[1 – Вступление]

- Добрый день, я Олег Самсонов, сегодня я представляю проект по интеллектуальному анализу текстов рабочих программ дисциплин университета.

[2 – Введение]

- О чем этот проект? У нас есть некий кейс автоматизации – в данном случае это процесс по составлению рабочих программ дисциплин, и мы хотим использовать методы интеллектуального анализа текста, также известного как текстмайнинг, чтобы помочь сотрудникам университета выполнять его быстрее.

- Целью данного проекта является программа, которая будет обрабатывать описание лекций из РПД и сравнивать их при помощи интеллектуального анализа.

- Таким образом, мы заменяем ручную работу на автоматическую. Представьте, вам нужно прочесть два документа, которые могут содержать 20-30 страниц и сравнить их между собой. А если документов сотня или больше, сколько бы времени у вас бы на это ушло? Теперь мы заменяем чтение на автоматизированное решение, нужно подождать несколько минут, и мы получаем отчет, который говорит нам то же, чего мы добивались и до этого. Далее подробно на том, что нужно сделать.

[3 – Теория]

- Рабочая программа дисциплины, или РПД, это такой справочник с описанием учебного предмета, он основан на нормативах установленных Министерством Образования. В них прописана суть дисциплины, что ожидается от учащихся по окончании обучения, темы которые преподаются, используемые программы и рекомендуемая литература, контрольные вопросы к экзамену, и какие-либо комментарии.

- Логично предположить, что желательно чтобы по каждому предмету рассказывали что-то своё. Если предметы похожие, то это просто один предмет с разными названиями, и если несколько таких даются одному потоку учащихся, то мы в пустую тратим время и студентов и преподавателей.

- По этому после заполнении тем в специальном редакторе для РПД нужно проверить, не сильно ли сильно наш предмет похож на какой-то другой. Как уже говорилось, присвистывать все другие РПД уже тяжело в рамках кафедры, а нужно это сделать для всего университета и учитывать предыдущие года. По этому мы хотим как-то это сделать быстро.

[4 – Набор данных]

- С чем мы можем работать? Вполне очевидно что это будут содержания занятий. В РПД они представлены как таблица или список тем сортированных по порядку чтения. Раз это текст, его можно изьять программно.

- Также для классификации нам понадобится информация о дисциплине, например к какому направлению она относится и какая кафедра её читает. Это расписано на титульном листе РПД.

- Если у нас РПД находятся не локально, то их нужно сначала загрузить. Это можно сделать по FTP протоколу, и мы можем считать сервер как источних данных.

- Нужно учесть, что несколько РПД могут быть сгруппированы в один общий документ, а также документы могут хранится в разных папках.

[5 – Алгоритм]

- Для анализа проходим через следующие шаги. У меня всё происходит через Питон скрип.

- Сначала нужно преобразовать РПД в датасет. Для этого мы создаём JSON структуру, в одной части которой находится справочная информация, а в другой содержание лекций списком, так РПД легче обрабатывать. Технически, doc файлы имеют древовидную структуру, нам нужно только для себя скопировать текст из определенных мест в документе. Я использую библиотеку Python-Docx.

- Затем мы проводим токенизацию, мы убираем всю пунктуацию, заменяем заглавные буквы на строчные, и делим слова в пробелах. Таким образом у нас есть большой список слов для каждой лекции.

- Затем проводится лемматизация, мы приводим все слова к нормальной форме. Я для этого использую PyMorphy3, который основывается на словаре славянских языков. Теперь в наших токенах есть только понятия, но не род или падежи.

- Дальше, мы добавляем к токенам весам. Можно использовать мешок слов, или просто посчитать количество каждого токена в наборе, или обратную частоту, мы сравниваем пропорции токенов с их встречаемостью во всех документах. Также мы можем составить индекс взаимности, посчитать в какой мере два токена встречаются вместе. Результат у каждого метода похожий, но может мы захотим использовать разные алгоритмы чтобы посмотреть что они скажут.

- Дальше мы убираем стоп-слова, это не-информативные токены, которые только искажают нам картину. Мы удаляем из датасета все союзы, частицы и предлоги, полный список можно получить из библиотеки NLTK. Альтернативно, мы убираем токены, веса которого находятся в топ 5%, если они встречаются гораздо чаще чем все другие, наверное это не информативные слова. Также мы убираем токены которые встречаются всего раз, они либо опечатки, либо слишком специфичные чтобы показать что-то наглядное.

- У нас получилась гигантская таблица, в столбцах которых прописаны токены, в строках лекции, а в ячейках веса. Таблица с сотнями значений наверное будет не выразительна, по этому лекции складываются вместе для каждого РПД, а сами РПД группируются по направлениям и кафедрой. Для выбора группы, сделаем оконный интерфейс, для предыдущих шагов также.

- В конце мы делаем корреляционный анализ, мы сравниваем все РПД в выборке с каждым РПД. Это математическая процедура и у неё есть несколько вариантов, и библиотека Pandas сделает её за нас. В итоге получится диагональная матрица в которой все дисциплины мы выбрали сравниваются между собой по содержанию.

[6 – Результат]

- Матрица окрашена методом «тепловая карта». Смысл данной матрицы заключается в том, чтобы найти похожие РПД, чем ярче ячейка, тем выше совпадение.

- Преподавателю нужно выяснить, есть ли у нас дисциплины в однов направлении, которые слишком похожи, или наоборот, у нас есть межпотокавая дискиплина, но в каких-то направлениях есть различия.

- Например, в данном рисунке видно, что «Технологии хранения и интерактивной обработки данных» и «Математическое обеспечение систем управления данными» очень похожи, а «Проектирование систем управления данными» и «Системы управления данными» практически совпадают, кафедре стоит пересмотреть эти РПД.

[7 – Выводы]

- Таким образом, мы использовали интеллектуальный анализ текста чтобы обнаружить РПД, которые очень похожи по содержанию, хотя написаны для разных предметов. Это заняло около трёх минут, если бы преподаватель искал бы это вручную, это бы заняло несколько месяцев, и не факт что некоторые совпадения были бы вообще заметны. Это аргумент для внедрения алгоритмов обработки больших данных для работы с документами.

- Также, так как у нас есть датасет со всеми темами, его возможно скормить какой-нибудь нейронной сети, возможно в будущем не нужно будет и составлять РПД, а только предоставить лекции для описания.

[8 – Выход]

- На этом у меня всё, благодарю за внимание.