

**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:	Введение в автоматическую обработку текстов
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Физтех-школа Прикладной Математики и Информатики кафедра анализа данных
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 1

Программу составил: Л.Л. Ромашкова, канд. физ.-мат. наук

Программа обсуждена на заседании кафедры анализа данных 06.03.2020

Аннотация

NLP (Natural Language Processing) - это подмножество более широкой области AI, которая пытается научить компьютер понимать и обрабатывать сырые данные на естественном языке.

В курсе мы постараемся дать вам понять и почувствовать, что происходит в мире. Какие задачи решаются, как это происходит; как некоторые статистические подходы (которые полностью занимали собой предыдущие курсы) получают новую жизнь и новую интерпретацию в нейросетях, а какие постепенно отмирают. Мы покажем, что NLP это не набор пар (задача, решение), а общие идеи, которые проникают в разные задачи и отражают некоторую общую концепцию. Вы также узнаете, что происходит на практике, когда какие подходы более применимы.

1. Цели и задачи

Цель дисциплины

- введение в основы автоматической обработки текстов, знакомство с основными понятиями, алгоритмами, существующими библиотеками обработки текстов.

Задачи дисциплины

- без углубления в детали, с сугубо инженерным взглядом на задачи и алгоритмы, познакомить студентов с основными вопросами обработки текстов, дать мотивацию разобраться в теме более глубоко;
- научить делать простые решения характерных задач на Python. Вывести студентов на уровень понимания предмета, позволяющий им в последующих семестрах с высокой эффективностью включиться в работу курса по анализу и автоматической обработке текста;
- дать представление о существующих библиотеках для обработки текстов;
- дать представление о том, что будет на курсе «Анализ текстов» в магистратуре, и какие сейчас есть актуальные задачи и последние достижения в обработке естественного языка.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников

работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре)	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценить качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- подходы к задачам классификации, кластеризации и аннотирования текстов. Иметь представление о существующих библиотеках для обработки текстов.

уметь:

- использовать средства языка программирования Python для решения задач тематического моделирования, извлечения словосочетаний и ключевых слов, тегирования последовательностей слов, поиска похожих текстов, аннотирования, извлечения признаков.

владеть:

- средствами разработки и тестирования программного кода на языке Python. Пакетами nltk, sklearn, gensim. Уметь работать с корпусами текстов (НКРЯ, OpenCorpora, Brown, 20newsgroups, reuters).

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	Алгоритмы извлечения ключевых слов из текста.	4	2		2
2	Аннотирование (unsupervised алгоритмы).	4	2		2

3	Введение в тематическое моделирование (общая идея, вероятностная модель).	4	4		4
4	Классификация текстов.	4	2		2
5	Кластеризация текстов.	4	4		4
6	Краткий обзор последних достижений.	2	4		4
7	Обзор задач.	4	4		4
8	Тегирование.	2	4		4
9	Языковые модели.	2	4		4
Итого часов		30	30		30
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 7 (Осенний)

1. Алгоритмы извлечения ключевых слов из текста.

Unsupervised алгоритмы извлечения ключевых слов из текста. Поиск коллокаций. Реализация поиска коллокаций.

2. Аннотирование (unsupervised алгоритмы).

Графовые алгоритмы. Алгоритмы на основе тематического моделирования и кластеризации. Multi-document summarization. Простое аннотирование на кластеризации.

3. Введение в тематическое моделирование (общая идея, вероятностная модель).

Word2vec. Близость текстов по смыслу. Cosine similarity и другие меры близости. Близость текстов в пространстве LSA, NMF, PLSA, LDA. Разбор примеров из tutorial gensim. Поиск новостей о том же событии и новостей на ту же тему: различия в функции близости и предобработке текста.

4. Классификация текстов.

Особенности работы с разреженными признаками, выбор алгоритмов. Классификация текстов по теме. Задача определения автора. Задача анализа тональности текста. Переобучение нелинейных классификаторов на разреженных признаках (пример из документации sklearn). Простой sentiment-анализ твитов. Sentiment-анализ с отбором признаков. Использование sklearn из nltk. Сравнение эффективности отбора признаков при использовании разных алгоритмов классификации.

5. Кластеризация текстов.

Сравнение разных алгоритмов кластеризации на нескольких темах из 20newsgroups или reuters по метрикам, использующим и не использующим разметку. Использование кластеризации для снижения пространства признаков.

6. Краткий обзор последних достижений.

Краткий обзор последних достижений (the-state-of-the-art алгоритмы. Обзор неохваченных и не раскрытых полностью вопросов. Обзор изученных на курсе вопросов, консультация к зачету.

7. Обзор задач.

Неформальное описание и примеры использования: классификации текстов (по теме, автору, тональности и т.д.), кластеризации текстов и тематического моделирования, извлечения словосочетаний и ключевых слов, тегирование последовательностей слов, поиск похожих текстов, аннотирование. Извлечение признаков. Tf*idf, n-граммы, нормализация токенов. Пакеты nltk, sklearn, gensim. Извлечение признаков из текстов, документация и примеры: sklearn tutorial, nltk-book. Корпусы текстов (НКРЯ, OpenCorpora, Brown, 20newsgroups, reuters).

8. Тегирование.

Тегирование последовательностей слов: POS-tagging, Named Entity Recognition. HMM, MEMM, CRF (общая идея, без детального вывода, обоснование –на курсе магистратуры). Задача Named Entity Recognition. Пример: решение с использованием обычных классификаторов (например, линейных) и признаками, содержащими контекст. Сравнение по качеству мультиклассовой классификации и работы двух последовательных классификаторов ("сущность/не сущность" и "тип сущности").

9. Языковые модели.

Генерация текстов с помощью языковой модели. Классификация спама: сравнение оценки вероятности возникновения текста в униграммной и в биграммной модели, сравнение качества классификации.

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. Обработка списков [Текст] : [учеб. пособие для вузов] / Дж. Фостер ; пер. с англ. В. В. Мартынюка под ред. Э. З. Любимского .— 364727 .— М. : Мир, 1974 .— 72 с.

Дополнительная литература

1. Автоматическая обработка, хранение и поиск информации [Текст]/Г. Сэлтон , -М., Советское радио, 1973

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

<http://elibrary.ru/defaultx.asp> Научная электронная библиотека;
<http://www.twirpx.com> Все для студента

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

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

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

Методические рекомендации позволяют студенту оптимальным образом организовать процесс обучения. В рабочей программе приведено примерное распределение часов аудиторной и внеаудиторной нагрузки по различным темам данной дисциплины.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника

профиль подготовки: Физтех-школа Прикладной Математики и Информатики
кафедра анализа данных

курс: 4

квалификация: бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Разработчик: Л.Л. Ромашкова, канд. физ.-мат. наук

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
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-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценить качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

2. Показатели оценивания компетенций

В результате изучения дисциплины «Введение в автоматическую обработку текстов» обучающийся должен:

знать:

- подходы к задачам классификации, кластеризации и аннотирования текстов. Иметь представление о существующих библиотеках для обработки текстов.

уметь:

- использовать средства языка программирования Python для решения задач тематического моделирования, извлечения словосочетаний и ключевых слов, тегирования последовательностей слов, поиска похожих текстов, аннотирования, извлечения признаков.

владеть:

- средствами разработки и тестирования программного кода на языке Python. Пакетами nltk, sklearn, gensim. Уметь работать с корпусами текстов (НКРЯ, OpenCorpora, Brown, 20newsgroups, reuters).

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

Перечень примерных заданий для промежуточного контроля:

1. Извлечение признаков из текстов, документация и примеры: sklearn tutorial, nltk-book. Пакеты nltk, sklearn, gensim.
2. Корпусы текстов (НКРЯ, OpenCorpora, Brown, 20newsgroups, reuters).
3. Разбор примеров из tutorial gensim. Поиск новостей о том же событии и новостей на ту же тему: различия в функции близости и предобработке текста.
4. Переобучение нелинейных классификаторов на разреженных признаках (пример из документации sklearn).
5. Простой sentiment-анализ твитов. Sentiment-анализ с отбором признаков. Использование sklearn из nltk.
6. Сравнение эффективности отбора признаков при использовании разных алгоритмов классификации.
7. Сравнение разных алгоритмов кластеризации на нескольких темах из 20newsgroups или reuters по метрикам, использующим и не использующим разметку.
8. Использование кластеризации для снижения пространства признаков.
9. Простое аннотирование на кластеризации.
10. Генерация текстов с помощью языковой модели.
11. Классификация спама: сравнение оценки вероятности возникновения текста в униграммной и в биграммной модели, сравнение качества классификации.
12. Реализация поиска коллокаций.
13. Задача Named Entity Recognition. Решение с использованием обычных классификаторов (например, линейных) и признаками, содержащими контекст.
14. Сравнение по качеству мультиклассовой классификации и работы двух последовательных классификаторов ("сущность/не сущность" и "тип сущности").

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

Перечень контрольных вопросов:

1. Обзор задач. Неформальное описание и примеры использования: классификации текстов (по теме, автору, тональности и т.д.), кластеризации текстов и тематического моделирования, извлечения словосочетаний и ключевых слов, тегирование последовательностей слов, поиск похожих текстов, аннотирование. Извлечение признаков. Tf*idf, n-граммы, нормализация токенов.
2. Введение в тематическое моделирование (общая идея, вероятностная модель). Word2vec. Близость текстов по смыслу. Cosine similarity и другие меры близости. Близость текстов в пространстве LSA, NMF, PLSA, LDA.
3. Классификация текстов. Особенности работы с разреженными признаками, выбор алгоритмов. Классификация текстов по теме. Задача определения автора. Задача анализа тональности текста.

4. Кластеризация текстов. Сравнение разных алгоритмов кластеризации на нескольких темах из 20newsgroups или reuters по метрикам, использующим и не использующим разметку.
5. Аннотирование (unsupervised алгоритмы). Графовые алгоритмы. Алгоритмы на основе тематического моделирования и кластеризации. Multi-document summarization.
6. Языковые модели. Генерация текстов с помощью языковой модели.
7. Unsupervised алгоритмы извлечения ключевых слов из текста. Поиск коллокаций.
8. Тегирование последовательностей слов: POS-tagging, Named Entity Recognition. HMM, MEMM, CRF (общая идея, без детального вывода, обоснование –на курсе магистратуры).
9. Краткий обзор последних достижений (the-state-of-the-art алгоритмы).

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

При проведении дифференцированного зачета обучающемуся предоставляется 30 минут на подготовку.

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