

Zadolba.li

Корпус текстов

Кривчанский Н., Куликов А.

ABBYY

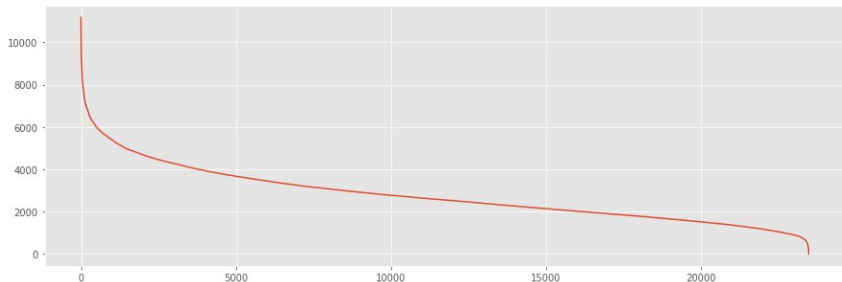
Москва, 2017

- Русскоязычный корпус;
- Широкий спектр тематик;
- Легко выделить несколько стилей;
- Имеется премодерация и теги историй.

- 23558 истории;
- С 10.09.2009 по 01.11.2017;
- Id, название, дата публикации, теги, текст, лайки, url, ссылки на другие истории;
- SQLite размер 91,2 MB.

- 23558 истории;
- 4041730 слов (nltk Regexp(\w+) tokenizer);
- Размер истории $L \in [10; 1399]$ слов, $\bar{L} \approx 270$ слов;
- 464750 предложений (nltk PunktSentenceTokenizer with russian model);
- Размер истории $S \in [1; 173]$ предложений, $\bar{S} = 20$ предложений.

Распределение лайков

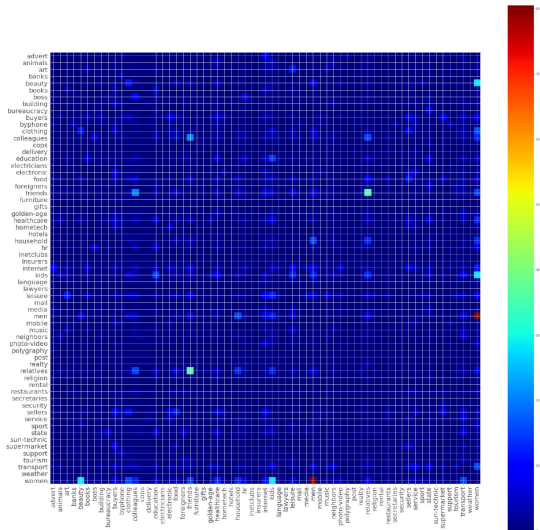


Распределение тегов

Самые частые теги:

women	2793
transport	2055
men	1566
friends	1544
internet	1518
healthcare	1516
education	1461
leisure	1440
kids	1415
relatives	1303

Корреляция тегов



Самые частые пары тегов

Самые частые пары тегов:

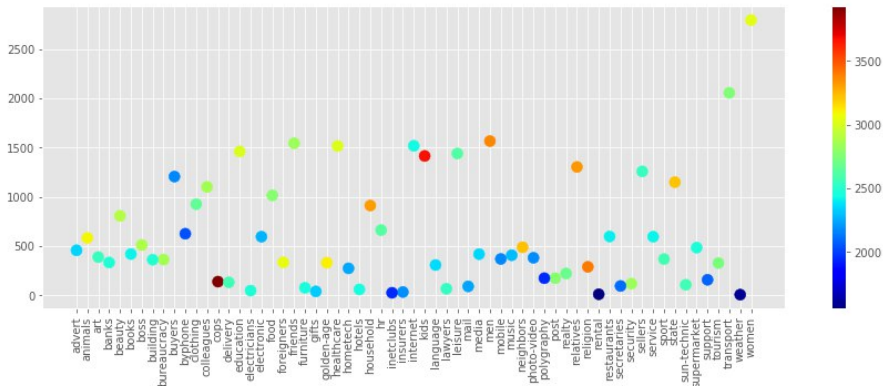
men	women	805
friends	relatives	360
beauty	women	284
kids	women	284
colleagues	friends	206
clothing	women	166
colleagues	relatives	157
household	men	156
transport	women	155
friends	women	149

Самые лайкаемые теги

Самые лайкаемые теги:

cops	3927.01
kids	3668.98
religion	3404.57
men	3380.20
relatives	3329.33
household	3322.93
neighbors	3223.52
state	3217.17
golden-age	3111.95
animals	3088.71

Самые лайкаемые теги



Цвет – среднее кол-во лайков у тега.

Высота – кол-во историй у тега.

- Хранится в отдельной табличке;
- Лемма + Morpho-tag + id истории + координаты в тексте.

Спасибо