

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
М.В.ЛОМОНОСОВА»**

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

«УТВЕРЖДАЮ»

Декан экономического факультета МГУ

профессор _____ А.А.Аузан

«___» _____ 202_ год

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины:

Анализ больших текстовых данных и информационный поиск

Уровень высшего образования:

МАГИСТРАТУРА

Направление подготовки:

38.04.01. ЭКОНОМИКА

Форма обучения:

ОЧНАЯ

Рабочая программа рассмотрена и одобрена
Учебно-методической комиссией экономического факультета
(протокол № _____, дата)

Москва 2023

Рабочая программа дисциплины разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки магистратуры 38.04.01. Экономика

ОС МГУ утвержден решением Ученого совета МГУ имени М.В.Ломоносова от 28 декабря 2020 года, протокол №7

Год (годы) приема на обучение: 2023 и последующие

1. Место и статус дисциплины в структуре основной профессиональной образовательной программы подготовки магистра

Статус дисциплины: *факультатив*

Триместр: 4

2. Входные требования (реквизиты) для освоения дисциплины

Учащиеся должны владеть знаниями по дискретной математике, математической логике, языкам программирования, теории вероятностей и математической статистике в объеме, соответствующем основным образовательным программам бакалавриата.

3. Планируемые результаты обучения по дисциплине, соотнесенные с требуемыми компетенциями выпускников

Компетенции выпускников (коды)	Индикаторы (показатели) достижения компетенций	Планируемые результаты обучения по дисциплине (модулю), сопряженные с компетенциями
ОПК-2. Способен применять продвинутые инструментальные методы экономического анализа в прикладных и (или) фундаментальных исследованиях.	ОПК-2.И-1. Владеет современными методами экономического анализа, математической статистики, анализа данных и эконометрики для решения теоретических и прикладных задач	ОПК-2.И-1.У-2. Умеет использовать современные методы экономического анализа, математической статистики, анализа данных и эконометрики для решения прикладных задач
ПК-8. Способен использовать различные инструментальные методы расчета и анализа социально-экономических показателей	ПК-8.И-2. Применяет современные инструментальные методы расчета и анализа социально-экономических показателей при решении практических и (или) исследовательских задач	ПК-8.И-2.У-1. Умеет применять инструментальные методы расчета и анализа показателей, в том числе с использованием различных платформ анализа данных и с применением языков программирования R и Python
МПК-1. Способен разрабатывать эконометрические модели и модели машинного обучения исследуемых экономических процессов и явлений, интерпретировать полученные результаты	МПК-1.И-1. Применяет продвинутый эконометрический инструментарий и методы машинного обучения для построения моделей экономических процессов и явлений. Интерпретирует результаты проведенного моделирования, делает выводы и дает рекомендации на его основе	МПК-1.И-1.У-1. Умеет разрабатывать эконометрические модели и модели, основанные на методах машинного обучения
МПК-3. Способен использовать соответствующий математический	МПК-3.И-1. Использует современные программные и аналитические средства для	МПК-3.И-1.У-1. Умеет применять современные программные и аналитические способы сбора,

аппарат и инструментальные средства для сбора, обработки, анализа и систематизации информации	сбора, первичной обработки и анализа данных	анализа и систематизации информации
---	---	-------------------------------------

4. Объем дисциплины по видам занятий

Объем дисциплины составляет 3 зачетных единицы: 108 академических часов, в том числе 54 академических часов составляет контактная работа с преподавателем, 54 академических часов составляет самостоятельная работа магистранта.

5. **Формат обучения:** очный, с использованием обучающей среды Op.Ecop (при необходимости допускается применение дистанционных образовательных технологий).
6. **Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий**

Наименование и краткое содержание разделов и тем дисциплины (модуля), Форма промежуточной аттестации по дисциплине (модулю)	Всего (часы)	В том числе						
		Контактная работа (работа во взаимодействии с преподавателем) <i>Виды контактной работы, часы</i>				Самостоятельная работа обучающегося <i>Виды самостоятельной работы, часы</i>		
		Занятия семинарского типа	Групповые консультации	Индивидуальные консультации	Всего	Выполнение домашних заданий	Подготовка рефератов и т.п..	Всего
Тема 1. Введение: Определение сферы информатического поиска, задачи информатического поиска, Информатическо-поисковые системы различной направленности. Архитектура информатическо-поисковых систем	10	6			6	4		4

Тема 2. Модели информационного поиска. Оценка качества информационного поиска	24	16			16	8		8
Тема 3. Методы расширения информационно-поисковых запросов. Вопросно-ответные системы. Диалоговые системы	16	8			8	8		8
Тема 4. Учет различных факторов. Анализ ссылок, логи запросов, анализ кликов, персонализация выдачи, поисковые сессии, словосочетания и близость расположения, тематические вероятностные модели. Комбинирование факторов. Модели Learning to-rank.	12	8			8	4		4
Тема 5. Автоматическая классификация и кластеризация текстов. Типы классификации. Тематическая классификация и анализ тональности. Методы классификации. Особенности методов классификации. Методы автоматической кластеризации. Автоматическое аннотирование.	28	12			12	8	8	16
Промежуточная аттестация: контрольная работа	4					14		
Итого	108	54				54		

Краткое содержание тем дисциплины

Тема 1.

Введение в информационный поиск. Задачи информационного поиска
Архитектура информационно-поисковых систем. Информационно-поисковые системы разного уровня и направленности

Тема 2.

Модели информационного поиска: булева модель, векторная модель, индекс информационно-поисковой системы
Тестирование систем информационного поиска. Точность, полнота, F-мера. Оценка качества ранжированного поиска.

Тема 3.

Методы расширения запросов при поиске. Тезаурусы, алгоритмы Relevance feedback. Автоматическое порождение тезаурусов для информационного поиска.

Тема 4.

Модели информационного поиска на основе вероятностей: вероятностная модель, языковая модель
Методы учета ссылок между документами. Алгоритмы PageRank, Hits
Логи запросов и кликов в информационном поиске. Классификация запросов. Поисковые сессии. Кликовые модель
Методы учета словосочетаний в информационном поиске. Позиционный индекс. Понятие окна
Методы сокращения пространства представления документов: латентный семантический анализ, вероятностное тематическое моделирование
Задача обучения ранжированию (learning to rank) в информационном поиске.

Тема 5.

Автоматическая классификация текстов. Классические методы классификации: метод KNN, наивный Байесовский классификатор, линейные классификаторы, метод опорных векторов.
Автоматическая классификация текстов в реальных условиях
Автоматический анализ тональности текстов.
Автоматическая кластеризация текстов. Особенности кластеризации новостей.
Векторные представления слов (word embeddings). Модель word2vec, fastText
Задачи и методы автоматического аннотирования текстов. Нейронные сети в задачах автоматического аннотирования
Вопросно-ответный информационный поиск.
Чат-боты.

7. Фонд оценочных средств для оценивания результатов обучения по дисциплине

7.1. Примеры оценочных средств:

Результаты обучения по дисциплине	Виды оценочных средств
ОПК-2.И-1.У-2.	Домашние задания

	Контрольные работы
ПК-8.И-2.У-1.	Домашние задания Контрольные работы
МПК-1.И-1.У-1.	Домашние задания Контрольные работы
МПК-3.И-1.У-1.	Домашние задания Контрольные работы

7.2. Критерии оценивания (баллы) по дисциплине:

Виды оценочных средств	Баллы
Домашние задания (5 штук)	100
Контрольные работы (2 штуки)	50
Итого	150

7.3. Оценка по дисциплине выставляется, исходя из следующих критериев:

Оценка	Минимальное количество баллов	Максимальное количество баллов
<i>Отлично</i>	127,5	150,0
<i>Хорошо</i>	97,5	127,0
<i>Удовлетворительно</i>	60,0	97,0
<i>Неудовлетворительно</i>	0,0	59,5

Примечание: в случае, если магистрант за триместр набирает менее 20% баллов от максимального количества по дисциплине, то уже на промежуточном контроле (и далее на пересдачах) действует следующее правило сдачи: «магистрант может получить только оценку «Удовлетворительно», и только если получит за промежуточный контроль, включающий весь материал дисциплины, не менее, чем 85% от баллов за промежуточный контроль».

7.4. Типовые задания и иные материалы, необходимые для оценки результатов обучения:

ПДЗ ТК1. Написание программы, реализующей векторную модель информационного поиска. Проверка на основе нескольких статей Википедии. В качестве документов для поиска выступают предложения выбранных статей Википедии

- Запросы – это факты из Википедии

- Коллекция – это все упомянутые статьи из всех фактов (не менее 2 из каждого факта)
- Документы – это предложения из статей Википедии, указанных в этих фактах, т.е. объединенная коллекция предложений статей всех фактов
- Все должно быть обработано морфологическим анализатором
- Нужно найти наиболее релевантные предложения
 - По векторной модели без idf
 - По tf.idf (idf в данном случае – это количество предложений, в которых встречалось слово)
 - Нормализация запроса и предложения
 - Т.е. выстроить все предложения из статей по мере сходства с запросом по векторной модели.
 - В отчете должны быть показаны веса выдаваемых предложений

ПДЗ ТК5. Реализация системы классификации цитат из новостных статей по тональности

- Выдаются данные для задачи анализа тональности: цитаты
 - Обучающая коллекция с ответами
 - Тестовая коллекция с ответами
 - Нужно обучиться на обучающей коллекции, затем провериться на тестовой
 - Классификация на три класса (позитивный, негативный, нейтральный)
 - Нужно выбрать метод из пакета (<http://scikit-learn.org/stable/>) и проверить качество классификации на разных типах векторизации: булевские вектора, частоты, tf.idf
 - Встроенный механизм векторизации:
 - http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
 - Попробовать методы: наивный Байес, SVM, разные параметры, что-то еще
 - Оценить результаты (F-мера), представить отчет.

1. Список вопросов для первой контрольной работы

2. Основные понятия информационного поиска
3. Виды поисковых систем по охвату и направленности. Особенности разных типов поисковых систем
4. Особенности научного поиска
5. Основные этапы обработки текстов в поисковой машине

6. Основные этапы обработки запроса в поисковой машине
7. Что такое графематический анализ? Что такое лемматизация
8. Как работает словарный морфологический анализ?
9. Как морфологические анализаторы обрабатывают слова, отсутствующие в словаре
10. Что такое постморфологический анализ. Основные методы.
11. Булевская модель информационного поиска. Преимущества и недостатки булевой модели поиска
12. Как измеряется качество булевского поиска
13. Алгоритм сопоставления запроса с документами (Алгоритм Merge)
14. Что такое векторная модель информационного поиска?
15. Поясните смысл показателей *idf* и *tf.idf*. Способы вычисления *tf*.
16. Классическая процедура оценки качества информационно поиска
17. Что такое РОМИП, какие задачи в нем решаются?
18. Что такое кривая полнота-точность?
19. Что такое 11-точечный график TREC?
20. Что такое пулинг в информационном поиске? Сложности, связанные с пулингом
21. Оценка качества в поисковых машинах
22. Шкалы оценок. Мера NDCG
23. Что такое информационно-поисковые тезаурусы? Зачем они нужны? Где применяются сейчас
24. Назовите методы расширения запросов пользователей при информационном поиске.
25. Что означает термин *relevance feedback*? Поясните основные принципы работы
26. Алгоритм Роккио для *relevance feedback*
27. Порождение и применение автоматического тезауруса для расширения запросов
28. Вероятностная модель информационного поиска: основная идея, различие с векторной моделью
29. Что такое языковые статистические модели?
30. Языковая модель информационного поиска
31. Вопросно-ответные системы: постановка задачи. основные компоненты, особенности тестирования.
32. Классификация вопросов в вопросо-ответных системах. Типы вопросов и типы ответов
33. Особенности обработки фактоидных вопросов
34. Особенности обработки нефактоидных вопросов.
35. Что такое PageRank? Зачем нужен, как вычисляется
36. Алгоритм HITS
37. Особенности использования кликов пользователя в качестве фидбека от пользователя

38. Классификация запросов по цели. Зачем нужна. Особенности обработки разных типов запросов
39. Поисковые сессии в интернет. Как выделять, зачем.
40. Особенности использования кликов пользователя в качестве фидбека от пользователя. Каскадная модель при обработке кликов.

Список вопросов для второй контрольной работы

1. Укажите основные методы автоматической рубрикации (классификации) текстов.
2. Что такое инженерный метод классификации текстов? Плюсы и минусы инженерных методов классификации
3. Укажите плюсы и минусы ручного рубрицирования.
4. Метод Байеса для автоматической классификации текстов
5. Метод Роккио для автоматической классификации текстов
6. Метод Knn для автоматической классификации текстов
7. Поясните основной принцип метода SVM для автоматической рубрикации текстов
8. Плюсы и минусы методов машинного обучения для рубрикации текстов
9. Особенности применения методов машинного обучения при классификации текстов в зависимости от размера обучающей коллекции
10. Что такое кластеризация текстов? Чем она отличается от классификации (рубрикации) текстов?
11. Метод K-means для кластеризации текстов
12. Агломеративная кластеризация – основной принцип и подвиды
13. Методы тестирования автоматической кластеризации
14. Особенности кластеризации потока новостей в реальном времени
15. Анализ тональности. Проблемы. Виды задач.
16. Анализ тональности как задача классификации: Методы, Признаки, Меры качества
17. Автоматическое аннотирование. Виды автоматических аннотаций.
18. Методы и признаки для отбора предложений в экстрактивном методе автоматического аннотирования
19. Метод MMR автоматического аннотирования
20. Метрика Rouge для тестирования автоматических аннотаций
21. Метод пирамид для тестирования автоматических аннотаций
22. Исправление несловарных ошибок на основе применения правила Байеса
23. Исправление ошибок перехода в другое словарное слова на основе применения правила Байса
24. Учет контекста при исправлении ошибок написания
25. Методы приближительного вычисления сходства документов в реальных поисковых системах

26. Обработка фразовых запросов и запросов с указанием близости слов в поисковых системах
27. Позиционный индекс в поисковой системе. Зачем нужен, как обрабатывается
28. Какие факторы помимо веса $tf.idf$ учитываются в поисковых моделях, как создаются многофакторные модели в информационно поиске
29. Метод шинглов для определения дубликатов документов
30. Краулинг в Интернет
31. Обучение ранжированию. Зачем нужно? Основные подходы
32. Понятие перестановок в задаче обучения ранжированию
33. Латентный семантический анализ (LSA). Зачем нужен? Основная идея. Преимущества и недостатки
34. Что такое тематические модели. Виды тематических моделей. Основной подход к моделированию
35. Оценки качества тематических моделей
36. Традиционные диалоговые системы: архитектура, основные этапы работы
37. Чат-боты: классификация, основные методы

7.5. Методические рекомендации и требования к выполнению заданий:

Самостоятельная работа учащихся состоит в изучении лекционного материала, учебно-методической литературы, выполнении домашних заданий и подготовки к промежуточной аттестации.

Письменные контрольные работы включают ответы на теоретические вопросы и выполнение практических заданий на применение моделей информационного поиска и методов оценки их качества.

8. Ресурсное обеспечение

8.1. Перечень основной и дополнительной литературы

Основная литература:

- 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. место издания *Изд-во НИУ ВШЭ Москва*, ISBN 978-5-9909752-1-7, 269 с.
- 2) Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
- 3) Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
- 4) Турнбулл Д., Берримен Дж. Релевантный поиск с использованием Elasticsearch и Solr, DMK-press, 2018.

- 5) Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval, Addison Wesley, 1999.
- 6) Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

Дополнительная литература:

- 1) Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
- 2) Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. Изд-во ИНТУИТ, 2009.
- 3) Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
- 4) Ландэ Д. В., Снарский А. А., Безсуднов И. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. — М.: Либроком (Editorial URSS), 2009.
- 5) Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
- 6) Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000.

8.2. Перечень лицензионного программного обеспечения

В процессе обучения используются инструментальные системы морфологического анализа русского языка (mystem, pymorphy), инструментальная система машинного обучения sci-kit learn (<http://scikit-learn.org/stable/>).

8.3. Перечень профессиональных баз данных и информационных справочных систем

8.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» (при необходимости)

- 1) <http://scikit-learn.org/stable/index.html>
- 2) <https://tech.yandex.ru/mystem/>
- 3) <https://nlp.stanford.edu/IR-book/>

8.5. Описание материально-технической базы

Для преподавания дисциплины требуется компьютерный класс, оборудованный проектором, а также маркерной или меловой доской.

9. Язык преподавания: русский

10. Преподаватель (преподаватели): д.т.н., доцент Лукашевич Н.В.

11. Разработчики программы: д.т.н., доцент Лукашевич Н.В.