***Введение***

Наш проект называется Автотьютор

Наша задача - рекомендовать иностранцам русские тексты которые будут не слишком сложные и не слишком простые для них

Классический подход - это сдача теста Тест Русского как Иностранного, получение уровня и далее чтение текстов, которым был присвоен соответствующий уровень.

Я топлю за более детальный и индивидуализированный подход.

=======

ПУСТЬ 30 ВОПРОСОВ НА ТЕКСТ И 300 НА УЧЕНИКА

Для англ порядка 300 фич

в англ языке спец корпуса и общие корпуса 3000 лемм покрывают 50-60%

NGSL корпус английского

задача обучения на определенных корпусах в опред ситуациях

сначала убедиться что преодолевает хотя бы половину

частотность не канает (10 40)

как приходят ученики -- могут быть продуктами

призадуматься до 15.00

Взять док и переписать в таймлайне (месяц-три итд) и настроить процесс так чтобы тестировалось как можно раньше

Тестирование на пользователях все переворачивает

сделать слепок знаний руками и сфокусироваться на полученном пространстве параметров -- учитель смотрит и оценивает рекомендацию

отложить задачу слепка ученика

взять лояльных пользователей и или учителей и составить подробную карту знаний учащегося

по словам - попарсить учебники и п

=======

**Часть 1. Фичи**

***Секция 1. Фичи слов***

**Параметры сложности слова не зависящие от его положения в тексте и предложении**

Каждому слову заранее присваивается вектор сложности (он не зависит от положения слова в предложении и присваивается ему скажем так навсегда)

Вектор включает в себя следующие параметры

*Частотность* - может быть присвоено значение 0 1 2 (чем больше значение тем реже слово встречается)

*Многозначность* - может быть присвоено значение 0 1 2 ( 0 - 1 значение слова, 1 - 2 значения слова, 2 - 3 и более значений слова)

*Заимствование* - может быть присвоено 0 (если слово заимствовано) и 1 (если слово не заимствовано

*Абстрактность* - может быть присвоено 0 если слово не абстрактно и 1 если абстрактно

То есть увеличение того или иного показателя демонстрирует увеличение сложности соответствующего параметра слова

=====

заимствование +

однозначное слово - редкое (их очень мало)

значение слов в контексте меняются

омонимы (плита)

абстрактные могут быть и глаголы

====

**Параметры сложности слова зависящие от его положения в тексте и предложении**

Каждому слову присваивается значение называемое tf\_idf. которое отображает важность слова в контексте всего текста

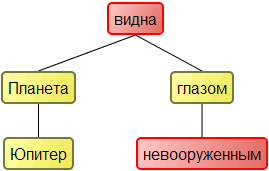
<https://ru.wikipedia.org/wiki/TF-IDF>

В качестве документов тут рассматриваются предложения

tf\_idf - это классический показатель используемый много где, дальше идет уже импровизация, которая пока что успешно прошла некоторую критику от лингвистов

Следующий показатель - это количество зависимых слов в синтаксическоим дереве

Для понимания это вот что



то есть у слова видна - этот показатель будет равен 2 а у слова глазом 1 а у невооруженным 0

Соответственно я делаю предположение о том,что удельную смысловую нагрузку конкретного слова в предложении можно сделать перемножением показтеля tf\_idf и количества зависимых слов, назовем этот показатель Sence\_value (SV)

В итоге каждое предложение условно можно представить вот в таком виде (Nwi - кол-во зависимых слов от слова с индексом i)

=========

КОНТРОЛИРОВАТЬ СЛОЖНОПОДЧИНЕННЫЕ ПРЕДЛОЖЕНИЯ

список незнакомых слов одинаково равны в незнакомости

глагол - царь предложения правильно ли это?

для тематически специфических корпусов тфидф норм

но человек может не знать

========

|  |  |  |
| --- | --- | --- |
| Слово 1 | Слово 2 | Слово 3 |
| SV1 = tfidf1\*Nw1 | SV2 = tfidf2\*Nw2 | SV3 = tfidf3\*Nw3 |

Каждый показатель нормируется (делится на сумму соответтсвующих показателей в конкретном предложении

|  |  |  |
| --- | --- | --- |
| Слово 1 | Слово 2 | Слово 3 |
| SV1 /SUMM | SV2/SUMM | SV3/SUMM |
| SVnorm1 = 0.2 | SVnorm2 = 0.5 | SVnorm3 = 0.3 |

***Секция 2 Фичи предложений***

На основе опыта преподавателей русского как иностранного и опроса иностранцев были выделены следующие фичи в рамках предложения

Отношение слов-отрицаний (не ни) к общему кол-ву слов в предложении

Отношение кореферентных слов (бы который) к общему кол-ву слов в предложении

Отношение смысловой нагрузки слова (показатель объяснен выше) к суммарной смысловой нагрузке слов в предложении для следующих явлений: деепричастия, причастия, возвратные глаголы, существительные с чередующимеся падежами (маме брата)

Средняя длина синтаксических связей

Усредненный вокабулярный вектор

***Секция 3 Фичи текста***

Из всех доступных формул мы предварительно исследовали и признали наиболее устойчивыми формулы LIX и TTR

Кроме них также рассматриваются усредненные вектор вокабуляра и вектор фич предложения

**Часть 2. Создание датасета знаний пользователя**

***Разметка предложений***

Далее основной этап алгоритма - разметка предложений посредством ответа на вопросы

При ответе на каждый из вопросов человек формирует три датасета, на основании которого будет в дальнейшем будет строиться модель его знаний

Допустим человек дал правильный ответ на вопрос “Куда шел человек”

В тексте есть предложение, которому соответствует ответ на поставленный вопрос, а именно “Человек шел домой”

Соответственно мы будем размечать это предложение в соответствие с полученным правильным/неправильным ответом

*Формирование датасета вокабуляра*

Пусть для слова “человек” параметры будут следующими (многозначность = 2 (у слова 2 и более значений), частотность = 0 (слово довольно частотное), заимствование = 1 (слово не заимствовано из английского языка), абстрактность = 0 (слово не абстрактно)

Таким образом вектор слова будет равен 2,0,1,0

Преобразовываем этот вектор в one hot encoding вид (имея в виду что для параметров многозначности и частотности есть три возможных значения 0,1,2 , а для заимствования и абстрактности 0 и 1) и получаем [0,0,1,1,0,0,0,1,1,0]

Аналогично для двух других слов смотрим вектора

Пусть vector (шел) = [1,0,0,1,0,0,0,1,1,0] vector (домой) = [0,0,1,0,0,1,0,1,1,0]

Как упоминалось ранее каждому слову в предложении соответствует SVnorm удельная смысловая нагрузка. Допустим здесь она будет распределена следующим образом

человек = 0.2

шел = 0.5

домой = 0.3

Проходимся итеративно по каждому слову в предложении рассматривая слово не в отдельности а в контексте левого и правого слова

Если слово первое по порядку то отображаем этот факт добавление слева искусственного вектора начала предложения [-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,]

Если слово последнее по порядку то отображаем этот факт добавлением справа искусственного вектора конца предложения [3,3,3,3,3,3,3,3,3,3]

Таким образом для каждого рассматриваемого слова получается вектор вокабулярных условий полученный конкатенацией вектора элемента слева, вектора самого слова и вектора элемента справа

В данном случае это будет выглядеть следующим образом

Человек [-1,-1,-1,-1,-1,-1,-1,-1,-1,-1, 0,0,1,1,0,0,0,1,1,0, 1,0,0,1,0,0,0,1,1,0]

шел [0,0,1,1,0,0,0,1,1,0 1,0,0,1,0,0,0,1,1,0 0,0,1,0,0,1,0,1,1,0]

домой [ 1,0,0,1,0,0,0,1,1,0 0,0,1,0,0,1,0,1,1,0 3,3,3,3,3,3,3,3,3,3]

======

в англ значение слова коллокации

работать со словосочетаниями хорошо но только триграммы не ок

======

Далее проставляется целевая переменная, которая рассчитывается по следующей логике

Принимаем тот факт, что показатель SVnorm отображает удельную важность слова в предложении

Далее вспоминаем тот факт, что вопрос был задан к предложению и мы можем делать выводы относительно знания/незнания каждого слова в предложении только с определенной долей вероятности, которую в данном случае как раз и может отобразить нормализованный (поделенный на сумму аналогичных показателей внутри предложения) показатель SV

В связи с этим для вектора каждого слова можно проставить целевую переменную, которая будет оставаться положительной при условии верного ответа на вопрос и будет умножаться на -1 в случае неверного ответа на вопрос (то есть с этой долей вероятности мы говорим, что данное слово человеком не было понято)

В итоге например для случая неверного ответа на вопрос в словарный датасет будут добавлены следующие вектора

Человек [-1,-1,-1,-1,-1,-1,-1,-1,-1,-1, 0,0,1,1,0,0,0,1,1,0, 1,0,0,1,0,0,0,1,1,0 -0.2]

шел [0,0,1,1,0,0,0,1,1,0 1,0,0,1,0,0,0,1,1,0 0,0,1,0,0,1,0,1,1,0 -0.5]

домой [ 1,0,0,1,0,0,0,1,1,0 0,0,1,0,0,1,0,1,1,0 3,3,3,3,3,3,3,3,3,3 -0.3]

*Формирование датасета предложений*

Напоминаю фичи предложений. которые формируют вектор предложения

Отношение слов-отрицаний (не ни) к общему кол-ву слов в предложении

Отношение кореферентных слов (бы который) к общему кол-ву слов в предложении

Отношение смысловой нагрузки слова (показатель объяснен выше) к суммарной смысловой нагрузке слов в предложении для следующих явлений: деепричастия, причастия, возвратные глаголы, существительные с чередующимеся падежами (маме брата)

Средняя длина синтаксических связей

Усредненный вокабулярный вектор

Целевой переменной для данного слова будут являться 1 или 0 в зависимости от того, верен или неверен ответ на вопрос

То есть вектор предложения будет выглядеть подобным образом

0, 0, 0.05026839632431031, 0.05026839632431031, 0, 0, 2, 0.375, 0.1875, 0.4375, 0.5625, 0.0625, 0.375, 0.0, 1.0, 0.9375, 0.0625, 0

При разборе каждого вопроса добавляется один подобный вектор

*Формирование датасета текстов*

Опять же, тексту соответсвуют показатели LIX TTR а также усредненные параметры предложений и вокабуляра

Целевой переменной является процент правильно отвеченных вопросов к тексту

9.33, 0.53, 0.4166666666666667, 0.15, 0.43333333333333335, 0.6666666666666666, 0.2, 0.13333333333333333, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.02254057389804958, 0.0, 0.0, 0.02563227477020398, 1.4173809523809524, 0.6

=====

если протестированные слова буду покрывать большую часть текстов корпуса

3000 слов англа встречается ….

=======

**Часть 3. Создание рекомендательных моделей**

После ответа на вопросы к 5-10 текстам у класса пользователя формируется три датасете, каждый со своей целевой переменной

=====

КОМПЕТЕНТУМ - ОБРАЗОВАТЕЛЬНОЕ + NLP

тематическая вариация задач

определение

SKY ENG самые крутые по оборотам

та часть рынка

ODS - career

больше всего в диалоговые системы

за рубежом умный поиск инфы

в области образвания - чат бот

=====

df\_вокабуляр

df\_предложения

df\_тексты

***Самый большой вопрос состоит в том как проводить обучение модели в данном случае***

***Для слов получается более-менее нормальный датасет. если пройти тест к 10-15 текстам, но для предложений и тем более текстам количество наблюдений довлльно мало***

В данной реализации я обучаю три разные модели

Далее подаю размеченный изначальным алгоритмом текст на вход в функцию рекомендации

Функция рекомендации делает три вида предсказаний. используя три ранее полученных модели

*Предсказание по вокабуляру*

Мы подаем модели на вход слова в таком же контексте (одно слово слева и справа) и получаем положительное или отрицательно значение

При отрицательном значении делаем вывод, что слово не будет понято человеком, причем величину отклонения предсказанного значения от нуля будем интерпретировать, как меру уверенности confidence(i) в своем предсказании

Для каждого слова, которому было предсказано положительно понимание мы рассчитываем произведение его SV (ненормированного, то есть tf\_idf \* [кол-во зависимых элементов +1]) на эту меру уверенности. Назовем полученное произведение SV\_conf+(i) = SV(i) \* confidence(i)

На основании этого показателя мы можем рассчитать совокупную уверенность в данных нами прогнозах относительно понимания указанных слов.Она будет равняться

∑SV\_conf+(i)/∑ SV(потенциально понятых пользователем)

то есть по сути это сумма (важность понятного слова \* уверенность в предсказании) / сумма (важность понятого слова)

Также для отображения количества потенциально понимаемой лексики мы будем использовать следующее значение ∑SV(потенциально понятых пользователем)/∑SV(всех слов в тексте)

*Предсказание по фичам предложения*

Для каждого предложения в тексте также формируем вектор состоящий из соответствующих вышеописанных фич И делаем предсказание, значение которого будет колебаться между 0 и 1, отображая вероятность понимания каждого из предложений человеком

Для всех предложений собираем список предсказаний понимания и ищем среднее значение этих пониманий, которое будет отображать вероятность понимая всех предложений в тексте

*Предсказание по текстовы фичам*

Аналогично для каждого текста рассчитаны текстовые фичи, которые подаются на предикт и выдается число от 0 до 1, которое будет отображать вероятность количества правиьно отвеченых вопросов к этому тексту

**Расчет рекоммендационных параметров**

Подразумевается. что комфортный уровень понимания всех параметров находится где-то в районе 85%

*Для вокабулярной рекомендации*

∑SV(потенциально понятых пользователем)/∑SV(всех слов в тексте) - 0.85

(разность потенциально понятой лексики и 0.85)

∑SV\_conf+(i)/∑ SV(потенциально понятых пользователем)

(важность понятного слова \* уверенность в предсказании) / сумма (важность понятого слова) --- увереность в данном предскзаании

*Для рекомендации по предложениям*

Доля понятых предложений - 0,85

*Для рекомендаций по тексту*

Доля вопросов с верными ответами на вопросы - 0.85

**Лидером рекомендации** будет текст, который будет иметь минимальные вышеперечисленные показатели при учете максимальности уверености в данном вокабулярном предсказании