Системы искусственного интеллекта (ИИ)

Это документ с материалами по ИИ. Более подробную информацию можно узнать у старост или у лектора: liliya@bmstu.ru Лилия Леонидовна Волкова

Почта Юрия Владимировича Строганова: stroganovyv@bmstu.ru

Согласно Э.У., у Вас 9 ЛР и 2 РК. Зачёт ставится, когда сданы все ЛР и РК.

Лабораторные и РК по СИИ нужно защищать, описав датасет, содержание лабораторной, принятые проектные решения и принципы работы кода (по коду). Также подготовьте примеры, на которых Вы продемонстрируете работу ПО. Обратите внимание, необходимо описать, какие меры-стратегии-подходы были применены и почему; как устроено решение и почему оно даёт конкретные результаты. См. раздел ЧаВо.

Язык всех лабораторных — русский.

Обновления

Всем крепкого иммунитета!
11 октября: выдано задание на ЛР4 и РК1.

Оглавление

Часть 1. Рекомендательные системы.

1.1. Материалы лекций

Материалы по лекции 1:
введение в ИИ и обзор прикладных задач: лекция Хохловой и вводный материал из учебника Павлова (стр. стр. 9-31, 36-43) [смотрите личный кабинет в электронном университете, там есть материалы к дисциплине].

Лекция 2:
Меры близости двух объектов как точек (векторов) в N-мерном пространстве (3 признака => трёхмерное, 4 признака => четырёхмерное, и т.д.) см.
1) стр. 171-172
http://clschool.miem.edu.ru/uploads/swfupload/files/011a69a6f0c3a9c6291d6d375f12aa27e349cb67.pdf
2) типы признаков и обзор ассоциативных мер близости — с. 421-427
https://docs.google.com/file/d/0B-SgBwisInUESG05WTRRYzM5Vkk/edit

Лекции 3-4: Рекомендательные системы
См. обзор тут, вместе с хорошей подборкой литературы:
https://drive.google.com/file/d/0B-SgBwisInUENWJnLXZGZ3NDZkU/view (полный текст на сайте конференции http://nps.itas.miem.edu.ru)

1.2. Задания на первый блок лабораторных

ЛР1: датасет для рекомендательной системы.
Выбрать предметную область (варианты уникальны и хранятся в google-документе: https://docs.google.com/spreadsheets/d/1fu9SMIn04DfbuD4bxNRTP0iVaDq8PMxmSxoZ18gsZN4/edit?usp=sharing), описать иерархию данных (в форме дерева, которое описывает классификацию данных) – для всех объектов или только для атрибута объектов (например, атрибут “жанр”). Рекомендуемый минимальный объём: 25 узлов, высота дерева 5. Добавить листьям дерева атрибуты, не менее 5 шт. Минимальные требования:
1 бинарный,
1 количественный,
1 категорийный (часто это признак, приводимый к численной шкале, например, категории расстояния: малое-среднее-дальнее),
1 не переводимый в количественные (например, тэг или аннотация).
Описать дерево можно в произвольном виде, оно демонстрируется на защите. Иерархия данных нужна, чтобы понять, как устроены данные в выбранной предметной области, на какие категории подразделяются данные — объекты, которые Вы будете анализировать и рекомендовать в рекомендательной системе, которую разработаете в будущих лабораторных.

Поясняющий пример к ЛР1. Есть дерево, в нём жанры являются потомками узла с категорией числа выпусков (для предметной области периодических изданий: сколько лет выпускается издание либо количество выпусков). Это пример ошибки проектирования дерева. Количество

выпусков — скорей признак объекта, а жанр — это полезный узел в дереве, который позволяет разделить всё множество изданий по некоторому смысловому признаку. Зная иерархию жанров, можно будет сравнивать отдельные издания по близости в дереве, где такая информация отражена. Не следует добавлять в дерево те узлы, что не являются смыслообразующими.
Вопрос: в теории кол-во выпусков тоже может являться смыслообразующим? Например, кто-то может любить маленькие серии, кто-то — большие.
Ответ: это может решаться фильтром при поиске (см. ЛР5) и/или при сравнении двух объектов (согласно выбранной комплексной мере близости (см. ЛР2)). Например, если количество выпусков сходно (и пользователю важен этот критерий), это даст вклад в меру сходства. Возможно, стоит брать не количество выпусков, а категорийный признак (короткая серия, средняя серия, длинная серия, очень длинная серия), впрочем, это вопрос проектного решения.
Итак, числовые признаки следует анализировать на стадии фильтров или в мере близости или расстояния, при сравнении. А разделение на жанры — это как раз вопрос древесной структуры, описывающей устройство данных в предметной области. Древесная структура предметной области будет затем использована древесной мерой близости, которая войдёт в состав обобщающей меры близости (ибо сравнивать объекты нужно, обладая знанием этой древесной структуры).
Итак, чтобы формализовать знания о предметной области, понять, как разбиваются наши данные на группы и подгруппы, чтобы в дальнейшем можно было точными методами определить сходство или расстояние между двумя сравниваемыми объектами с учётом знаний о предметной области. Часть знаний содержится в признаках, часть знаний Вы формализуете в виде дерева. Обе части анализируются в ЛР2.

ЛР2:
2.1. Выбрать не менее 3 мер близости для сравнения сходства двух объектов путем оценки сходства (расстояние и сходство могут рассматриваться как взаимно обратные величины в общем случае), включая одну ассоциативную мерублизости. Реализовать меры оценки близости (сходства) двух объектов. Объекты в лабораторных, посвящённых рекомендательным системам, — это объекты утверждённого датасета.
Примеры ме: евклидово расстояние, расстояние городских кварталов, косинусная мера близости, расстояния Чебышёва, Минковского.
Реализованные меры могут оценивать все признаки объектов или же их часть.
2.2. Мера близости по дереву (древесная). Создайте меру оценки расстояния между узлами в дереве, которое Вы составили в ЛР1. Мера расстояния по дереву — топологическая; на некоторых датасетах может возникнуть проблема неравномерности расстояний между поддеревьями и/или внутри них. Так, например, в предметной области кухонь мира для среднестатистического пользователя расстояния больше между группами кухонь (например, на 1м уровне дерева расположены азиатская и европейская кухни, на 2м — кухни стран регионов), а кухни в пределах регионов примерно равноудалены. Для пользователя же продвинутого уровня (например, разбирающегося в кухнях некоторого региона) различия между кухнями одной страны могут быть столь же велики, как и между кухнями других регионов: например, две отдельных китайских кухни могут для такого пользователя быть не менее далёкими, чем для среднестатистического пользователя — итальянская и китайская.
2.3. Проведите эксперименты по сравнению объектов. Выберите ту или те меры, которые лучше всего позволяют сравнивать объекты Вашей предметной области. Также на этом этапе можно отсеять неподходящую меру.
2.4. Реализуйте обобщающую меру, которая учитывает ВСЕ признаки Ваших объектов. Используйте предыдущие результаты.

Особенное внимание обратите на те признаки, которые отмечены в задании на ЛР1 как не приводимые к числовым. Одна из возможностей сравнения тегов — использование библиотеки Word2Vec, которая позволяет оценивать семантическую близость слов на основании машинного обучения на материале совместной встречаемости слов в текстах обучающей выборки (т.е. контекстной близости слов); другая возможность — формирование экспертной оценки (например, в форме матрицы смежности, в которой приведена попарная близость возможных значений признаков).
Обобщающая мера может быть линейной комбинацией частных мер близости.
2.5. Возможное [дополнительное] задание: используйте меру корреляции. Корреляция используется для выявления взаимосвязанных данных, особенно важно это для многомерных данных при сокращении размерности пространства признаков. Введение в корреляцию см. https://nafi.ru/upload/spss/Lection_6.pdf Оцените взаимную зависимость Ваших признаков.

ЛР3. Контент-ориентированная рекомендательная система
3.1. Вход: 1 объект (затравочный). Выход: список рекомендаций, ранжированный по убыванию близости с затравкой. Примените Вашу обобщающую меру близости.
3.2. Вход: массив объектов (лайков). Выход: сформированный ранжированный список рекомендаций.
3.3. Вход: массив затравочных объектов и массив дизлайков. Выход тот же. Реализуйте механизм дизлайков.
3.4. У рекомендательной системы должен быть пользовательский интерфейс (консольный или графический), который позволяет пользователю задать свои предпочтения и выполнить поиск по каждому из трех сценариев. Примечание: передача параметров методу (в т.ч. через консоль) пользовательским интерфейсом не считается. Проверка корректности вводимых значений полностью лежит на разработчике, то есть на Вас.

ЛР4. Параметрический поиск.
4.1. Следует реализовать интерфейс (консольный или графический, см. п. 3.4) для параметрического поиска по фильтрам (для примера можно взять фильтры в крупных электронных магазинах или музыкальных и кино-ресурсах). Пользователь может задать ограничения по фильтрам (или не задать их для части фильтров) и получить выборку подходящих объектов.
4.2. Если поисковая выдача пуста, реализуйте механизм, формирующий выборку с формулировкой “не найдено точного соответствия, однако, возможно, Вам понравится”. Возможно, Вы используете механизм формирования рекомендаций, а возможно, несколько отодвинете слишком строгие границы фильтров.

РК1. Объедините ЛР3 и ЛР4 в единую систему с общим интерфейсом.
Возможно, Вам понадобится функция полезности, в которой Вы будете начислять очки за совпадения и/или накладывать штраф за несовпадения.
Требуется добавить функциональность фильтрации полученных рекомендаций. Следовательно, требуется хранить историю запросов, чтобы можно было уточнять результаты предыдущего поиска/запроса рекомендаций либо по запросу начать поиск заново.

Часть 3. Диалоговые системы

3.1. Введение

Искусственный интеллект использует различные человеко-машинные интерфейсы[1]. Традиционный интерфейс — клавиатура для ввода текста и визуальный канал отображения информации. При этом более естественным для человека считается взаимодействие на естественном языке[2], посредством естественно-языковых интерфейсов. Примеры: голосовые помощники Яндекс Алиса [1, 12], Apple Siri [8, 9], Microsoft Cortana [10, 11], Microsoft Xiaoice [6], Amazon Alexa [13], Google Assistant [14]. В частности, для русского языка есть решение Яндекс SpeechKit [18]: можно получить ключ для бесплатного использования с лимитом запросов в день.

Обработке текстовых данных посвящена дисциплина “компьютерная лингвистика”, она же (с точностью до нюансов) “машинная лингвистика”, в более общем смысле — обработка естественного языка, или natural language processing (NLP). Выделяют [15] следующие ключевые этапы обработки текста[3].

1. Графематический анализ (также токенизация) — поток символов разбивается на токены, предложения, абзацы. Здесь решаются задачи склейки слов, разделённых переносами, расшифровки сокращений, и т.п.
2. Морфологический анализ — для словоформ определяются начальные формы и морфологические свойства. Ключевая проблема этапа — омонимия, для русского языка — омография (например, словоформа “стали” может относиться к слову с начальной формой “сталь”, сущ. (“марки стали”), или к слову с начальной формой “стать”, гл. (“птицы стали на крыло”). Упрощённая альтернатива этапу — стемминг, или выделение неизменяемой основы слова (это лучше работает для английского языка, т.к. в русском языке есть слова с беглыми и изменяющимися буквами в корне, а также совсем короткие неизменяемые части слова, или стемы, например, для “быть”, “буду” стемом будет “б”. Одно из лучших средств для морфологического анализа для русского языка — библиотека *pymorphy* [16], обученная на корпусе OpenCorpora [17].
3. Синтаксический анализ — определяются связи между словами в предложении. Наиболее полную информацию предоставит дерево зависимостей. Этот этап имеет свои неоднозначности разбора, которые множатся из-за омонимии. Альтернативы — частичный семантический анализ и выделение всех связей между словами в пределах N-граммы [15], обычно N-граммы выделяются в пределах предложения. Полученные связи используются на основании предположения, что близко расположенные слова связаны между собой, что можно подвергнуть критике. Так, в некоторых задачах можно проводить фильтрацию служебных частей речи, если они не несут смысла в конкретной задачи (например, чтобы связать сказуемое с дополнением, достаточно провести прямую связь между соотв. словами в пределах N-граммы, а предлог можно использовать для уточнения типа связи).

---

[1] Определение см. 1[ю] главу книги “Речевой и многомодальный интерфейсы” из библиотеки кафедры, спрашивайте у старост, см. раздел 3.2 данного документа

[2] Естественный язык – термин, противопоставляемый формальному языку, искусственно созданному

[3] См. главу Э.С. Клышинского из учебного пособия на сайте http://clschool.miem.edu.ru/%D0%9C%D0%B0%D1%82%D0%B5%D1%80%D0%B8%D0%B0%D0%BB%D1%8B-%D1%88%D0%BA%D0%BE%D0%BB%D1%8B.html

4. Семантический анализ — определение смысла текста, или его прагматики. Здесь часто используются грамматики[4], как и в синтаксическом и синтактико-семантическом анализаторах. По итогам формируется некоторое внутреннее представление смысла, или прагматики, входного текста. Если ставится такая задача, то внутреннее представление программой смысла преобразуется (например, ищется ответ на вопрос, фраза адаптируется к иностранному языку, формируется некий иной смысловой ответ на входящее воздействие) и затем проходит в обратном порядке этапы синтеза текста: семантический, синтаксический (строится дерево зависимостей), морфологический (слова обретают форму и параметры, ставятся в нужную форму), графематический (получается текст).
Примечание: для введения в этапы анализа текста настоятельно рекомендуется глава III “Начальные этапы анализа текста” учебного пособия [15].

Ключевые семантические анализаторы и вопросно-ответные системы для русского языка — ABBYY Compreno [3], IBM Watson [2], Яндекс Алиса [1], Ф-2 [4] (робот не отвечает на вопросы, но скорее представляет собой собеседника, который комментирует произносимые человеком фразы) — часто ограничивают домен анализируемой информации (IBM Watson адаптирована для медицинской рекомендательной системы, Яндекс Алиса ограничена заложенными в неё сценариями взаимодействия) либо проводят семантический анализ на основании грамматик и/или словарей семантических маркеров, что служит естественным ограничением области их знаний. Для таких систем нужны эксперты для составления грамматик и сценариев, наполнения словарей.
Часто используются семантические признаки и семантические валентности (роли, по аналогии с химией, когда объекты могут вступать только в определённые типы связей), например, валентности из [5] используются в [4].

Источники литературы:

1. Как устроена Алиса. Лекция Яндекса [эл. ресурс]. Режим доступа: https://m.habr.com/en/company/yandex/blog/349372/ (дата обращения 17.06.2019).
2. Когнитивная система IBM Watson [эл. ресурс]. Режим доступа: https://m.habr.com/en/company/ibm/blog/266015/ (дата обращения 17.06.2019).
3. Anisimovich K. V. et al. 2012. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11(18): В 2 т. Т.2: Доклады специальных секций. М.: Изд-во РГГУ. С. 91–103.
4. Робот Ф-2 [эл.ресурс]. Режим доступа: http://f2robot.com/robot/ (дата обращения: 13.11.2020).
5. Wierzbicka A. 1980. Lingua Mentalis: The semantics of natural language. New York: Academic Press.
6. Microsoft Xiaoice и тест Тьюринга [эл. ресурс]. Режим доступа: https://rb.ru/story/Xiaoice/ (дата обращения: 13.11.2020).
7. Voice assistants (обзор) [эл. ресурс]. Режим доступа: https://en.wikipedia.org/wiki/Virtual_assistant (дата обращения: 13.11.2020).

[4] Грамматики, см. дисциплины “Дискретная математика” [Великая синяя книга А.И. Белоусова], бакалавриат, и “Теория формальных языков”, магистратура [там же]

8. The Story of Siri, by its founder Adam Cheyer [эл. ресурс]. Режим доступа: https://medium.com/wit-ai/the-story-of-siri-by-its-founder-adam-cheyer-3ca38587cc01 (дата обращения: 13.11.2020).
9. Siri for developers [эл. ресурс]. Режим доступа: https://developer.apple.com/siri/ (дата обращения: 13.11.2020).
10. Introduction to Cortana intelligence suite [эл. ресурс]. Режим доступа: https://social.technet.microsoft.com/wiki/contents/articles/36688.introduction-to-cortana-intelligence-suite.aspx (дата обращения: 13.11.2020).
11. Principles of Cortana skills design — MSDN [эл. ресурс]. Режим доступа: https://docs.microsoft.com/en-us/cortana/skills/design-principles (дата обращения: 13.11.2020).
12. Как создать навык для Алисы с нуля — Академия Яндекса [эл. ресурс]. Режим доступа: https://academy.yandex.ru/posts/kak-sozdat-navyk-dlya-alisy-s-nulya (дата обращения: 13.11.2020).
13. How Amazon Alexa works: your guide to natural language processing (AI) [эл. ресурс]. Режим доступа: https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3?gi=d983f99812f3 (дата обращения: 13.11.2020).
14. Google Assistant: The complete history of the voice of Android / Digital Trends [эл. ресурс]. Режим доступа: https://www.digitaltrends.com/mobile/google-assistant/ (дата обращения: 13.11.2020).
15. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
16. Морфологический анализатор pymorphy2 [эл. ресурс]. Режим доступа: https://pymorphy2.readthedocs.io/en/stable (дата обращения: 13.11.2020).
17. OpenCorpora — открытый корпус [эл. ресурс]. Режим доступа: www.opencorpora.org (дата обращения: 13.11.2020).
18. Яндекс SpeechKit API [эл. ресурс]. Режим доступа: http://api.yandex.ru/speechkit/ (дата обращения: 10.04.2017).

3.2. Диалоговые системы и мультимодальные интерфейсы

Как было сказано выше, искусственный интеллект использует человеко-машинный интерфейс для коммуникации с человеком. Это могут быть текстовые интерфейсы, голосовые (например, [18]), жестовые, мультимодальные (например, [5]).

Лекция:
1. Главы 1 и 3 книги: А.Л. Ронжин, А.А. Карпов, И.В. Ли. Речевой и многомодальный интерфейсы [выслано старостам].
2. Продолжение — две статьи тех же авторов, см. ветка из 3 сообщений в твиттере: http://twitter.com/iu_bublik/status/1332205347117936640/
3. Некоторые полезные примеры Вы найдёте на страницах 61-75 (про семантику) и 133-146 (последние можно просмотреть по диагонали ради примеров) в книге: Терри Виноград. Программа, понимающая естественный язык.
Книга классическая, находится в открытом доступе, например, тут: https://www.studmed.ru/terri-vinograd-programma-ponimayuschaya-estestvennyy-yazyk_d485ebb2c21.html

3.3. Примеры использования автоматической обработки естественного языка для формирования рекомендаций

Пример автоматического анализа описаний ароматов для формирования рекомендаций о парфюме на основании краткого описания назначения парфюма:
Анастасия Бодрова. Чат-бот подбирает парфюм [эл. ресурс]. Режим доступа: https://sysblok.ru/nlp/chat-bot-podbiraet-parfjum/ (дата обращения: 19.11.2021).

3.4 Доп.материалы к диалоговой системе для вдохновения

Могут быть прочитаны и/или просмотрены даже за чаем.

1. Рассказ “ALDAN M.A.G. 3,14”, или ради какой прекрасной цели стоит писать ИИ вида диалоговая система. По мотивам "Понедельник начинается в субботу".
2. Видео из доп.материалов к Wall-E про озвучку мультфильма и про удивительные решения в озвучке на студии Дисней (и немного в Звёздных войнах). youtube: Wall-E Animation Foley and Sound Design
3. В продолжение пред.пункта: м/ф студии Pixar “Burn-E” и “Smash and Grab” (короткометражки[5]). Обратите внимание на озвучку и на то, что именно создаёт “характер” роботов, на их особенные чёрточки.

3.4. Задания на второй блок лабораторных

Все задания выполняются в привязке к русскому языку.

ЛР6. Тематика и сценарии бесед. Вам нужно выбрать тему диалога с ИИ (Вашим собственным мини-ИИ), это по умолчанию может быть та же тема, что и в первой части лабораторных. Если Вы меняете тему, отметьте это в том же гугл-документе, НЕ стирая старой темы (ибо это разные блоки лабораторных). Допустим, это погода. Вам нужно выделить возможные сценарии бесед на эту тему, по возможности исчерпывающий перечень: например, возможно спросить про погоду (перечень параметров запроса, которые можно извлечь из фразы на естественном языке, варьируется в зависимости от того, указал ли человек на дату/время, город/район), спросить, пойдёт ли дождь/брать ли зонт, задать вопрос в форме “сегодня жарко?” или “сколько градусов жары?”, а также произнести фразу, подразумевающую ответ ИИ и, возможно, требующую промежуточного умозаключения (например, “холодно, не правда ли?”, “пора брать водные лыжи! когда это кончится?”, “что у нас с погодой [где] [время — сейчас (определить время) либо требуется уточнить]?”). Также возможен сценарий “английской” общей беседы на тему погоды, возможно извлекать анекдоты про погоду из памяти, ассоциировать события с такой же погодой (“в последний раз, когда так лило,...”), перевести разговор на климат и пр.
Что на выходе: документ с описанием темы, набора сценариев беседы на выбранную тему с примерами, выделение шаблонов фраз (используйте местозаменители и расшифровывайте их).
Минимум: 25 вопросов к диалоговой системе; постарайтесь каждый переформулировать 2-3 способами. Выделите возможные типы/классы вопросов и те способы, которыми их можно задать. Результат, к которому мы идём, — сценарии для диалоговой системы.

[5] М/ф студии Pixar “Wall-E” (полный метр) оставим на каникулы

Пример: краткая форма выжимки из анализа способов, которым может быть задан запрос на выборку ноутбуков с заданными параметрами из базы (псевдокод регулярного выражения[6]). Пусть описаны нетерминалы [А.И. Белоусов, С.Б. Ткачёв. Дискретная математика]:

ge={от|с|начиная с|минимум|больше|более|выше|мощнее|не меньше|не менее|не ниже|не слабее}

le={до|вплоть до|максимум|не больше|не более|не выше|не мощнее|меньше|менее|ниже|слабее}

qe={около|примерно|порядка|в районе}

vs={ge|le|qe}

avail1i=[имеющиеся [у {В|в}ас][в наличии [у {В|в}ас]][в продаже [у {В|в}ас]]]

avail2i=[поступившие [к {В|в}ам][в продажу][{в|на} {<срок>|<дата>|<интервал времени>}]]

avail3i={[, имеющиеся][в наличии][в продаже]|, поступившие [к {В|в}ам][в продажу][{в|на} {<срок>|<дата>|<интервал времени>}]}

avail1t=[имеющимися [у {В|в}ас][в наличии [у {В|в}ас]][в продаже [у {В|в}ас]]]

avail2t=[поступившими [к {В|в}ам][в продажу][{в|на} {<срок>|<дата>|<интервал времени>}]]

avail3t={[, имеющимися][в наличии][в продаже]|, поступившими [к {В|в}ам][в продажу][{в|на} {<срок>|<дата>|<интервал времени>}]}

Тогда упомянутый вопрос из списка может выглядеть так (псевдокод регулярного выражения, про регулярные выражения рекомендуется [Джеффри Фридл. Регулярные выражения. Глава 1]):

{{{М|м}не нужны|{П|п}окажи[те]|{М|м}еня интересуют} {[все][avail1i|avail2i]ноут{ы|буки}|[все]ноутбуки[avail3i]}|{И|и}нтересуюсь [всеми]{[avail1t|avail2t]ноут[бук]ами|ноут[бук]ами[avail3t]}|{К|к}ак [у {В|в}ас]{с наличием ноут[бук]ов|с ноут[бук]ами|насч{ё|т}т[наличия] ноут[бук]{а|ов}}} с <название параметра> {[vs] <значение>|ge <значение> [le <значение>]|le <значение> [ge <значение>]}[{и |или |и ещё |а также |; |, | }с <название параметра> {[vs] <значение>|ge <значение> [le <значение>]|le <значение> [ge <значение>]}]*[?]

Регулярные выражения — не единственный способ описать один вопрос, но он, возможно, наиболее наглядный.

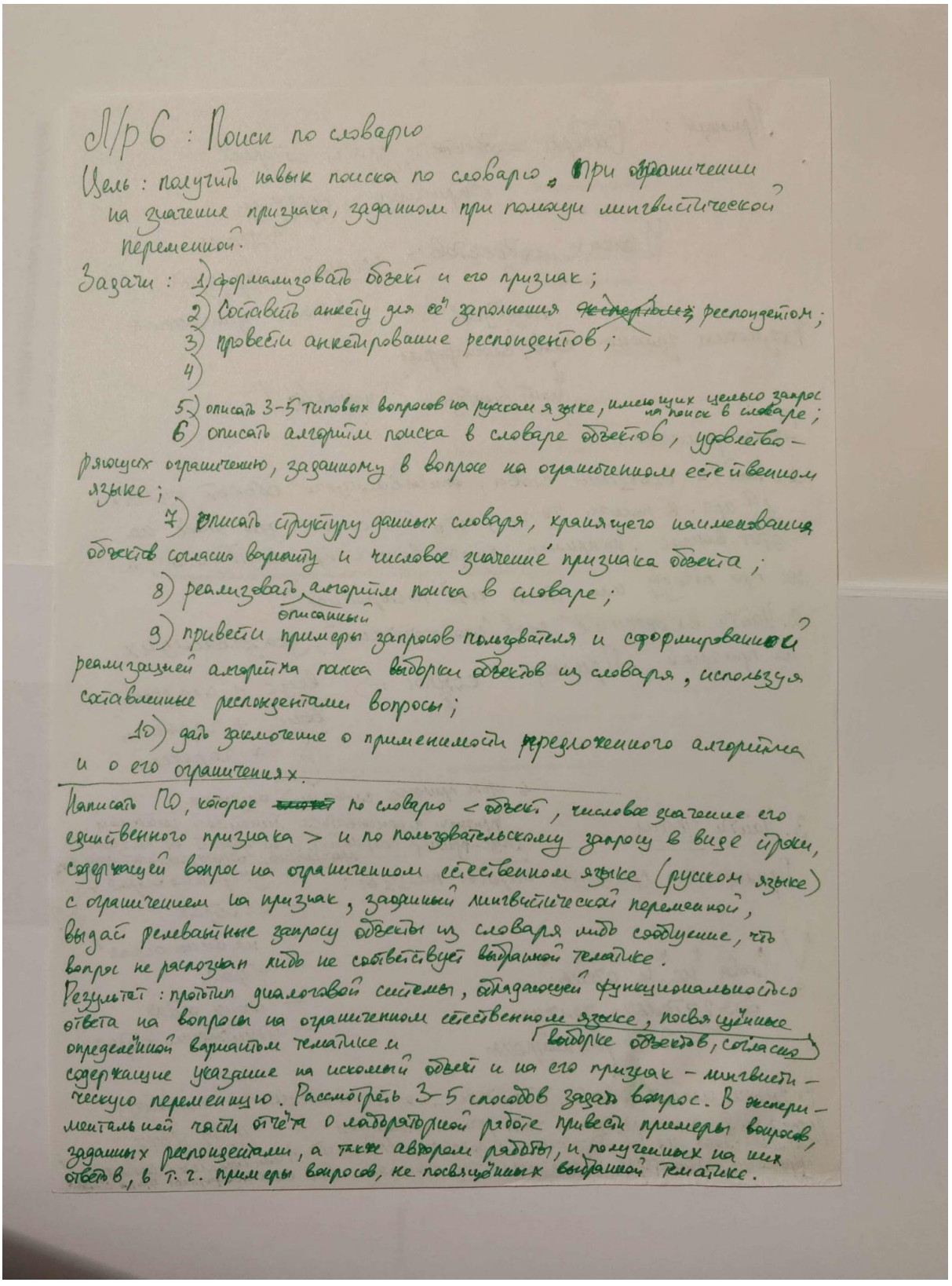
ЛР7. <u>Поиск по словарю при ограничении на значение признака, заданном при помощи лингвистической переменной.</u>

Лабораторная связана с нечёткими функциями принадлежности значений некоторого параметра термам – категориальным словам, описывающим значения признака, например, “большая”, “средняя”, “малая” скорость. За счёт того, что пользователь вводит в текстовом запросе словесное описание значения(й) признака в виде термов, признак здесь задаётся через лингвистическую переменную.

См. фото 1, фото 2 (для просмотра в полном качестве нужно сохранить фотографию из этого документа). Пропущена 4я задача: построить функцию принадлежности термам числовых

[6] Рекомендуется 1 глава книги: Джеффри Фридл — Регулярные выражения

значений признака, описываемого лингвистической переменной, на основе статистической обработки мнений респондентов, выступающих в роли экспертов.



Л/р 6 : Поиск по словарю

Цель : получить навык поиска по словарю при ограничении на значение признака, заданном при помощи лингвистической переменной.

Задачи : 1) формализовать объект и его признак;
2) Составить анкету для её заполнения ~~экспертом~~ респондентом;
3) провести анкетирование респондентов;
4)
5) описать 3-5 типовых вопросов на русском языке, имеющих целью запрос на поиск в словаре;
6) описать алгоритм поиска в словаре объектов, удовлетворяющих ограничению, заданному в вопросе на ограниченном естественном языке;
7) описать структуру данных словаря, хранящего наименования объектов согласно варианту и числовое значение признака объекта;
8) реализовать описанный алгоритм поиска в словаре;
9) привести примеры запросов пользователя и сформированной реализацией алгоритма поиска выборки объектов из словаря, используя составленные респондентами вопросы;
10) дать заключение о применимости предложенного алгоритма и о его ограничениях.

Написать ПО, которое ~~имеет~~ по словарю <объект, числовое значение его единственного признака> и по пользовательскому запросу в виде строки, содержащей вопрос на ограниченном естественном языке (русском языке) с ограничением на признак, заданный лингвистической переменной, выдаёт релевантные запросу объекты из словаря либо сообщение, что вопрос не распознан либо не соответствует выбранной тематике.

Результат : прототип диалоговой системы, обладающей функциональностью ответа на вопросы на ограниченном естественном языке, посвящённые определённой вариантом тематике и выборке объектов, согласно содержащие указание на искомый объект и на его признак - лингвистическую переменную. Рассмотреть 3-5 способов задать вопрос. В экспериментальной части отчёта о лабораторной работе привести примеры вопросов, заданных респондентами, а также авторами работы, и полученных на них ответов, в т. ч. примеры вопросов, не посвящённых выбранной тематике.

Примеры: Выбери <объекты>, для/у которых <признак> средний

↑ терм

У каких <объектов> средний <признак>?

↑ ноутбуков ↑ диагональ

Возможны замены слов/словоформ

ноут/ноутбук → ноутбов/…/

Этапы решения:

1. Найти вхождение слова, описывающего объект => это 1й арг. в пользу того, что ДС сможет дать ответ на этот вопрос. Например, задан вопрос про ноутбуки, а не про погоду или что-то ещё постороннее.
2. Найти вхождение слов/словосочетаний, указывающих на признак. Пример: компот => вкус + сладкий; вкус + горький; λ + горький

в этом примере слово м.б. опущено

3. Найти терм

Признак д. описываться линейной шкалой. Вкус - пример спорный, т.к. он зависит от разных признаков; острота, сладость - допустимы.

Пример для признака остроты пиццы:

- умеренно острая
- не острая
- не очень острая (≡? умеренно острая)
- острая
- очень острая
- убийственно острая

! Слова "не" и "очень" обязательно должны быть рассмотрены

Возможно использовать библиотеку морфологического анализа pymorphy2 для определения части речи и начальной формы слова, но не обязательно: достаточно выделить 1 или несколько словоформ/фрагментов словоформ: диагонал без окончания или все словоформы этого слова

9.12.2022

Список термов утверждается вместе с лингвистической переменной и объектом, который ее описывает, а также нужно указать единицы измерения признака. Пример: объект – *компот*, числовой признак – *сладость* как уровень сахара в веществе, в % или единицах Bitrix. Термы:

— не сладкий/очень не сладкий (*)

— не очень сладкий

— средне сладкий/средний

— сладкий

— очень сладкий

— приторно сладкий

(*) У одной категории могут быть два разных обозначения. Для компота термы "не сладкий" и "очень не сладкий" совпадают, но для других данных описания свойства Х признака У могут не совпадать: "не Х" и "очень не Х" по отношению к значению признака У.

Если оценка респондентом значения (например, сладость) не очевидна без специального прибора, следует предоставить вместе с анкетой примеры для респондентов, например, клюквенный бауманский компот, <подставьте число> %. В этом случае следует привести по 2 примера на каждый терм.

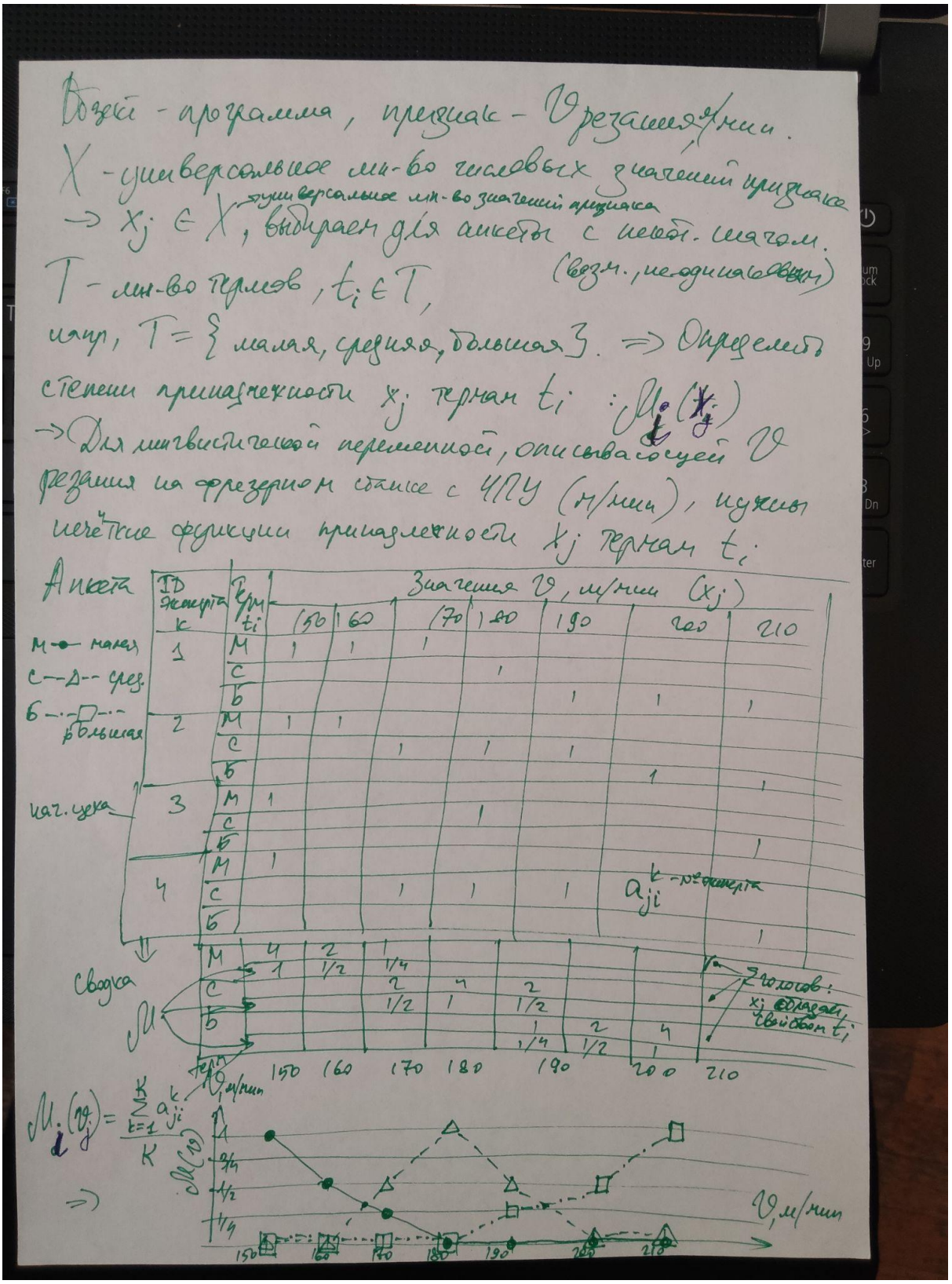
Результаты содержания сахара в соках

Название сока	Содержание сахара %	Заявленное производителем содержание
«Добрый» Вишня	10,9	Содержит сахар
«Фруктовый сад» Апельсин (нектар)	9,1	Содержит сахар
«Привет» Мультифрукт	8,9	Содержит сахар
«Агуша» Груша с мякотью	1,8	Без сахара
«Гранатовый» (натуральный осветленный) "Goycay-Sud" ATSC/ Азербайджан	8,6	Без сахара
«Моя семья» Виноград. (Нектар)	11,3	Сахарный сироп
J7 (яблочный) (Нектар)	7,9	Содержит сахар

Требуется построить семейство нечётких функций принадлежности значений признака x_j термам t_i – $\mu_i(x_j)$.

Потребуется собрать анкеты с респондентов (в лабораторной работе – не менее 4 шт.): подписи к столбцам – значения признака x_j, к строкам – термы t_i. В ячейках расположены a_{ji}^k – бинарные

значения, каждое из которых соответствует высказанному респондентом k мнения о том, что значение признака x_j имеет свойства терма t_i, k принадлежит отрезку от 1 до K, K – количество респондентов.



Респондент получает ОДНУ анкету. Анкеты могут быть представлены в отчёте по одной или в склеенном виде, как на иллюстрации. Затем формируется сводная таблица, в примере для каждого терма указаны две строки: в верхней приведена сумма голосов, в нижней – значение функции принадлежности. По каждой нижней строке терма строится функция принадлежности. Теперь можно представить все функции на графике. Диапазоны значений признака, соответствующие заданному терму, определяются по набору функций. Так, в примере графики функций принадлежности числовых значений термам “малая” и “средняя” пересекаются в точке, находящейся на отрезке между значениями 160 и 170. Значение скорости в точке пересечения служит границей между значениями, соответствующими терму “малая”, и значениями, соответствующими терму “средняя”.

Примеры решения с 3 курса:

https://drive.google.com/file/d/1oUKU5QHjRE8XE4kRYeUvBSBb0U3W__7y/view?usp=sharing

https://drive.google.com/file/d/121pZnjJizC7c0u6cMi0N6n7S1O35s8y0/view?usp=sharing

ЛР8. Поиск по словарю: анкетирование респондентов
ЛР засчитывается, если участвовать в опросах минимум 4 человек.

РК2. Эссе
Требуется написать эссе, около 1 стр. текста, в котором Вы по материалам курса лекций опишете, что такое ИИ (какие есть определения и как они помогают понять, что есть ИИ), какими свойствами и какой функциональностью обладают ИИ. Также нужно привести области применения ИИ с примерами.

ЧаВо

>> Каков формат сдачи лабораторных?
Лабораторные и РК подлежат защите. На сдаче требуется описать датасет, содержание лабораторной, принятые проектные решения и принципы работы кода (по коду). Также подготовьте примеры, на которых Вы продемонстрируете работу ПО. Обратите внимание, необходимо описать, какие меры-стратегии-подходы были применены и почему; как устроено решение и почему оно даёт конкретные результаты.

>> Как взвешивать альтернативы?
Возможно, Вы возьмёте несколько отдельных функций полезности для разнотипных аргументов и для каждой подставите свой коэффициент важности. Возможно, Вы будете накладывать штрафы за неудовлетворение запросу или попадание по списку отказных альтернатив/отписки. Мерилом успешности любых мер, в том числе Ваших составных мер (если Вы их делаете такими) будет результат — релевантность рекомендаций. Её Вы можете оценить это, потому что в рамках ЛР Вы сами проводите апробацию РС.

>> Как сдать задолженность?
Сдать всю заявленную практику: 9 ЛР, а также РК1 и РК2 (см. выше “Кому сдавать лабораторные…”). Сначала работы защищаются очно или, в случае иногородних и иностранных студентов, в дискорде. Уважительной причиной для сдачи в дискорде являются уважительные причины с точки зрения деканата: болезнь (с больничным), карантин, невозможность приехать в МГТУ иногородним и иностранным студентам.
Затем для защиты работ будут заданы вопросы, см. титульную страницу.
Внимание: перечень **тем** лабораторных блока 1 **уникален**, как и блока 2. Ссылка на гугл-документ с темами в части 1 данного документа. Если студент не зафиксировал уникальную тему, он(а) не может сдать лабораторные.
Направление на сдачу дисциплины (для того, чтобы оформить зачёт) студент самостоятельно получает у зам.декана, как всегда.

P.S. Прошу Вас пройти небольшой опрос по курсу СИИ, чтобы сделать этот курс лучше: (ссылка будет в декабре)