ПРОТОКОЛ УЧАСТИЯ

1. Направление олимпиады _____
2. Профиль олимпиады _____ код _____
3. Дата _____
4. Город участия _____
5. Место проведения _____
6. Аудитория _____

7. Сведения об участнике:
Регистрационный номер_____
ФИО _____
Дата рождения _____
Контактный телефон _____

8. Правила участия в очных олимпиадных состязаниях

8.1. На рабочем месте участник должен иметь: документ, удостоверяющий личность; согласие на обработку персональных данных; титульный лист (распечатывается из личного кабинета); канцелярские принадлежности (ручка, карандаш, линейка и пр.); допускается наличие питьевой воды в прозрачной бутылке, шоколада, очков.

8.2. Участники состязаний по направлению **«Математические методы анализа экономики»** имеют право пользоваться Таблицей стандартного нормального распределения, Таблицей распределения Стьюдента, Таблицей распределения Фишера, Таблицей распределения хи-квадрат, предоставляемые Оргкомитетом.

8.3. Участники состязаний по направлению **«Градостроительство»** имеют право пользоваться личным Градостроительным кодексом печатного издания (электронные издания строго запрещены).

8.4. Участники по направлению **«Инноватика»** имеют право пользоваться справочным материалом без ограничений типа источника информации (электронные издания строго запрещены);

8.5. Участники по направлениям **«Измерения в психологии и образовании», «Менеджмент», «Востоковедение и африканистика», «Журналистика»** и по профилям **«Консультативная психология. Персонология» , «Корпоративный юрист», «Право информационных технологий и интеллектуальной собственности», «Политика. Экономика. Философия», «Менеджмент в СМИ»** имеют право пользоваться личным англо-русским словарем печатного издания (электронные издания строго запрещены);

8.6. Участники по направлению **«Политический анализ и публичная политика»** и по профилю **«Население и развитие»** имеют право пользоваться личным русско-английским и/или англо-английским словарем печатного издания (электронные издания строго запрещены;

8.7. Участники по направлениям **«Филология»** и **«Лингвистика»** (*не более двух бумажных словарей*) имеют право пользоваться личным англо-русским, французско-русским и немецко-русским словарем (электронные издания строго запрещены).

8.8. Участники по направлениям **«Градостроительство», «Политический анализ и публичная политика», «Прикладная математика и информатика», «Мировая экономика», «Математические методы анализа экономики», «Финансы и кредит», «Электроника и наноэлектроника», «Инноватика»** по профилям **«Бизнес-информатика», «Электронный бизнес», «Управление в сфере науки, технологий и инноваций/Governance of Science, Technology and Innovation»** могут пользоваться личным бухгалтерским (простой, +/–) калькулятором.

8.9. **Не разрешается использовать** и даже **иметь на рабочем месте**: справочные материалы, кроме перечисленных в пп. 8.2-8.7; калькуляторы, кроме случаев, перечисленных в п. 8.8; карманные компьютеры и любые иные электронно-вычислительные устройства; мобильные телефоны и иные средства связи; плееры. Пользование указанными материалами и средствами запрещено как в аудитории, так и во всем здании на протяжении всего состязания до момента окончания времени, отведенного на выполнение олимпиадного задания.

8.10. **Не разрешается**:

- задавать вопросы другим участникам и отвечать на вопросы других участников;
- вставать с мест и пересаживаться, создавать помехи другим участникам и организаторам;
- списывать и позволять списывать у себя другим участникам;
- обмениваться любыми материалами и предметами;
- продолжать выполнение задания после окончания времени, отведенного на состязание;
- выходить из аудитории в случае, если состязание длится 120 и менее минут. По истечении второго часа состязания допускается выход участника из аудитории по уважительной причине, но не более чем на 5-7 минут и в сопровождении одного из организаторов.

8.11. В случае нарушения правил участия в олимпиаде участник отстраняется от дальнейшего участия в этом состязании, а его работа аннулируется.

8.12. **Участник имеет право** обратиться к организаторам с вопросами по организации состязания и оформлению работы, внести по окончании состязания в протокол проведения состязания замечания и претензии к олимпиадным заданиям, обратиться за медицинской помощью, досрочно сдать работу, получить по окончании состязания текст олимпиадного задания.

8.13. Досрочная сдача работ участников прекращается за 15 минут до окончания состязания.

8.14. По окончании состязания участники должны оставаться на своих местах до разрешения организаторов покинуть аудиторию.

ОЛИМПИАДА НИУ ВШЭ ДЛЯ СТУДЕНТОВ И ВЫПУСКНИКОВ

9. Правила оформления работы

9.1. Участник должен выполнить работу на бланке ответов. Первый лист бланка ответов – это Протокол участия, второй лист – Протокол проверки работы, далее – листы для записи решений и ответов на задания олимпиады, на этих листах можно делать записи с обеих сторон. Текст задания в качестве бланка ответов использовать запрещается.

9.2. Участник должен заполнить титульный лист, заполнить и подписать Протокол участия (первый лист бланка ответов), записать название направления, профиля и кода в Протоколе проверки работы (второй лист бланка ответов), и более ничего на первом и вторых листах бланка ответов не писать.

9.3. Участники Олимпиады по направлениям **«Математика», «Математические методы анализа экономики», «Финансы и кредит»** и **«Экономика»** имеют право поменять профиль, выбранный ими при регистрации.

9.4. Работа должна быть выполнена ручкой с чернилами синего или черного цвета. Запрещается использование ручки с чернилами красного и зеленого цветов, использование карандаша для записи решения, ответов.

9.5. Бумага для черновиков и дополнительные листы к бланку ответов выдаются организаторами по просьбе участников.

9.6. Черновики работ, как правило, не проверяются. Жюри может принять решение о проверке черновиков работ участников Олимпиады по направлениям: **«Градостроительство», «Политический анализ и публичная политика», «Прикладная математика и информатика», «Математика», «Социология», «Демография» и «Прикладная математика»** и по профилю **«Лингвистическая теория и описание языка»** в случае указания в работе «смотреть черновик».

9.7. В бланке ответов и в черновиках, предъявляемых к проверке, нельзя указывать ФИО, делать какие-либо записи, указывающие на авторство работы.

9.8. В бланке ответов можно делать исправления, которые должны быть понятными и однозначно трактуемыми. Если необходимо внести исправления, то следует аккуратно зачеркнуть неправильный ответ и написать правильный.

9.9. Почерк участника должен быть понятным. Жюри может отказать участнику в проверке работы в случае «нечитаемого» почерка.

9.10 Разрешается замена ручки, титульного листа, бланка ответов.

9.11 Участники олимпиады по направлению **«Программная инженерия»** выполняют часть задания на компьютере.

9.12 Участники олимпиады по направлению **«Журналистика»** могут выбрать задание (блок 2) выполняемое на компьютере.

С ПРАВИЛАМИ ОЗНАКОМЛЕН, достоверность предоставленной информации подтверждаю _____ _____
(подпись) (ФИО участника)

АКТ
отстранения участника

Акт составлен в связи с отстранением участника _____от состязания.
(ФИО участника)

Причины отстранения:
_____
_____
_____

Акт составлен _____ _____
(подпись) (ФИО организатора в аудитории)
_____ _____
(подпись) (ФИО организатора в аудитории)

С актом ознакомлен _____ _____
(подпись) (ФИО участника)

линия отреза

Олимпиада студентов и выпускников

Направление Фундаментальная и прикладная лингвистика

Профиль Компьютерная лингвистика код 310

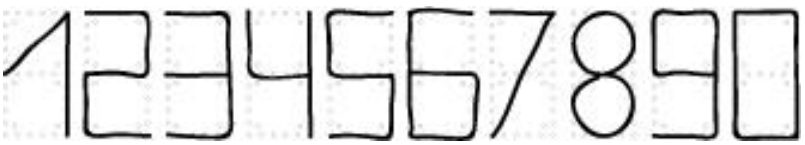
Заполняется организатором в аудитории			
Количество доп. листов	Количество черновиков	Время выхода	Замена ручки
		с _____ до _____	☐ да

Протокол проверки

Заполняется членами жюри. Пометки участников не допускаются.

Итоговый балл

Пример заполнения



шифр

Олимпиада для студентов и выпускников – 2016 г.

Направление «Фундаментальная и прикладная лингвистика»

Профиль:
«Компьютерная лингвистика» **КОД – 310**

Время выполнения задания – 180 мин.

Вопрос 1

Решите задачу

Алиса загадала одно из следующих слов: «яблоко», «крокодил», «дирижабль», «груша», «бегемот», «машина», «зелёный». Боб хочет угадать слово Алисы, задавая вопросы, на которые она отвечает «да» или «нет». Какое минимальное количество вопросов ему придётся задать, чтобы гарантированно угадать загаданное слово? Объяснить, какие вопросы в каком порядке вы будете задавать и доказать, что меньшим числом вопросов обойтись нельзя.

Ответ

Вопрос 2

Задание

Перед вами стоит задача разработать систему, автоматически анализирующую отзывы посетителей о ресторанах. Система должна не просто определить положительное или отрицательное отношение автора отзыва, но выявить его мнение по конкретным критериям оценки ресторана.

Прочитайте **приложение** с отзывами о посещении ресторана и выполните следующие задания:

А) Составьте список критериев («аспектов»), по которым может быть охарактеризован ресторан.

Б) Составьте инструкцию разметчика корпуса отзыва в соответствии с выделенными аспектами. Укажите списки объектов разметки, принципы включения текста в разметку, сложные случаи и их разрешение.

В) Разметьте тексты в приложении согласно составленной вами инструкции.

Г) Сформулируйте классы проблем, с которыми вы столкнулись при разметке текстов отзывов.

Ответ

Приложение к вопросу 2

1. Ресторан в полне хорош. Но в меню все старое
2. Прекрасное место!!!!! Ниавкуснейшие блюда! Цены соответвуют качеству полностью! Персонал в стиле Новикова!) Спасибо большое, только положительные впечатления!!! Негативные странные отзывы не имеют ничего общего с правдой!
3. Ужасный ресторан! Цены просто заоблачные, ладно было бы вкусно, но это невозможно есть, маленькие порции, официанты тупят, в зале грязно! Вам бы программу "Ревизорро" поставили бы Вас на место! Это не дело, Вы кормите там людей!
4. Все очень круто и еда и обслуживание и очень быстро готовят 10 мин и люля готов!! Очень вкусно кстати!!!!
5. Уважаемые рестораторы!Мне не пришлось побывать в вашем ресторане,но я часто проезжаю мимо.Вы находитесь в самом центре Москвы ,вывеску видно издалека , но написана она с ошибкой.Предложение "Страна , которой нет"пишется с запятой.Ваша вывеска режет глаз своей неграмотностью.МОжет быть,что-то можно исправить? ...

6. Удивительно, что в центре Москвы есть такие заведения. Еда отвратительная, все блюда жутко пересолены, морсы и чаи пить невозможно. Мясо жесткое, ко всему прочему баранина очень сильно напоминает свинину.
7. отметила вчера в этом заведении свой день рождения. Помимо того, что официанты забыли заказы двоих из гостей и принесли их только после того, как им напомнили об этом дважды, мои букеты роз (один из которых стоил 100 долларов и был куплен за космические деньги в Цветном) подарили другой имениннице вместе со всеми поздравительными открытками, которые я даже не успела прочитать. Менеджер ресторана...
8. Отмечала день рождения. Очень понравилось. Кухня отличная. Обслуживание превосходное. Огромное спасибо за отличный вечер.
9. На бизнес-ланче принесли хлеб по твердости схожий с кирпичом! Он не ломался ни руками, ни зубами! на просьбу заменить - принесли такой же!!!!
10. 25 апреля заказала стол,через ресторан ру.В 19 часов пришли,но нашего заказа не оказалось.Попросили посадить за свободный стол,но администратор сказала,что всё заказано,надо было заказать.Не аргументом для неё были даже показанные ей

смс с подтверждением заказа.Вечер был испорчен ,Так что с удовольствием получили бы бонус,обещанный рестораном.ру

11. Добрый день! Были в вашем кафе вчера, нам понравилось; отличное местоположение, красивый интерьер, в том числе приглушенный свет, приятная музыка, довольно громкая, но разговаривать не мешала. Нас обслуживала приятная девушка- официант, видно, что блюда знает, уверенно дает рекомендации, отвечает на вопросы и тд и вообще все официанты в зале как бы на подхвате, не успеешь салфетку положить,...
12. Блюда вкусные, но слишком настойчивые официанты. Хочу заказать одно блюдо, мне сразу рекомендуют замену и после моего отказа от рекомендаций официанта говорят, что "зря, зря....ну ладно". С моей точки зрения весьма странный подход. Когда принесли чай, то чашка была грязная-неприятный осадок (правда, десерт от заведения преподнесли абсолютно бесплатно, этот момент приятен).
13. К сожалению, на первый взгляд замечательное место, в самом центре столицы, оставило о себе самые неприятные впечатления. И виной тому официанты, которые в какой то момент после подачи основных блюд просто про нас забыли. Не менее 20 минут мы пытались докричаться хоть до кого нибудь, неоднократно

просили посчитать, все спрашивали есть ли у вас скидочные карты и после этого исчезали, пока...

14. Посетил случайно, с коллегой, очень понравилось, настолько всё вкусно и по деньгам не так уж дорого, учитывая, что это центр Москвы. Обстановка уютная, приглушённый свет.

Вопрос 3.

Задание

Прочтите пост из блога, посвященного автоматической обработке естественного языка (http://nlpers.blogspot.com/). Составьте краткое резюме этого поста на русском языке (объемом в один абзац, но не более 1000 знаков), отразив в нем основной тезис автора поста. Выскажите ваши собственные соображения по проблемам, затронутым в посте и их решениям.

NLP as a study of representations

Ellen Riloff and I run an NLP reading group pretty much every semester. Last semester we covered "old school NLP." We independently came up with lists of what we consider some of the most important ideas (idea = paper) from pre-1990 (most are much earlier) and let students select which to present. There was a lot of overlap between Ellen's list and mine (not surprisingly). . The whole list of topics is posted as a comment. The topics that were actually selected are here.

I hope the students have found this exercise useful. It gets you thinking about language in a way that papers from the 2000s typically do not. It brings up a bunch of issues that we no longer think about frequently. Like language. (Joking.) (Sort of.)

One thing that's really stuck out for me is how much "old school" NLP comes across

essentially as a study of representations. Perhaps this is a result of the fact that AI -- as a field -- was (and, to some degree, still is) enamored with knowledge representation problems. To be more concrete, let's look at a few examples. It's already been a while since I read these last (I had meant to write this post during the spring when things were fresh in my head), so please forgive me if I goof a few things up.

I'll start with one I know well: Mann and Thompson's rhetorical structure theory paper from 1988. This is basically "the" RST paper. I think that when a many people think of RST, they think of it as a list of ways that sentences can be organized into hierarchies. Eg., this sentence provides background for that one, and together they argue in favor of yet a third. But this isn't really where RST begins. It begins by trying to understand the communicative role of text structure. That is, when I write, I am trying to communicate something. Everything that I write (if I'm writing "well") is toward that end. For instance, in this post, I'm trying to communicate that old school NLP views representation as the heart of the issue. This current paragraph is

supporting that claim by providing a concrete example, which I am using to try to convince you of my claim.

As a more detailed example, take the "Evidence" relation from RST. M+T have the following characterization of "Evidence." Herein, "N" is the nucleus of the relation, "S" is thesatellite (think of these as sentences), "R" is the reader and "W" is the writer: relation name: Evidence constraints on N: R might not believe N to a degree satisfactory to W constraints on S: R believes S or will find it credible constraints on N+S: R's comprehending S increases R's belief of N the effect: R's belief of N is increased locus of effect: N

This is a totally different way from thinking about things than I think we see nowadays. I kind of liken it to how I tell students not to program. If you're implementing something moderately complex (say, forward/backward algorithm), first write down all the math, then start implementing. Don't start implementing first. I think nowadays (and sure, I'm guilty!) we see a

lot of implementing without the math. Or rather, with plenty of math, but without a representational model of what it is that we're studying. The central claim of the RST paper is that one can think of texts as being organized into elementary discourse units, and these are connected into a tree structure by relations like the one above. (Or at least this is my reading of it.) That is, they have laid out a representation of text and claimed that this is how texts get put together. As a second example (this will be sorter), take Wendy Lehnert's 1982 paper, "Plot units and narrative summarization." Here, the story is about how stories get put together. The most interesting thing about the plot units model to me is that it breaks from how one might naturally think about stories. That is, I would naively think of a story as a series of events. The claim that Lehnert makes is that this is not the right way to think about it. Rather, we should think about 3 stories as sequences of affect states. Effectively, an affect state is how a character is feeling at any time. (This isn't quite right, but it's close enough.) For example, Lehnert presents the following story: When John tried to start his care this morning, it wouldn't turn over. He asked his neighbor Paul for help. Paul did something to the carburetor and got it going. John thanked Paul and drove to work.

The representation put forward for this story is something like: (1) negative-for-John (the car

won't start), which leads to (2) motivation-for-John (to get it started, which leads to (3) positivefor-John (it's started), when then links back and resolves (1). You can also analyze the story from Paul's perspective, and then add links that go between the two characters showing how things interact. The rest of the paper describes how these relations work, and how they can

be put together into more complex event sequences (such as "promised request bungled"). Again, a high level representation of how stories work from the perspective of the characters.

So now I, W, hope that you, R, have an increased belief in the title of the post. Why do I think this is interesting? Because at this point, we know a lot about how to deal with structure in language. From a machine learning perspective, if you give me a structure and some data (and some features!), I will learn something. It can even be unsupervised if it makes you feel better. So in a sense, I think we're getting to a point where we can go back, look at some really hard problems, use the deep linguistic insights from two decades (or more) ago, and start taking a crack at things that are really deep. Of course, features are a big problem; as a very wise man once said to me: "Language is hard. The fact that statistical association mining at the word level made it appear easy for the past decade doesn't alter the basic truth. :-)." We've got many of the ingredients to start making progress, but it's not going to be easy!

How long'll it take to say that?

tl;dr: Given a string of text in some language you might want to know how long it would take to speak it. Here are some perl/python "one-liners" to estimate that. Currently supports English (US), German, French, Italian, Spanish and Japanese. The estimates are all in seconds.

Brave are those who read the source code. I promise it's not intentionally obfuscated -- there's just a lot of unicoding going on, and then copious use of the right-handed saturn operator in perl.

Why would you want code for this rather than a dictionary? Dictionaries are limited to their vocabulary, which is also sensitive to things like tokenization. Having a completely parametric solution seemed much more generalizable.

Why would I possibly want this?

We've been doing a bunch of work recently on simultaneous machine interpretation (aka "real time machine translation"). None of us is a speech person, either in the "recognition" or "synthesis" sense. This unfortunately means that to date, all of our models treat each word as "equally long" when training and, perhaps more importantly, *evaluating* models. For instance, if we want to measure the décalage (aka time-lag, aka ear-voice-span) between when a Japanese word is "heard" and the corresponding English translation is "spoken", we've been assuming all words take precisely 1 second to speak. This is obviously ridiculous.

One quick and dirty alternative is to use a text-to-speech (aka speech synthesis) system to synthesize a sentence and use its length as an estimate of how long it would take a person to

speak it. This is a totally plausible approach since decent open source synthesis software exists (we use such software to crease these one-liners), but it's slow and bloated and can't be easily distributed. This would be more accurate, but I was after a quick and filthy solution.

How do these scripts work?

There are two functions: one for estimating the amount of time it would take to speak a single word (sayWord in the code), and another for estimating the amount of time it would take to speak a sentence (sayit in the code). I'll first describe how sayWord works; sayit is pretty straightforward.

sayWord works by extracting a bunch of features from the word to be spoken (each of these is a particular regular expression) and evaluating a linear function of the counts of the matches of those regular expressions. The process by which coefficients were generated is explained later. The features are things like: number of characters, number of vowels, number of consonant groups, number of non-letters, number of vowel-consonant switches, number of digits of various lengths, and whether the word starts or ends with a vowel; and then, for each "reasonable" unicode character for European languages, the count of those characters. Not all of these features appears for each language because I used l1 regularization to prune down the feature set.

[Note: Japanese is different because it uses a different character set. The feature structure is basically the same, but replace "consonant" with "kana" and "vowel" with "kanji" and "reasonable character" with "each of the 100 most frequent Japanese characters" in Japanese Wikipedia.]

sayit attempts to pronounce a sentence by pronouncing each word individually (via sayWord) and then rescaling the resulting estimate because we ... don't ... pause ... between ... words. This rescaling is linear, and again estimated from data (explained below).

How are the coefficients estimated for words?

Basically I take a vocabulary of the 50k-100k most frequent words for each language, use a speech synthesis program to say them (for the European languages, I used MaryTTS; for Japanese I used Open JTalk).

I then extract all the features mentioned above, create a regression problem regressing on the number of seconds it takes to speak (after removing quiet time around the word) and throwing in

some l1 regularization with vw to make sure that it didn't use too many features. I optimized quantile loss (aka absolute value loss) rather than squared loss.

One thing I did *not* do was weight the words by their frequency. I could have done this and the resulting regression weights change a bit but not too much. At the end it spends a lot of energy making sure it estimates the speaking time of "a" and "the" and "an" correctly. Because short words are typically high frequency (and vice versa) this meant that it tended to underestimate all other words. And because my relatively frequencies were from Wikipedia, they might not match yours. Keeping a uniform distribution felt like a better solution.

I also tried not regularizing and also using things like character ngram features to get a better fit. In the end I didn't include this either. You can about halve the error rate by doing these things, but I felt like having simpler, smaller models made more sense here. After all, the thing we're regressing on (speaking time from a TTS system) is sort of an artificial benchmark anyway, so getting a bit lower error is not obviously that meaningful.

Overall the mean absolute errors in prediction are:

- German: 51ms
- English: 55ms
- French: 59ms
- Italian: 41ms
- Japanese: 33ms

If you were to just predict the median on each, you would get mean absolute errors in the 550ms (Japanese) to 980ms (English) range, so this is quite an improvement.

How are the coefficients estimated for sentences?

For sentences, I fit a model of the form:

- time = const + a * [summed time for words] + b * [# of words]

The three parameters (const, a and b) were estimated by having the TTS systems speak entire sentences (about 40k from each language, taken randomly from Wikipedia) and then using the total time from the TTS (unioned across all languages) and the corresponding features. These are simply hardcoded and shared across languages. You could of course do a bit better by doing this on a language-by-language basis, but I didn't do this again for simplicity.

Conclusion

Chances are no one except me wants this. But if you do, please feel free to use it. I'd appreciate some sort of credit though :). And if you really want the dictionaries, you can find them here: de, en-US, fr, it, ja.

Thanks to Graham Neubig for providing the tokenized Japanese Wikipedia text and pointing me to Open JTalk!

Ответ