Классификация фильмов на жанры по текстовому описанию, используя методы тематического моделирования

Шокоров Вячеслав Александрович

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

Отчет по заданию курса "Вероятностные тематические модели"

Москва,
2022 г.

Тематическая модель

Тематическая модель — это представление наблюдаемого условного распределения $p(w|d)$ терминов (слов или словосочетаний) w в документах d коллекции D. При этом принимается *гипотеза условной независимости* — вероятность появления слова w, относящегося к теме t в документе d не зависит от документа $p(w|d,t) = p(w|t)$. Следовательно вероятность слова в заданном документе моделируется как:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

где T — множество тем. $\varphi_{wt} = p(w|t)$ — неизвестное распределение терминов в теме t и $\theta_{td} = p(t|d)$ —– неизвестное распределение тем в документе d являются параметрами модели.

Оптимизируемый функционал

Оптимальные значения матриц φ и θ являются точками максимума логарифма правдоподобия коллекции:

$$\log L(\varphi, \theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \varphi_{wt}\theta_{td} \to \max$$

Итоговый классификатор

Оптимизируемый функционал

Максимизация логарифма правдоподобия с регуляризатором:

$$\log L(\varphi, \theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\varphi, \theta) \to \max;$$

$$R(\varphi, \theta) = \sum_{i} \tau_i R_i(\varphi, \theta)$$

Итоговый классификатор

На полученной матрице документов θ обучается классификатор SVM (support vector machine), задача которого предсказывать жанр фильма.

Базовое решение

TF-IDF

Для построения векторного пространства, которое будет в дальнейшем использовано для классификации текстов, воспользуемся метрикой TF-IDF:

$$\mathrm{TF}-\mathrm{IDF}(t,d) = \frac{n_i}{n} \log \frac{|D|}{|\{d : t_i \in d, d \in D\}|}.$$

Итоговый классификатор

Также используется SVM, который обучается на матрице TF-IDF.

Описание данных

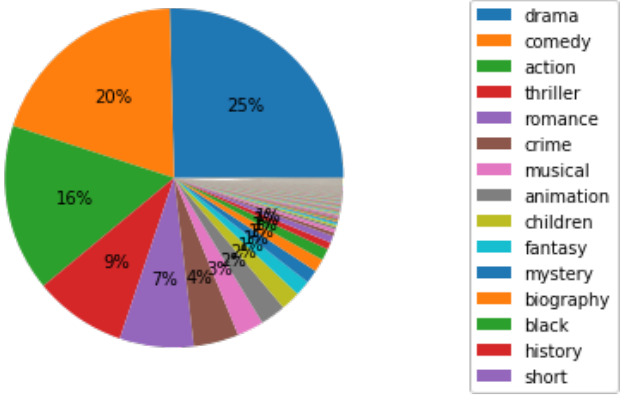
Датасет

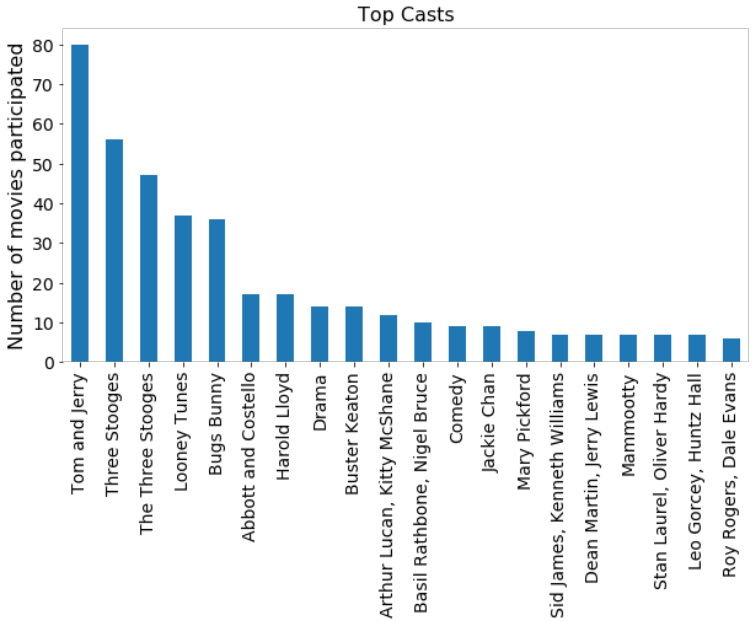
Датасет состоит из текстовых описаний сюжета взятых из Википедии. Набор данных содержит описания 34886 фильмов со всего мира. Также даны год, название, происхождение(страна), режисер, главные актеры, жанр и сюжет.

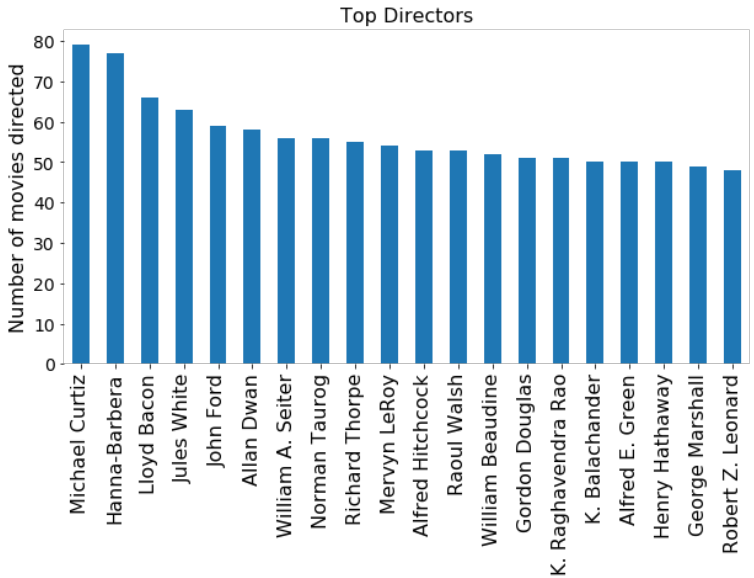
Предобработка

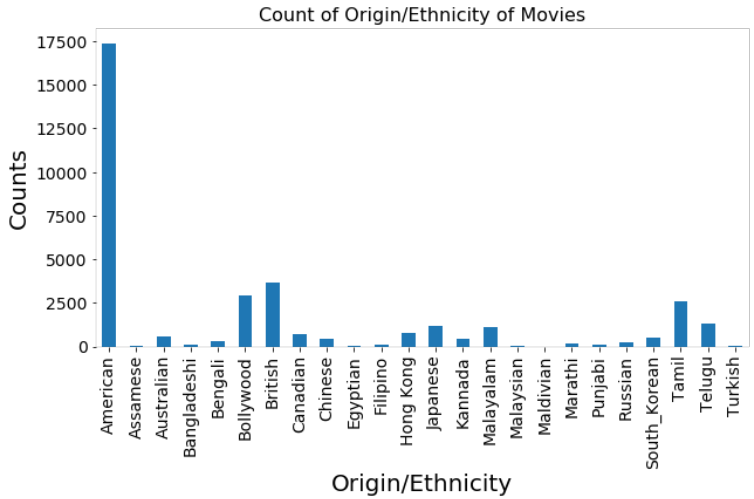
- Чистка жанров. Было много объединенных, специфичных, редких.
- Чистка пустых, малоописанных фильмов.
- Предобработка текста: лемантизация, приведение к нижнему регистру, удаление редких слов.

Описание данных



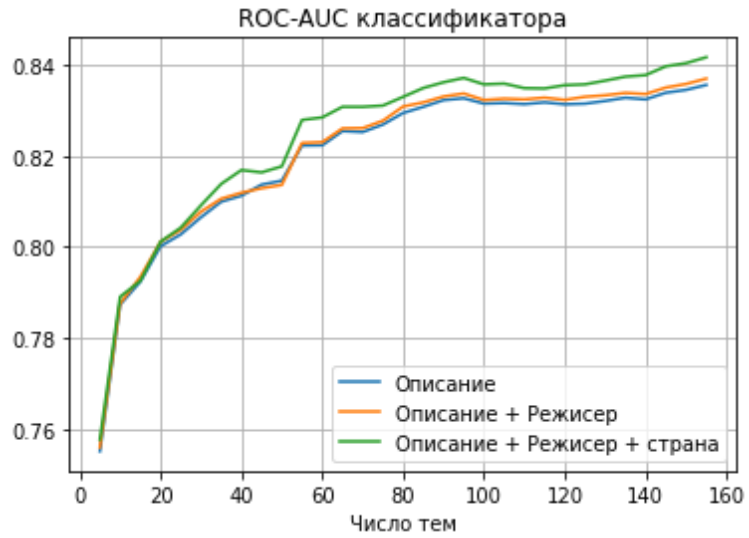


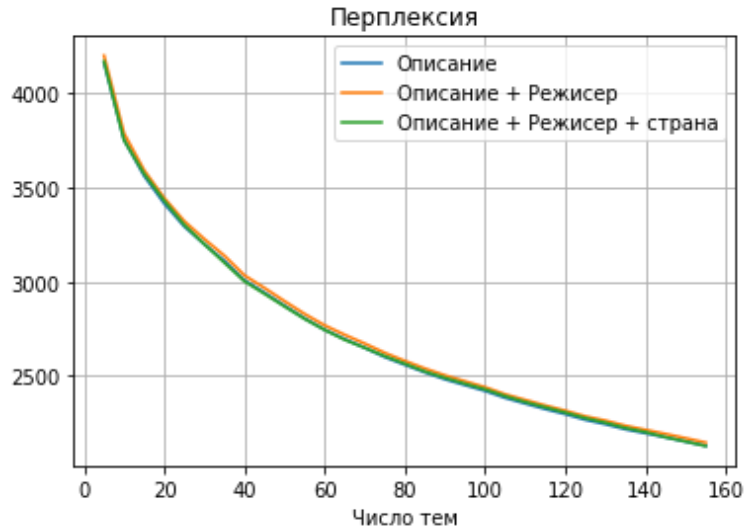




Вычислительный эксперимент

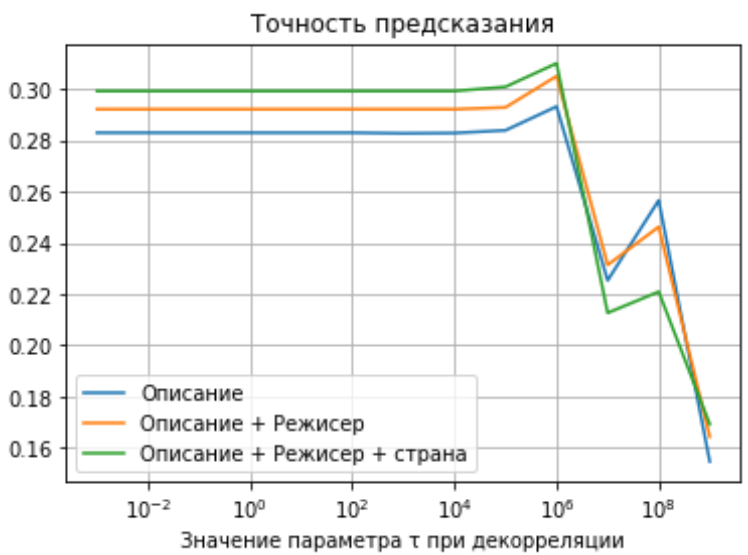
Зависимость качества классификатора от числа тем в тематической модели.

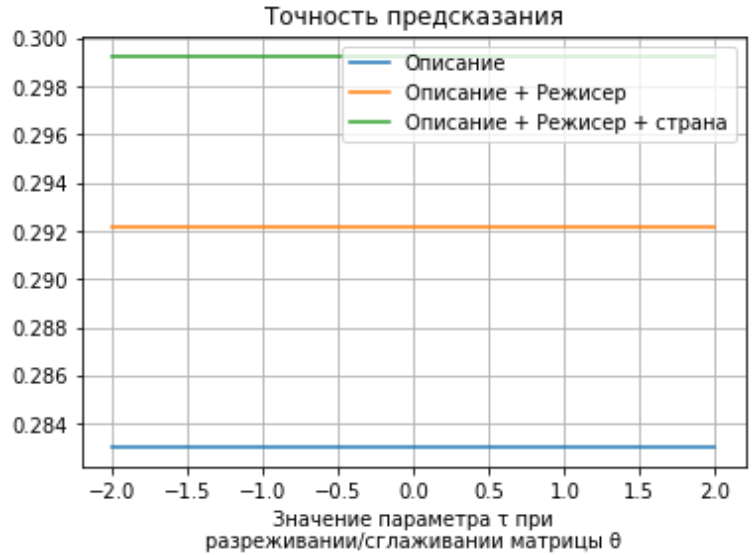


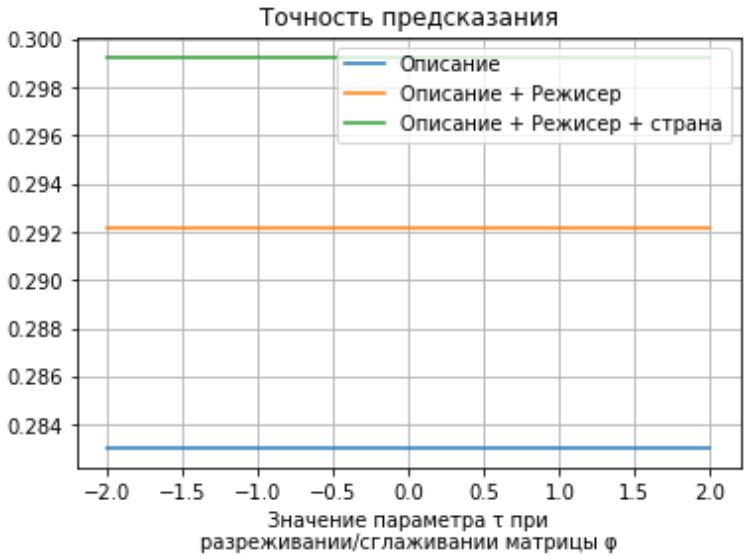


Вычислительный эксперимент

Зависимость качества классификатора от различных регуляризаторов в тематической модели.







Сравнение моделей

Модель	Точность	ROC-AUC
TF-IDF + SVM	51.2	87.8
BigARTM (без регуляризаторов)+ SVM	33.4	81.9
BigARTM (с регуляризаторами)+ SVM	45.2	85.7

Выводы

- С ростом тем в тематической модели качество классификации текстов растет
- Регуляризаторы сглаживания/разреживания матриц не дают прироста
- Тематическая модель плохо подходит для задачи классификации текстов
- Тематическая модель способна выделять отдельные жанры фильмов. Например, была найдена тема, которая полностью принадлежит в анимационному жанру и имеет ключевые слова: tom, back, jerry, get, dog, nick, go, see.