Анализ веб-документов

Команда eshkere
Котлов Артём
Райков Михаил
Замотаев Родион

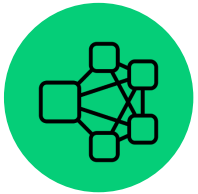
Обзор проекта

Задача

В каждой группе веб-документов выделить подмножество **документов одной темы** и найти **аномалии**.

Web-страница

Представлена в виде dat файла



Группа

Веб-страницы, объединенные некоторой общей тематикой

Аномалия

Веб-страница, выбивающаяся из общей тематики группы

Обзор данных

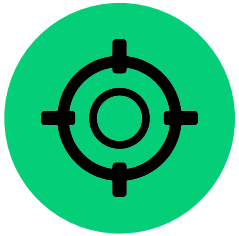
28026
Web-страниц

Train

129
Групп

Test

180
Групп



Target
1 – документ
соответствует теме группы
0 - аномалия

Объект исследования

Web-страница ⇄ Группа

Обзор данных

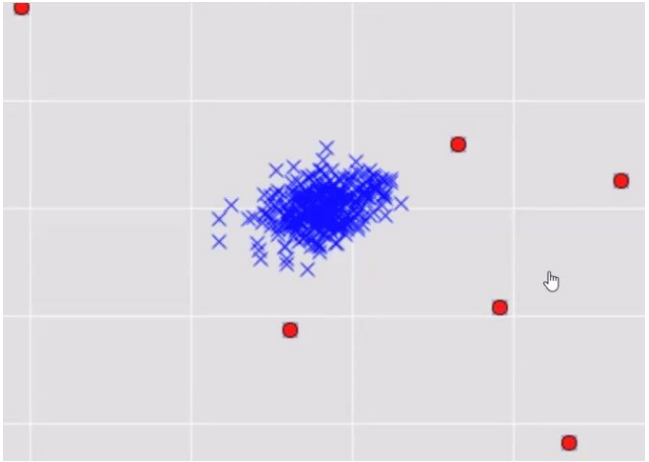
	A	B	C	D
1	pair_id	group_id	doc_id	target
2	1	1	15731	0
3	2	1	14829	0
4	3	1	15764	0
5	4	1	17669	0
6	5	1	14852	0
7	6	1	15458	0
8	7	1	14899	0
9	8	1	16879	0
10	9	1	16310	0
11	10	1	15440	0
12	11	1	16242	0

Train таблица принадлежности документа к группе и его таргет

Doc_id – идентификатор web-страницы

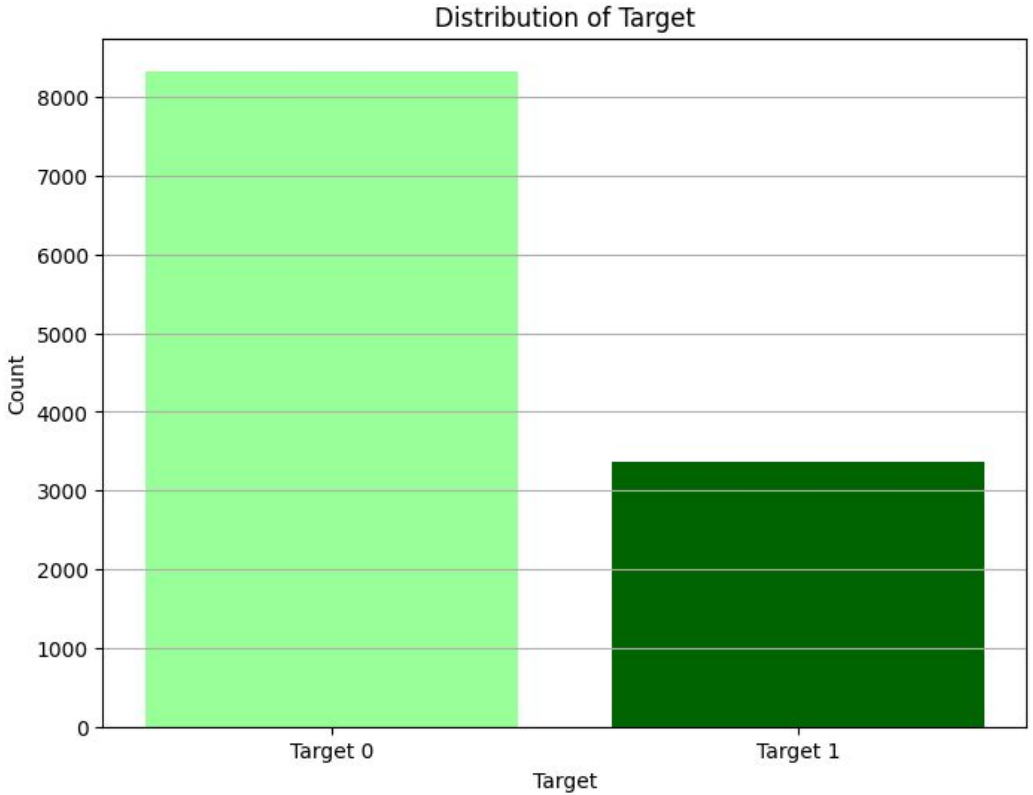
Group_id – идентификатор группы

Pair_id – идентификатор пары (web-страница + группа)



Пример визуализации группы

Распределение классов



Обработка данных
Парсинг html

Dat файлы содежат url веб-страницы + html
Вход: dat-файл веб-страницы
Выход: Основная информация веб-страницы

Задача: Какая информация будет потенциально полезной для определения связи между веб-страницей и группой?

	doc_id	title
0	15731	ВАЗ 21213	Замена подшипников ступицы	Нива
1	14829	Ваз 2107 оптом в Сочи. Сравнить цены, купить п...
2	15764	Купить ступица Лада калина2. Трансмиссия - пер...
3	17669	Классика 21010 - 21074
4	14852	Ступица Нива — замена подшипника своими руками

Парсинг заголовков

	pair_id	group_id	doc_id	target	text
0	1	1	15731	0	Замена подшипников ступицы, руководство по рем...
1	2	1	14829	0	Ваз 2107 оптом. Продажа, поиск, поставщики и м...
2	3	1	15764	0	Продажа запчастей ступица для легковых и грузо...
3	4	1	17669	0	Классика 21010 - 21074 Поставки по низким цена...
4	5	1	14852	0	Передняя ступица Нива имеет свои особенности в...

Парсинг основного текста страницы

Обработка данных

Токенизация и лемматизация

Вход: сырые тексты
Выход: обработанные токены
Задачи:

1. Обработка пропусков
2. Приведение текста к нижнему регистру
3. Удаление спецсимволов
4. Удаление шума и пунктуации
5. Токенизация nltk.tokenize
6. Лемматизация pymorphy

	doc_id	title_processed
0	15731	[ваз, замена, подшипник, ступица, нива]
1	14829	[ваз, оптом, сочи, сравнить, цена, купить, пот...
2	15764	[купить, ступица, лада, калина2, трансмиссия, ...
3	17669	[классика]
4	14852	[ступица, нива, замена, подшипник, свой, рука]

Токенизация заголовков

	doc_id	text_processed
0	15731	[замена, подшипник, ступица, руководство, ремо...
1	14829	[ваз, оптом, продажа, поиск, поставщик, магази...
2	15764	[продажа, запчасть, ступица, легковой, грузово...
3	17669	[классика, поставка, низкий, цена, гарантия, к...
4	14852	[передний, ступица, нива, иметь, свой, особенн...

Токенизация основного текста страницы

Обработка данных
Векторизация текстов

Вход: токены текстов

Выход: векторизованные тексты

Задачи:

1. Выбор векторайзера (Bert, Tf-Idf)
2. Учитывая особенности выбранного векторайзера, векторизовать текст

	doc_id	title	title_embeddings
0	15731	ВАЗ 21213 | Замена подшипников ступицы | Нива	[-0.76475143, -0.15909283, -0.131504, -0.45721...
1	14829	Ваз 2107 оптом в Сочи. Сравнить цены, купить п...	[-0.5048249, 0.0954747, 0.007827005, -0.387454...
2	15764	Купить ступица Лада калина2. Трансмиссия - пер...	[-0.5446935, -0.027800601, -0.17422892, -0.588...
3	17669	Классика 21010 - 21074	[-0.9462604, 0.012580608, -0.1167638, -0.28440...
4	14852	Ступица Нива — замена подшипника своими руками	[-0.47577825, 0.100058936, -0.30862555, -0.178...

Bert эмбеддинги заголовков

	tfidf_0	tfidf_1	tfidf_2	tfidf_3	tfidf_4
11685	0.0	0.0	0.0	0.0	0.000000
11686	0.0	0.0	0.0	0.0	0.000000
11687	0.0	0.0	0.0	0.0	0.707107
11688	0.0	0.0	0.0	0.0	0.000000
11689	0.0	0.0	0.0	0.0	0.000000

Tf-Idf матрица заголовков

Гипотезы

Заголовки документов, соответствующих тематике группы, имеют большое количество высоких значений парных косинусных сходств с заголовками других документов группы, в отличие от аномалий, которые либо имеют низкие значения косинусных сходств, либо имеют несколько больших значений с другими аномалиями группы

Заголовки документов, соответствующих тематике группы, имеют большое количество попарных общих слов с заголовками других документов группы, в отличие от аномалий

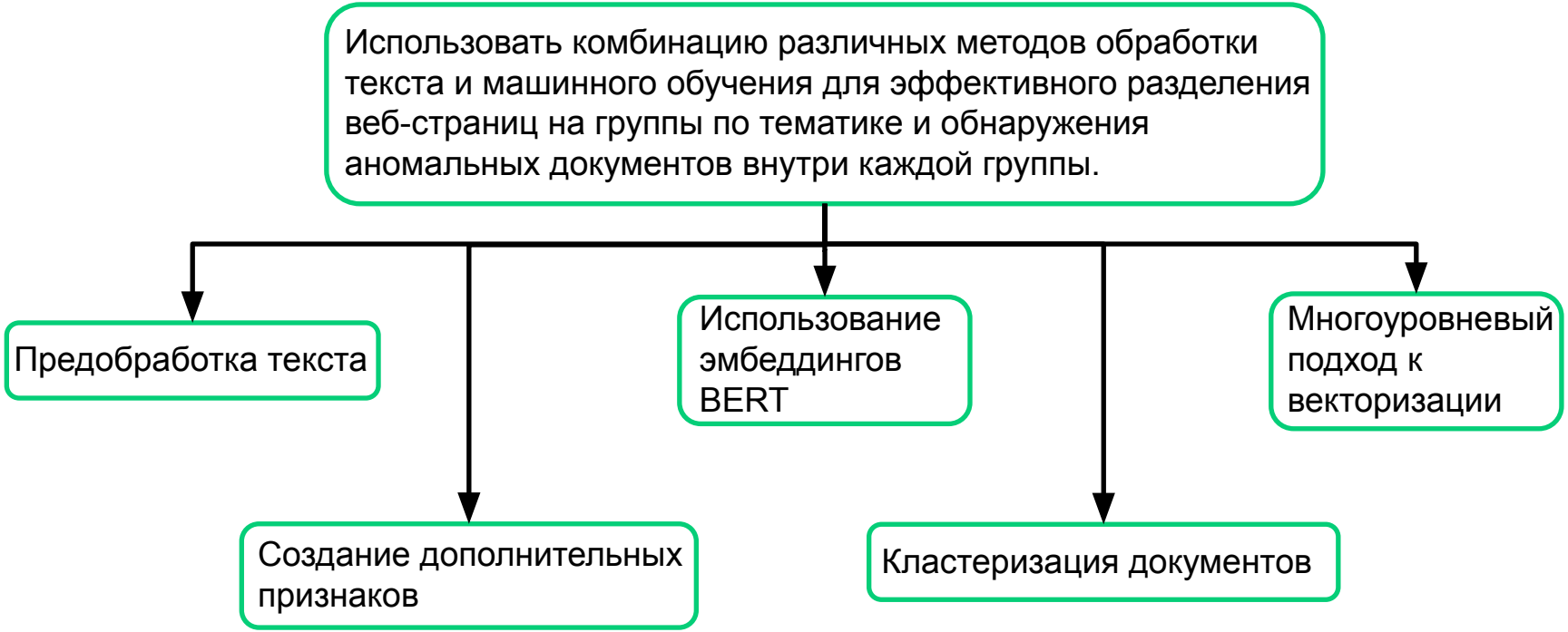
Гипотезы

Кластеризация в каждой группе может отделить аномалии (выбросы) от документов, которые соответствуют тематике группы

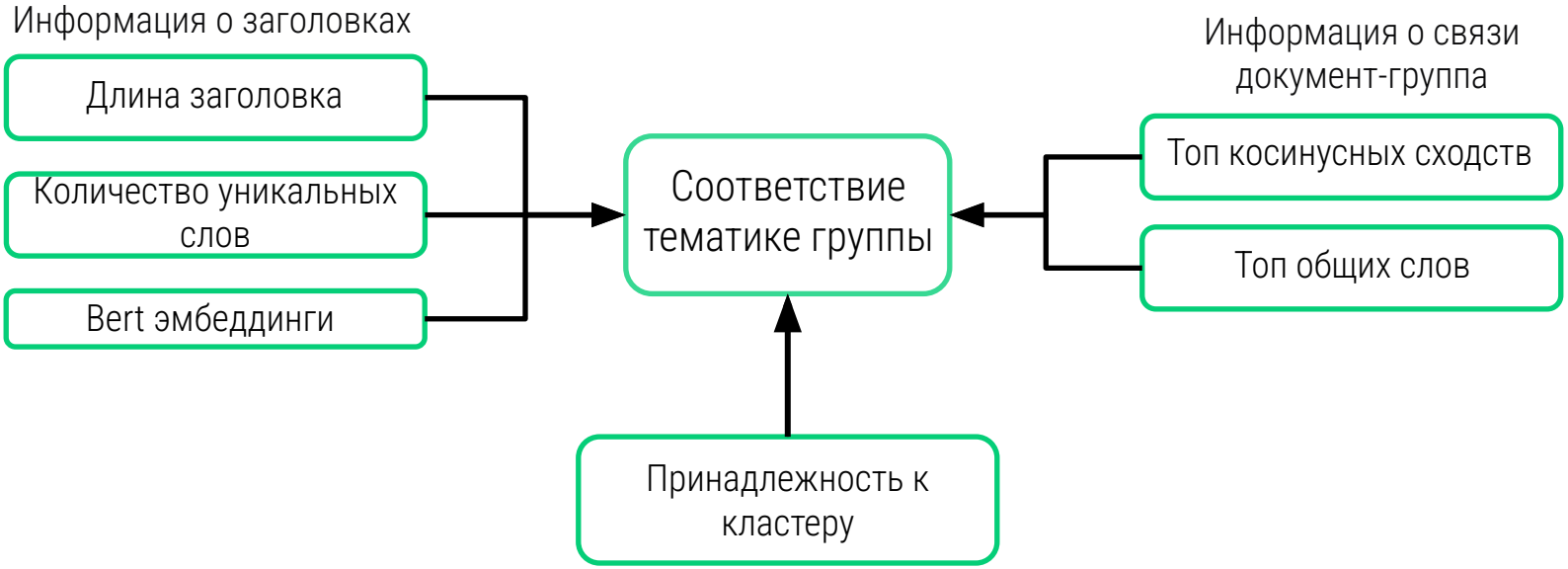
Документы, соответствующие тематике группы, в своём основном тексте имеют топ популярных слов, схожий с групповым топом, в отличие от аномалий

Документы, соответствующие тематике группы, имеют более низкие значения WMD попарных расстояний по заголовкам с другими документами группы, в сравнении с аномалиями

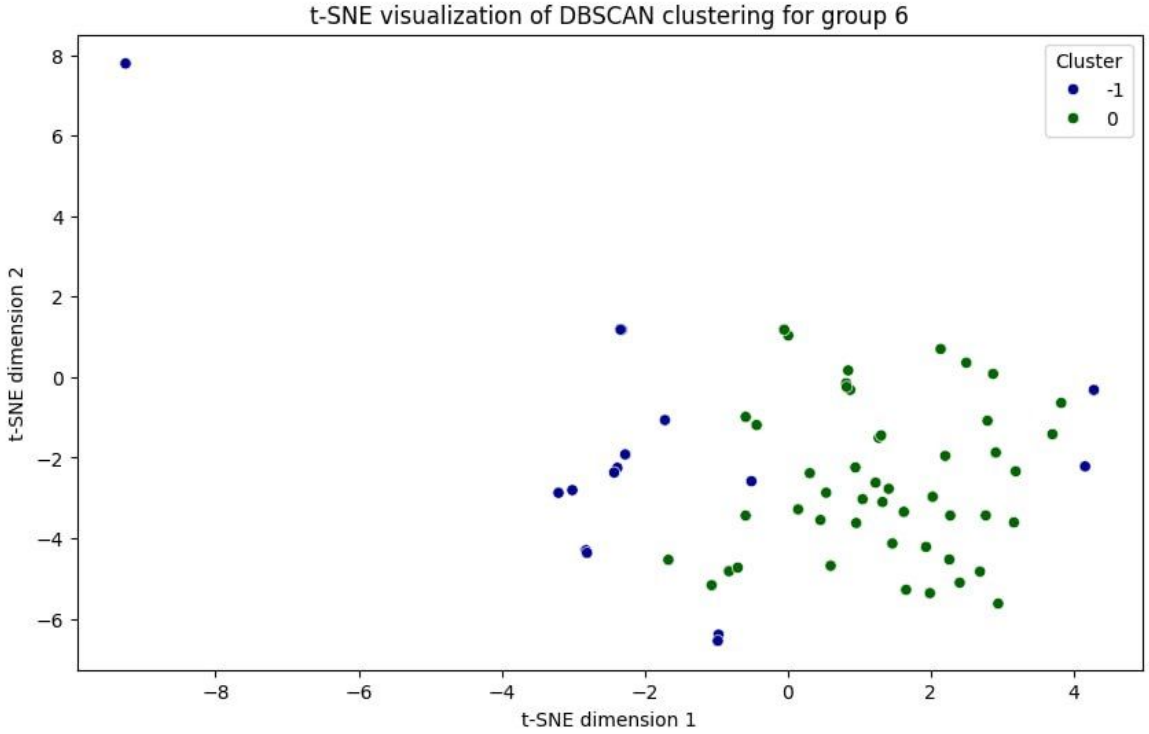
Ключевая идея



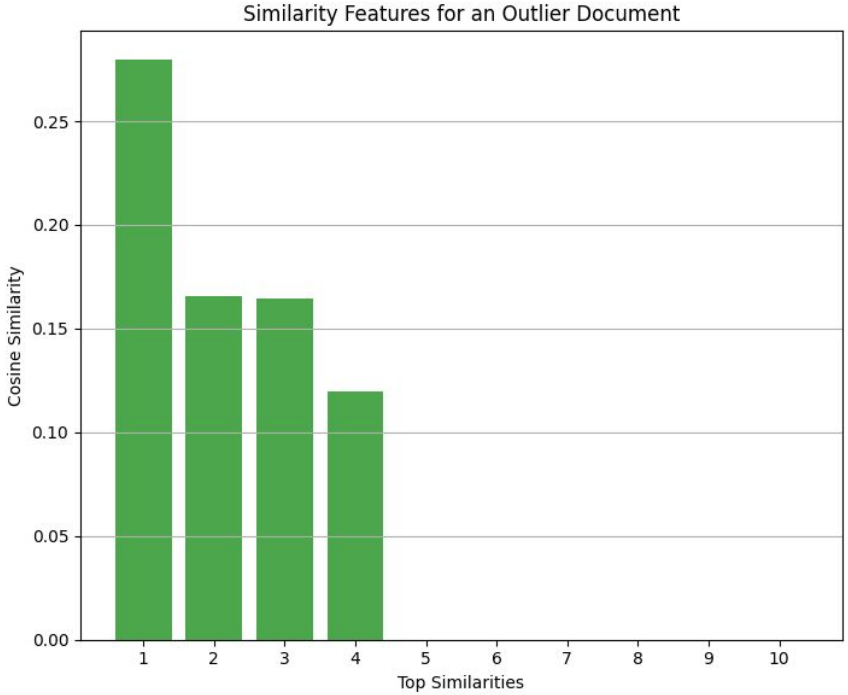
Модель исследования

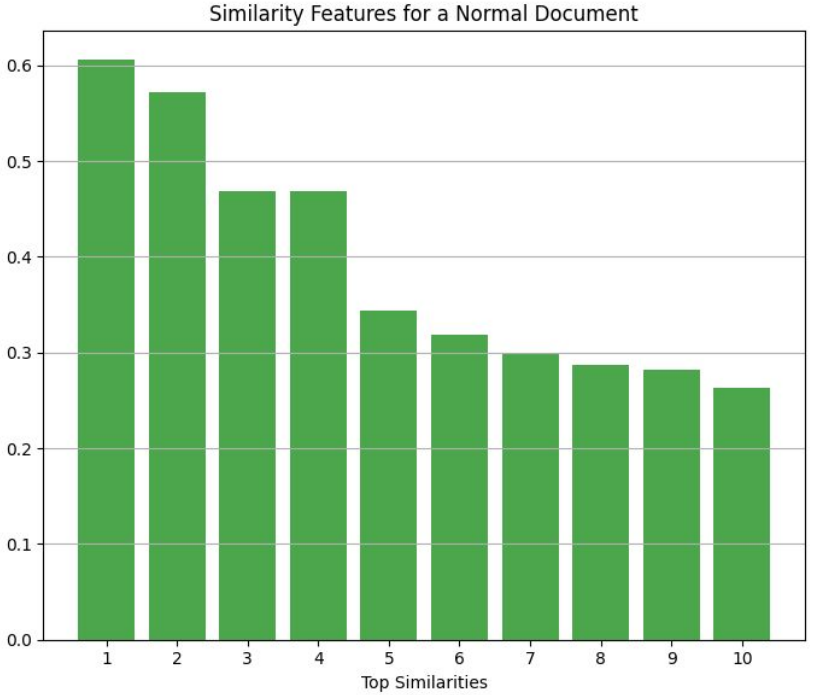


Кластеризация группы

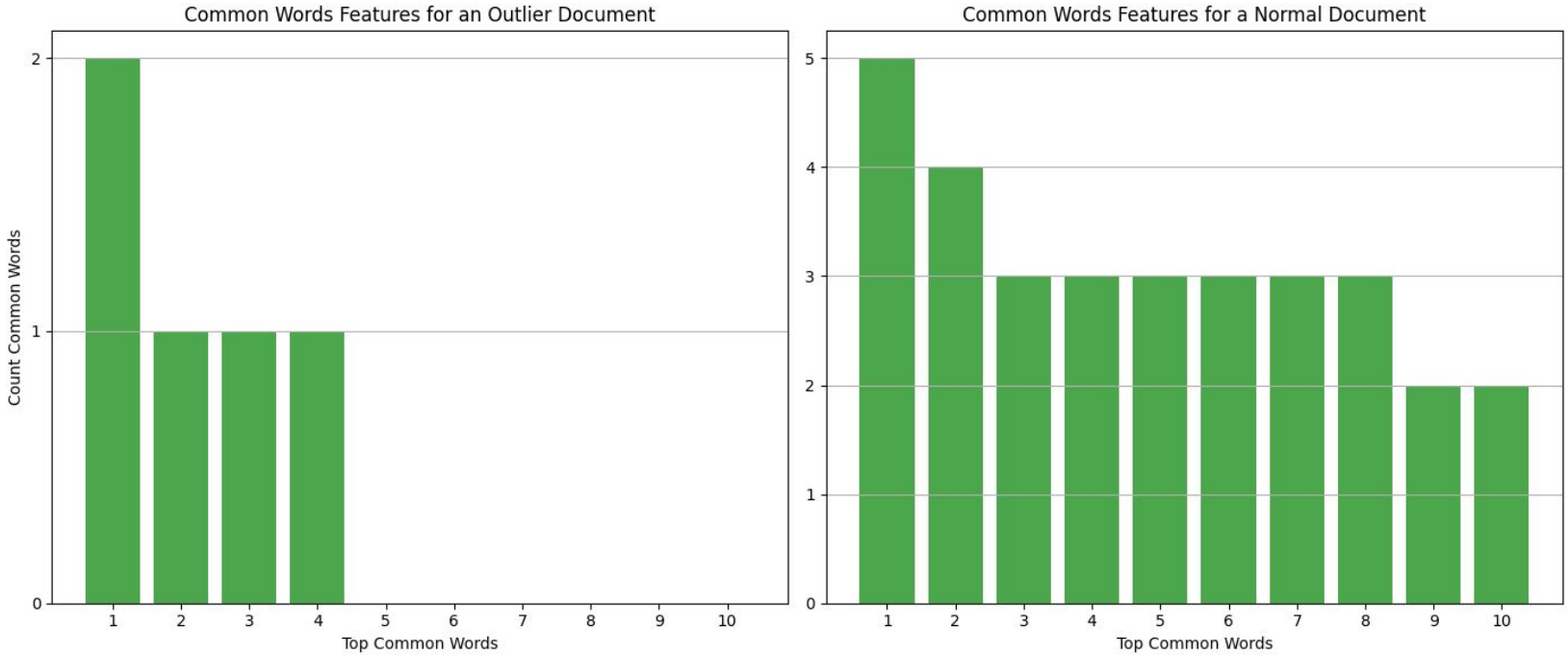


Косинусные сходства

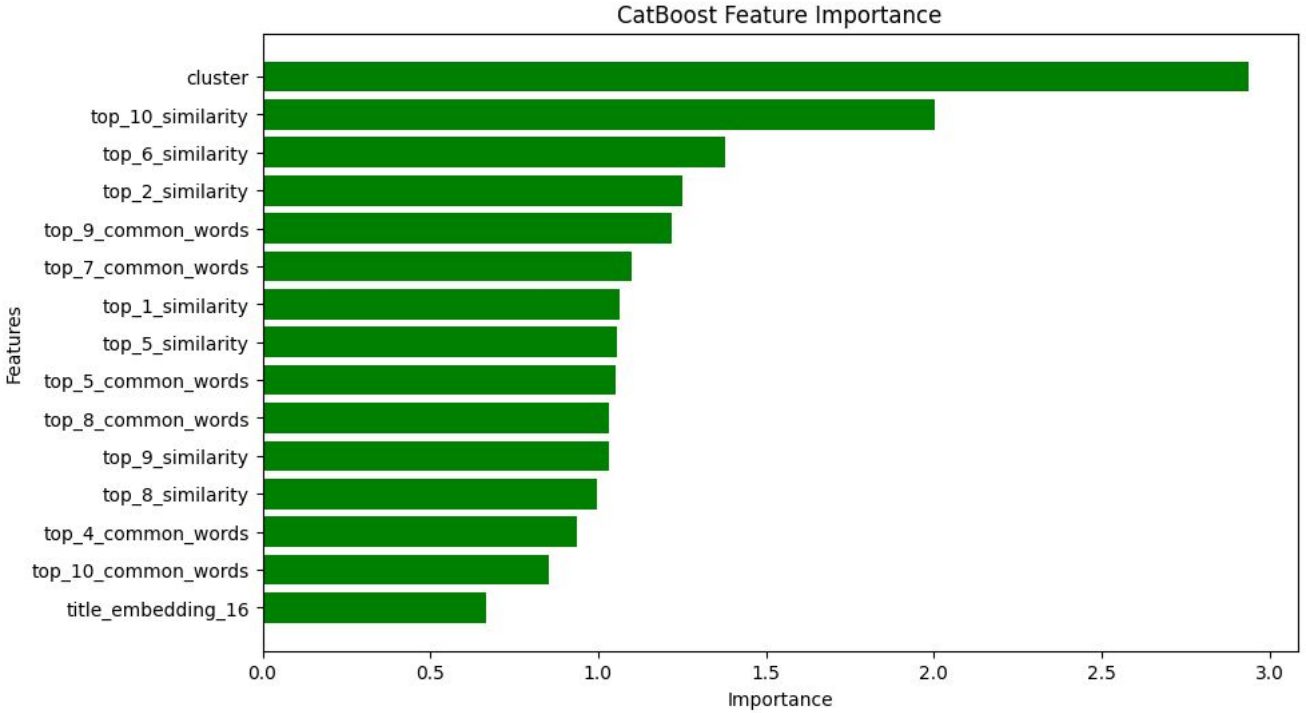




Общие слова


Common Words Features for an Outlier Document
Count Common Words
Top Common Words
0
1
2
1
2
3
4
5
6
7
8
9
10
Common Words Features for a Normal Document
Top Common Words
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10

Важность признаков



Эволюция решения

Модель	Accuracy	Kaggle
NLTK+TFIDF-COS+DBSCAN+CatBoost	0.84	0.70
NLTK+BERT+TFIDF-COS+DBSCAN+CatBoost	0.85	0.717
NLTK+BERT+TFIDF-COS+DBSCAN+WMD+CatBoost	0.85	0.711
NLTK+BERT+BERT-COS+DBSCAN+WMD+CatBoost	0.76	0.69
NLTK+BERT+STAKING	0.66	0.62
LDA+word2vec+LightGBM+Doc2Vec+STAKING	0.70	0.69
NLTK+BERT+TFIDF-COS+DBSCAN+MCW+CatBoost	0.85	0.718