



VK Cloud

Задача классификации новостных текстов с присвоением тегов

НОУ ТСЖС ПЛЕМЯ



бизнес

Команда



Шахмин Павел

Разработчик



Салимова Алина

девушка-вдохновитель



Беликов Даниил

ML-инженер



Злотин Григорий

ML-инженер

Задача

Улучшить выдачу
рекомендательной
системы новостей VK



Существующие решения

1

TF-IDF

2

Bag of
words

3

Word2Vec

1

K-means

2

Hierarchical
clustering
BIRCH

3

Gaussian
Mixture
Models (GMM)

4

OPTICS, DIANA,
Fuzzy analysis
clustering, Affinity
Propagation

Архитектура решения

Извлечение
keywords

YAKE! is a light-weight unsupervised automatic keyword extraction method which rests on text statistical features extracted from single documents to select the most important keywords of a text. Our system does not need to be trained on a particular set of documents, neither it depends on dictionaries, external-corpus, size of the text, language, or domain. And it is free under GPL 3 license.

YAKE
features extracted
text statistical features text
important keywords
light-weight unsupervised automatic
statistical features
automatic keyword extraction
statistical features extracted
keyword extraction method
text statistical single documents
keyword extraction

YAKE!

Определение тега

1

Tf-IDF

2

UMAP

3

HDBSCAN

4

BERT +
HDBSCAN

Health



Science



Television



Travel



Movies



Dance



Real Estate



Economy



Sports



Theater



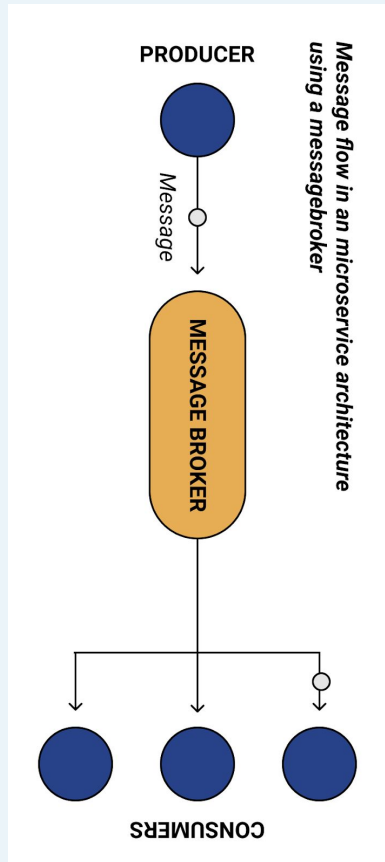
Opinion



Music



Особенности реализации:



Развертывание через Kubernetes cluster

Распределяет нагрузку между несколькими нодами. Автоскейлинг в зависимости от нагрузки



Очередь сообщений RabbitMQ

Накапливает запросы и равномерно распределяет между подами с ML



Асинхронный веб-сервер FastAPI

Принимает запросы пользователей, демонстрирует начальную страницу и документацию к API



Продолжительное развертывание через Github Actions

Автоматически собирает docker-образы и отправляет в кластер на развертывание

Особенности решения:

1

Скорость работы

2

Качество классификации
благодаря последовательному
применению моделей

3

Выделение keywords

4

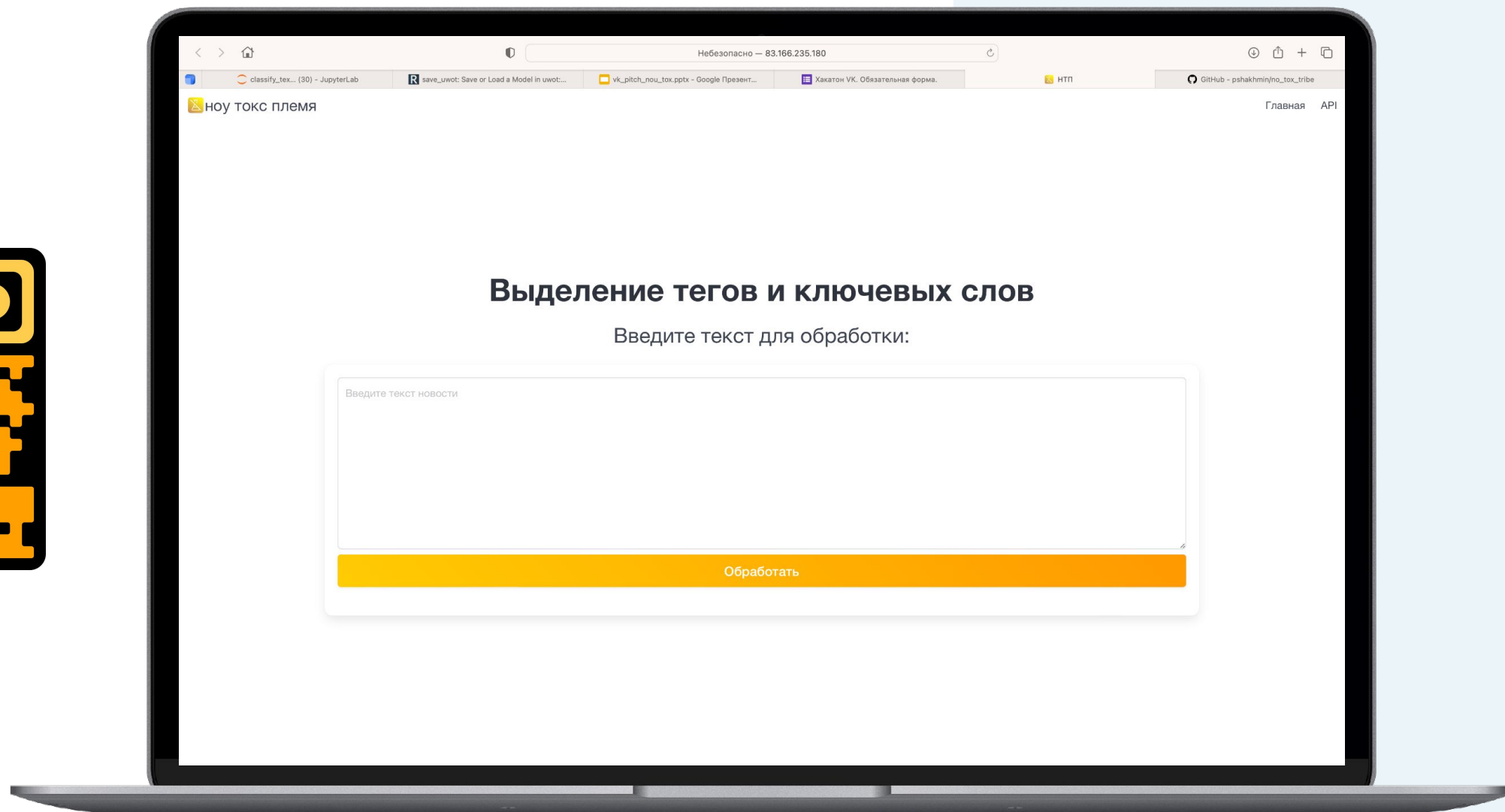
Глубина классификации (число
значимых кластеров)

Полученные метрики

0.72

accuracy

Демонстрация решения



Развитие решения

1

Обогащение
кластеров

2

Создание
дополнительных
уровней
кластеризации

3

Использование
keywords в качестве
фичей



Спасибо за внимание!

