2019/20 учебный год

Всероссийская олимпиада студентов «Я – профессионал»

Задания заключительного (очного) этапа по направлению **«Большие данные»**

Задача 1.

Листая ленту постов социальной сети, исследователь Андрей заметил, что его друзья увлекаются разными вещами: спортом, археологией, программированием, машинным обучением и т.д. Собрав посты всех сообществ своих друзей за последнее время, Андрей решил провести анализ данных социальной сети среди своих друзей и определить характерные увлечения друзей по их группам.

К сожалению, у него нет размеченных данных с правильными темами. Помогите Андрею построить тематическую модель сообществ пользователей социальной сети и определить характерные темы.

Участникам предоставлены тексты постов социальной сети по сообществам (группам, пабликам) пользователей. Необходимо выделить кластеры в сообществах и сформировать интерпретируемые темы. Тема в данном случае — это набор статистически связанных слов. Известно также, что всего таких тем выделяется 16.

Входные данные: `vk_clients_text.csv` (https://yadi.sk/d/N1c-yJHA92z-6g) с колонками ‘owner_id, text`, где ‘owner_id это идентификатор группы пользователя, `text` это агрегированный текст всех постов в группе пользователя за определенное время (одна строка - одна группа). Длина текста и количество слов для каждой группы отличаются.

Выходные данные: после выделения тем в ответ следует указать в каждой строке через пробел первых пять слов (первых по возрастанию вероятности принадлежности к теме), характеризующих тему. Всего должно получиться 16 строк, одна строка - одна тема. Строки (темы) упорядочить между собой по возрастанию количества попавших в эту тему групп из исходного файла. Все слова можно писать с маленькой буквы и так, как они указаны в документе.

Верный ответ:

вода мясо лист добавить соль
рука нога сказать день поднять
вода кожа масло день мёд
рубль музей улица адрес стоимость
упражнение мышца организм тело тренировка
ребёнок год деньга работа родитель
фильм год сша хороший самый
человек мир год должный другой
петербург год место дом город
игра уровень друг страница очки
масло яйцо сыр приготовление минута
год русский война россия день

мама ребёнок дом собака папа
друг год репост сделать новый
человек жизнь любить любовь женщина
день хотеть любить новый год

Критерии проверки:

Участник получает 1 балл за строку, если в его ответе совпало хотя бы 2 слова (>= 2) из этой строки. Итого, максимально он может получить 16 баллов за задачу.

Решение:

Решением является разработанная тематическая модель текстов. Например, LDA или BigARTM с настроенными гиперпараметрами.

Задача 2.

Разочаровавшись в прогнозе погоды на выходные, метеоролог Семён решил сделать свою модель предсказания погоды. В качестве первого этапа для создания такой модели, он решил исследовать прогнозирование температуры воздуха. Данные атмосферных показаний собираются метеорологическими станциями по всему миру, каждая станция фиксирует историю показаний в данной географической точке за время своей работы. Он собрал временные ряды с показаниями величины температуры для нескольких станций за некоторый промежуток времени с часовым шагом, но обнаружил, что в них есть пропуски по времени различной длительности, от часа до месяца. Пропуском считается отсутствие значения в столбце. Помогите Семёну восстановить пропуски в наблюдениях для входных файлов.

Входные данные: пронумерованные csv файлы (https://yadi.sk/d/uoAvbPax7LMbTQ) с наблюдениями в определенной точке, со столбцами 'Date', 'T', каждая строка содержит время в формате 'YYYY-MM-DD hh:mm:ss' и значение температуры воздуха в градусах Цельсия или его пропуск. В каждом файле ‘station_id.csv’ может быть разное количество показаний и разное количество пропусков. Пропуски могут быть как в начале, так и в конце файла. Также в файлах могут содержаться выбросы (экстремальные значения), являющимися ошибками измерения.

Выходные данные: ваша задача выписать для первых семи файлов по два первых восстановленных значения в этом файле. У вас должна получиться последовательность из 14 чисел через пробел. Числа записываются с точностью до одного знака после запятой.

Верный ответ:

18.0 18.0 0.0 -1.0 -11.1 -11.6 17.0 15.0 -10.6 -10.0 -6.0 -6.0 18.0 17.0

Критерий проверки:

Следует рассчитать среднюю квадратичную ошибку = $\sqrt{(\sum(ai - bi)^2)14}$, где ai - ответ верный, bi - ответ участника. Если эта ошибка при округлении до двух знаков после

запятой получается <= 1.03, то за задачу присваивается максимальный балл = 15 баллов, иначе - 0 баллов.

Решение:
Решением является модель интерполяции данных. Данная модель может быть основана на выборе средних соседних значений или на аппроксимации какой-то функцией, например, линейной регрессией. Надо понимать, что ни одна модель не даст 100% точного предсказания пропусков, среднюю квадратичную ошибку можно лишь минимизировать до определённого уровня.

Задача 3.

Дата-саентист Кеша решил проверить, насколько названия глав «Алисы в стране чудес» соответствуют реально происходящим в тексте событиям. Он решил сделать всю необходимую предварительную обработку текста, включая приведение слов к нижнему регистру, удаление стоп-слов, цифр/неалфавитных символов и т. д. После формирования корпуса документов (глав книги) методом TF-IDF он сформировал ранжированные списки наиболее важных для каждой главы слов.

Какое слово имеет максимальное значение метрики TF- IDF для двенадцатой главы под названием «Alice’s Evidence» ?

Входные данные:

Книга «Alice in Wonderland» Льюиса Кэрролла на английском языке доступна по ссылке http://www.gutenberg.org/files/11/11-0.txt

Выходные данные:

Необходимо записать слово на английском языке, которое имеет максимальное значение метрики TF- IDF для двенадцатой главы книги «Alice in Wonderland».

Верный ответ:

king

Критерий проверки:
За верно решённую задачу участник набирает 15 баллов.
Решение:
Решением является алгоритм подсчёта метрики TF- IDF для каждого слова с предварительным удалением служебных слов.

Задача 4.

Вы являетесь разработчиком рекомендательной системы, которая предлагает туристам пешеходные маршруты по городским достопримечательностям. Ваша задача - построить наиболее интересный маршрут между заданными пользователем точками так, чтобы время прохождения маршрута укладывалось в определенные временные рамки. Время прохождения включает в себя как время на посещение каждой локации, так и время на дорогу. Под интересностью маршрута понимается количество посещений пользователей в социальных сетях, а также рейтинг мест. В приложенном к заданию файле находится набор из 5 тысяч локаций в Санкт-Петербурге. Ваша задача - проложить маршрут от Технологического института ("title": "Tekhnologichesky Institut") до улицы Некрасова ("title": "Ул. Некрасова"), который возможно пройти за 10 часов.

Входные данные:

Набор локаций города Санкт-Петербург - http://bit.ly/360xLYd

Выходные данные:

Набор локаций итогового маршрута. Имена локаций должны быть указаны так же, как в исходном наборе данных через запятую без пробела, каждое название в кавычках.

Верный ответ:

"Tekhnologichesky Institut", "Nevsky Prospekt", "The St. Petersburg Philharmonia (Small Hall)", "St.Isaac's Square", "Church of the Savior on Spilled Blood", "Peter and Paul Fortress", "Mikhailovsky Garden", "National Pushkin Museum", "Набережная Фонтанки", "Anichkov Bridge", "Ул. Некрасова"

Критерий проверки:

За каждую верно указанную локацию участник получает 1 балл. Максимум 11.

Решение:

В этой задаче решением является модель оптимизации с граничными условиями, где под граничными условиями понимается время преодоления маршрута.

Задача 5.

Известно, что значение показателя Y зависят от набора признаков X = (x1,x2,...x10). Требуется разработать модель, которая позволит верно предсказывать значения Y на основании значений X.

Также известно, что Y получается из X следующим образом:

$Y = Z^T * A2 + b2 + w;\ \ Z = cos(\ X * A1 + b1);\ \ w \sim N(0, \sigma^2)$ - *некий шум.*

где A1 - матрица размерности 10x2; Z и b1 - вектора длины 2; A2 - вектор длины 2, b2 - скаляр; ф-ция cos применяется к матрице поэлементно.

Обучите модель на тренировочной выборке и определите значение среднеквадратичной ошибки (MSE) на тестовой выборке.

Входные данные:

Тренировочный датасет (train.csv), тестовый датасет (test.csv).

Выходные данные:

Наилучшее из полученных вами значений MSE на тестовой выборке. Ответ укажите просто числом.

Стоит отметить, что из-за наличия шума значение MSE фактически ограничено снизу, поэтому нереалистично малые значения MSE не будут оцениваться.

Верный ответ:

18 < MSE < 24

Критерий проверки:

Если ответ укладывается в заданный интервал, то 15 баллов, иначе - 0 баллов.

Решение:

Регрессионная модель, имеющая наименьшее возможное значение метрики MSE.

Задача 6.

Имеется следующая конфигурация кластера:

10 одинаковых по производительности узлов с объемом оперативной памяти 64 Гб (на каждом узле), пропускная способность всей сети 10 Гбит/с. Скорость последовательного чтения жёстких дисков - 100 Мб/с, последовательной записи - 90 Мб/с.

Имеется набор данных на кластере, объёмом 2 Тб (2048 Гб), равномерно распределённых по узлам. Известно, что если бы обработка этих данных производилась одним сервером, то на это понадобилось бы 100 000 сек. Обработка предполагает выполнения операции shuffle, в которой участвуют 50% данных каждого

из узлов. Можно считать, что все задействованные данные в результате операции распределяются по узлам равномерно.

В качестве упрощения допускаем, что цикл действий узла при отправке данных: прочитать последовательно 100 Мб, сжать, отправить. Прием данных может осуществляться параллельно и происходит в специальный буфер в оперативной памяти (размер буфера). Цикл при получении данных: взять из буфера блок в 100Мб, разжать, записать на диск. Узел начинает новый цикл, не важно приема или передачи, только по окончанию предыдущего. Размер буфера 20% от общего объема памяти.

Вам необходимо подобрать такой кодек сжатия, чтобы добиться наименьшего времени выполнения обработки данных. Также можно отказаться от использования сжатия.

Имеется возможность использовать следующие виды кодеков:

Кодек №0:

Сжатие отсутствует

Кодек №1:

Степень сжатия - 2

Скорость сжатия - 100 Мб/с

Скорость разжатия - 120 Мб/с

Кодек №2:

Степень сжатия - 3.5

Скорость сжатия - 40 Мб/с

Скорость разжатия - 100 Мб/с

Кодек №3:

Степень сжатия - 5

Скорость сжатия - 30 Мб/с

Скорость разжатия - 100 Мб/с

Кодек №4:

Степень сжатия - 7.5

Скорость сжатия - 20 Мб/с

Скорость разжатия - 70 Мб/с

1) Какой кодек сжатия позволит добиться наименьшего времени выполнения задачи?

2) Аналогичный вопрос, только в случае пропускной способности сети 1 Гбит/с?

Ответ введите в формате номер кодека, пробел, время выполнения (с точностью до одного знака после запятой). Ответы для вопросов разделяются точкой с запятой. Пример: 1 12345.6; 0 6789.9

Решение:

1 случай.

На 10 узлах равномерно хранится 2048 Гб, т.е. 204.8 Гб на 1 узел. Передать необходимо 45%, т.е. 92.16 Гб (94371,84 Мб), т.к. данные передаются равномерно. Оставшиеся 5% идут на тот же самый узел с которого и отправляются. Соответственно в этом случае передачи и цикла сжатия/разжатия не будет.

10 Гбит/с = 1250 Мб/с

Каждый из узлов сможет использовать только 1 / 10 пропускной способности, т.е. 125 Мб/с.

Так как последовательность действий процессора - "прочитать 100 Мб, сжать, отправить и принять данные по сети, разжать, записать", то:

Т.к. у нас будет обмен 94371.84 Мб, то количество таких циклов 943.7184. Значит, необходимо найти минимальное время цикла.

В случае кодека 1:

Считывание 100 Мб - 1 сек

Сжатие - (100/100) - 1 сек

Отправка 50 Мб - (50/125) 0.4 сек

Разжать 50 Мб - (50/120) 0.417 сек

Записать 100 Мб - (100/90) 1.11 сек

Итого полный цикл займёт - 3.927 сек

В случае кодека 2:

Считывание 100 Мб - 1 сек

Сжатие - (100/40) - 2.5 сек

Отправка (100/3.5) 28.57 Мб - (28.57/125) 0.22856 сек

Разжать 28.57 Мб - (28.57/100) 0.2857 сек

Записать 100 Мб - (100/90) 1.11 сек

Итого полный цикл займёт - 5.12426 сек

Другие кодеки не имеет смысла считать, т.к. тут понятно, что мы упираемся в скорость сжатия, а не сеть.

В случае отсутствия сжатия:

Считывание 100 Мб - 1 сек

Отправка 100 Мб - (100/125) 0.8 сек

Записать 100 Мб - (100/90) 1.11 сек

Итого полный цикл займёт - 2.9 сек

Всего таких циклов - 943.7184, 943.7184*2.9 = 2736.78336 сек

Т.е. лучше отказаться от сжатия. Т.к. одной машиной обработка длилась бы 100 000 сек, то итоговое время - 10 000 + 2 736.78336 = 12 736.8

---

2 случай.

1 Гбит/с = 125 Мб/с

В случае кодека 1:

Считывание 100 Мб - 1 сек

Сжатие - (100/100) - 1 сек

Отправка - (50/12.5) 4 сек

Разжать - (50/120) 0.417 сек

Записать 100 Мб - (100/90) 1.11 сек

Итого полный цикл займёт - 7.527 сек

В случае кодека 2:

Считывание 100 Мб - 1 сек

Сжатие - (100/40) - 2.5 сек

Отправка - (28.57/12.5) 2.2856 сек

Разжать - (28.57/100) 0.2857 сек

Записать 100 Мб - (100/90) 1.11 сек

Итого полный цикл займёт - 7.1813 сек

Кодек 3:

Считывание - 1 сек

Сжатие - (100/30) - 3.33 сек

Отправка - (20/12.5) 1.6 сек

Разжать - (20/100) 0.2 сек

Записать - (100/90) 1.11 сек

Итого полный цикл займёт - 6.24 сек

Кодек 4:

Считывание - 1 сек

Сжатие - (100/20) - 5 сек

Отправка - (13,33/12.5) 1.0664 сек

Разжать - (13,33/70) 0.19 сек

Записать - (100/90) 1.11 сек

Итого полный цикл займёт - 7.3664 сек

Отсутствие сжатия:

Считывание 100 Мб - 1 сек

Отправка 100 Мб - (100/12.5) 8 сек

Записать 100 Мб - (100/90) 1.11 сек

Итого полный цикл займёт - 10.11 сек

Лучший цикл будет с кодеком №3.

Таким образом, лучшее время - 943.7184*6.24 = 5888.8028 сек

Итого: 15 888.8

Верный ответ: 0 12736.8; 3 15888.8

Критерий проверки:

За каждую верно решенную часть участник получает по 14 баллов. Итого максимально = 28 баллов.