.Г. Анастасьев

Программа обсуждена на заседании кафедры компьютерной лингвистики 04.06.2020

Аннотация

NLP – очень динамично развивающаяся область. За последние несколько лет имели место очень существенные продвижения в области, появились универсальные методы, позволяющие получить неплохой результат для большого количества задач NLP. Тем не менее, остается достаточно широкий класс «сложных» задач, а также набор особенностей корпуса (язык корпуса, его доменная специфичность, размер и т. п.) таких, что универсальные методы дают недостаточное для практического применения качество. В этих условиях необходимо использование лингвистических признаков и/или нетривиальных архитектур.

В первой части данного курса будет подробное знакомство с базовыми задачами и подходами NLP (эмбединги, языковые модели, пайплайн NLP, задачи NER, текстовой классификации и машинного перевода, seq2seq и attention).

Для большинства задач будет дан небольшой исторический обзор (бегло рассмотрены классические методы их решения) и подробно разобраны современные модели их решения.

Курс состоит из чередующихся теоретических и практических занятий.

1. Цели и задачи

Цель дисциплины

дополнение уже существующих, но делающий акцент на практической применимости получаемых студентами знаний.

Задачи дисциплины

развитие у студентов навыков, необходимых для работы с задачами в сфере natural language processing.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

| Код и наименование компетенции | Индикаторы достижения компетенции |
|--|---|
| ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре) | ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников |

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- классические методы машинного обучения, используемые в задачах классификации текстов;
- методы глубинного обучения, применяемые для анализа текста;
- основные подходы к теггированию последовательностей;
- современные методы, применяемые для создания систем машинного перевода;
- способы создания системы саммаризации текста;
- подходы к реализации машинно-обучаемой части персональных помощников.

уметь:

- работать с основными задачами в сфере natural language processing;
- анализировать результаты, получаемые при применении модели;
- реализовать произвольную систему, описанную в современных статьях.

владеть:

- методами машинного обучения в задачах natural language processing;
- средствами для разработки моделей для обработки текста и способами анализа ошибок модели.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

| № | Тема (раздел) дисциплины | Трудоемкость по видам учебных занятий, включая самостоятельную работу, час. | | | |
|-----------------------|--|---|----------|-----------------|----------------|
| | | Лекции | Семинары | Лаборат. работы | Самост. работа |
| 1 | Введение. | 3 | 3 | | 3 |
| 2 | Векторное представление слов. Часть 1. | 4 | 4 | | 4 |
| 3 | Векторное представление слов. Часть 2. | 4 | 4 | | 4 |
| 4 | Свёрточные нейронные сети. | 3 | 3 | | 3 |
| 5 | Рекуррентные нейронные сети. Часть 1. | 4 | 4 | | 4 |
| 6 | Рекуррентные нейронные сети. Часть 2. | 4 | 4 | | 4 |
| 7 | Языковые модели. Часть 1. | 4 | 4 | | 4 |
| 8 | Языковые модели. Часть 2. | 4 | 4 | | 4 |
| 9 | Модели Seq2seq. | 5 | | | 3 |
| 10 | Примеры применения моделей Seq2seq. | 5 | | | 2 |
| 11 | Трансформеры (transformers) и реферирование текстов. | 5 | | | 3 |
| 12 | Диалоговые (вопросно-ответные) системы. Часть 1. | 5 | | | 2 |
| 13 | Диалоговые (вопросно-ответные) системы. Часть 2. | 5 | | | 2 |
| 14 | Предварительно обученные модели. | 5 | | | 3 |
| Итого часов | | 60 | 30 | | 45 |
| Подготовка к экзамену | | 0 час. | | | |
| Общая трудоёмкость | | 135 час., 3 зач.ед. | | | |

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 7 (Осенний)

1. Введение.

Анализ тональности на материале базы кинокритик (IMDB). Краткий обзор применения машинного обучения для обработки естественного языка. Краткое введение в библиотеку keras.

2. Векторное представление слов. Часть 1.

Знакомство с векторным представлением слов (эмбедингами): "обучение без учителя", выявление связей слов друг с другом. Анализ близости словосочетаний. Машинный перевод, основанный на анализе слов (при использовании векторных представлений слов библиотеки MUSE).

3. Векторное представление слов. Часть 2.

Введение в библиотеку PyTorch.

Применение линейной регрессии в numpy и pytorch.

Применение методов CBoW, skip-gram, negative sampling и структурированных моделей Word2vec.

4. Свёрточные нейронные сети.

Знакомство со свёрточными нейронными сетями (convolutional neural networks). Связь между свёрткой и n-граммами. Простая реализация поиска определения фамилии при помощи свёрточной модели (символьной). Визуализация результатов.

5. Рекуррентные нейронные сети. Часть 1.

Применение рекуррентных нейронных сетей для задачи текстовой классификации. Реализация "теста на память". Детектор фамилий в мультязыковом варианте: символьный LSTM-классификатор.

6. Рекуррентные нейронные сети. Часть 2.

Применение рекуррентных нейронных сетей для определения последовательностей (sequence labelling). Реализация PoS-тэгера, основанного на словных и символьных эмбедингах.

7. Языковые модели. Часть 1.

Модель для русского языка символьного уровня. Её применение для генерации "твитов троллей": реализация модели с фиксированным окном при помощи свёртки и рекуррентных нейронных сетей.

Реализация простой условной модели: генерация фамилий (данные для задания: язык модели)

Решение задачи по классификации негативных комментариев.

8. Языковые модели. Часть 2.

Языковая модель уровня слов. Её применение для генерации поэтических текстов. Примеры применения transfer learning, мультизадачного обучения к языковым моделям.

Семестр: 8 (Весенний)

9. Модели Seq2seq.

Применение моделей Seq2seq для задач машинного обучения и создания подписей к изображениям. Кодирование байтовых пар (byte-pair encoding), лучевой поиск (beam search) и другие методы, использующиеся в машинном переводе.

10. Примеры применения моделей Seq2seq.

Подробные примеры применения Seq2seq в машинном обучении и создания подписей к изображениям.

11. Трансформеры (transformers) и реферирование текстов.

Реализация модели трансформатора (?) для задачи реферирования текста. Обсуждение модели Pointer-Generator Networks для реферирования текстов.

12. Диалоговые (вопросно-ответные) системы. Часть 1.

Целеориентированные диалоговые системы. Реализация мультизадачной модели: "намеренный" классификатор (intent classifier) и тэгер слов для диалогового менеджера.

13. Диалоговые (вопросно-ответные) системы. Часть 2.

Системы диалогов на бытовые темы и глубокие структурированные семантические модели (deep structured semantic models, DSSM). Реализация диалоговой системы на Стэнфордском вопросно-ответном датасете (Stanford Question Answering Dataset, SQuAD) и модели (чат-бота) на датасете субтитров (OpenSubtitles).

14. Предварительно обученные модели.

Применение предварительно обученных моделей для различных задач: использование модели Universal Sentence Encoder для оценки сходства предложений; использование ELMo для выставления тэгов последовательностей (с использованием метода условных случайных полей = conditional random field, CRF); использование модели BERT на датасете SWAG для выявления возможных продолжений (как состояний программы).

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6. Перечень рекомендуемой литературы

Основная литература

1. Нейронные сети : полный курс = Neural Networks. A Comprehensive Foundation, [учебное пособие] / Саймон Хайкин ; [перевод с английского]. Санкт-Петербург, Диалектика, 2019

Дополнительная литература

1. Нейронные сети на персональном компьютере [Текст]/А. Н. Горбань, Д. А. Россиев , -Новосибирск, Наука, 1996

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

на практических занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения дисциплины, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

- посещения всех занятий, предусмотренных учебным планом по дисциплине;
- ведения конспекта занятий;
- напряжённой самостоятельной работы студента.

Самостоятельная работа включает в себя:

- чтение рекомендованной литературы;
- проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;
- решение задач, предлагаемых студентам на занятиях;
- подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями преподавателю.

Возможен промежуточный контроль знаний студентов в виде решения задач в соответствии с тематикой занятий.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

| | |
|---|--|
| по направлению: | Информатика и вычислительная техника |
| профиль подготовки: | Физтех-школа Прикладной Математики и Информатики кафедра компьютерной лингвистики |
| курс: | 4 |
| квалификация: | бакалавр |
| Семестры, формы промежуточной аттестации: | |
| 7 (осенний) - Дифференцированный зачет | |
| 8 (весенний) - Дифференцированный зачет | |
| Разработчик: | Д.Г. Анастасьев |

1. Компетенции, формируемые в процессе изучения дисциплины

| Код и наименование компетенции | Индикаторы достижения компетенции |
|--|---|
| ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре) | ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников |

2. Показатели оценивания компетенций

В результате изучения дисциплины «Основные задачи и модели NLP» обучающийся должен:

знать:

- классические методы машинного обучения, используемые в задачах классификации текстов;
- методы глубинного обучения, применяемые для анализа текста;
- основные подходы к теггированию последовательностей;
- современные методы, применяемые для создания систем машинного перевода;
- способы создания системы саммаризации текста;
- подходы к реализации машинно-обучаемой части персональных помощников.

уметь:

- работать с основными задачами в сфере natural language processing;
- анализировать результаты, получаемые при применении модели;
- реализовать произвольную систему, описанную в современных статьях.

владеть:

- методами машинного обучения в задачах natural language processing;
- средствами для разработки моделей для обработки текста и способами анализа ошибок модели.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Анализ тональности на материале базы кинорецензий (IMDB).
2. Машинный перевод, основанный на анализе слов (при использовании векторных представлений слов библиотеки MUSE).
3. Введение в библиотеку PyTorch.
4. Применение линейной регрессии в numpy и pytorch.
5. Применение методов CBOW, skip-gram, negative sampling и структурированных моделей Word2vec.
6. Связь между свёрткой и n-граммами.
7. Простая реализация поиска определения фамилии при помощи свёрточной модели (символьной).
8. Применение рекуррентных нейронных сетей для задачи текстовой классификации.
9. Детектор фамилий в мультиязыковом варианте: символьный LSTM-классификатор.
10. Применение рекуррентных нейронных сетей для определения последовательностей (sequence labelling).
11. Реализация PoS-тэгера, основанного на словных и символьных эмбедингах.
12. Реализация простой условной модели: генерация фамилий (данные для задания: язык модели)
13. Языковая модель уровня слов.
14. Применение моделей Seq2seq для задач машинного обучения и создания подписей к изображениям.

15. Кодирование байтовых пар (byte-pair encoding), лучевой поиск (beam search) и другие методы, использующиеся в машинном переводе.
16. Подробные примеры применения Seq2seq в машинном обучении и создания подписей к изображениям.
17. Обсуждение модели Pointer-Generator Networks для реферирования текстов.
18. Целеориентированные диалоговые системы.
19. Системы диалогов на бытовые темы и глубокие структурированные семантические модели (deep structured semantic models, DSSM).
20. Реализация диалоговой системы на Стэнфордском вопросно-ответном датасете (Stanford Question Answering Dataset, SQuAD) и модели (чат-бота) на датасете субтитров (OpenSubtitles).

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Задача лемматизации.
2. Реализация логистической регрессии.
3. Задача распознавания именованных сущностей (NER).
4. Классификация с использованием "мешка слов" (bag-of-words).
5. Классификация с использованием TF-IDF.
6. Машинный перевод.
7. Реализация CBoW модели.
8. Реализация Negative Sampling.
9. Реализация структурированной модели Word2vec (structured Word2vec).
10. Реализация свёрточной модели нейронной сети: визуализация эмбеддингов и свёрток модели, улучшение модели.
11. Рекуррентные нейронные сети: тестирование модели, визуализация работы, улучшение модели.
12. Рекуррентные нейронные сети: тестирование модели с предобученными эмбеддингами + инференс с использованием всей матрицы эмбеддингов. Дополнительное обучение эмбеддингов, регуляризация предобученных эмбеддингов.
13. Рекуррентные нейронные сети: модель с символьными эмбеддингами, визуализация эмбеддингов. Альтернативные функции над символьными эмбеддингами.
14. Реализация модели со словными и символьными эмбеддингами, реализация модели "энкодер-декодер".
15. Улучшения языковой модели: SGD (stochastic gradient descent), dropout, большее количество юнитов и слоёв LSTM (long short-term memory).
16. Реализация условной модели генерации фамилий. Обучение для задачи классификации комментариев.
17. Реализация языковой модели уровня слов: перенос с "порошков" на "пирожки", единая модель на "порошков" и "пирожков". Реализация обработки лемм и грамматических значений, метрических шаблонов.
18. Реализация языковой модели с векторами LDA.
19. Реализация модели Seq2seq: scheduled sampling, кодирование байтовых пар, лучевой поиск, генерация подписей к изображениям.
20. Реализация модели для реферирования текстов: оценка, визуализация, анализ эмбеддингов. Реализация LabelSmoothing, модели Pointer-Generator.
21. Реализация диалоговой модели: мультизадачная модель, асинхронное обучение (asynchronous training).
22. Реализация PoS-тэгера.
23. Реализации hard negatives mining, semi-hard negatives mining для диалоговой модели.
24. Задача по реализации чат-бота.
25. Предварительно обученные модели: реализация модели "одного ближнего соседа" (1NN) на представлениях из USE.
26. Реализация тэгера на предобученных эмбеддингах.
27. Реализация тэгера на ELMo, с CRF.
28. Обучение модели: BERT, обучение с накоплением градиентов.

Билет 1:

1. Реализация логистической регрессии;
2. Реализация Negative Sampling.

Билет 2:

1. Машинный перевод;
2. Задача по реализации чат-бота.

Критерии оценивания

отлично (10) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

отлично (9) - выставляется студенту, показавшему свободное оперирование знаниями учебной программы дисциплины, выполнение заданий творческого характера.

отлично (8) - выставляется студенту, показавшему владение программным учебным материалом с наличием несущественных ошибок в действиях, самостоятельно исправляемых учащимся.

хорошо (7) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускается в ответе или в решении задач некоторые неточности.

хорошо (6) - выставляется студенту если он осознает воспроизведение программного учебного материала, в том числе и различной степени сложности, с несущественными ошибками, затруднения в применении отдельных навыков.

хорошо (5) - выставляется студенту если теоретическое содержание освоено не полностью, некоторые практические навыки сформированы недостаточно, в некоторых случаях были допущены ошибки.

удовлетворительно (4) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации.

удовлетворительно (3) - выставляется студенту в случае большого количества недочетов и неправильных ответов, а также пассивной работе в ходе занятий, многие учебные задания не выполнены.

неудовлетворительно (2) - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

неудовлетворительно (1) - выставляется студенту, который не освоил теоретическое и практическое содержание курса, все выполненные учебные задания содержат грубые ошибки.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, конспектами лекций или другими материалами.