

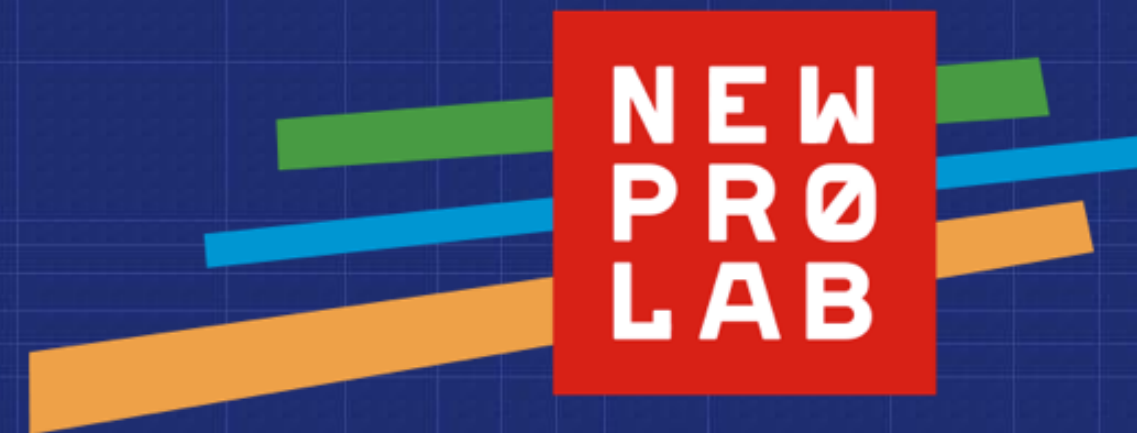


NEW
PRO
LAB

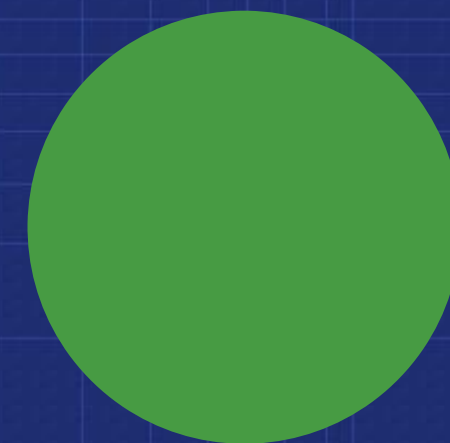
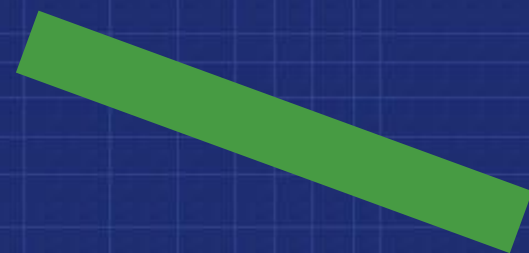
Рекомендательные системы

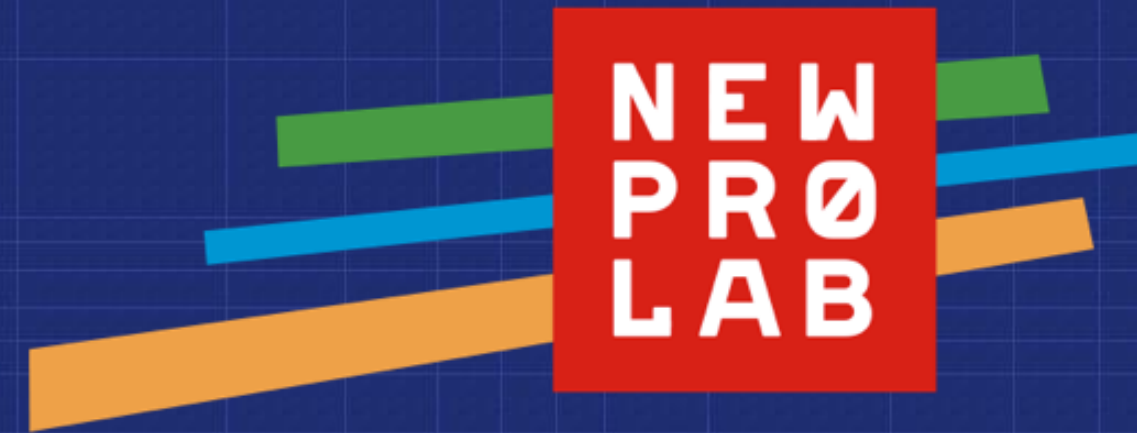
Гриша Сапунов

NEWPROLAB.COM

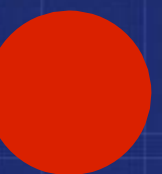
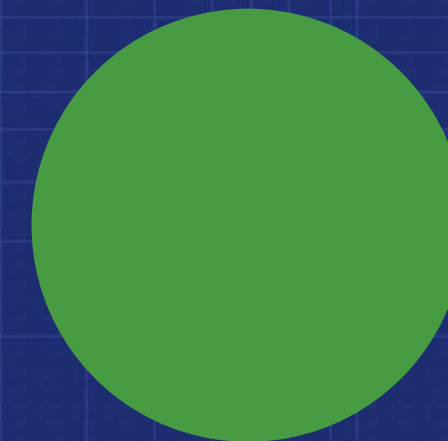
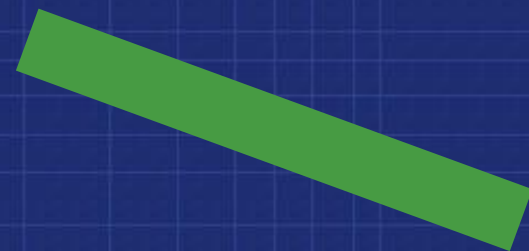


Гриша Сапунов: grigory.sapunov@gmail.com
Дмитрий Игнатов: dmitrii.ignatov@gmail.com





Content-based recommenders – 2

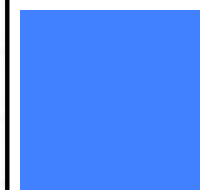




бренд	olivegray
цвет	желтый
фасон	классика
страна	Россия
сезон	демисезон
длина	103 см
вырез горловины	округлый
цена	6 110 □

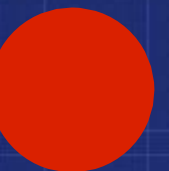
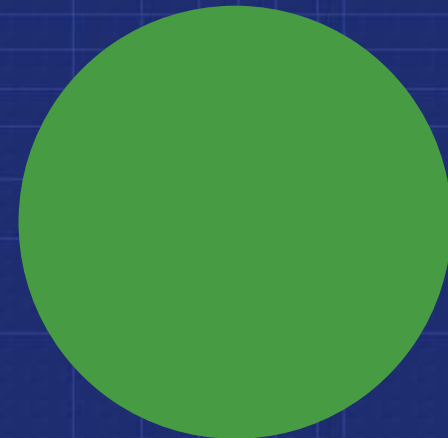
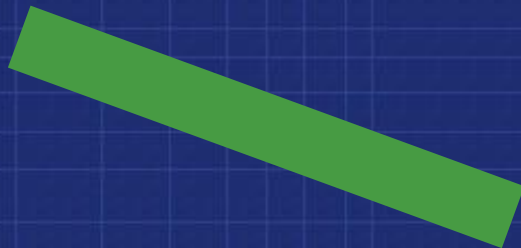
Это платье абсолютно идеальный вариант для любого повода. Сделанное из плотного трикотажа, оно имеет формообразующие рельефы, которые создают поддерживающий эффект и скрывают все ваши недостатки. Юбка-карандаш со шлицей и длинный рукав идеально дополняют это платье, делая его еще более восхитительным.

ПЛАТЬЕ
ИДЕАЛЬНЫЙ
АБСОЛЮТНО
ПОДДЕРЖИВАЮЩИЙ
ЭФФЕКТ
СЛУЖИТ
СДЕЛАННОЕ
ВОСХИТИТЕЛЬНЫЙ
ДЛЯ
ПЛОТНОГО
ДЛИННЫЙ
РУКАВ
ЮБКА-КАРАНДАШ
ПОВОДА
ФОРМООБРАЗУЮЩИЕ
СКРЫВАЮТ
ДОПОЛНЯЮТ
ШЛИЦЕЙ
НЕДОСТАТКИ
ИДЕАЛЬНО
ВАРИАНТ





TF-IDF

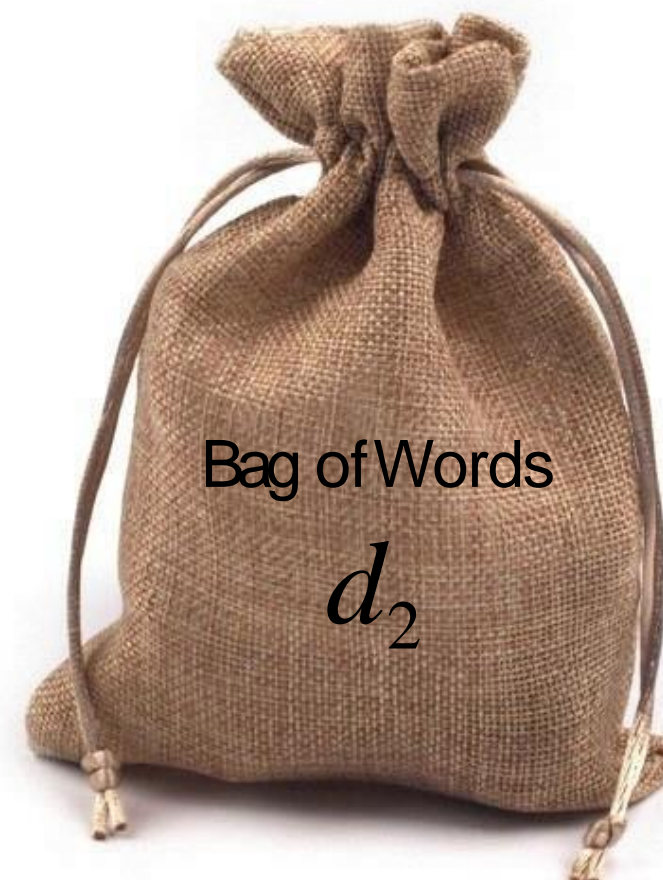
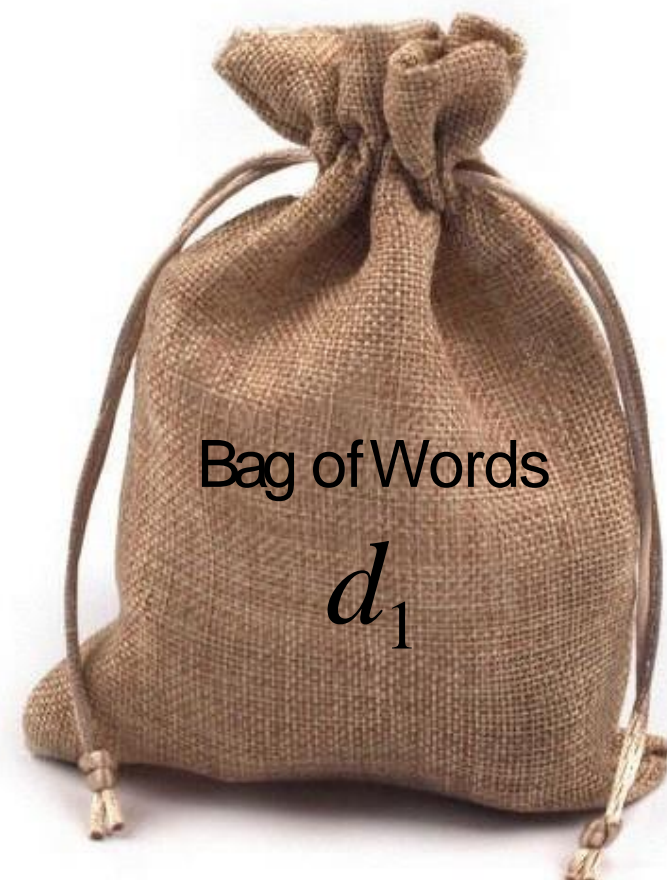




Представление документа



Представление документа



Программы анализа и лингвистической обработки текстов

История изменений

Название	Автор(ы), Организация	Комментарий
Apache OpenNLP	The Apache Software Foundation, Incubator	OpenNLP - это организационный центр "open source" проектов, связанных с машинной обработкой естественного языка под эгидой Apache. OpenNLP предлагает обширный Java-инструментарий обработки текстов на основе методов машинного обучения. Он включает средства токенизации, выделения предложений, разметки частей речи, выделения имен собственных, разбора текста и разрешения перекрестных ссылок. Имеется документация на английском языке. Для скачивания доступен исходный код и бинарные компоненты (для запуска требуется установка Java VM).
Link Grammar Parser	John Lafferty Daniel Sleator Davy Temperley Carnegi Melon University, USA	Link Grammar Parser – это синтаксический парсер английского языка. Работает со словарем, включающем около 60000 словарных форм. Реализован на C для Unix. Есть также версия для Windows API32. Имеет консольный интерфейс. Исходные предложения для разбора могут вводиться вручную с клавиатуры или задаваться в ASCII-файле для пакетной обработки. Программа распространяется бесплатно.
Проекты Cíbola/Oleada	Computing Research Laboratory (CLR) New-Mexico State University, USA	Проекты Cíbola/Oleada реализуют обширные компьютерные системы лингвистического анализа текстов, представленных в Unicode. Компоненты системы включают средства работы с мультязыковыми текстами (MUTT), построения конкорданса (XConcord) для текстов на более чем 16 языках, статистического анализа, автоматического перевода, различные словари и тезаурусы. Некоторые версии этих компонентов доступны для бесплатной загрузки после процедуры формальной регистрации. Все компоненты реализованы в среде X11 Window System для SunOs и Solaris.
Russian Morphological Dictionary	Sergey Sikorsky	Программа для синтаксического и морфологического анализа русскоязычных текстов. Работает с входным ASCII-текстом. Используется морфологический словарь, включающий 120000 слов. Реализована на SWI-Prolog для Windows. Программа распространяется бесплатно.
Mystem	Илья Сегалович, Виталий Титов компания Яндекс	Компактный, очень быстрый и бесплатный морфологический парсер русскоязычных текстов, реализованный на основе словаря Зализняка. Доступны для загрузки версии для Windows и Linux. Работает как консольное приложение и имеет различные режимы представления результатов.
Лингвоанализатор	Д.В.Хмелев	On-line версия программы математического анализа структуры текста. Целью анализа является определение близости любого из предлагаемых пользователем текстов к одному из авторских эталонов, определенных заранее. (Авторский эталон - это набор текстов данного автора, взятый из ресурсов Русской Фантастики). Программа анализирует входной текст и выдает имена трех писателей, которые могли бы быть его наиболее вероятными авторами. Кроме этого, программа находит три произведения каждого из авторов, которые наиболее близки данному тексту.
Программные продукты фирмы LingSoft	LingSoft, Финляндия	Компоненты грамматического разбора, морфологического анализа и лемматизации (нормализации) для английского, немецкого, финского, датского, норвежского, шведского, эстонского и русского языков. Это коммерческие продукты, которые могут быть использованы при разработке других систем.
		СУБД StarLing, позволяющая работать с мультязычными текстами большой длины, с транскрипционными знаками, с удобным поиском, с анализом и синтезом словоформ по словарю Зализняка, с переводом по словарю Мюллера. Есть функции для сравнительно-исторических исследований (глотнохронология). Для загрузки

- редкое слово корпуса ↑
- частое слово документа ↑
- размер документа ↓



Какие слова важнее?

TF-IDF



TF: Term Frequency



частота термина k
в документе j

$$TF_{t,d} = \frac{f_{td}}{\max_z f_{zd}}$$

максимальная частота
по всем терминам z
в документе j

Variants of TF weight

weighting scheme	TF weight
binary	$\{0,1\}$
raw frequency	$f_{t,d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \frac{f_{t,d}}{\max f_{t,d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max f_{t,d}}$

wikipedia.org



IDF: Inverse Document Frequency



$$IDF_t = \log \frac{N}{df_t}$$

число документов
в корпусе

число документов
с термином t

Variants of IDF weight

weighting scheme	IDF weight
unary	1
inverse frequency	$\log \frac{N}{n_t}$
inverse frequency smooth	$\log \left(1 + \frac{N}{n_t} \right)$
inverse frequency max	$\log \left(1 + \frac{\max_t n_t}{n_t} \right)$
probabilistic inverse frequency	$\log \frac{N - n_t}{n_t}$

$$TF_{t,d} * IDF_t = \frac{f_{td}}{\max_z f_{zd}} \log \frac{N}{df_t}$$

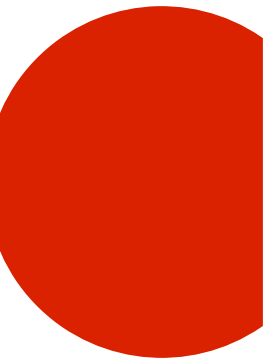


- выбрать n слов с наибольшим весом
- 100-200 обычно достаточно
- резать по threshold

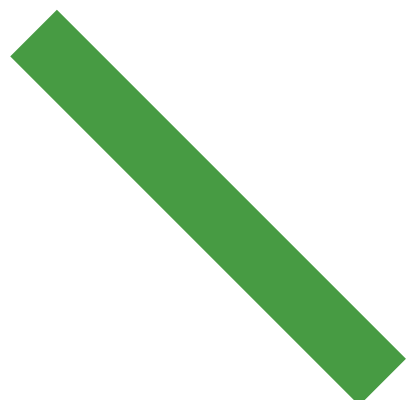
Нормализованные веса

$$w_{t,d} = \frac{TF_{t,d} * IDF_t}{\sqrt{\sum_z (TF_{z,d} * IDF_z)^2}}$$

делим на длину
вектора

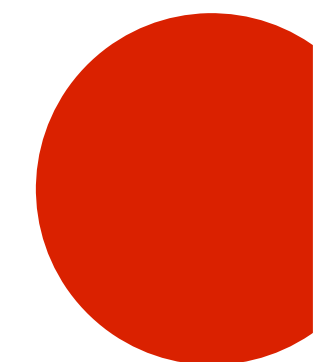


TF-IDF + cos





NEW
PRO
LAB



качество текста



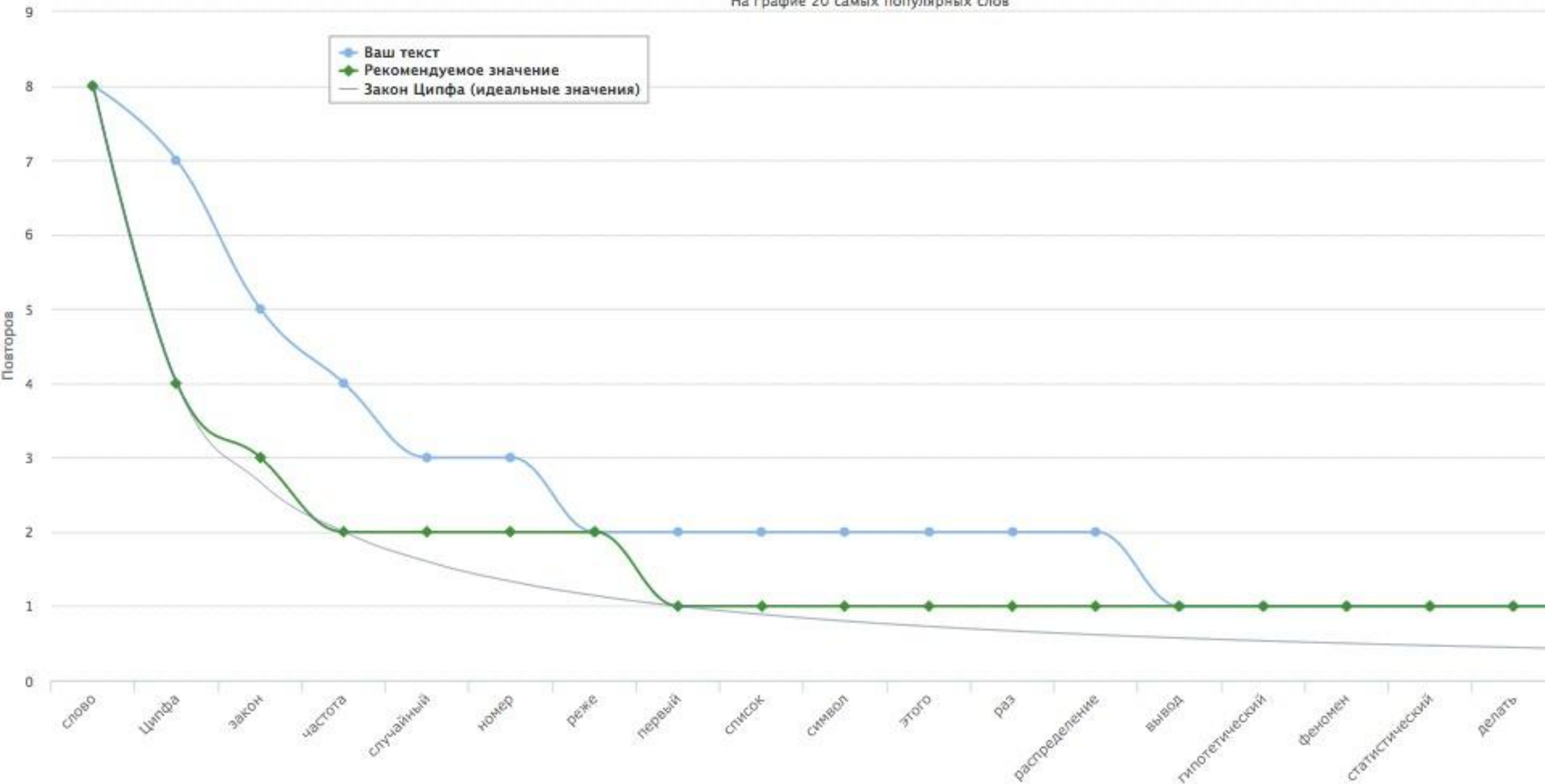
Закон Ципфа

- эмпирическая закономерность распределения слов естественного языка
- частота слова обратно пропорциональна порядковому номеру

1y.ru

Анализ контента по закону Ципфа

На графике 20 самых популярных слов



Оценка качества: 71% (удовлетворительно)

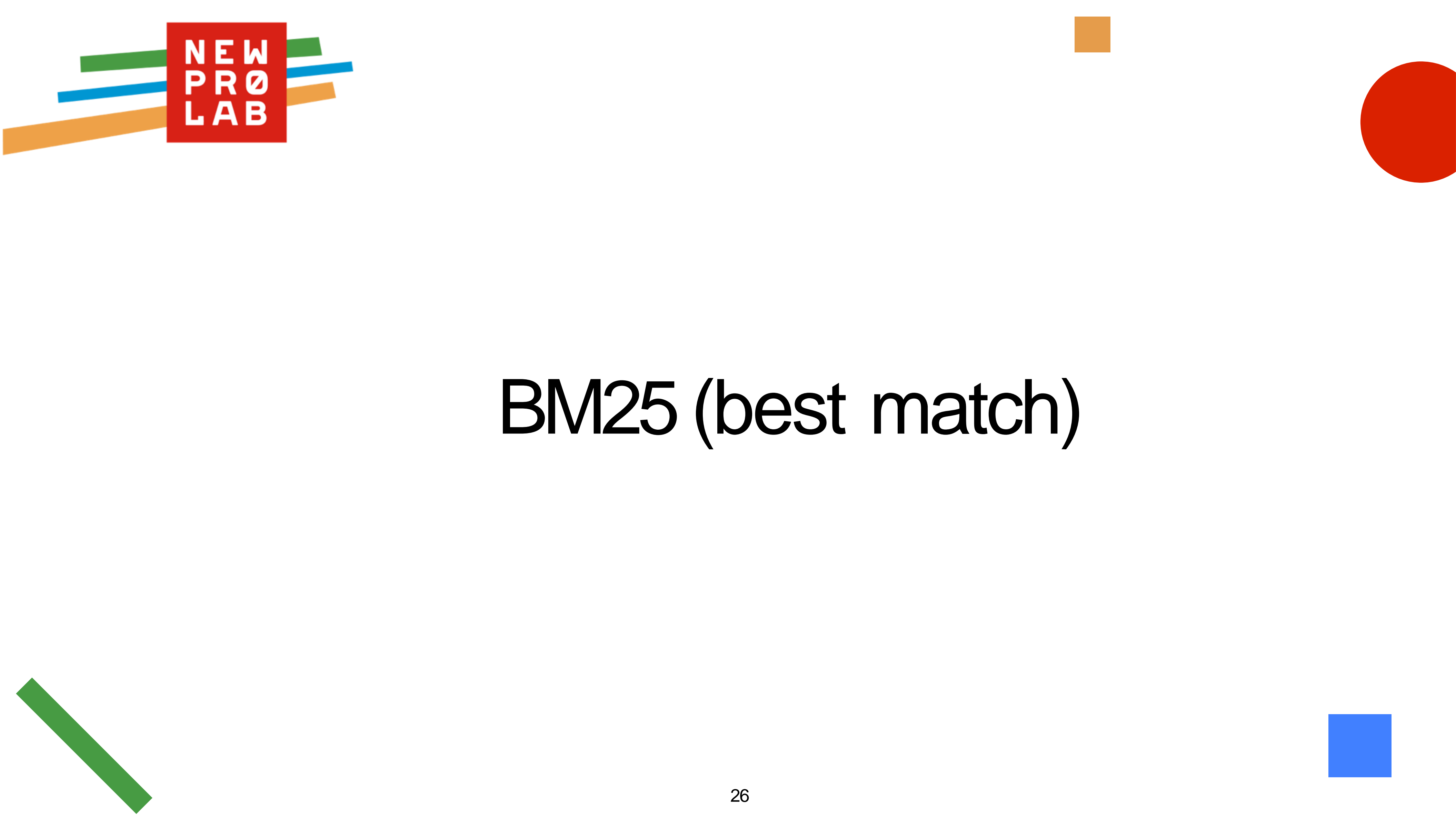
Всего слов на странице: 188

Время выполнения: 0.25 сек

При проверке были игнорированы [стоп-слова](#)

Описание распределения слов доступно на [википедии](#)

Слово	Повторов
<div><div></div> слово</div>	8
Цифра	7
<div><div></div> закон</div>	5
<div><div></div> частота</div>	4
<div><div></div> номер</div>	3
<div><div></div> случайный</div>	3
этого	2
<div><div></div> раз</div>	2
<div><div></div> распределение</div>	2
<div><div></div> символ</div>	2
<div><div></div> список</div>	2



BM25 (best match)

BM25 (best match)

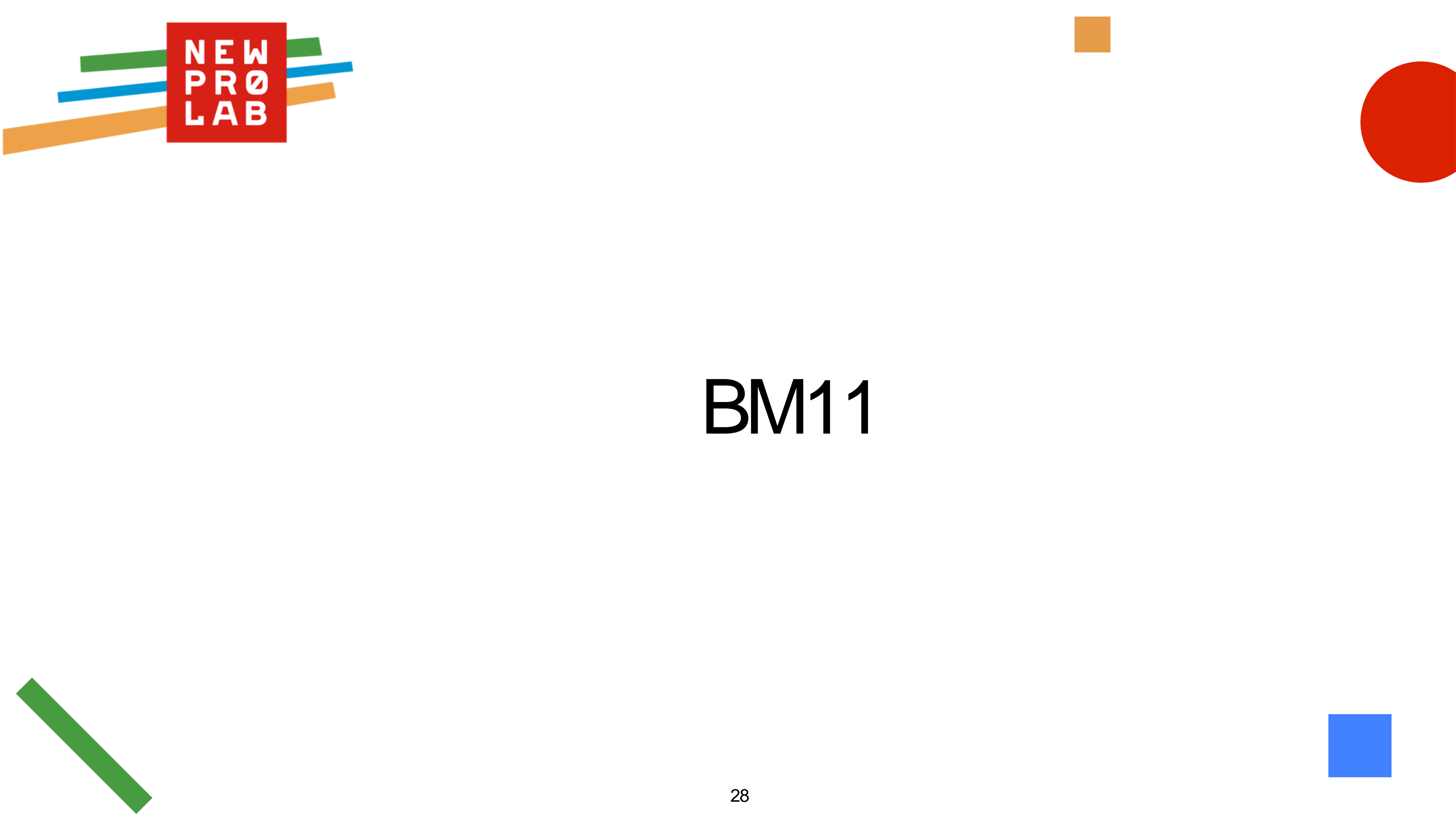
$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1(1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

query

$$k_1 \in [1.2, 2], b = 0.75$$

свободные
параметры

средняя длина
документа в
корпусе

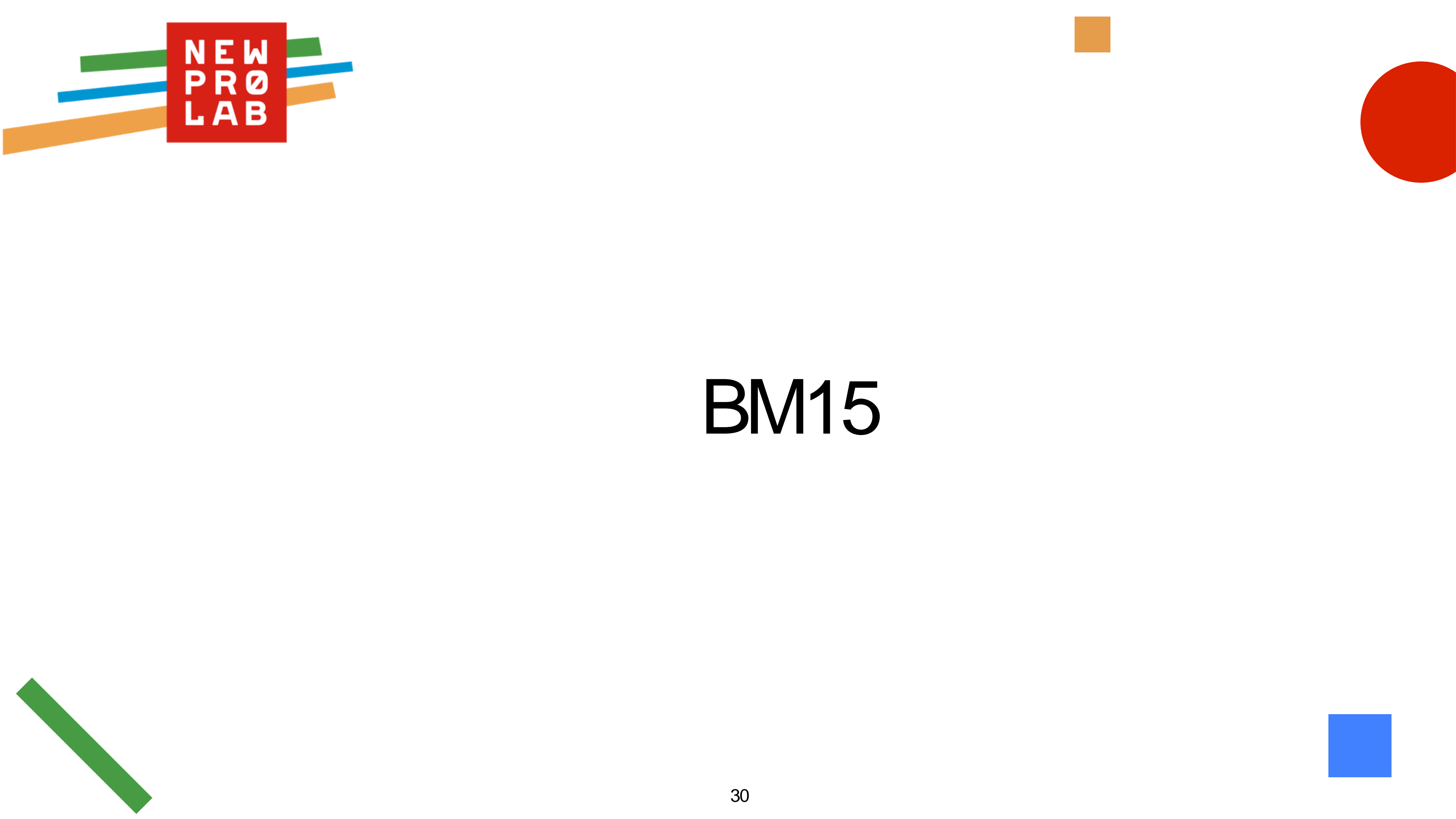


BM11

BM11

$$\textit{score}(D, Q) = \sum_{i=1}^n \textit{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \frac{|D|}{\textit{avgdl}}}$$

$$k_1 \in [1.2, 2], b = 1$$



BM15

BM15

$$\textit{score}(D, Q) = \sum_{i=1}^n \textit{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1}$$

$$k_1 \in [1.2, 2], b = 0$$

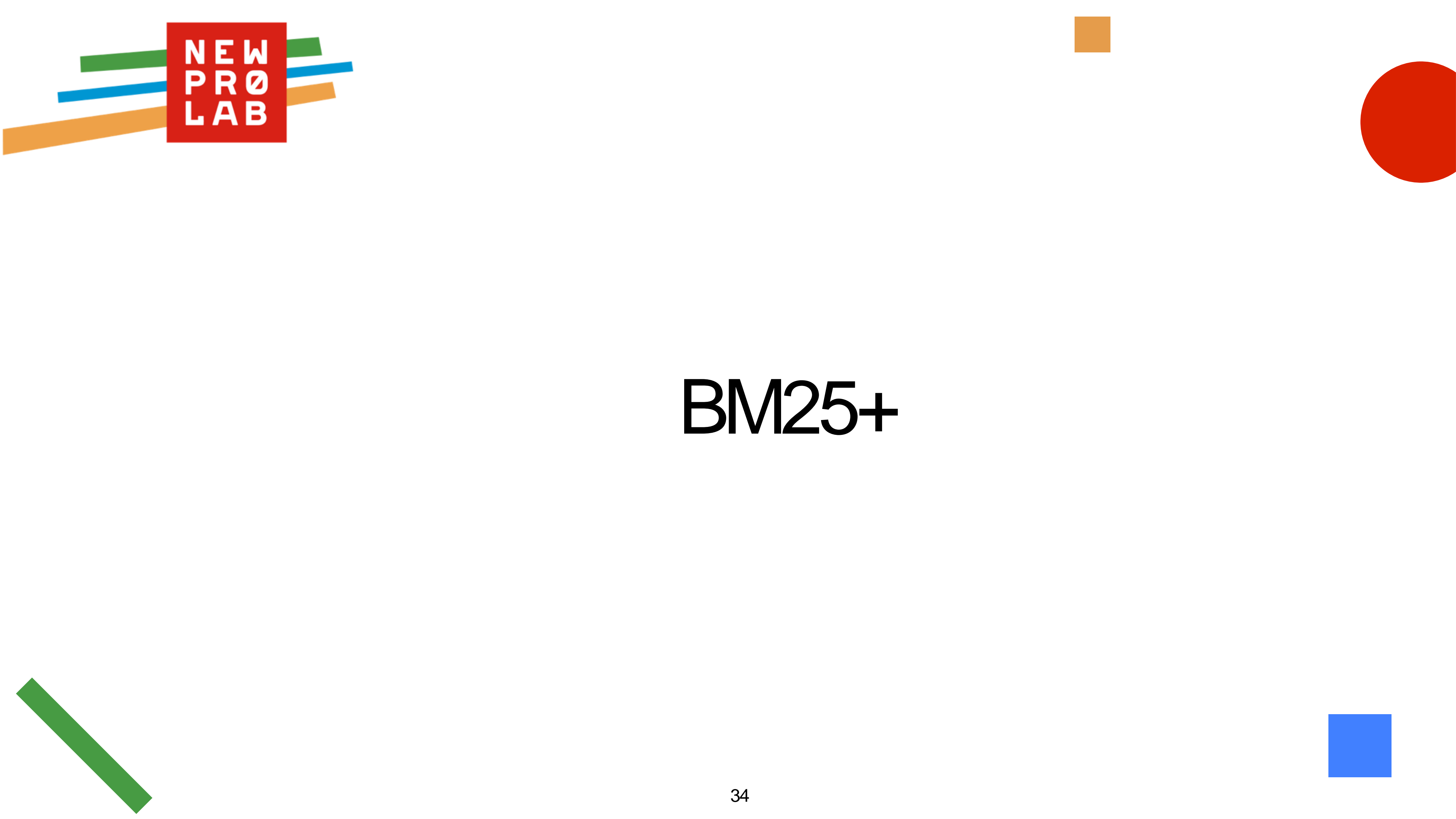
свободные
параметры



NEW
PRO
LAB

BM25F

- ТЕКСТ СОСТОИТ ИЗ РАЗНЫХ УЧАСТКОВ
- ВАЖНОСТЬ КУСКОВ РАЗНАЯ
- К ТАКИМ УЧАСТКАМ ОТНОСЯТ ТЕГ Title, метатеги, заголовки и подзаголовки, околоосылочный текст.

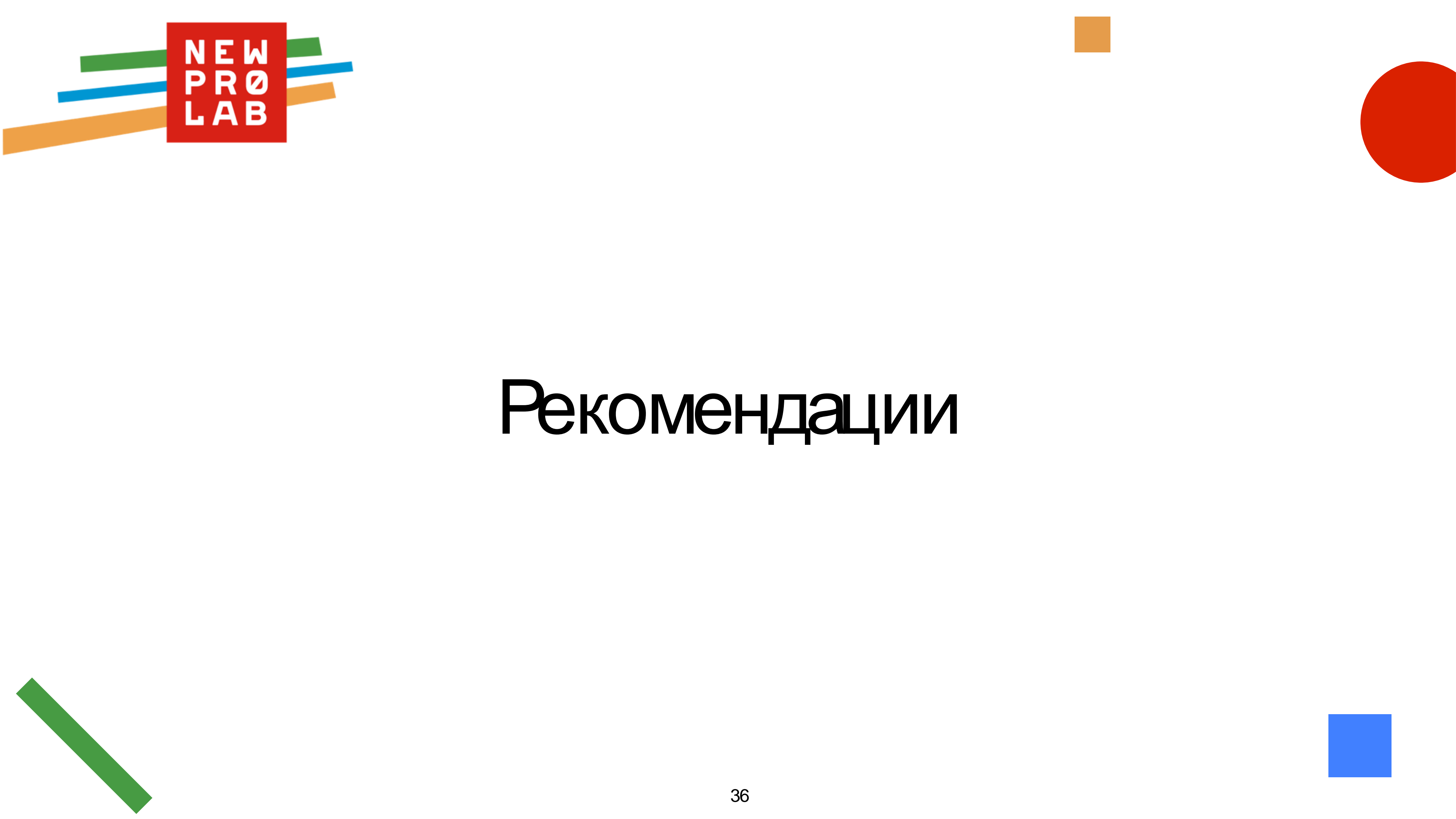


BM25+

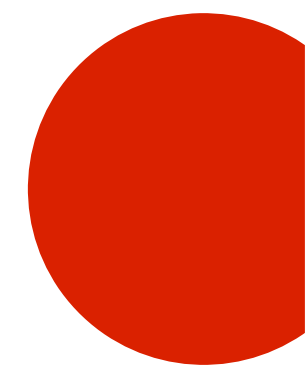
$$\text{score}(D, Q) = \sum_{D \cap Q} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta$$

$$k_1 = 2, b = 0.75, \delta = 1$$

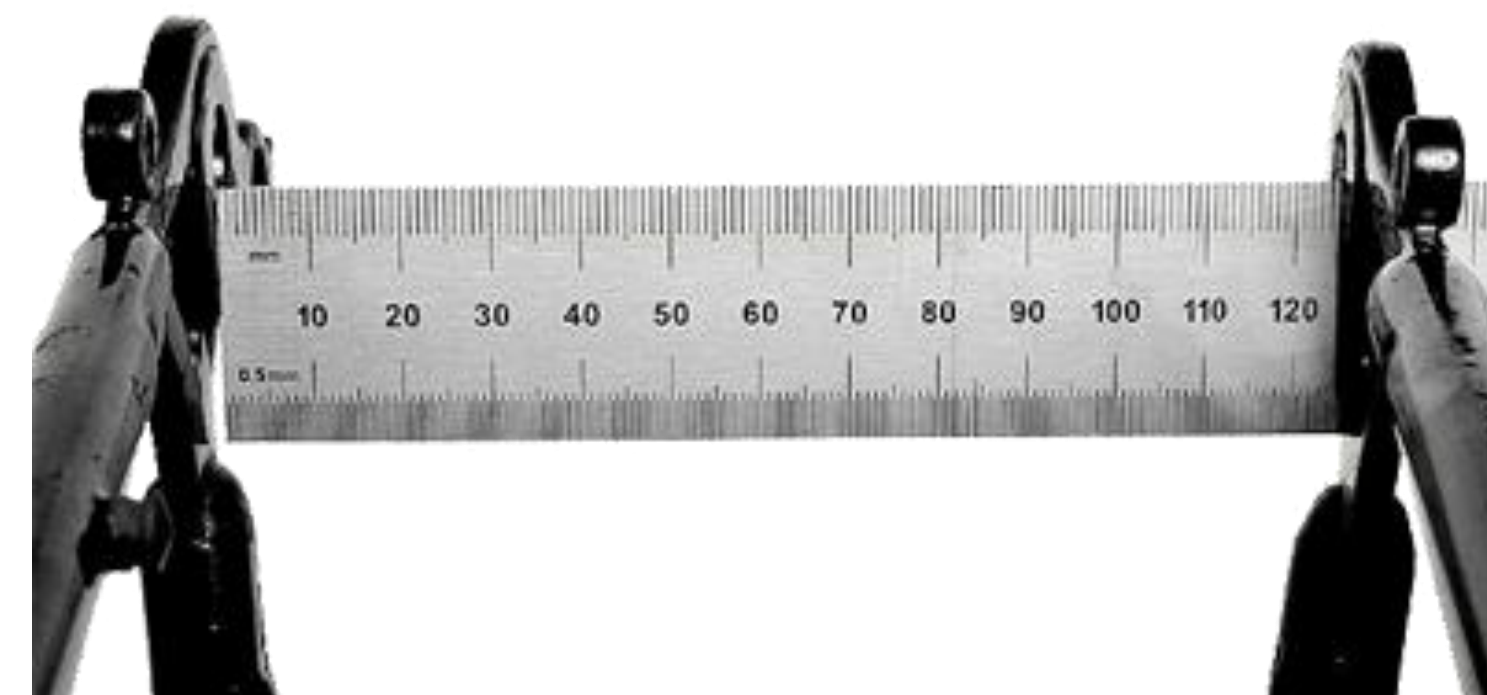
свободные
параметры



Рекомендации



упорядочивать по





классификатор kNN



- вычислить расстояние до объектов
- отобрать k ближайших
- взять класс, наиболее часто встречающийся среди kNN

4



5

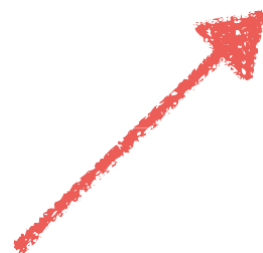
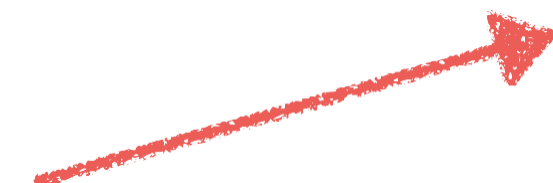


5



k-nearest neighbors (kNN)

3



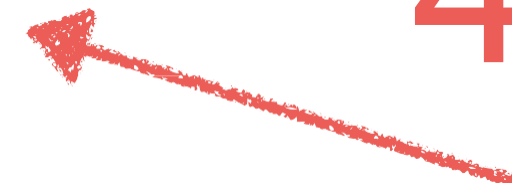
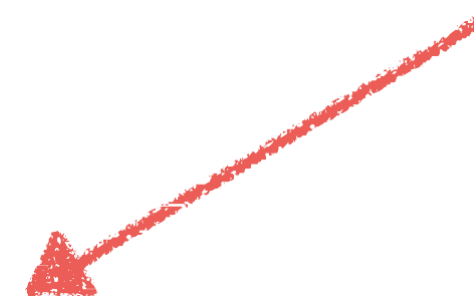
3

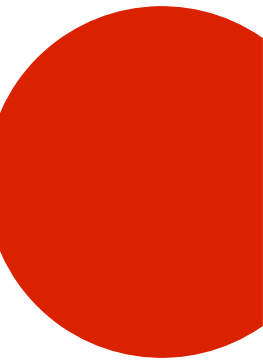


4



4





наивный Байесовский классификатор



$$c = \arg \max_c P(C | O)$$

строка текста

классы

вероятность
такого текста
в классе C

вероятность
класса C

$$P(C | O) = \frac{P(O | C)P(C)}{P(O)}$$

вероятность
текста O



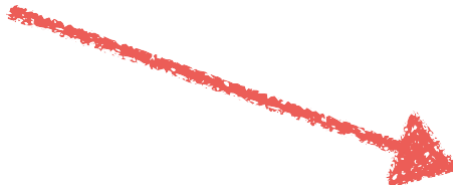
вероятность
класса C



$P(C)$



вероятность
такого текста
в классе C


$$P(O \mid C) = P(o_1 o_2 \dots o_n \mid C)$$

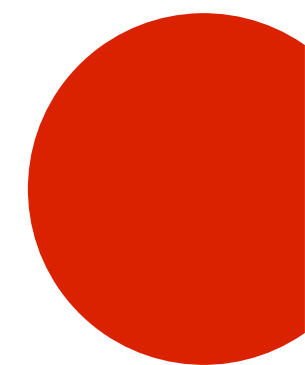


$$P(o_1 | C), P(o_2 | C), \dots, P(o_n | C)$$



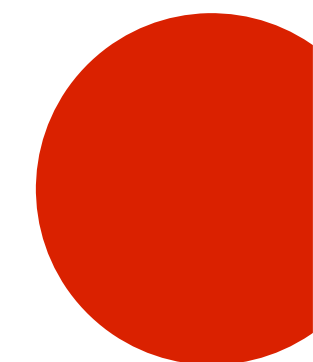


- не нужны данные о других пользователях
- нет проблемы холодного старта у товаров



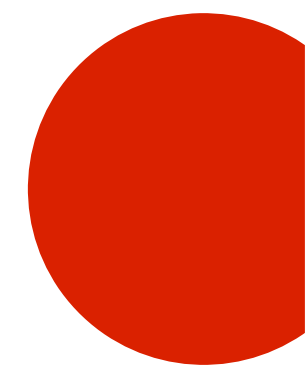
нет проблемы рекомендовать
пользователям с уникальным вкусом



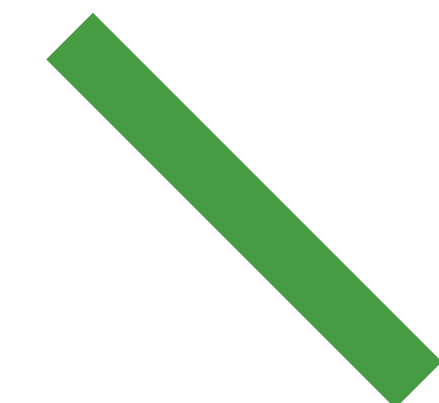


может рекомендовать новые или
непопулярные товары





легко объяснить



требуется превратить контент в
осмысленные признаки товара

- СЛОЖНО для аудио, видео, картинок



тяжело учитывать мнение других о
качестве товара





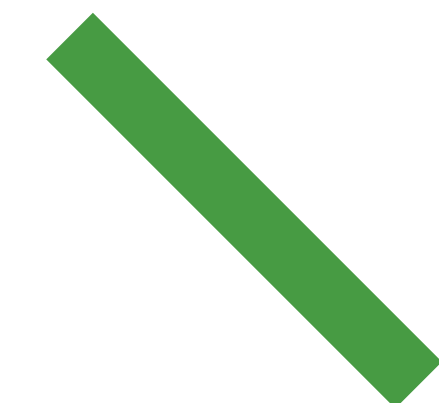
иногда теряется смысл



Вегетарианцу будет
нечего съесть
в этом ресторане.



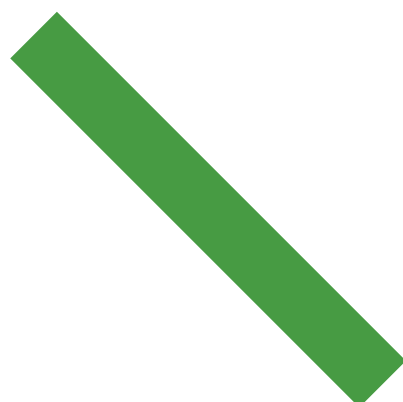
СИНОНИМЫ



казистый
хороший пышный видный
нарядный смазливый
щегольской изящный
пленительный благовидный
разубранный обворожительный
разукрашенный благообразный чудный
благотворный восхитительный художественный пригожий
привлекательный миловидный распрекрасный
блестящий бесподобный хорошенький
великолепный взрачный божественный
роскошный прелестный
живописный картинный
дивный



фразеологические обороты

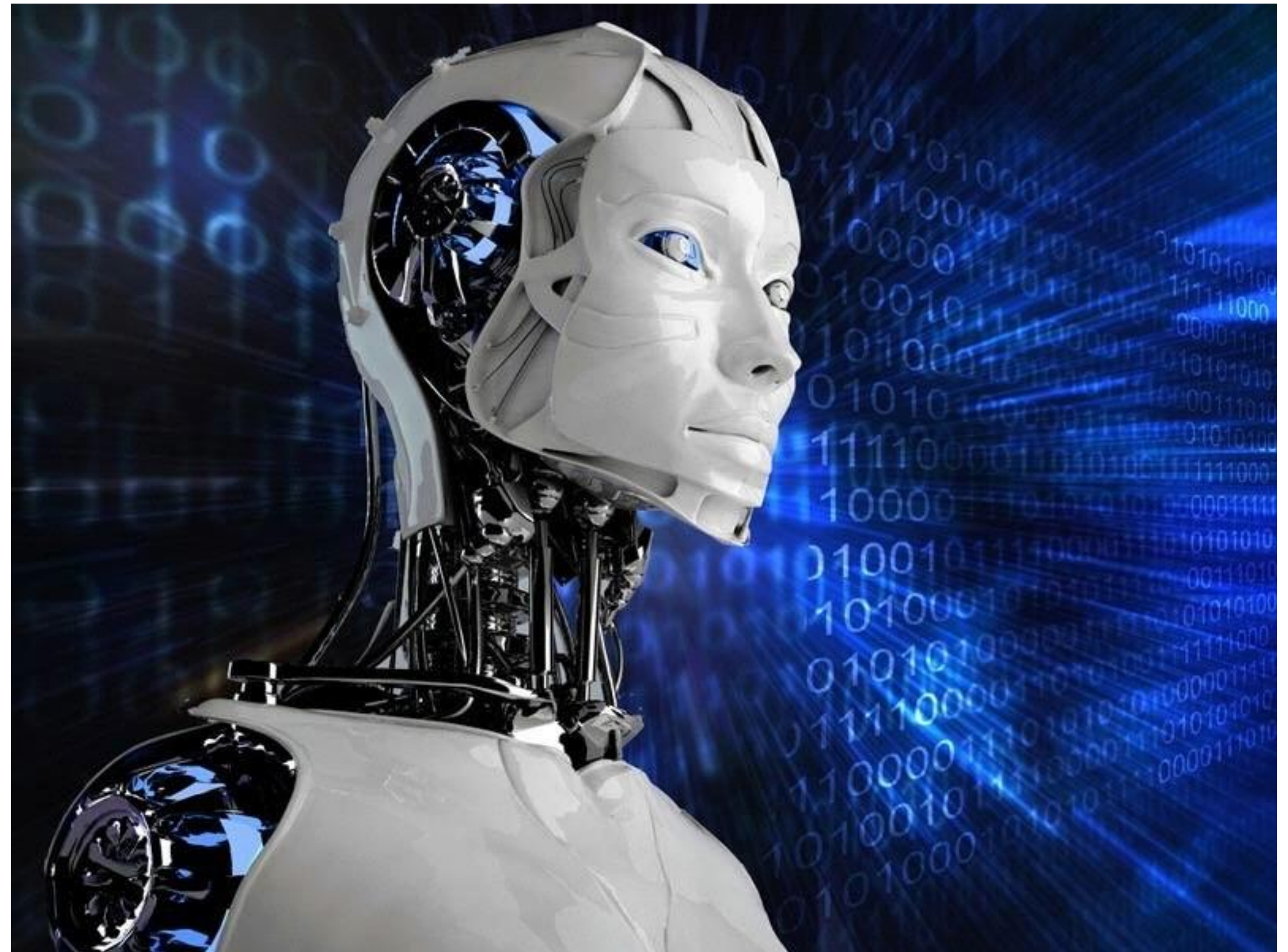


разбит
ь
сердце



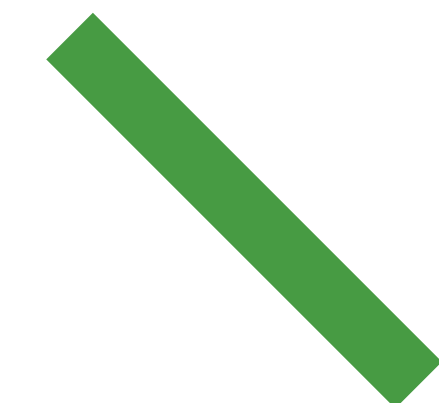


ИСКУССТВЕННЫЙ
РАЗУМ





МНОГОЗНАЧНОСТЬ





NEW
PRO
LAB

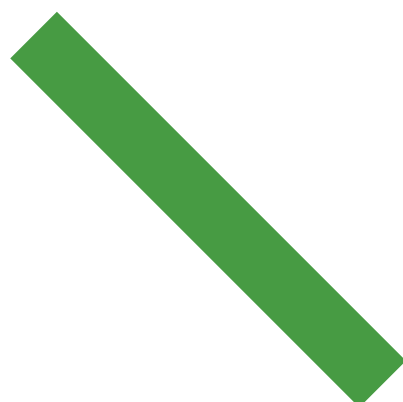


Мне это уже не надо! (overfit)





~~ЧТО-ТО НЕОЖИДАННОЕ~~



Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy *

Dept. of Measurement and Information Systems
Budapest University of Technology and
Economics

Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk *†

Dept. of Telecom. and Media Informatics
Budapest University of Technology and
Economics

Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@tmit.bme.hu

ABSTRACT

The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization (MF) based CF approaches assign low dimensional feature vectors to users and items. We link CF and content-based filtering (CBF) by finding a linear transformation that transforms user or item descriptions so that they are as close as possible to the feature vectors generated by MF for CF.

We propose methods for explicit feedback that are able to handle 140 000 features when feature vectors are very sparse. With movie metadata collected for the NP movies we show that the prediction performance of the methods is comparable to that of CF, and can be used to predict user preferences on new movies.

We also investigate the value of movie metadata compared to movie ratings in regards of predictive power. We compare our solely CBF approach with a simple baseline rating-based

1. INTRODUCTION

The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and metadata about users and items, if available.

There are two basic strategies that can be applied when generating recommendations. Collaborative filtering (CF) methods are based only on the activity of users, while content-based filtering (CBF) methods use only metadata. In this paper we propose hybrid methods, which try to benefit from both information sources.

The two most important families of CF methods are matrix factorization (MF) and neighbor-based approaches. Usually, the goal of MF is to find a low dimensional representation for both users and movies, i.e. each user and movie is associated with a feature vector. Movie metadata (which are mostly textual) can also be represented as a vector, using