Постановка задачи. Актуальность.

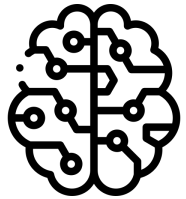
- Постановка задачи
  - Исследовать применение тематических моделей к транскрибациям телефонных звонков. Проанализировать интерпретируемость полученных тематик

- Актуальность
  - Поставлена в рамках работы в крупном банке
  - Применимость тематических моделей к данным траскрибаций не изучена

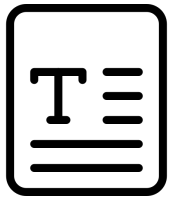
Схема работы решения

Существующий функционал

Колл-центр



Перевод аудиозаписей в текст



База транскрибаций диалогов

Разделение диалогов на тематические группы

Выделение слов, характеризующих группу

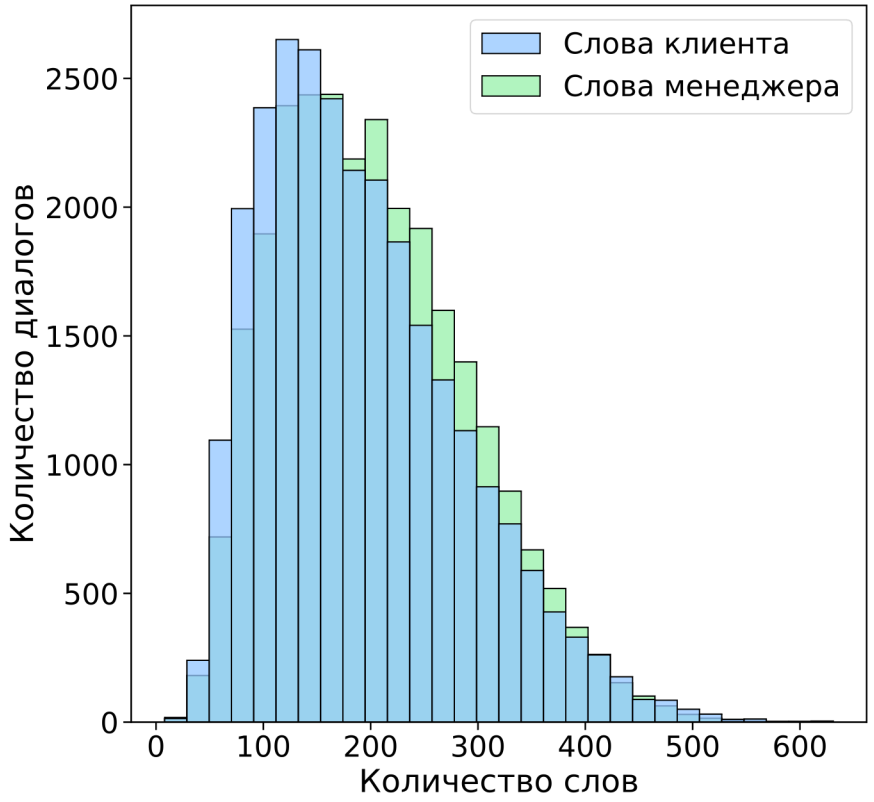
- Анализ тем диалогов
- Оценка размеров тем
- Выявление всплеска звонков по определенной теме
- Возникновение новых тематик

Снижение затрат

Методы решения. Подготовка данных

- Удаление повторов слов или фраз
- Замена несуществующих слов (квазислов) на наиболее похожие по звучанию существующие слова
- Удаление шаблонных фраз, в которых не содержится полезной информации
- Лемматизация (Pymorphy2)
- Выделение устойчивых словосочетаний
- Приведение синонимов к одному слову

Моделирование. Описание набора данных



Распределение длины диалога

- 27281 диалогов по кредитной тематике за один месяц
- Оптимальное количество тем - 6

Сравнение моделей

- «Классические» модели:
  - LDA
  - BigARTM
- Модели автокодировщика:
  - NVDM
  - AVITM
  - CTM
- Модели с использованием семантических векторов документов:
  - BERTopic
  - Top2Vec

BERTopic

- Состоит из нескольких этапов:
  - Векторизация текста (doc2vec, BERT)
  - Снижение размерности (UMAP)
  - Кластеризация HDBSCAN
  - Извлечение тем кластеров

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j},$$

Где m – обще количество слов в теме, t – частота слова, w – общее количество слов

Оценка качества моделей
Когерентность

$$PMI(u,v) = \log \frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)},$$

- где $P(w_i, w_j)$ – совместная вероятность токенов w_i и w_j, $P(w_i)$ и $P(w_j)$ – вероятность отдельно токена w_i и токена w_j.
- Метрика вычисляется по формуле:

$$C = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{PMI}(w_i, w_j)$$

- Также существуют модификации метрики

Оценка качества моделей Разнообразие тем

- Тематическое разнообразие

$$td = \frac{|\{W_1,\dots,W_n\}|}{N \cdot M},$$

где M – количество тем, $|\{W_1, \dots, W_n\}|$ - количество уникальных среди наиболее вероятных слов всех тем

- Inverted rank biased overlap

$$IRBO\left(W_i, W_j, p, d\right) = 1 - \left(\frac{W_i}{N} \cdot p^N + \frac{1-p}{p} \sum_{i=1}^{N} \frac{w_i}{i} \cdot p^i\right),$$

где d – количество слов в одной теме, p – параметр силы влияния порядка слов на метрику.

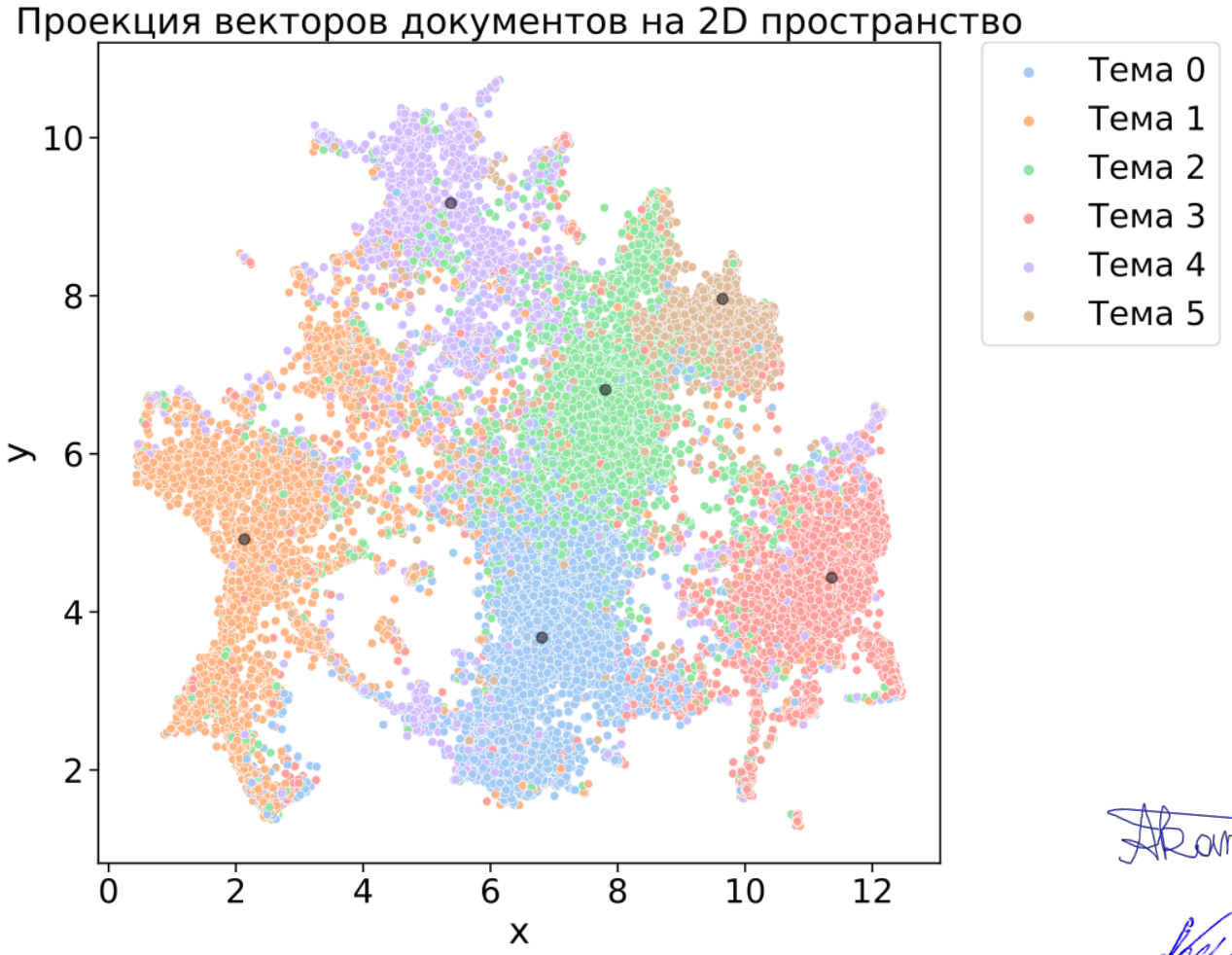
В отличие от тематического разнообразия, учитывает порядок слов в темах.

Оценка качества

Модель	Когерентность	Тематическое разнообразие	Inversed rank biased overlap
Top2Vec	**0,7796**	**1**	**0**
BERTopic doc2vec	0,7401	**1**	**0**
BERTopic BERT	0,7185	**1**	**0**
AVITM с удалением стоп слов	0,6808	0,8667	0,0575
AVITM без удаления стоп слов	0,6358	0,8222	0,0662
NVDM с удалением стоп слов	0,6182	0,9444	0,0123
CTM с удалением стоп слов	0,6173	**1**	**0**
NVDM без удаления стоп слов	0,6165	0,8556	0,029
CTM без удаления стоп слов	0,6121	**1**	**0**
LDA	0,5471	0,8833	0,0458
BigARTM	0,4819	0,8667	0,0369

Визуализация тем BERTopic

- Doc2Vec векторизация документов
- Сжатие размерности вектора с помощью UMAP



Экспертная оценка

- 5 экспертов
- Оценка каждой темы по шкале «Отлично», «Хорошо», «Плохо»
- Темы сгруппированы по моделям, эксперты могут оценить разнообразие тематик
- Нет информации о названии модели

Экспертная оценка. Результаты

Модель	Средняя оценка
BERTopic doc2vec	**1.375**
BERTopic BERT	1.125
NVDM с удалением стоп слов	1.125
Top2Vec doc2vec	1.0
CTM с удалением стоп слов	0.917
CTM без удаления стоп слов	0.917
NVDM без удаления стоп слов	0.917
BigARTM	0.625
AVITM	0.583
LDA	0.417
Случайная	0.167

- Максимальное значение оценки = 2
- Оценка экспертов коррелирует с метрикой когерентности

Оценка времени обучения модели

Модель	Время обучения (в минутах)	Устройство
LDA	**2,8**	CPU
BigARTM	6,58	CPU
NVDM	30,28	CPU
AVITM	33,45	CPU
CTM	10,15	GPU
Bertopic BERT	15,74	GPU
Bertopic doc2vec	18,46	CPU
Top2Vec	**7,18**	GPU

- Модели с использованием семантических векторов требуют GPU
- Все модели применимы на практике

Результаты работы

- Показана применимость тематических моделей к транскрибированным текстам
- Разработан алгоритм подготовки текстов, улучшающий качество тем
- Проведено сравнение различных архитектур
- Подтверждена связь между когерентностью и оценками экспертов
- На основе результатов разработан и внедрен WEB-сервис, позволяющий анализировать диалоги колл-центра

Подготовленные публикации

- «Применение тематического моделирования для анализа транскрибаций телефонных разговоров» XLVII Международная молодёжная научная конференция «Гагаринские чтения - 2021