Sberbank Data Science Contest 2017

Задача А

Определение релевантности вопроса параграфу

Желубенков Александр
9 декабря, 2017

Данные

«Специально для данного соревнования был собран первый в своем роде набор данных для вопрос-ответных систем на русском языке. Данные были собраны из русскоязычных статей, лежащих в открытом доступе. Совместными усилиями более тысячи человек удалось собрать 100 543 пары вопросов и ответов по 18 334 уникальным параграфам.»

«В двух представленных нами задачах мы предоставим участникам 50 365 пар вопросов и ответов с их параграфами для анализа и построения моделей. Оставшиеся пары вопросов и ответов будут скрыты и использоваться в качестве тестовых множеств двух задач.»

Задача A: определение релевантности вопроса

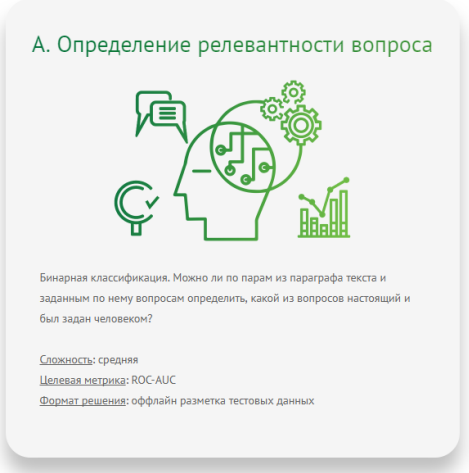
Требуется: построить алгоритм, определяющий релевантность поставленных вопросов к параграфу текста. Для решения этой задачи требуется не только понимать, относится ли вопрос к параграфу, но и насколько корректно он поставлен.

Задача бинарной классификации, в которой целевая переменная target принимает два значения: 0 и 1.

Класс 1 - релевантные вопросы, заданные к параграфу человеком.

Класс 0 - вопросы, либо заданные человеком к другим параграфам, **либо были составлены компьютером**.

Метрика: ROC-AUC



Данные для задачи A

Тренировочная выборка: 119 399 пар (параграф, вопрос);

Тестовая выборка: 74 295 пар (параграф, вопрос);

Формат: paragraph_id, question_id, paragraph, question, target.

Замечания:

- Релевантные вопросы класса 1 были случайно выбраны из собранных вопросов и ответов.
- Нерелевантные примеры класса 0, составленные человеком, были получены случайным выбором вопроса к другому параграфу по той же теме.
- Нерелевантные вопросы класса 0, сгенерированные компьютером, в тренировочных данных отсутствуют.

Примеры из данных

Пары из тренировочной выборки

Параграф: «Более фантастический характер имеют аллегории и пророчества: в первых Леонардо да Винчи использует приемы средневековых энциклопедий и бестиариев; вторые носят характер шутливых загадок, отличающихся яркостью и меткостью фразеологии и проникнутых язвительной, почти вольтеровской иронией, направленной по адресу знаменитого проповедника Джироламо Савонаролы. Наконец, в афоризмах Леонардо да Винчи выражена в эпиграмматической форме его философия природы, его мысли о внутренней сущности вещей. Художественная литература имела для него чисто утилитарное, подсобное значение.»

Релевантный вопрос: «Какой характер носят пророчества Леонардо да Винчи?»

Нерелевантный вопрос: «К отрицанию каких наук приходит Леонардо да Винчи?»

Примеры синтетических вопросов из тестовой выборки:

- «Что происходило с сайтами по зерна на почве религиозного фанатизма?»
- «Общим предком Льва Толстого и формой крыльев подражают южноамериканские парусники papilio bachus, papilio zagreus, белянка dismorphia astynome?»
- «Когда ввели различие между естественными и социальными условиями от направлений развивали теоретические обоснования отличия мира природы от реальности, допустил несколько направлений?»

Подготовка данных

Морфологический анализ: pymystem3

- Нормализация;
- Выделение граммем;

Частоты:

- Частоты слов(буквенных триграмм) в параграфах/вопросах;
- Idf слов(буквенных триграмм) в параграфах/вопросах;

Количество вопросов и параграфов:

	Train	Test	Train & Test	Train | Test
Параграфы	9.078	1.627	0	10.705
Вопросы	38.868	43.340	7	82.201
Пары	119.398	74.286	0	193.684

Факторы вопроса (1-2)

(1) **Количество повторений вопроса**

Было замечено, что синтетические вопросы НЕ повторяются, т.е. вопрос может быть синтетическим только если он один раз встретился в данных.

(2) **Простые текстовые факторы**

Было замечено, что если вопрос заканчивается на однобуквенное слово, то вероятно он синтетический.

Пример:

- «Что рабочий добавляет своим трудом стоимость, б?»

Факторы:

- Окончание вопроса на однобуквенное слово
- Количество слов в вопросе, доля уникальных слов в вопросе
- Количество запятых, скобок

Факторы вопроса (3-4)

(3) **Встречаемость граммем**

Преобразуем тексты вопросов, используя граммемы, выделенные pymystem3:

- Части речи; биграммы частей речи (текст заменяется на последовательность частей речи слов)
- Пары "Падеж сущ. + ближайший предлог, встретившийся ДО сущ."
- Падежи сущ. и прилаг; биграммы "Падеж сущ/прилаг + Падеж сущ/прилаг"
- Число(Род) сущ, прилаг, глаг, мест-сущ, мест-прилаг; биграммы последовательных чисел (мн или ед), (муж, ср, жен)
- Прочие обозначения, выделяемые pymystem3 (гео, имя, фам, ...); биграммы последовательных обозначений

Факторы: количество и доля выделенных граммем, которые встречаются не менее чем в 50 различных вопросах, **число факторов**: 1484.

(4) **Встречаемость граммем в параграфах**

Вместо отдельных факторов для каждой выделенной граммемы(пары граммем), можно, на основании того «насколько часто граммемы(биграммы граммем) встречаются среди всех параграфов», посчитать агрегированные статистики по всем выделенным для вопроса граммемам:

- средняя встречаемость в параграфах;
- минимальная встречаемость в параграфах (т.е. насколько редко можно встретить граммему в нормальных текстах)

Факторы вопроса (5)

(5) **Количество повторов слов/словосочетаний**

Было замечено, что в синтетических вопросах достаточно часто друг за другом могут идти одинаковые слова/словосочетания.

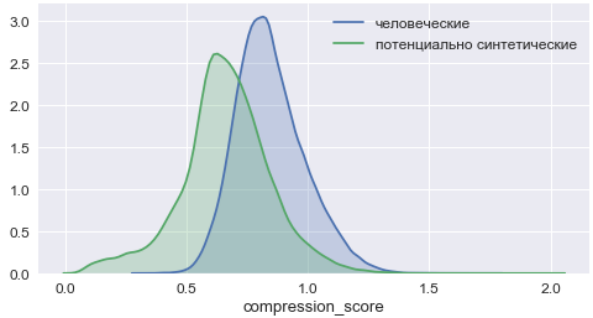
В параграфах и человеческих вопросах такие повторения НЕ встречаются.

Примеры:

- «Сколько рек будет немного медленней, то некоторые виды, например, некоторые виды, например некоторые виды, например некоторые формы?»
- «Как авторов книг и фрагменты стилистически сырого текста, а историки историки историки историки?»
- «Что действует на любое на любое тело в газе в поле тяготения?» - но это человеческий вопрос

Факторы:

- Количество найденных повторов
- Степень сжатия строки (zlib.compress)



Факторы вопроса (6)

(6) **Слова с большой буквы**

Было замечено, что в синтетических вопросах слова "с большой буквы" встречаются редко(если отбросить первое слово). Иногда они все же встречаются, но располагаются ближе к началу вопроса.

Это наблюдение можно использовать двумя способами:

- если в вопросе встретилось слово "с большой буквы", тогда вопрос скорее НЕ может быть синтетическим
- если в вопросе встречаются слова, которые в параграфах часто пишутся "с большой буквы"(например, слово Леонардо), то вопрос скорее всего синтетический

Примеры:

- «Какой характер носят пророчества Леонардо да Винчи?»
- «Кто сравнивает леонардо да винчи в базовых принципов гражданского права, которое начало курсировать в эпохе неолита?»

Факторы:

- наличие слова с большой буквы
- относительная позиция слова с большой буквы
- сумма по всем словам "вероятности" написания слова с большой буквы
- максимум по всем словам "вероятности" написания слова с большой буквы

Факторы вопроса (7)

(7) **Проверка грамматики с помощью LanguageTool**

LanguageTool is Open Source style and grammar checker (http://wiki.languagetool.org/)

На выходе для каждого вопроса получим набор правил, которые срабатывают на данном вопросе.

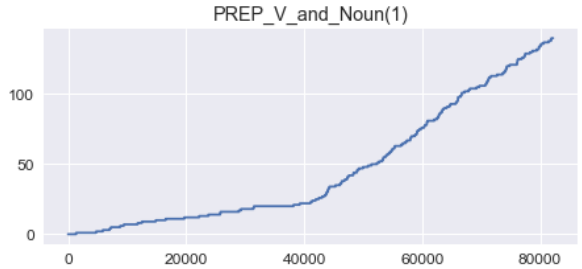
Примеры:

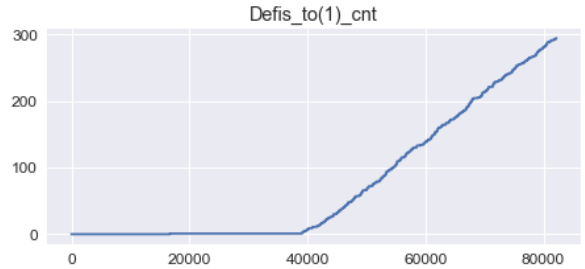
- «Когда роли посредников в отношений нарушений конвенции?»

Правило PREP_V_and_Noun(1): "Предлог «в» предполагает употребление существительного в винительном или предложном падеже"

- «Кто то таковые становятся известны окружающим , логические аргументы и различную статистическую информацию?»

Правило Defis_to(1): "Слова с частицей «-то» пишем через дефис"





Факторы вопроса (8)

(8) **Наибольшая общая подстрока вопроса и всех параграфов**

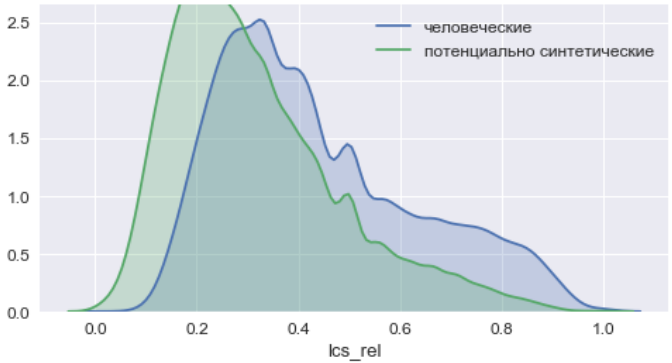
Часто в человеческом вопросе может содержаться достаточно длинная подстрока **одного из параграфов**.

Этот факт может нести сигнал о том, что вопрос НЕ является синтетическим.

Пример:

Параграф: «… Аппарат Гольджи асимметричен — цистерны располагающиеся ближе к ядру клетки (цис-Гольджи) содержат наименее зрелые белки, к этим цистернам непрерывно присоединяются мембранные пузырьки — везикулы, отпочковывающиеся от эндоплазматического ретикулума...»

Вопрос: «Что содержат цистерны располагающиеся ближе к ядру клетки ?»



Факторы вопроса (9)

(9) **Однородность вопроса**

В синтетических вопросах могут встречаться слова из вообще НЕ связанных тематик. Наличие пары таких слов может нести сигнал о том, что вопрос синтетический.

Связь слов можно считать на основании схожести разных векторных представлений слова w:

- **word-paragraphs**: бинарный вектор по параграфам (в i-й компоненте будет стоять 1, если слово w содержится среди слов i-го параграфа)
- **word-questions**: бинарный вектор по вопросам (в i-й компоненте будет стоять 1, если слово w содержится среди слов i-го вопроса)
- **word-words(q)**: целочисленный вектор по словам (в i-й компоненте будет стоять число параграфов(вопросов), в которых содержится и слово w, и слово с номером i)
- **w2v**: предобученные w2v-embeddings (Источник «CoNLL 2017, Russian/ru.vectors»: word embeddings of dimension 100 computed from lowercased texts by word2vec)

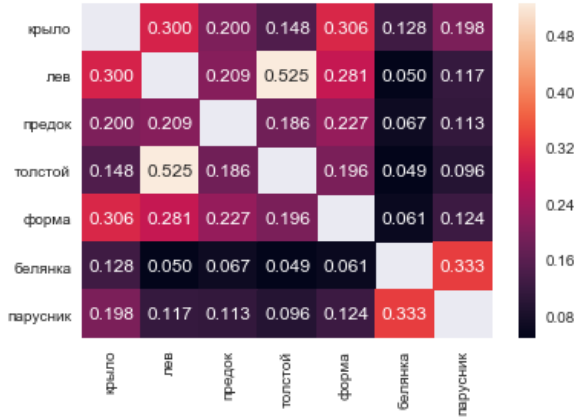
В итоге факторами будут некоторые статистики (min, minimax), вычисленные для матрицы слов вопроса, значения которой вычислены на основании схожести векторных представлений слов.

Факторы вопроса (9)

(9) **Однородность вопроса**

Примеры вопросов:

- «Что происходило с сайтами по зерна на почве религиозного фанатизма?»: (min_score = 0.034)
- «Почему речной окунь наименее популярная рыба для разведения, чем судак?»: (min_score = 0.188)
- «Общим предком Льва Толстого и формой крыльев подражают южноамериканские парусники papilio bachus, papilio zagreus, белянка dismorphia astynome?»: (min_score = 0.049)



Факторы вопроса (10)

(10) **Повторы суффиксов, префиксов**

Было замечено, что в синтетических вопросах могут использоваться одни и те же префиксы, суффиксы человеческих вопросов, к которым добавляется оставшаяся часть вопроса.

Факторы: частоты встречаемости суффиксов/префиксов разной длины среди всех уникальных вопросов.

Примеры префиксов:

- Сколько продолжается жизненный цикл организации взгляд финансиста?
- Сколько продолжается жизненный цикл и выступали перед публикой со своим отцом и продолжается с главными соперниками клуба?
- Сколько продолжается жизненный цикл личинки развиваются в проливе китера?
- Сколько продолжается жизненный цикл её клетка?

Примеры суффиксов:

- Как признан морально устаревшим , долгое время кризис, рецессия, депрессия?
- Когда потерял её и обслуживании высококвалифицированных инвесторов долгое время долгое время кризис, рецессия, депрессия?
- Когда то здесь получали высокие оценки со стороны инвесторов долгое время долгое время кризис, рецессия, депрессия?

Факторы пары (параграф, вопрос) (1)

(1) **Схожесть вопроса параграфу**

Схожесть вопроса параграфу можно вычислять на основании мощности пересечения слов(частей слов, триграмм) вопроса и параграфа. Слова можно взвешивать, например, используя idf для слов.

Схожесть должна зависеть от близости найденных слов, для этого параграфы могут разбиваться на предложения или на перекрывающиеся кусочки слов. В решении использовалось скользящее окно длиной 30 слов с шагом 15 слов.

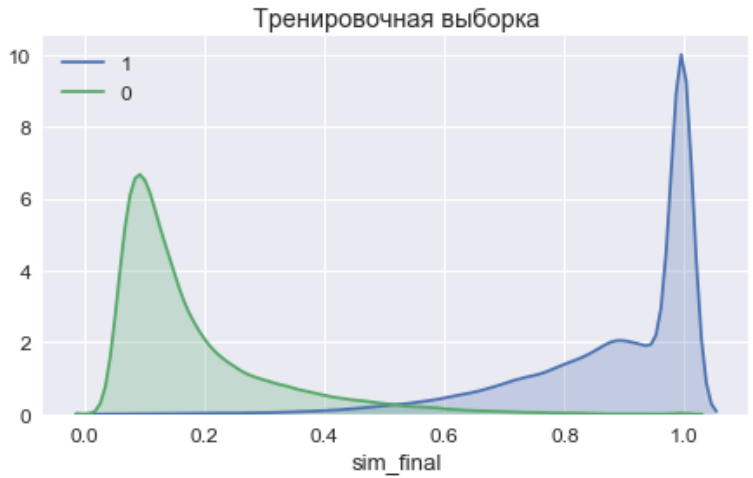
Т.к. в вопросах нередко встречаются опечатки, их можно исправлять (в решении использовался простой алгоритм, найденный в интернете).

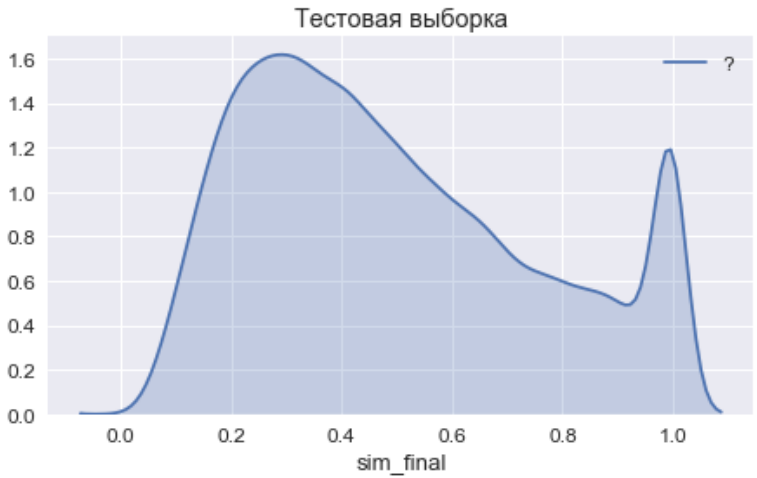
Факторы:

- Триграммы + скользящее окно
- Нормализованные слова (скользящее окно, только существительные)
- Нормализованные слова + разбиение на предложения
- Нормализованные слова + стемминг

Факторы пары (параграф, вопрос) (1)

(1) **Схожесть вопроса параграфу**





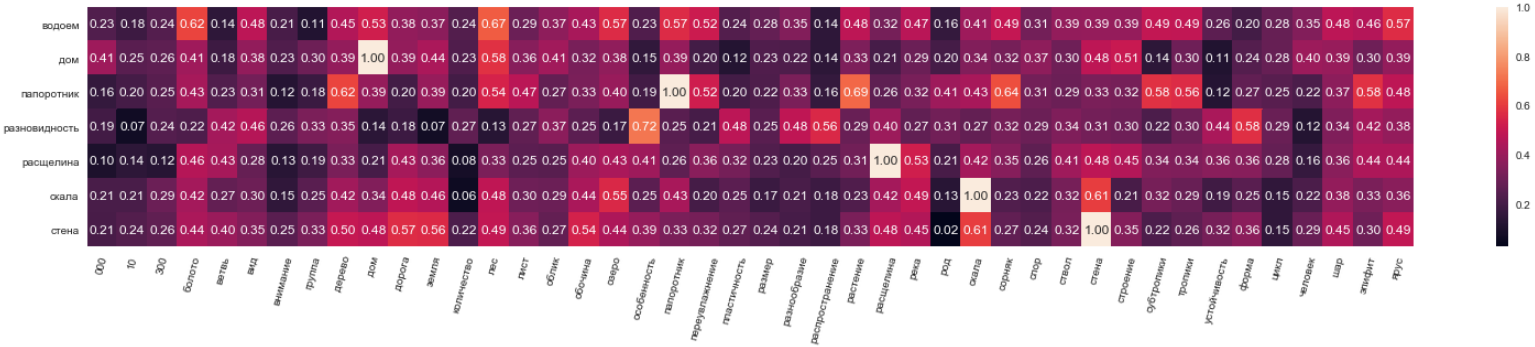
Факторы пары (параграф, вопрос) (2)

(2) **Однородность слов вопроса и параграфа**

Аналогично тому, как на основании векторных представлений слов вычислялась однородность вопроса, можно вычислять однородность слов вопроса словам параграфа. **Статистики:** minimax.

Параграф: «Современные **папоротники** — одни из немногих древнейших растений, сохранивших значительное разнообразие, сопоставимое с тем, что было в прошлом. Папоротники сильно различаются по размерам, жизненным формам, жизненным циклам, особенностям строения и другим особенностям... Папоротники встречаются в лесах — в нижнем и верхнем ярусах, на ветвях и стволах крупных деревьев — как эпифиты, в **расщелинах скал**, на болотах, в реках и озёрах, на **стенах** городских **домов**, на сельскохозяйственных землях как сорняки, по обочинам дорог…»

Вопрос: «Какая разновидность **папоротников** произрастает вблизи водоемов, в **расщелинах скал**, а также на **стенах** городских **домов**?»



Факторы пары (параграф, вопрос) (3)

(3) **Наибольшая общая подстрока вопроса и параграфа**

Часто в вопросе может содержаться достаточно длинная подстрока параграфа, этот факт может нести сигнал о том, что вопрос релевантен параграфу.

Пример:

Параграф: «… Аппарат Гольджи асимметричен — цистерны располагающиеся ближе к ядру клетки (цис-Гольджи) содержат наименее зрелые белки, к этим цистернам непрерывно присоединяются мембранные пузырьки — везикулы, отпочковывающиеся от эндоплазматического ретикулума...»

Вопрос: «Что содержат цистерны располагающиеся ближе к ядру клетки ?»

Факторы пары (параграф, вопрос) (4)

(4) **Ранги параграфа для вопроса и вопроса для параграфа**

Вычислим сильнейший из факторов схожести вопроса и параграфа (Нормализованные слова + скользящее окно) для всех возможных пар (параграф, вопрос).

Для каждого параграфа на выходе оставим топ-250 вопросов по схожести.

На основании вычисленных схожестей, будем вычислять факторы, опираясь на следующие гипотезы:

- для пары (параграф, вопрос) может быть важно, что параграф является самым похожим на вопрос, т.е. могут быть важны не только абсолютные значения схожести, но и ранги параграфов внутри вопроса.
- т.к. человеческий вопрос задан к какому-то из параграфов, то он должен быть максимально похож хотя бы на один из параграфов (в предположении идеального фактора схожести и предположения, что параграф, к которому задавался этот вопрос присутствует в открытых предоставленных данных)
- распределение схожестей на параграфы для синтетических вопросов должно отличаться от распределения схожестей для человеческих вопросов

Факторы пары (параграф, вопрос) (4)

(4) **Ранги параграфа для вопроса и вопроса для параграфа**

Факторы для пары (параграф, вопрос):

- ранг вопроса для параграфа;
- ранг параграфа для вопроса;
- произведение "ранга вопроса для параграфа" и "ранга параграфа для вопроса";
- дельта между максимальной схожестью вопроса с одним из параграфов и схожестью данного вопроса данному параграфу (сколько не хватает до максимума).

Факторы для вопросов:

- минимальный ранг вопроса по всем параграфам;
- среднее, дисперсия схожестей вопроса по всем параграфам(для которых данный вопрос попал в топ-250), количество таких параграфов;
- максимальная схожесть вопроса по всем параграфам;
- 2(5, 10, 20)-я максимальная схожесть вопроса по всем параграфам (аналог квантилей).

Факторы пары (параграф, вопрос) (5)

(5) **Вопросительные/последние слова + LabelEncoding на основании макс. схожести**

Было замечено, что вопросительные слова/словосочетания часто повторяются, при этом:

- Вопросительные слова могут задавать структуру вопроса и от них могут зависеть вычисленные схожести.
- Распределение схожестей внутри пар (параграф, вопрос) для конкретных вопросительных слов может различаться.

LabelEncoding вопросительных слов с преобразованием их в среднее значение максимальной схожести.

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
qWord2		в вид	к кто	в зависимость	из сколька	мочь	из что	в результат	в чей	представитель	на основа	первый	среди	зачем	у какой	в ход
max_qsim	len	38	42	30	31	25	195	66	72	27	49	22	36	31	166	20
	mean	0.948891	0.944668	0.934427	0.93044	0.92405	0.916684	0.915969	0.914102	0.91402	0.912969	0.912882	0.911569	0.909929	0.909844	0.905904

		83	84	85	86	87	88	89	90	91	92	93	94	95	96	97
qWord2		насколько	год	о какой	в скольким	приводить	что	из скольким	дата	кто	как	когда	где	чем	сколько	вкак
max_qsim	len	31	20	26	27	20	12658	32	20	7607	8470	6533	5992	3958	9071	94
	mean	0.803583	0.795958	0.793886	0.789854	0.759028	0.755883	0.747257	0.720749	0.69506	0.685423	0.650181	0.625294	0.616653	0.569173	0.414015

Аналогично можно поступить с окончаниями вопроса(последними словами), однако вместо самих слов использовать их части речи, оставляя или не оставляя при этом предлоги и союзы.

Предсказательные модели: подходы 1 и 2

Подход 1.

$$score(p, q) = sim(p, q), \text{где}$$

$sim(p, q)$ – схожесть вопроса параграфу.

Подход 2.

$$score(p, q) = sim(p, q) * p_{human}(q), \text{где}$$

$p_{human}(q)$ – модель, предсказывающая «насколько вопрос похож на вопросы из Train выборки».

Цель модели $p_{human}(q)$: научиться определять человеческие вопросы.

Обучение модели $p_{human}(q)$: «вопросы из Train» vs «вопросы из Test».

Предсказательные модели: подходы 3 и 4

Подход 3.

$$score(p, q) = p_{rel}(p, q), \text{где}$$

$p_{rel}(p, q)$ – модель, предсказывающая релевантность вопроса параграфу.

Цель модели $p_{rel}(p, q)$: научиться отделять релевантные пары (параграф, вопрос) от нерелевантных пар и пар с синтетическими вопросами.

Обучение модели $p_{rel}(p, q)$: «положительные пары из Train» vs «остальные пары» (т.е. отрицательные пары из Train и вся Test выборка).

Подход 4.

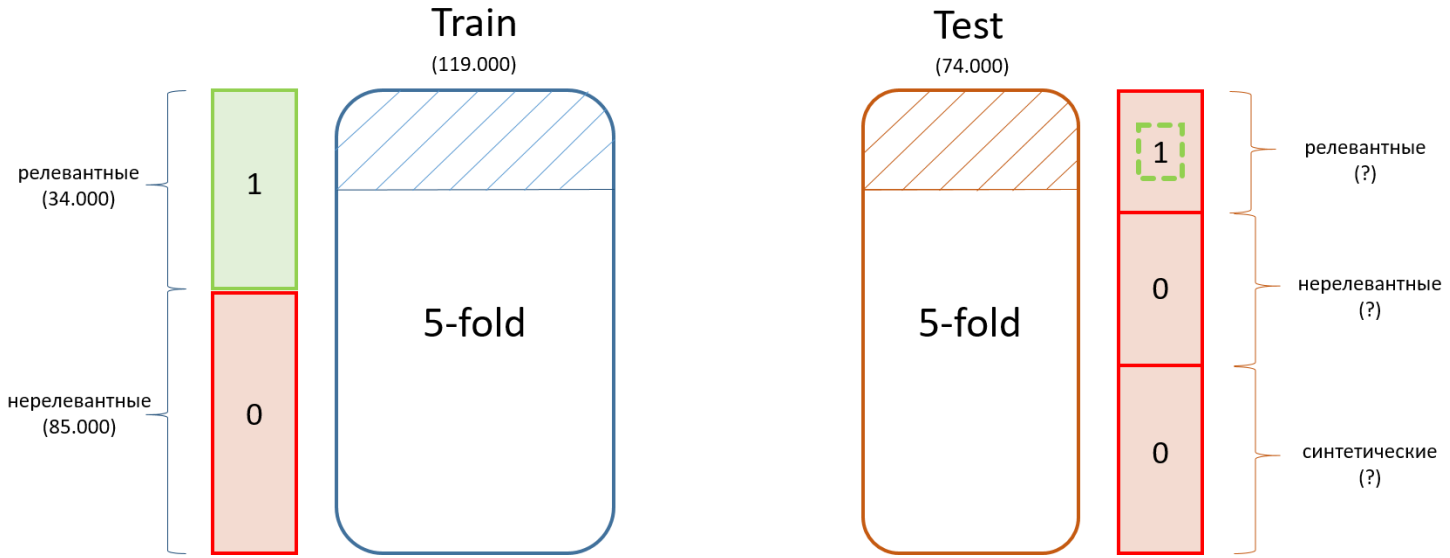
$$score(p, q) = p_{rel}(p, q) * p_{train}(p, q), \text{где}$$

$p_{train}(p, q)$ – модель, предсказывающая «насколько пара (параграф, вопрос) похожа на пары из Train выборки».

Цель модели $p_{train}(p, q)$: научиться отделять нормальные пары от пар с синтетическими вопросами.

Обучение модели $p_{train}(p, q)$: «пары из Train» vs «пары из Test».

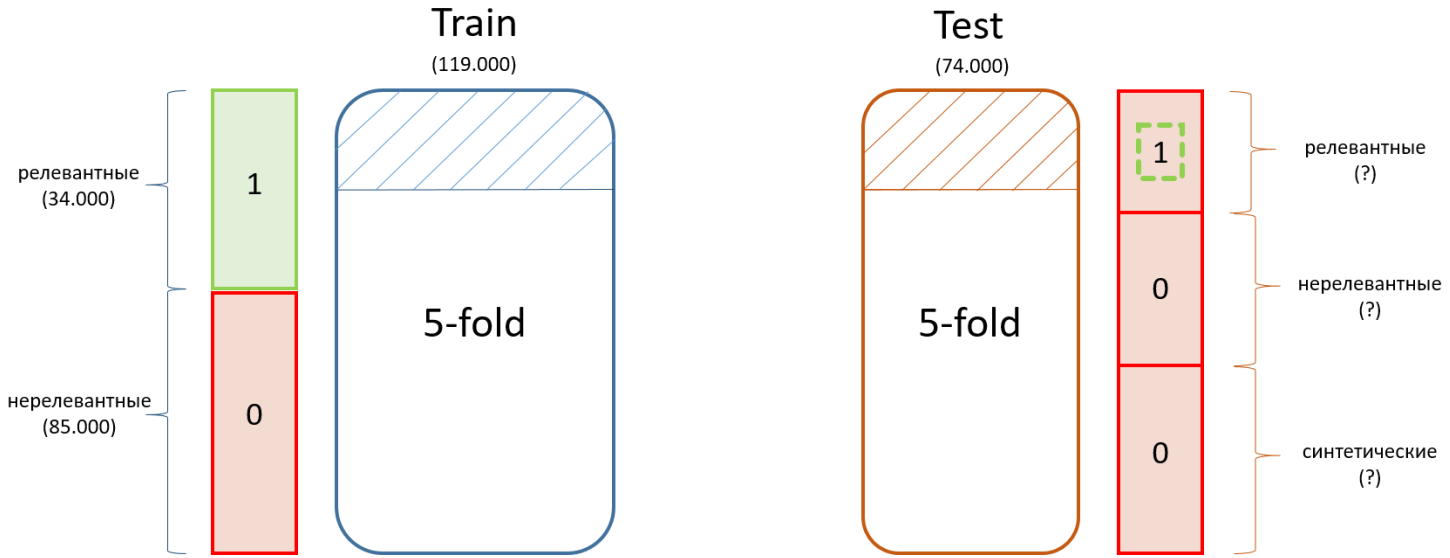
Обучение модели $p_{rel}(p, q)$



Цель модели $p_{rel}(p, q)$: научиться отделять релевантные пары (параграф, вопрос) от нерелевантных пар и пар с синтетическими вопросами.

Обучение модели $p_{rel}(p, q)$: «положительные пары из Train» vs «остальные пары» (т.е. отрицательные пары из Train и вся Test выборка).

Обучение модели $p_{rel}(p, q)$



Метрики, вычисляемые на кросс-валидации:

- ROC AUC;
- log loss на выборке Train;
- ROC AUC на выборке Train;

Обучение модели $p_{rel}(p, q)$

Отбор факторов:

- Число факторов вопроса: 1708 -> 298
- Число факторов пары (параграф, вопрос): 36 -> 31
- Число факторов параграфа: 0.

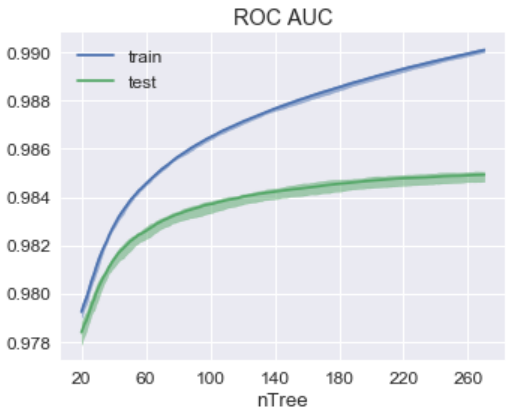
Обучение:

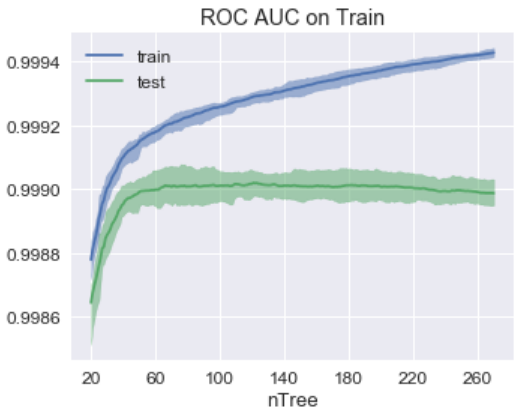
dmlc XGBoost

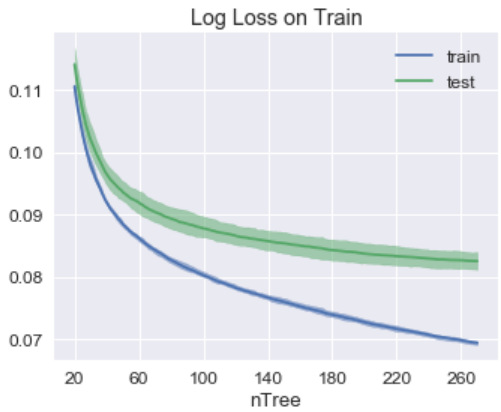
Параметр	Отбор признаков	Итоговая модель
max_depth	5	4
min_child_weight	30	50
n_estimators	80	270
learning_rate	0.4	0.2
colsample_bylevel	0.1	0.2
subsample	1	0.85

Обучение модели $p_{rel}(p, q)$

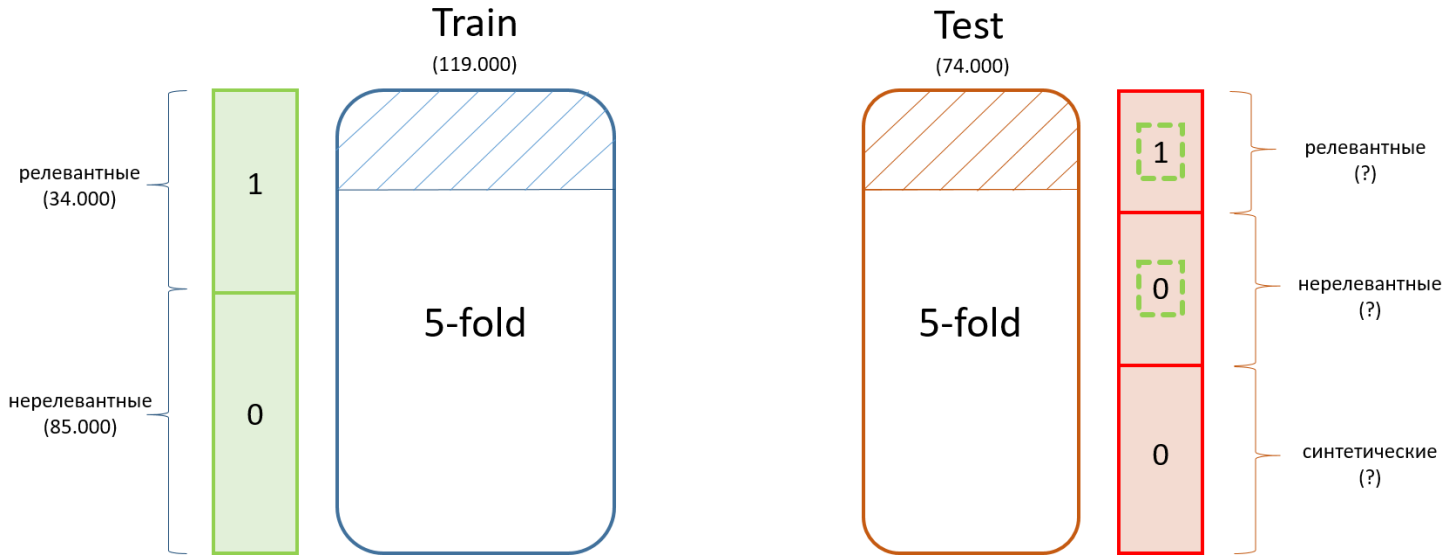
Кривые обучения:







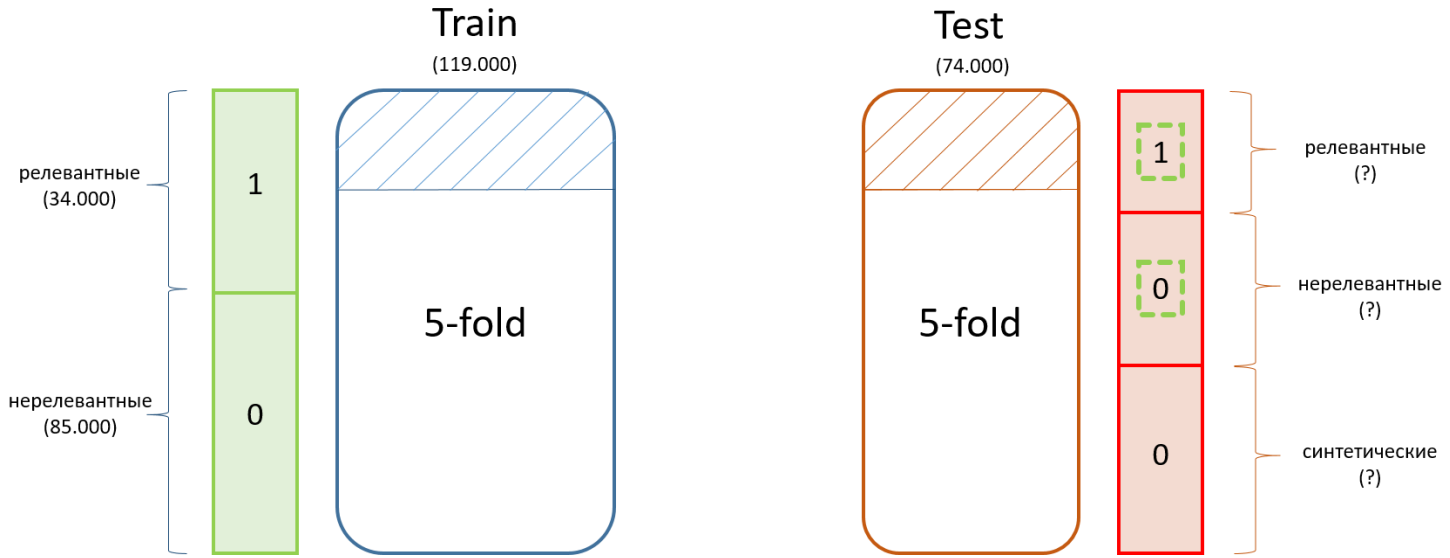
Обучение модели $p_{train}(p, q)$



Цель модели $p_{train}(p, q)$: научиться отделять нормальные пары от пар с синтетическими вопросами.

Обучение модели $p_{train}(p, q)$: «пары из Train» vs «пары из Test».

Обучение модели $p_{train}(p, q)$



Метрики, вычисляемые на кросс-валидации:

- ROC AUC;
- log loss на положительных примерах;

Обучение модели $p_{train}(p, q)$

Отбор факторов:

- Число факторов вопроса: 1708 -> 301
- Число факторов пары (параграф, вопрос): 36 -> 31
- Число факторов параграфа: 0.

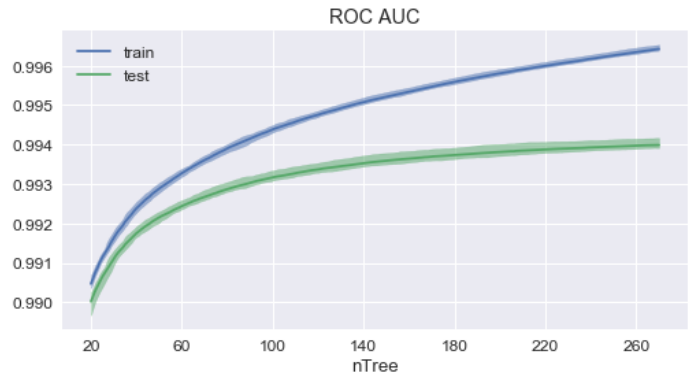
Обучение:

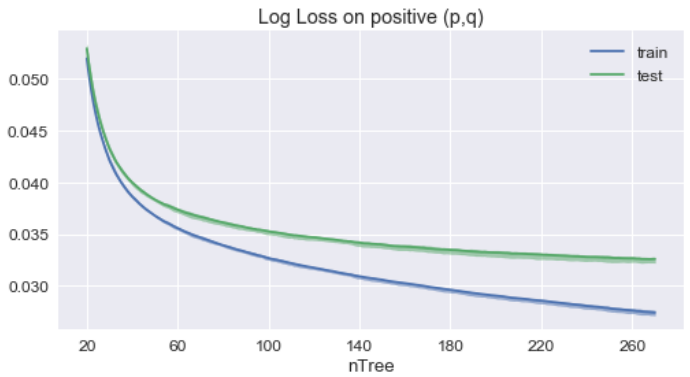
dmlc XGBoost

Параметр	Отбор признаков	Итоговая модель
max_depth	5	4
min_child_weight	30	50
n_estimators	80	270
learning_rate	0.4	0.20
colsample_bylevel	0.1	0.20
subsample	1	0.85

Обучение модели $p_{train}(p, q)$

Кривые обучения:



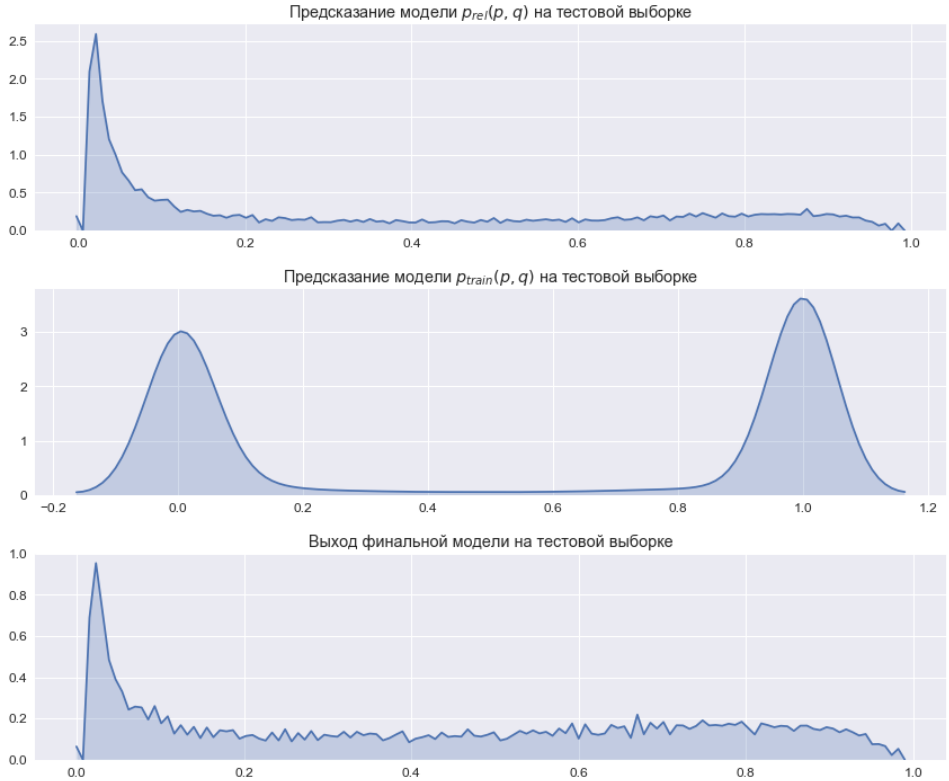


Финальная модель

Усреднение рангов финальных предсказаний 5 моделей по различным разбиениям test-а на фолды.

Финальная модель

Усреднение рангов финальных предсказаний 5 моделей по различным разбиениям test-а на фолды.



Результаты подходов на Public LB

Подход	Формула	Public Score (ROC AUC)
1	$sim(q,p)$	< 0.8
2	$sim(q,p) * p_{human}(q)$	< 0.983
3	$p_{rel}(q,p)$	< 0.988
4	$p_{rel}(q,p) * p_{train}(q,p)$	0.99+

Итоговые результаты

Public LB

#	Участник	Последняя загрузка	Всего загрузок	Рейтинг
1	topspin26	31 октября 2017 г. в 2:56	92	0.99438
2	kirillskor	31 октября 2017 г. в 1:29	201	0.99415
3	puryginm	31 октября 2017 г. в 0:06	14	0.99374
4	igordoynikov	31 октября 2017 г. в 2:01	131	0.99371
5	vlarine [ooc]	31 октября 2017 г. в 1:23	129	0.99365
6	therealroman	31 октября 2017 г. в 1:33	35	0.99349
7	prampampam	31 октября 2017 г. в 2:46	76	0.99343
8	antihype	29 октября 2017 г. в 12:55	49	0.98998
9	ambitious	31 октября 2017 г. в 0:36	73	0.98173
10	ln	30 октября 2017 г. в 21:29	46	0.97990

Private LB

#	Участник	Последняя загрузка	Всего загрузок	Рейтинг
1	topspin26	31 октября 2017 г. в 2:56	92	0.99466
2	vlarine [ooc]	31 октября 2017 г. в 0:00	129	0.99440
3	therealroman	31 октября 2017 г. в 1:33	35	0.99416
4	prampampam	31 октября 2017 г. в 2:46	76	0.99349
5	puryginm	30 октября 2017 г. в 20:16	14	0.99335
6	igordoynikov	31 октября 2017 г. в 2:01	131	0.99330
7	kirillskor	31 октября 2017 г. в 1:10	201	0.99304
8	antihype	29 октября 2017 г. в 12:55	49	0.99054
9	ambitious	27 октября 2017 г. в 16:08	73	0.98177
10	ln	30 октября 2017 г. в 19:21	46	0.98138

Технические детали

Инструменты:

- IPython notebook
- pymystem3, pandas, numpy, scipy, sklearn, xgboost
- Intel "Core-i7-7700K", 32GB RAM

Время:

- Генерация факторов: ~8 часов (без учета вычисления схожестей для всех пар (параграф, запрос) на Hadoop-кластере).
- Обучение финальной модели и получение предсказаний: ~3 часа.

Ссылка на код:

- https://github.com/Topspin26/SberbankDataScienceContest_2017

Другие решения и идеи

- Более качественная проверка грамматики (правописания);
- Другие word-embeddings для слов;
- Выделение тематик параграфов;
- Использование данных и моделей из задачи B.

Спасибо за внимание!

Желубенков Александр
zhelubenkovalexandr@gmail.com