Тематическое моделирование образовательных целей пользователей в системе дистанционного образования

Потапова Полина

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. К. В. Воронцов

Москва,
2020 г.

Цель работы

Актуальность

Пользователи систем дистанционного образования получают персонализированные рекомендации. Для построения ИОТ необходимо учитывать их образовательные цели. Нужно уметь оценивать конкретность целей.

SMART методология постановки целей

- Specific: Конкретная
- Measurable: Измеримая
- Achievable or Attainable: Достижимая
- Relevant: Значимая
- Time bound: Ограниченная по времени

Цель работы

Построить модель бинарной классификации для определения конкретности образовательных целей, написанных на естественном языке.

Примеры целей

Хорошо сформулированы относительно SMART

- Изучить историю дизайна, колористику, в совершенстве владеть графическими программами линейки Adobe: Ps, Ai, Ae, Xd, Lr, Id, Pr, а также историю и инструменты менеджмента для поступления в магистратуру Политехнического университета Петра Великого до июня 2020 года.
- Войти в категорию А1 по классификации призывников, что даст возможность попасть в Кремлёвский полк. Deadline: 2021 г.

Плохо сформулированы относительно SMART

- Счастья всем даром, и пусть никто не уйдёт обиженным!
- Хочу быть способным поддержать диалог на любую тему.

Постановка задачи оценки конкретности образовательной цели

Дано

Выборка $X = (x_i, y_i)_{i=1}^{l}$,
x_i - признаковое описание образовательной цели (документа),
$y_i \in \{0, 1\}$ - метка класса.

Найти

Алгоритм бинарной классификации $a(x) : X \to Y = \{0, 1\}$, предсказывающий, является ли образовательная цель конкретной.

Постановка задачи оценки конкретности образовательной цели

Варианты алгоритма $a(x)$

baseline:	Наивный байесовский классификатор, логистическая регрессия, простая тематическая модель классификации
основное решение:	Тематическая модель конкретности образовательных целей

Критерий

Площадь под ROC-кривой (roc_auc), accuracy, precision и recall.

Тематическое моделирование

Вероятностная модель порождения текстовой коллекции документов:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, \;\; d \in D, w \in W,$$

где $p(w|t) = \phi_{wt}$ – распределение терма $w \in W$ в теме $t \in T$,
$p(t|d) = \theta_{td}$ – распределение темы $t \in T$ в документе $d \in D$.
Принцип максимума регуляризованного правдоподобия:

$$\sum_{w \in W} \sum_{d \in D} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \sum_{w \in W} \phi_{wt} = 1, \phi_{wt} \geq 0, \sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0$$

EM-алгоритм:

$$p_{tdw} = \underset{t \in T}{norm}(\phi_{wt}\theta_{td});$$

$$\phi_{wt} = \underset{w \in W}{norm}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \qquad n_{wt} = \sum_{d \in D} n_{dw}p_{tdw};$$

$$\theta_{td} = \underset{t \in T}{norm}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w \in d} n_{dw}p_{tdw};$$

Мультимодальное тематическое моделирование

Документ $d \in D$ состоит из токенов w_m разных модальностей $m \in M$.
Принцип максимума регуляризованного правдоподобия:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

Условия неотрицательности и нормировки:

$$\sum_{w \in W_m} \phi_{wt} = 1, \phi_{wt} \geq 0, \sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0.$$

EM-алгоритм:

$$p_{tdw} = \underset{t \in T}{norm}(\phi_{wt} \theta_{td});$$

$$\phi_{wt} = \underset{w \in W}{norm} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \qquad n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw};$$

$$\theta_{td} = \underset{t \in T}{norm} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \qquad n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw};$$

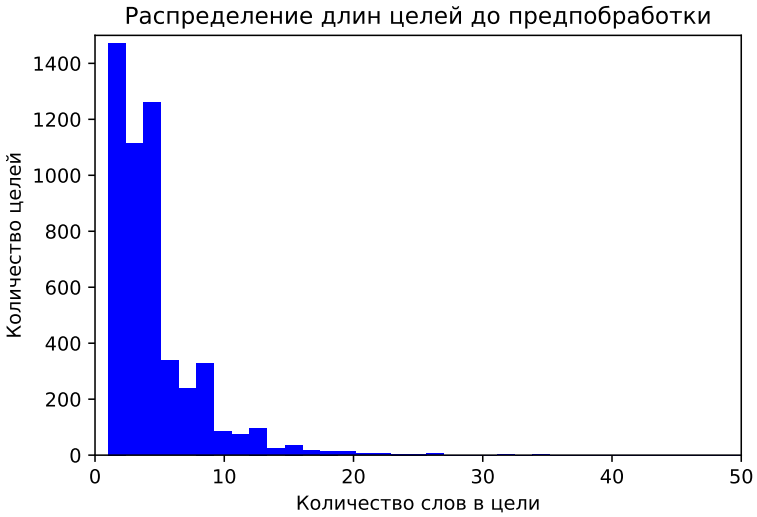
Описание исходных данных

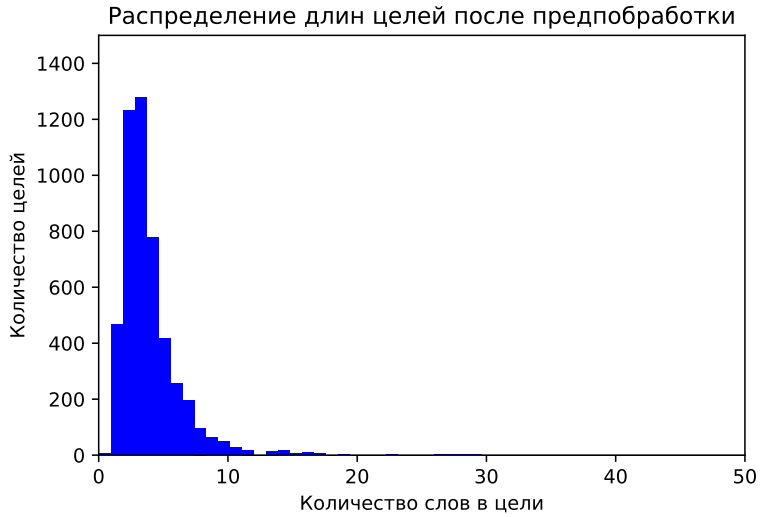
В анкетировании приняли участие около 11 тысяч человек. Каждый человек написал по 3 цели. Наиболее подробно люди описывали первую цель. Из полученных 33 тысяч целей было размечено 6 тысяч первых целей.

Данные о целях:

- Формулировка цели на естественном языке
- Уточнение цели (какой может быть первый шаг, конечный результат, преграды для достижения цели, . . .)
- Социально-демографическая информация (уровень образования, сфера занятости, пол, возраст, . . .)
- Экспертная разметка цели (Specificity, Achievable, Time bound, . . .)

Распределение длин целей

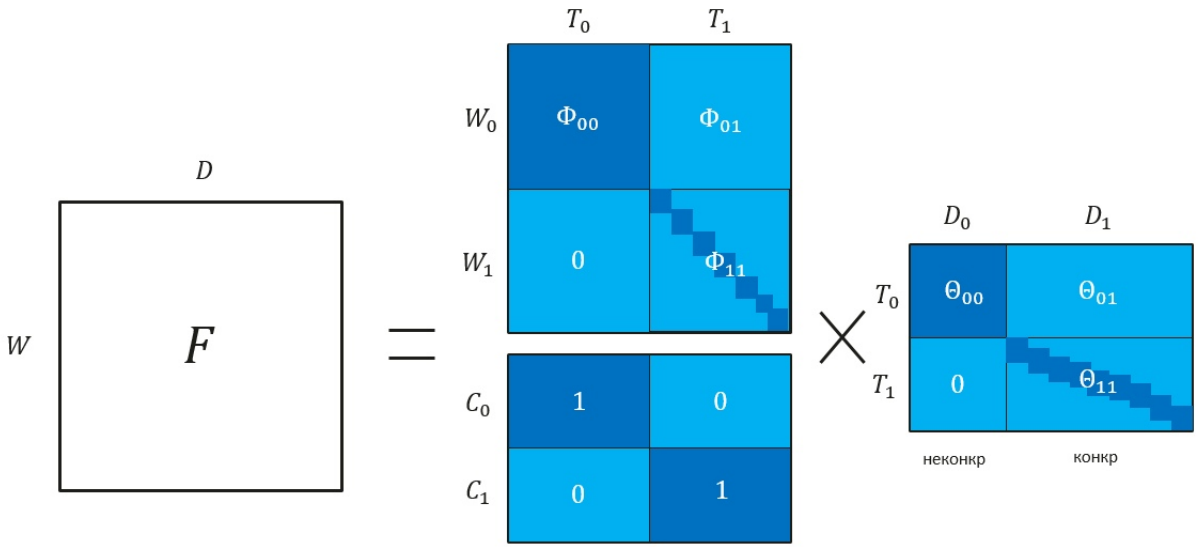




	до предобработки	после предобработки
count	5153	4955
min	1	0
50%	3	3
75%	6	4
max	220	139

Разработка модели специфичности документов

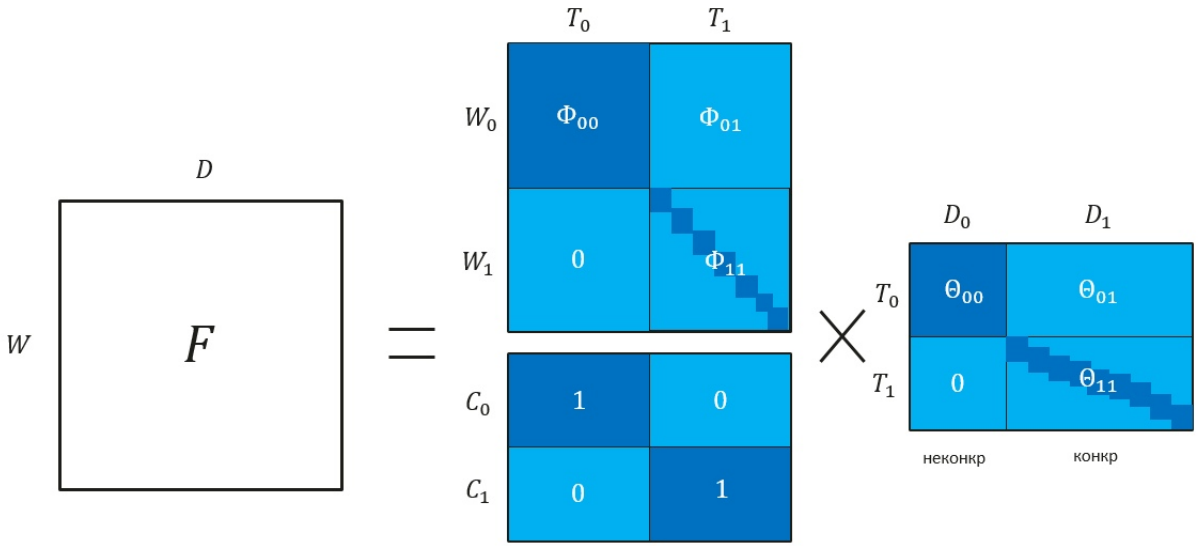
Имеем коллекцию документов $D = D_0 \cup D_1$.
D_0 - неконкретные документы, D_1 - конкретные, специфичные.
Предположение: Неконкретные слова входят во все документы, конкретные - только в конкретные.
$T = T_0 \cup T_1$, T_0 - неконкретные темы, T_1 - конкретные темы



Разработка модели специфичности документов

Инициализация

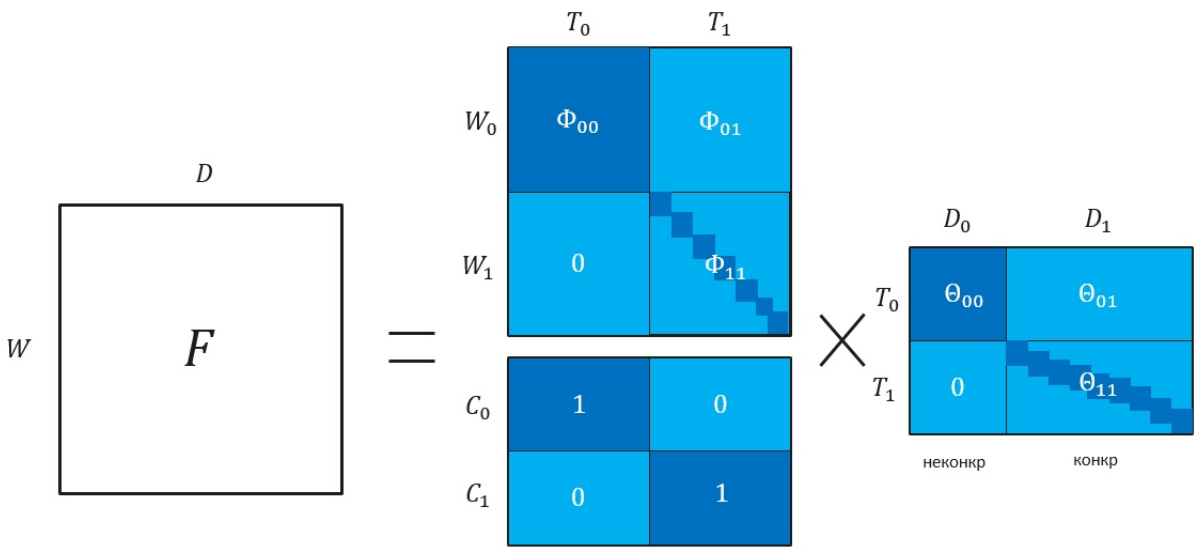
- Φ_{00}: строим вспомогательную тематическую модель неспецифичных целей, т.е. строим T_0 по D_0
- Φ_{01}: инициализируем нулями
- Φ_{10}, Φ_{11}: инициализируем случайными числами из $U(0,1)$ так, чтобы вероятности токенов в Φ_{11} были в 100 раз больше, чем в Φ_{01}
- Θ: инициализируем случайными числами из $U(0,1)$



Разработка модели специфичности документов

Обучение

- Φ_{00}: сглаживание
- Φ_{11}, Θ_{11} разреживание
- Φ_{11}, Θ_{11} декорреляция



Реализация модели специфичности документов

Параметры, которые можно оптимизировать

- количество неспецифичных тем T_0
- количество специфичных тем T_1
- веса τ_m модальностей
- коэффициенты регуляризации

Описание и анализ результытов

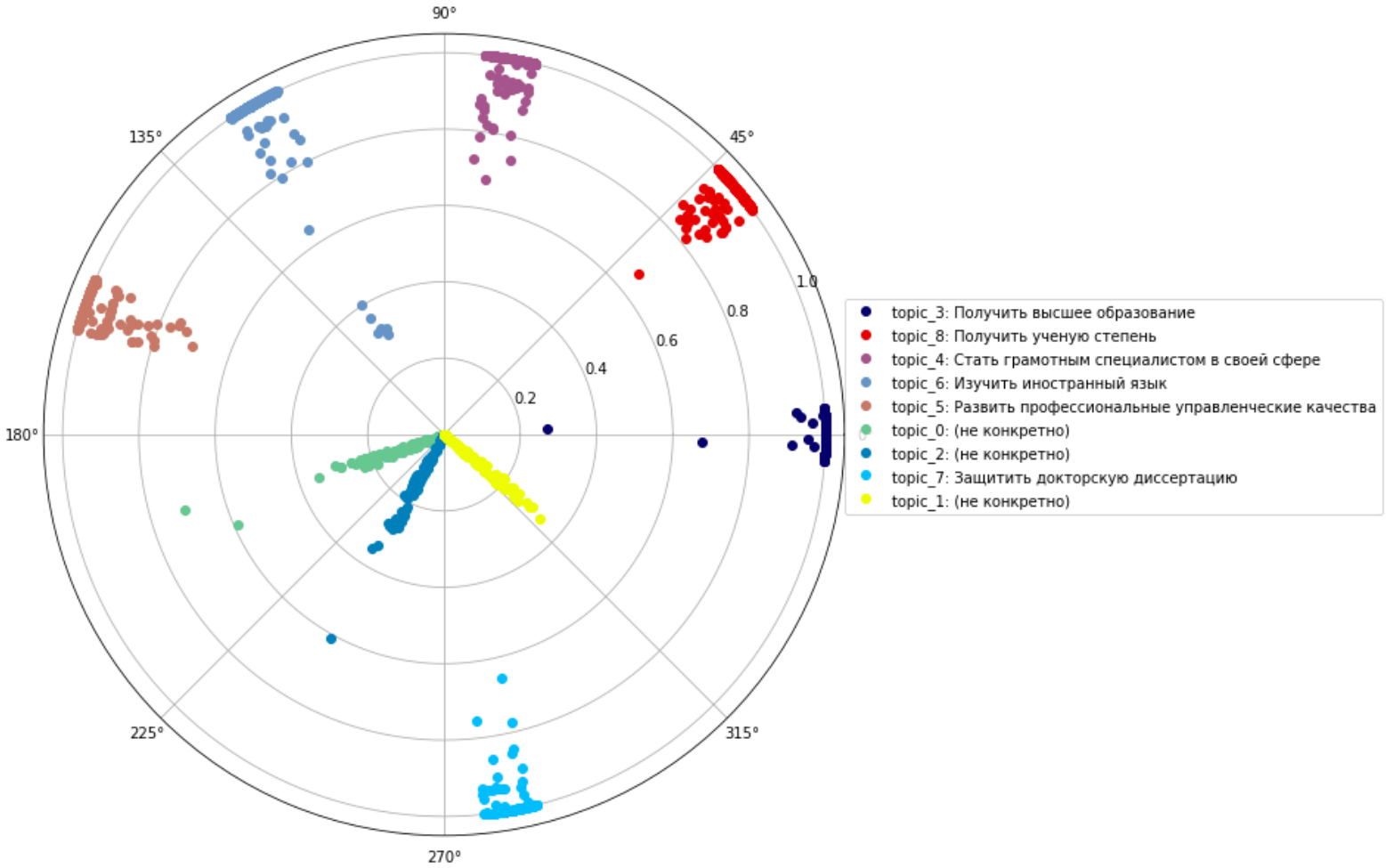
Таблица: Наиболее характерные слова для неконкретных тем

topic_0	topic_1	topic_2
знание	новое	повышение
навык	стать	уровень
получение	свой	новый
развитие	хороший	приобретение
новый	расширение	умение
получить	узнать	рост
новое	кругозор	большой
компетенция	человек	работа
новый_знание	узнать_новое	знание_умение
получение_новый	познание	повышение_уровень

Таблица: Названия конкретных тем

topic_3	Получить высшее образование
topic_4	Стать грамотным специалистом в своей сфере
topic_5	Развить профессиональные управленческие качества
topic_6	Изучить иностранный язык
topic_7	Защитить докторскую диссертацию
topic_8	Получить ученую степень

Рис.: Специфичность всех документов



Сравнение качества всех моделей

- MultNB — мультиномиальный наивный байесовский классификатор
- LogReg — логистическая регрессия
- ARTM_class — простая APTM для классификации
- Main_model — предлагаемая APTM

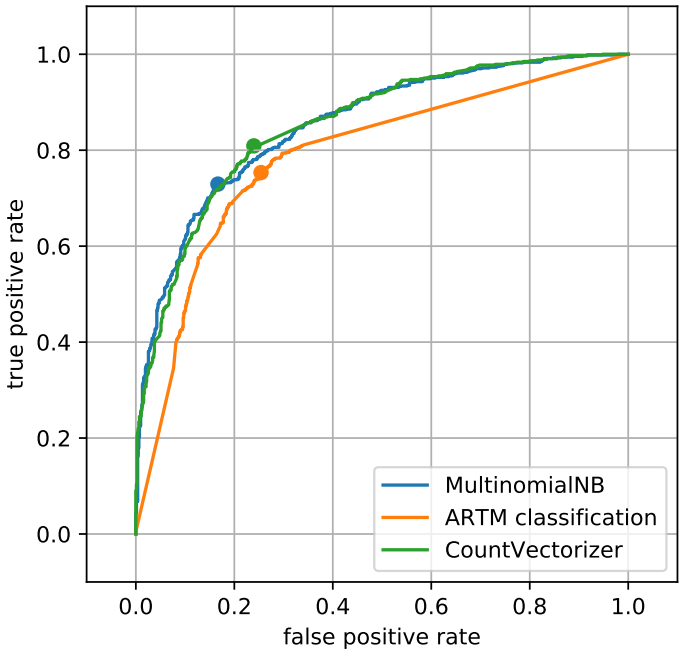
критерий	Main_model	MultNB	LogReg	ARTM_class
roc_auc	**0.871657**	0.843594	0.843779	0.787452
precision	0.752055	0.788851	**0.792321**	0.705634
recall	**0.825564**	0.702256	0.682707	0.753383
accuracy	**0.800269**	0.782784	0.778077	0.749159

Основные результаты, выносимые на защиту

1. Вероятностная тематическая модель мультимодальных анкетных данных с частичным обучением для оценивания тематики и конкретности образовательных целей пользователей в системе дистанционного образования.
2. Способ обучения тематической модели, предполагающий явное разбиение коллекции документов, множества тем и словаря слов на специфичные и неконкретные.
3. Способ графической визуализации тематической модели в виде круговой диаграммы «темы–специфичность».
4. Результаты экспериментов, показывающие, что предложенная модель классифицирует анкетные данные точнее, чем стандартные модели бинарной классификации.

Спасибо за внимание!

Базовое решение



	MultNB	LogReg	ARTM_classification
roc_auc	**0.843594**	**0.843779**	0.787452
precision	0.788851	0.792321	0.705634
recall	0.702256	0.682707	0.753383
accuracy	0.782784	0.778077	0.749159

Реализация модели специфичности документов

Подбор параметров модели

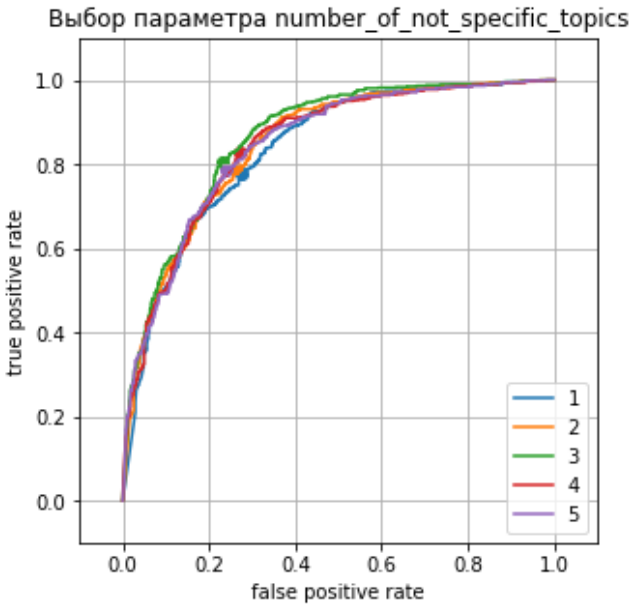


Рис.: Зависимость качества модели от количества неспецифичных тем

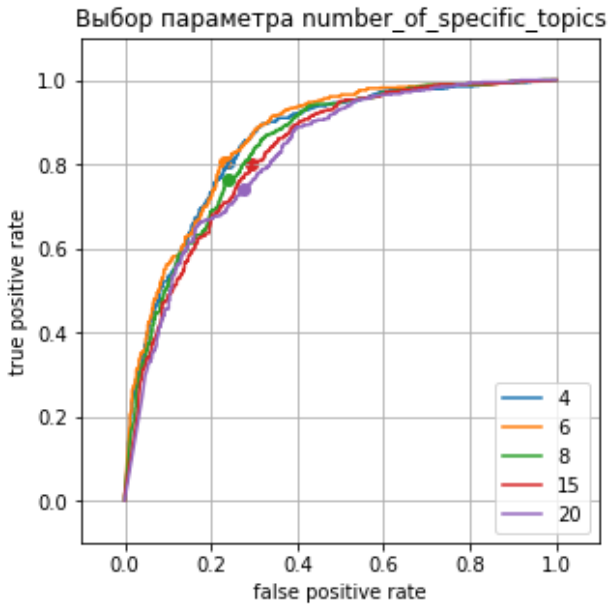


Рис.: Зависимость качества модели от количества специфичных тем