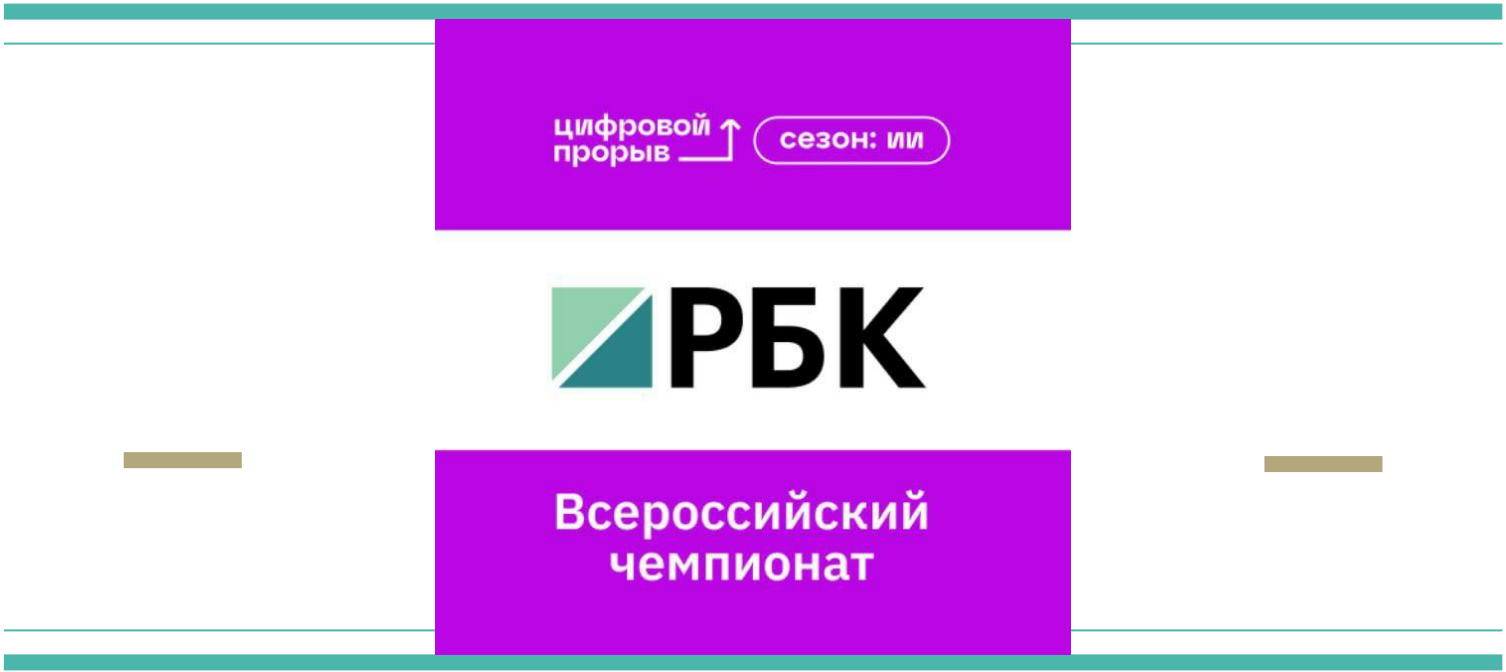
Всероссийский чемпИИонат



Александр • 22.07.2022

Задача

title	publish_date	session	authors	ctr	category	tags	views	depth	full_reads_percent
Европейский банк развития приостановил доступ ...	2022-04-04 10:29:44	IDE7mtH4RBqGn-8MXfGffQ	[]	1.580	5409f11ce063da9c8b588a18	['55928d339a794751dc8303d6', '542d1e28cbb20f86...	20460	1.134	35.850
Кремль назвал регулярным процессом учебные зап...	2022-02-18 10:00:39	KtVJsteHStO5oditt3Uvzw	['54244e01cbb20f03076b236d','5878a2ec9a7947e53...	1.853	5409f11ce063da9c8b588a12	['549d25df9a794775979561d2', '58abcf539a7947f1...	19038	1.142	38.355
Госсекретарь Швеции заявила о нежелании вступа...	2022-02-12 04:24:02	hk7puWJwSziw0m3sfTkKWA	[]	0.000	5409f11ce063da9c8b588a12	['5430f451cbb20f73931ecd05', '5409f15de063daa0...	51151	1.185	36.424
Песков назвал прагматичной выдачу лицензии Газ...	2022-04-22 13:24:55	7UKY2SSZTjCcjhwBzxw37w	[]	0.000	5409f11ce063da9c8b588a12	['5409f297e063daa0f408b11c', '545caa9ecbb20f36...	3782	1.053	30.169
В Хабаровске задержали главу филиала РАНХиГС п...	2022-04-25 10:42:23	wuMYES90REuV5YhrN75lXg	[]	0.000	5433e5decbb20f277b20eca9	['5409f42ae063daa0f408b5d7', '585c20e19a79470e...	3065	1.063	34.617

Тренировочная задача

Данные

title	publish_date	session	authors	views	depth	full_reads_percent	ctr
Какие места на Украине взяли под контроль росс...	2022-03-23 11:29:10	QGhF9EXPR0aUGgop3UoSxg	[]	616365	1.240	23.667	5.907
В России оценили инициативу Байдена изымать ак...	2022-04-28 14:58:07	Ref07c4BSDOyLxn3pENGRg	['5a8d11a39a7947c5e1550980']	10145	1.050	34.835	0.000
Bloomberg узнал о планах ФРГ нарастить займы д...	2022-04-24 04:06:29	Dg8Mc_DoTV6c0hik47bjVQ	['5e1ddee09a7947609de2f69b']	16916	1.070	35.889	1.740
Bloomberg узнал о планах США ввести санкции пр...	2022-04-23 13:11:09	V_SsaXpVT1S-UYwxlvvc7A	['5a8d11a39a7947c5e1550980']	11874	1.060	27.851	2.924
Лавров заявил Ле Дриану об отсутствии прогресс...	2022-02-19 13:46:11	uxm07DqbTP60j4iaffdjPQ	[]	47841	1.204	39.638	0.000

Алгоритм

Скрапинг

description	keywords	copyright	article_categories	article_authors	article_publication_date	article_word_count	text
В то же время гендиректор считает, что уход би...	Компания, Объявить, Российский, Фактор, Пообещ...	«РосБизнесКонсалтинг»	"Бизнес"	"Юлия Выродова"	Wed, 04 May 2022 21:31:54 +0300	256	Coca-Cola может «в какой-то момент полностью и...
Страны, которые заявляют об атаках русских хак...	Доказательство, Официальный, Государство, Безо...	«РосБизнесКонсалтинг»	"Политика"	""	Wed, 30 Mar 2022 23:39:47 +0300	366	Страны, которые заявляют об атаках «русских ха...
Полиция Ростовской области задержала двух рабо...	Сбербанк, Отделение, Сообщить, Банкомат, Задер...	«РосБизнесКонсалтинг»	"Общество"	""	Mon, 04 Apr 2022 17:48:13 +0300	193	Полиция Ростовской области задержала двух рабо...
Минстрой включил в перечень системообразующих ...	Компания, Строительство, Перечень, Системообра...	«РосБизнесКонсалтинг»	"Экономика"	""	Fri, 29 Apr 2022 12:21:38 +0300	298	Минстрой включил в перечень системообразующих ...
Премьер-министр Матеуш Моравецкий создаст рабо...	Вторжение, Моравецкий, Польский, Подготовка, Б...	«РосБизнесКонсалтинг»	"Политика"	""	Wed, 16 Feb 2022 15:24:16 +0300	225	Премьер-министр Матеуш Моравецкий создаст рабо...

Текстовые эмбеддинги

TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

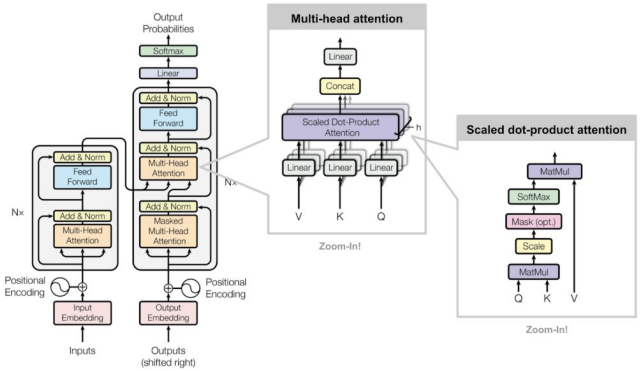
Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

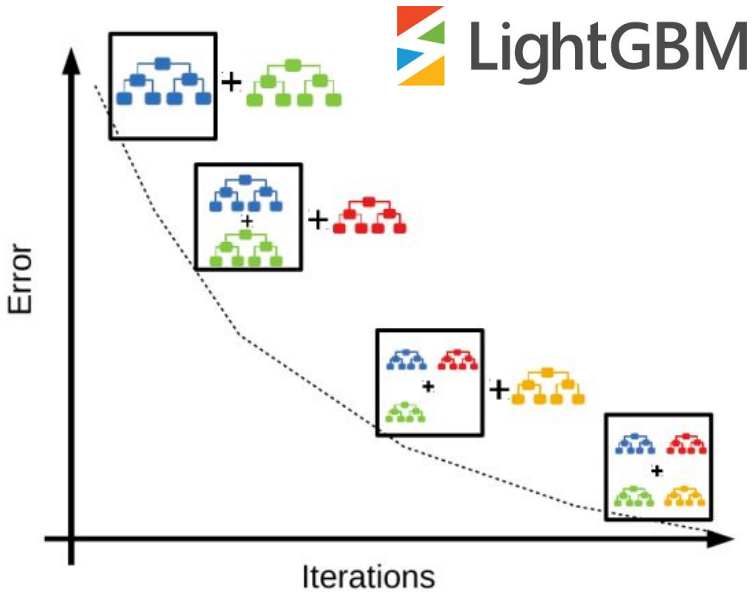
N = total number of documents

rubert-tiny2

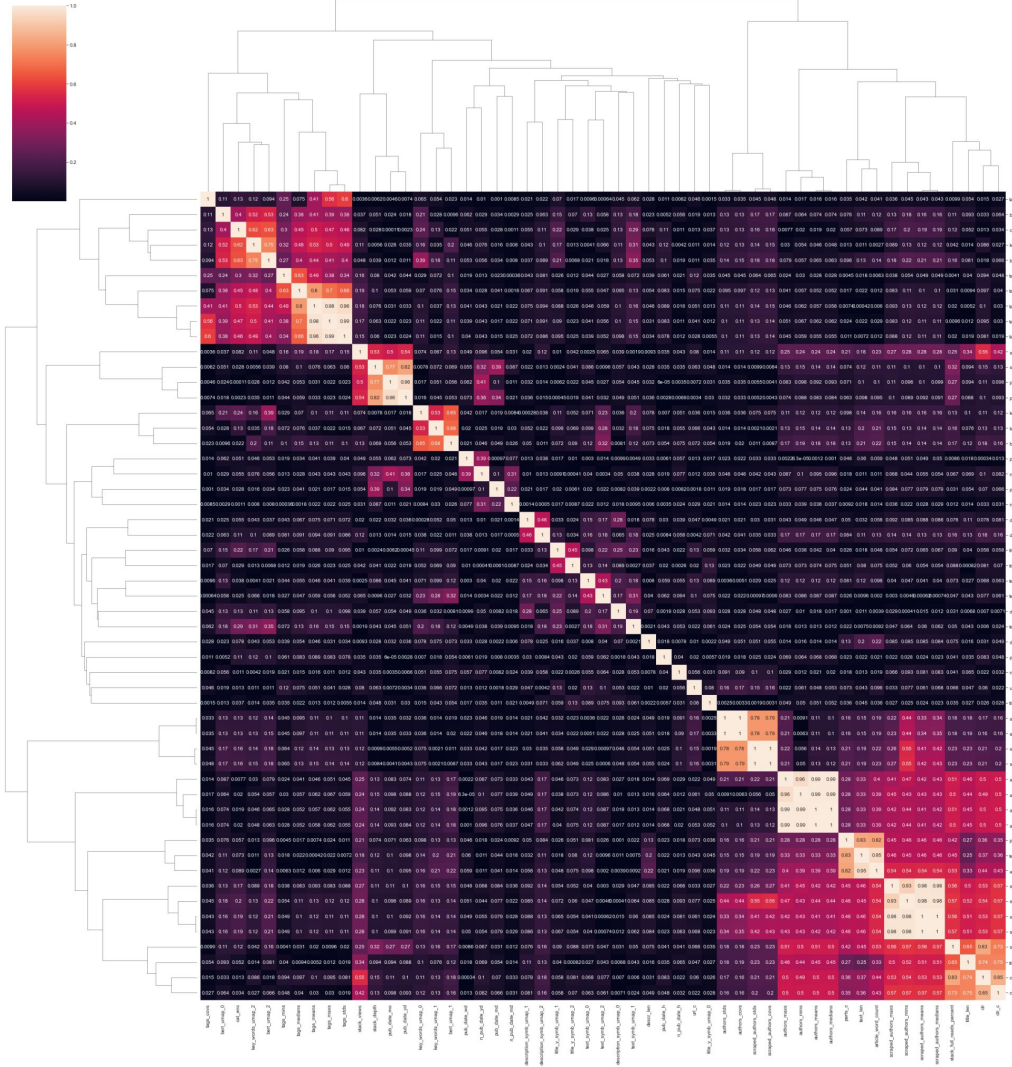


UMAP

Итоговое решение

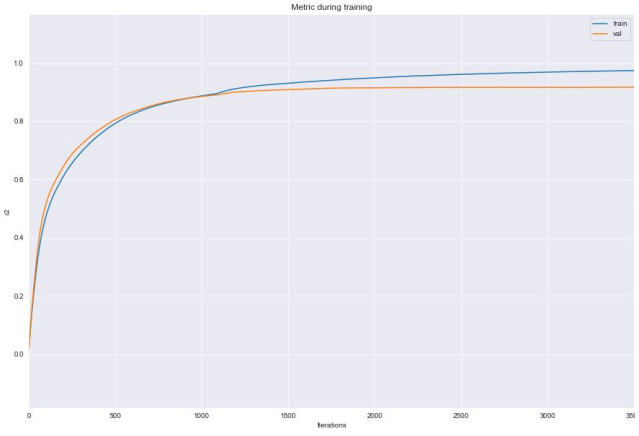


LightGBM

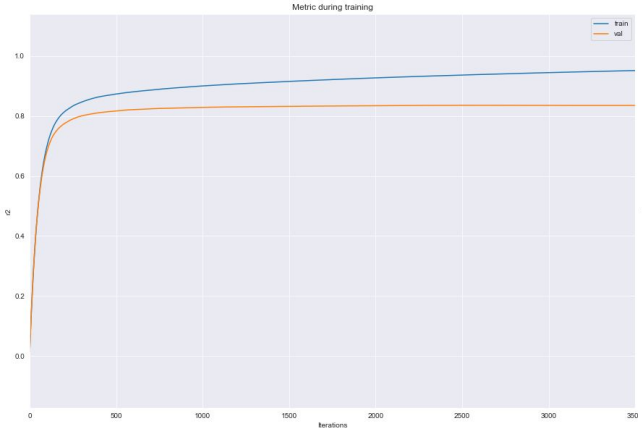


Результат

Views



Depth



Full Reads Percent