Олимпиада. Направление «Компьютерная лингвистика»

Критерии

1. Задание 1. – 20 баллов

Пункт а)

10 баллов. Дан верный ответ с доказательством.

5 баллов. Дан верный ответ без достаточного обоснования. Или дан неверный ответ, но предложен алгоритм, который находит нужного пользователя за не более чем c*n запросов, где n – число пользователей, а c – не зависящее от n число.

0 баллов. Верный ответ с неправильным обоснованием, или неверный ответ и неоптимальный алгоритм.

Пункт б).

10 баллов. Верное доказательство.

5 баллов. Доказательство содержит пробел. Или доказано, что любой алгоритм из некоторого специального класса (вроде того, что мы описали выше) может не справиться за 2016 запросов. Или текст решения может быть интерпретирован и как верное, и как неверное рассуждение.

Пример пробела в доказательстве: «Один запрос позволяет исключить не более одного кандидата, значит, 2016 запросов позволят исключить не более 2016 кандидатов». Чтобы сделать такой вывод, нужно убедиться в том, что нет такого способа делать запросы и рассуждать, при котором, накопив ответы на 2016 запросов и пользуясь знанием о существовании искомого пользователя, можно было бы исключить 2017 кандидатов.

0 баллов. Доказывается неверное утверждение. Доказательство содержит неверные утверждения и пробелы. Доказательство полностью неверное.

Правильный ответ: **2017.**

Решение.

Пункт а)

Если на запрос вида «подписан ли пользователь X на обновления пользователя Y?» получен ответ «да», X не может быть искомым пользователем (т.к. тот не подписан ни на кого). Если же получен ответ «нет», Y не может быть искомым пользователем (на искомого пользователя подписаны все).

Пусть программа работает следующим образом. Сначала есть 2018 кандидатов на роль искомого пользователя. Программа выбирает пару кандидатов (X, Y), делает один запрос и исключает, пользуясь описанным выше соображением, либо X, либо Y из числа кандидатов. Далее процедура повторяется, пока не останется один кандидат.

Т.к. один запрос позволяет исключить одного кандидата, после 2017 запросов останется единственный кандидат, который и есть искомый пользователь.

Замечание: можно сказать, что мы описали тут некоторый класс возможных алгоритмов, которые отличаются способом выбора пары (X,Y) из списка кандидатов.

Пункт б). Пусть на самом деле дела обстоят так: 2017 пользователей зовут Иванами и все они подписаны на пользователя Марию, а друг на друга не подписаны (тогда из условия следует, что Мария ни на кого не подписана).

Допустим, программа (любая, не обязательно из описанного выше класса) сделала какие-то 2016 запросов. Получив список этих запросов и ответов на них, мы можем для каждого запроса исключить либо «пользователя X», либо «пользователя Y», пользуясь описанным выше соображением. Т.к. Иванов 2017, а запросов 2016, найдется как минимум один Иван, которого мы не можем исключить из рассмотрения этим способом. Выберем одного такого Ивана и будем называть его далее Ваней, а не Иваном.
Покажем, что и программа, как бы она ни была устроена, тоже не смогла исключить возможность, что Ваня – искомый пользователь.
Заметим, что среди сделанных запросов не было запроса вида "подписан(а) ли X на Ваню", потому что в противном случае мы бы получили ответ "нет" и не стали бы тогда называть Ваню Ваней. Запрос про то, подписан ли Ваня на Марию, также не был сделан, потому что в противном случае мы бы получили ответ "да" и не стали бы называть Ваню Ваней. Рассмотрим возможность, что каждый Иван подписан только на Марию и на Ваню, Мария подписана только на Ваню, а Ваня не подписан ни на кого и является искомым пользователем. На любой запрос, кроме тех, про которые мы только что установили, что они не были сделаны, ответы для реальной и этой гипотетической сети совпадают. Таким образом, при помощи запросов, которые были сделаны, рассматриваемую возможность исключить нельзя.

Замечание: можно изменить это рассуждение так, чтобы не использовать дополнительных предположений о структуре сети.

Отсутствие проверки корректности задачи не считалось недостатком решения. Баллы за это не снижались.

2. Неверно, что вор должен был признаться в краже или свидетельствовать о невиновности кого-то другого. Он мог сделать любое правдивое заявление, например: "Я знаю, кто украл бульон".
3. Из условия не следует, что показания троих подозреваемых позволили установить, что бульон украл лишь один из подозреваемых и что только он дал правдивые показания.

2. Задание 2. – 45 баллов

Критерии

а) Правильная идентификация всех перифраз с учетом принципов работы вопросно-ответных систем. Так как некоторые случаи спорные, за немногочисленные ошибки в идентификации перифраз снималось немного (1 балл). При определении того, является ли пара перифразами, необходимо было учитывать, что в условии вопрос касался не информационно-поисковой системы, а вопросно-ответной системы (диалоговой системы)

б) Должны были обязательно быть упомянуты: 1) нормализация текста, в том числе обработка буквенно-цифровых комплексов в примере 13); 2) спелл-чекинг, в частности, в примере 8; 3) перевод или транслитерация слов, написанных не на русском языке и не в кириллической раскладке, в частности, в примере 10; 4) стемминг (или лемматизация с обоснованием); 5) обнаружение синонимов, в том числе выраженных не словами, а коллокациями; 6) обнаружение отношений гиперонимии-гипонимии (пример 14); в идеале – при указании лингвистического модуля необходимо пояснение, как его применение поможет детектированию парафразы (например, упоминание tf.idf никак не

проясняет, каким образом нужно его применить, чтобы мы смогли отделить пары – парарфразы и пары не парафразы, если они близки по лексическому составу).

NB: *Где находится* и *расположение* находятся в более сложном отношении, чем отношение синонимии
Баллы снимались за отсутствие одного или нескольких из вышеупомянутых параметров, отсутствие аргументации, неправильное употребление терминов.

в) Приведено больше одного правила. Необходимость приведенных правил обоснована. Все обозначения в правилах понятны или объяснены дополнительно. Правила не повторяют дословно написанное в предыдущем пункте.

г) Четко описаны объект машинного обучения, используемые признаки, приведен и аргументирован метод. Баллы снимались за отсутствие одного или нескольких из вышеупомянутых параметров, отсутствие аргументации, неправильное употребление терминов и непонимание принципов организации модели машинного обучения.

д) Приведено как минимум три примера. Примеры не должны быть однотипными. Предложенные решения для обработки этих примеров должны приводить к правильным результатам.

Задание 3. – 35 баллов

3.1 – максимальная оценка 15 баллов

Основания	Оценка
Нет задания	0
Объем реферата не позволяет оценить полноту и точность понимания текста	2
содержит слишком общие оценочные формулировки, главная мысль статьи не сформулирована с достаточной степенью конкретности (не указана задача сентимент-анализа, не указана необходимость исследования социальных переменных)	5
отражены несколько основных мыслей автора, но не сохранена структура аргументации, часть важных упомянутых проблем не упомянуто	8
в пересказе отражена большая часть проблем текста, сохранен порядок логического построения аргументации автора	10
текст хорошо структурирован и составлен системно, отсутствует личная оценка аргументов в пересказе, показано полное понимание текста и его статус	15

Пример реферата:

3.1

Автор блога высказывает свое мнение о применимости стандартных NLP методов в извлечении из текста мнений (сентимент-анализе).
Главный тезис автора – ограниченность существующих решений для этой задачи не в применяемых технологиях, это проблема того, что человеческие суждения неоднородны и зависят от разных социальных причин: пола, возраста, культурных и социальных параметров (религии, политических убеждений, расы и т.п.). Автор указывает, что для предсказания сентимента недостаточно анализировать только языковые выражения, необходимы предиктивные модели того, как устроена человеческая оценка в зависимости от конкретной предметной области, от социальных переменных - иначе в глобальном смысле проблема нерешаема (хотя и отмечается, что на узкопрофильных задачах чисто лингвистический подход дает хорошие результаты).
Автор предлагает два пути преодоления ограниченности существующих решений:

(а) глобальный – проводить социальные исследования того, как люди что-то оценивают; наиболее сложный способ;
(б) локальный – добавлять в NLP модуль модель того, как устроена оценка для конкретной тематики для разных групп людей с разными социальными параметрами.
В общем и целом, предлагается пересмотреть подход к архитектуре приложений для сентимент-анализа, если область его применения затрагивает несколько областей или разнородные данные одной области.

За 3.2 - 20 баллов

Основания	Оценка
Нет задания (пример отрицательного решения: я согласна с автором, что это очень сложная задача)	0
Высказанное мнение не позволяет оценить полноту и точность понимания проблематики текста	2
Содержит слишком общие оценочные формулировки	3
Упомянута одна из проблем, поднимающихся в тексте и приведены соображения по ее возможным решениям, либо упомянуты несколько проблем, но их решение описано очень поверхностно	8
выделено 2 основные проблемы и расписаны варианты их решения, однако предложенные решения не претендуют на системное решение проблемы, а лишь называют возможное направление работы	12
согласие или несогласие с автором поста аргументировано и подкреплено примерами; выделены основные проблемы текста и по ним расписаны конкретные варианты решения, способные решить проблему	20

Пример ответа:
Я считаю, что данная точка зрения имеет под собой основание, так как социолингвистические исследования показывают, что в употреблении языка представителями различных социальных групп существуют различия (в зависимости от пола, социального положения, национальности и т.п.). Также оценка и ее выражение зависит от социолингвистических переменных - так, например, существительное *патриот* в современном контексте может иметь и положительные, и отрицательные коннотации в зависимости от политических взглядов автора текста.
Автор поста утверждает, что для добавления поведенческой модели оценивания требуется анализ социологических данных, фрагментирование аудитории и изучение психологически пользовательских групп.
Я считаю, что характеристики автора текста возможно получить с использованием стандартных методов NLP (классификацией текстов по полу, социальной группы автора и т.п.), т.е. я предлагаю добавлять в систему сентимент-анализа автоматическое распознавание социальных и психологических параметров автора текста. Для этого можно использовать предыдущие тексты того же автора, поисковые запросы, другую доступную текстовую информацию - так сейчас работают системы таргетирования рекламы. В защиту такого чисто лингвистического метода можно сказать, что в узкотематических текстах или в более структурированных текстах эти методы, несмотря на их ограниченность, имеют хорошие результаты.
Если же все-таки пересматривать архитектуру приложения более кардинальным образом, то хорошим решением может служить такой пайплайн:
1) кластеризация пользователей относительно их текстов, с учетом социолингвистических особенностей
2) создание локальных моделей сентимент-анализа для каждой группы отдельно

3) для каждого нового отзыва пытаемся определить его группу и, соответственно, классифицировать сентимент отзыва

Тем не менее, хочется отметить, что социальный фактор может оказаться далеко не единственным влияющим на модель. Автор текста фокусируется только на социальных особенностях, тогда как дополнительными факторами могу оказаться соответствие ожиданий пользователей и реальных свойств продукта, на который получен отзыв, цена продукта (от более дорогих больше ожиданий, больше негативных отзывов, а для более дешевых - более лояльная оценка), назойливость и объем его рекламы.