Проект

Этап 0.

Первое задание по проекту состоит в следующем.

Вам нужно написать краткое резюме по проекту и заполнить на проект анкету.

Ссылку на анкету я кину в чат.

Пояснение к заполнению анкеты ниже.

Если Вы затрудняетесь с тем, чтобы поставить какую-то задачу, пишите. Но лучше обсудить тогда лично по скайпу или до (или после занятий). Если Вы сомневаетесь в выполнимости, тоже пишите – обсудим.

Пока можно заполнить только какие-то общие соображения – из какой области с каким языком, по какой тематике…

Если не знаете ответ на какой-то из вопросов – пишите.

Прикладываю несколько презентаций.

NB!!! Презентация по проекту «English with friends» - финальная – сейчас не нужно углубляться в детали, связанные с технологиями нейронных сетей и векторных моделей. В презентации хорошо расписано, какие списки использовались, какая предобработка делалась, какие программы использовались и т.п.

Вопросы можно задавать по почте toldova@yandex.ru (пожалуйста, в теме письма указывайте ДПО), в скайп – toldova. Что-то можно будет уточнить устно в скайпе.

0.1 Из какой области вы хотите решать задачу

(это какая-то очень большая область: например, социологическое исследование по социальным сетям, рекомендательная система по ресторанам, проверка орфографии, обучение языку – генерация упражнений…)

0.2. С каким языком/языками вы хотите работать

0.3. По какой тематике

1. Название, кто участвует в группе

Примеры:

(а)"Рекомендательная система по выбору места отдыха"

(б) "English with Friends"

(в) "Автоматизированное кодирование ответов на открытые вопросы"

(г) "Портрет района"

(д) "Наивная поэзия"

2. Пояснение: какая конечная задача у системы

Примеры:

(а) "В результате диалога с пользователем система должна предложить пользователю какой-то тур/несколько туров на выбор, отзывы на эти туры, подсветить в тексте позитивные и негативные моменты"

(б) "Создание чат-бота для приложения Telegram с заданиями по английскому языку на основе сериала “Друзья”"

(в) "По ответам респондентов на открытые вопросы выделить наиболее важные параметры, поределить, какими словами пользователи описывают те или иные свойства продуктов"

(г) "Есть новостной сайт, где публикуются новости разных районов. В новостях об одних районах чаще сообщается, что было соврешено какое-то преступление, о других - новости о каких-то клубах по интересам, мероприятиях... Представляется, что по новостям, можно построить "портрет" района"

(д) "Графоманы и любители пишут стихи. Они отличаются от стихов профессиональных поэтов. Хотелось бы, чтобы система умела отличать люительские стихи от профессиональных"

3. Область применения, зачем она нужна, почему это важно: где и кем такая система может применяться, оценить востребованность системы

(а) "У человека есть какие-то отдельные представления и ограничения, что бы ему хотелось (например, он бы хотел совсем недорого покататься на горных лыжах так, чтобы домик был недалеко от подъемника...), какие-то характеристики поехдки ему очень важны, а какие-то не очень; часто выбор очень большой, и не хочется просматривать все самому; хотелось бы, чтобы система предложила что-то на выбор + "подтянула" бы отзывы про эти места"

(б)"Система должна сама генерировать упражнения на основе лексики и реплик из сериала. Так учить язык проще и веселее. + в упражнениях можно учитывать лексический минимум соответствующий уровню владения языка и типичные ошибки в подборе лексики"

(в)"При опросе людей о качестве продуктов задают "открытые" вопросы (не да/нет, вопросы - на которые человек отвечает "произвольным текстом"), эти вопросы важны, потому что иногда заранее нельзя предсказать, какие именно качества продукта могут понравиться пользователю. Ответы на вопросы кодируются вручную: например, "Очень нравится, что товар Х продается по низкой цене" -> приемлемые цены. Нужно автоматизировать эту процедуру - выявить наиболее частотные параметры продукта и наиболее типичные слова и выражения, с помощью которых продукт характеризуется"

(г) составление социологического портрета района методом контент-анализа

(д) это нужно для диссертационного исследования "Особенности детской поэзии". В литературовелческих работах называются разные признаки такой поэзии: использование "избитых" оценочных слов, повышенное количество таких слов, большое количество наречий типа *very*... Хотелось бы проверить эти гипотезы и показать, что это так или не так

4. Есть ли похожие системы; решения каких систем хотелось бы взять за образец

Тут можно либо привести ссылку на сайт/сайты (может быть, такой системы нет, но есть какой-то образец для одного из компонентов системы, например, есть система определения настроения по твитам, а вы хотите генерировать разные стихи в соответствии с настроениями -))))

5. Описаны ли какие-то технологии создания системы, на которые Вам удалось посмотреть

Тут краткая характеристика технологии или ссылка на сайт, где вы про эту технологию читали

(можно попробовать поискать вот здесь: https://habr.com/ru/search/?q=%5B%D0%BC%D0%BE%D1%80%D1%84%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F%5D&target_type=posts), вот здесь:

https://medium.com/better-programming)

Вот здесь очень много всего на английском, но это самые последние достижения в NLP

https://www.aclweb.org/anthology/

6. Какие данные нужны для разработки, откуда планируется брать данные (тексты, словари...)

Примеры ответов:

(а) 1) тексты описания туров (trip-advizer – не знаю, можно ли оттуда что-то скачать); 2) какая-то уже готовая система типа trip-advizer, booking..., в которой можно взять основные параметры, которые интересуют пользователя; тексты отзывов 3) возможно, какой-то словарь тональной лексики (есть для русского языка списки rusentilex - https://www.labinform.ru/pub/rusentilex/index.htm)

(б) корпус субтитров сериала Станем друзьями, списки лексических минимумов для разных уровней владения английским языком (пока не знаю, где брать);

(еще нужны данные скорее для большого проекта: списки ошибок из learner corpus - корпуса текстов, которые писали не носители языка и делали ошибки; эти корпуса размечены по разным типам ошибок; нужны ошибки на глаголы - т.е. либо часто путающиеся формы глаголов, либо часто путающиеся глаголы)

(в) надо выбрать 2-3 продукта, нужен корпус ответов на вопросы про них; корпус беру свой (рабочий)

(г) надо выбрать 3-4 района и создать корпус новостных текстов про эти районы; новости про каждый район берем вот с этого сайта: …

(д) корпус профессиональной поэзии, корпус любительской поэзии, списки "простых" и частотных прилагателньых и наречий, морфологический анализатор для английского языка

7. Описание системы: как это примерно должно выглядеть во frontend: что пользователь (если таковой имеется) подает на вход, что получает в ответ. Разбор примера. Либо: что Вы хотите после обработки корпусов и т.п. получить на выходе:

Например: пользователю выдается предложение с пропуском и варианты заполнения, он выбирает один из вариантов…

Пользователь подает ответы на открытые вопросы, каждому ответу ставится в соответствие тег. Например: «мне понравилось что много разных товаров» -> “разнообразный ассортимент”

Например, см. слайд 3 из презентации по открытым вопросам или слайд 8 из “English with friends”

8. Предполагаемая постановка задачи для Вас

8.1. Какие данные берете для разработки системы:

Дано:

тексты: например, новостные тексты с сайта XXX,

дополнительные списки: например, списки лексического минимума для 3-его года изучения английского языка, готовые списки стопслов для английского языка (XXX)

8.2. Как выглядит результат после обработки текста:

Тут нужно какой-то прямо-таки пример текста/фрагмента списка и т.п.

(г) Получаем список наиболее частотных слов и словосочетаний для каждого из подкорпусов новостей для разных районов.

Арбат: ресторан, кафе,…

Гольяново: задержать, разбить…

(д) список наречий, которые встречаются в любительской поэзии; список наречий, который встречается в профессиональной поэзии

Список относительных частот частей речи… в корпусе любительской поэзии и в корпусе профессиональной поэзии

(е) Вот у меня текст человека, который учит русский:

“Это играет значение” – подсвечивается (выдается в строке выдачи: «играет значение» - стилистическая ошибка – играет роль или имеет значение)

То есть попробуйте подумать, как должен выглядеть результат.

8.3. Какие готовые модули Вы планируете брать:

например, морфологический анализатор PyMorphy; токенизатор из NLTK…;

не знаю, какие; хочу вот этот, но не знаю, как с ним работать