

Skill Factory. DST-10.

Никишин Андрей

---

## Дипломный проект: «Кластеризация поисковых запросов для рекламных кампаний»



# Бизнес-цели

## Выявление популярных запросов в семантическом ядре товарной группы

	1	2
1	письменный стол	122559
2	компьютерный стол	107513
3	угловой стол	28134
4	купить письменный стол	24047
5	купить компьютерный стол	23058
6	стол +для школьника	18430
7	стол офисный	17216
8	письменный стол +для школьника	12305
9	угловой компьютерный стол	10909
10	икеа письменный стол	9738
11	письменный стол москва	7463
12	компьютерный стол москва	7357
13	письменный стол белый	6753
14	купить компьютерный стол +в москве	5887
15	угловой письменный стол	5586
16	купить письменный стол +в москве	5550



- Письменные столы.
- Компьютерные столы.
- Угловые столы.
- Белые столы.
- Столы для двоих.
- Столы для школьников.
- Маленькие столы.
- Столы для офиса.
- Столы с надставками.
- Столы с ящиками.
- Столы со стеллажами и шкафами.

# Бизнес-цели

## Выявление популярных запросов в семантическом ядре товарной группы

---

- Анализ того, что необходимо производить.
- Оценка потенциального спроса.
- Оценка цветовых предпочтений.
- Оценка товаров конкурентов.
- Выявление тенденций спроса.
- Выявление трендов.
- Обновление информации о текущем ассортименте (1 раз в квартал).
- Актуализация рекламных кампаний.


Столы Домашний офис Стеллажи Прихожие Комоды Туалетные столики Тумбы Шкафы Полки Картины











# Бизнес-цели

## Распределение поисковых запросов по посадочным страницам

### Столы угловые

 	 	 
Стол угловой Рикс-8 3 499 руб.	Стол угловой Триан-5 8 499 руб.	Стол угловой Триан-5 ПРАВЫЙ 8 499 руб.
		
Стол угловой Краст-2 10 299 руб.	Стол угловой Краст-2 ПРАВЫЙ 10 299 руб.	Стол угловой Краст-3 14 399 руб.







### Столы белые

		
Стол письменный Ренцо-2 8 399 руб.	Стол письменный Нейт-3 9 299 руб.	Стол угловой Триан-1 5 199 руб.
 		 
Стол Слим-1, прямой 6 499 руб.	Стол угловой Триан-1 ПРАВЫЙ 5 199 руб.	Письменный стол Милан 7 199 руб.

### Столы для двоих

#### Столы для двоих Тандем

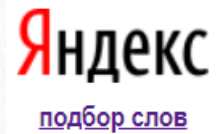
Сортировка по цене ^ | по популярности | по умолчанию

 	 	 
Стол для двоих Тандем-2 10 199 руб.	Стол для двоих Тандем-2Я, с ящиками 10 999 руб.	Стол для двоих Тандем-3 12 999 руб.

↑ конверсия рк

# Процесс работы

## Сбор семантического ядра



[Директ](#) [Справочник](#) [Метрика](#) [Рекламная сеть](#) [Маркет](#) [ещё](#)

Подобрать☒ По словам☐ По регионам☐ История запросов[Москва и область](#)ВсеДесктопыМобильныеТолько телефоныТолько планшеты

Последнее обновление:

09.04.2021

Что искали со словом «письменный стол» — 112 166 показов в месяц

Статистика по словам

Показов в месяц <sup>?</sup>

<a href="#">письменный стол</a>	112 166
<a href="#">купить письменный стол</a>	22 107
<a href="#">письменный стол +для школьника</a>	10 943
<a href="#">икеа письменный стол</a>	9 562
<a href="#">письменный стол москва</a>	7 103
<a href="#">письменный стол белый</a>	5 829

Запросы, похожие на «письменный стол»

Статистика по словам

Показов в месяц <sup>?</sup>

<a href="#">купить стол</a>	284 336
<a href="#">компьютерный стол</a>	95 789
<a href="#">детский стол</a>	29 858
<a href="#">мебель стол</a>	24 089
<a href="#">стол школьник</a>	18 678
<a href="#">угловой стол</a>	26 407
<a href="#">стол офисный</a>	17 928





# Процесс работы

## Удаление локальных и глобальных минус-слов

---

### Global minus

Запросы, которые исключаются из всех рекламных кампаний, т. к. не ведут к продаже товара:

- Собрать письменный стол своими руками.
- Купить компьютерный стол БУ.
- Обои для рабочего стола.
- Инструкция по сборке стола.

Запросы, которые содержат зарегистрированные торговые марки:

- Купить игровой стол в Ситилинк.
- Обеденные столы Хофф.
- Письменный стол Столплит.



# Процесс работы

## Удаление локальных и глобальных минус-слов

---

### Local minus

Запросы, которые исключаются из рекламных кампаний по данной товарной группе.

- Компьютерный стол красный.
- Стекланный письменный стол.
- Складные письменные столы купить.
- Стул детский для письменного стола школьника.
- Столы для школьника Тула.





# Процесс работы

## Удаление локальных и глобальных минус-слов

### Ручная работа на внимательность

- Списки локальных минус-слов по каждой товарной группе и список глобальных минус-слов собраны в компании.
- Фильтруем творчески и с энтузиазмом :)

**Красный, красного цвета, красная кромка и т. д. → «красн»**

1237	письменный стол красный	95
1417	красный компьютерный стол	83
1968	стол компьютерный черно красный	58
2589	компьютерный стол красногорск	43
3642	письменный стол красное дерево	29
4410	стол компьютерный красный купить	23
5192	стол офисный красный	16
6705	офисные столы краснодар	5





# Процесс работы

## Выявление популярных запросов

	1	2
1	письменный стол	122559
2	компьютерный стол	107513
3	угловой стол	28134
4	купить письменный стол	24047
5	купить компьютерный стол	23058
6	стол +для школьника	18430
7	стол офисный	17216
8	письменный стол +для школьника	12305
9	угловой компьютерный стол	10909
10	икеа письменный стол	9738
11	письменный стол москва	7463
12	компьютерный стол москва	7357
13	письменный стол белый	6753
14	купить компьютерный стол +в москве	5887
15	угловой письменный стол	5586
16	купить письменный стол +в москве	5550



- Письменные столы.
- Компьютерные столы.
- Угловые столы.
- Белые столы.
- Столы для двоих.
- Столы для школьников.
- Маленькие столы.
- Столы для офиса.
- Столы с надставками.
- Столы с ящиками.
- Столы со стеллажами и шкафами.

# Процесс работы

## Сегментация поисковых запросов

	1	2
1	письменный стол	122559
2	компьютерный стол	107513
3	угловой стол	28134
4	купить письменный стол	24047
5	купить компьютерный стол	23058
6	стол +для школьника	18430
7	стол офисный	17216
8	письменный стол +для школьника	12305
9	угловой компьютерный стол	10909
10	икеа письменный стол	9738
11	письменный стол москва	7463
12	компьютерный стол москва	7357
13	письменный стол белый	6753
14	купить компьютерный стол +в москве	5887
15	угловой письменный стол	5586
16	купить письменный стол +в москве	5550



	1	2	3
1	письменный стол	122559	письменный
2	компьютерный стол	107513	компьютерный
3	угловой стол	28134	угловой
4	купить письменный стол	24047	письменный
5	купить компьютерный стол	23058	компьютерный
6	стол +для школьника	18430	школьник
7	стол офисный	17216	офис
8	письменный стол +для школьника	12305	школьник
9	угловой компьютерный стол	10909	угловой
10	икеа письменный стол	9738	GM
11	письменный стол москва	7463	письменный
12	компьютерный стол москва	7357	компьютерный
13	письменный стол белый	6753	белый
14	купить компьютерный стол +в москве	5887	компьютерный
15	угловой письменный стол	5586	угловой
16	купить письменный стол +в москве	5550	письменный



# Цели дипломного проекта

---

## Бизнес-цели:

- Ускорить и упростить процесс распределения поисковых запросов по группам для сотрудников отдела маркетинга компании.
- Создать возможность быстрого анализа тенденций изменения спроса в существующих товарных группах и перспективных направлениях развития ассортиментной политики.
- Реализовать функцию мониторинга ассортимента конкурирующих организаций.
- Упростить процесс выявления трендов и тенденций спроса в рабочей нише.

## Технические цели:

- Разработать инструмент автоматической обработки, очистки, фильтрации и кластеризации семантического ядра поисковых запросов.
- Добиться точности алгоритма, сопоставимой с точностью при ручной обработке запросов.



# Цели дипломного проекта

---

## Основные тезисы

- Как показывает практика, человек, при ручной разметке данных, ошибается приблизительно в 10-15% случаев.
- Цена ошибки не высока, т. к. она нивелируется последующим автоматическим алгоритмом, регулирующим бюджет рекламных кампаний.
- Наша задача создать алгоритм, который максимально точно распределит данные по группам, также, как это сделал человек.

## Цели по качеству:

- Точность менее 80% - неудовлетворительно.
- Точность от 80 до 85% - удовлетворительно.
- Точность от 86 до 90% - хорошо.
- Точность свыше 90% - отлично.





# Процесс работы

## Подготовка данных

- Очистили данные от пропусков и дублей. Перевели в нижний регистр.
- С помощью библиотеки Mystem провели лемматизацию поисковых запросов.
- С помощью библиотеки NLTK очистили запросы от стоп-слов.
- С помощью библиотеки string.punctuation очистили запросы от знаков математических операций и пунктуации.
- Вынесли обработанные данные в отдельный столбец датафрейма.
- Создали еще один признак – списки слов поискового запроса в начальной форме.

7218	оформление стола +на двоих	3	GM	0	оформление стол двое	[оформление, стол, двое]
7219	варианты письменных столов +для двоих детей	3	тандем	0	вариант письменный стол двое ребенок	[вариант, письменный, стол, двое, ребенок]
7220	стол подоконник +для двоих детей	3	LM	0	стол подоконник двое ребенок	[стол, подоконник, двое, ребенок]
7221	письменный стол +для двоих детей размеры	3	GM	0	письменный стол двое ребенок размер	[письменный, стол, двое, ребенок, размер]
7222	тандем 3 стол письменный +для двоих	3	тандем	0	тандем 3 стол письменный двое	[тандем, 3, стол, письменный, двое]

# Процесс работы

## Очистка данных от минус-слов

- Проблема коротких минус-слов:
  - письменный стол цвет **бук**.
  - стол +для **бу**хгалтера офисный.
  - угловые столы санкт-петербург.
  - компьютерный стол +для ноут**бу**ка купить +в москве.
- Проблема технических тегов поисковых запросов «+, !, -, “”».
- Трехэтапная фильтрация запросов по минус-словам.



## Результат

- Датасет ключевых запросов сократился с **7223** до **3769** строк.
- Удалось сформировать универсальную систему фильтрации датасетов, не "подогнанную" под конкретные данные.
- С помощью 3-х этапной фильтрации удалось достичь **100%** точности обработки **48%** данных.
- Ни одна из релевантных ключевых фраз не была убрана из датасета, в тоже время все нерелевантные запросы были отфильтрованы.



# Процесс работы

## Рекомендации по выбору групп для распределения поисковых запросов

- Ограничения для алгоритма: Алгоритм не должен быть подстроен под конкретные входные данные (должен быть универсальным).
- Критерий качества данного этапа: Алгоритм в рекомендациях выдает все группы, которые были выбраны человеком при ручной обработке.

### Выбранные группы

- Компьютерные столы.
- Угловые столы.
- Письменные столы.
- Столы для офиса.
- Столы для детей и школьников.
- Столы для двоих детей (тандем).
- Столы со шкафом
- Белые столы.
- Маленькие столы.
- Столы с ящиками.
- Столы с надставками.

	words	frequency	count
1	стол	3775	790854
188	письменный	1175	335221
609	компьютерный	1392	302578
553	купить	583	133844
303	угловой	913	105511
293	школьник	486	80355
586	москва	183	57979
442	офисный	389	38763
213	недорого	119	35825
69	ящик	111	28481
201	белый	137	22226
552	полк	86	17403
610	надстройка	103	17302
45	двое	186	16050
492	полка	68	10268
102	шкафчик	37	7867
451	детский	90	7624
516	дом	72	7335
115	маленький	59	7297
479	магазин	55	6934

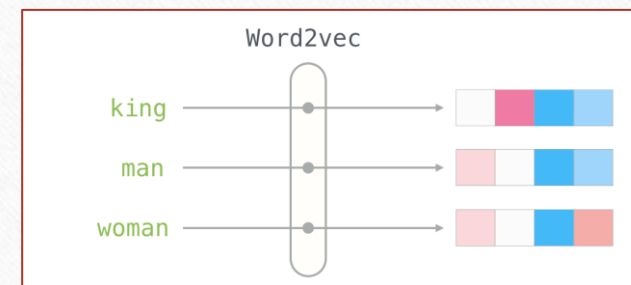
# Процесс работы

## Распределение запросов по выбранным группам

- Обучение модели **Word2Vec**, **Doc2Vec** (обучение на семантическом ядре).
- Библиотека **Ktrain** (кластеризация на топики, выбор кластеров для запросов).
- Близость векторов запросов. Пересечение ключевых слов поисковых запросах.
- Библиотека **Gensim**, обученная модель: **word2vec-ruscorpora-300**:
  - Проект RusVectōrēs.
  - Входит в стандартный api библиотеки gensim.
  - 223Мб
  - 184973 векторов.
  - Обучена на полном национальном корпусе русского языка (НКРЯ).

## Ссылки:

- НКРЯ: <https://ruscorpora.ru/new/>
- RusVectōrēs: <https://rusvectors.org/ru/>
- Gensim: <https://radimrehurek.com/gensim/index.html>





# Процесс работы

## Кластеризация (подготовка)

- Добавление частеречных тегов (теги, означающие часть речи).
- Установка центров кластеров из списка рекомендаций:
  - письменный
  - компьютерный
  - угловой
  - школьник, детский
  - офисный
  - белый
  - ящик
  - двое
  - надстройка
  - маленький
  - шкафчик

phrase_upos_list
[письменный_ADJ, стол_NOUN]
[компьютерный_ADJ, стол_NOUN]
[угловой_ADJ, стол_NOUN]
[купить_VERB, письменный_ADJ, стол_NOUN]
[купить_VERB, компьютерный_ADJ, стол_NOUN]

	words	frequency	count
1	стол	3775	790854
188	письменный	1175	335221
609	компьютерный	1392	302578
553	купить	583	133844
303	угловой	913	105511
293	школьник	486	80355
586	москва	183	57979
442	офисный	389	38763
213	недорого	119	35825
69	ящик	111	28481
201	белый	137	22226
552	полк	86	17403
610	надстройка	103	17302
45	двое	186	16050
492	полка	68	10268
102	шкафчик	37	7867
451	детский	90	7624
516	дом	72	7335
115	маленький	59	7297
479	магазин	55	6934

# Процесс работы

## Кластеризация

- Преобразуем таргеты (центры кластеров) и поисковые запросы в эмбединги.
- Используем функцию косинусного сходства между вектором каждой фразы и центрами кластеров, которая возвращает вероятность отношения фразы к каждому кластеру.  
[Подробнее о косинусном сходстве](#)
- Относим поисковую фразу к кластеру с максимальной вероятностью попадания.

**Accuracy: 77.47%**





# Процесс работы

Приоритет отношения фразы к той или иной группе

Купить угловой письменный стол для школьника

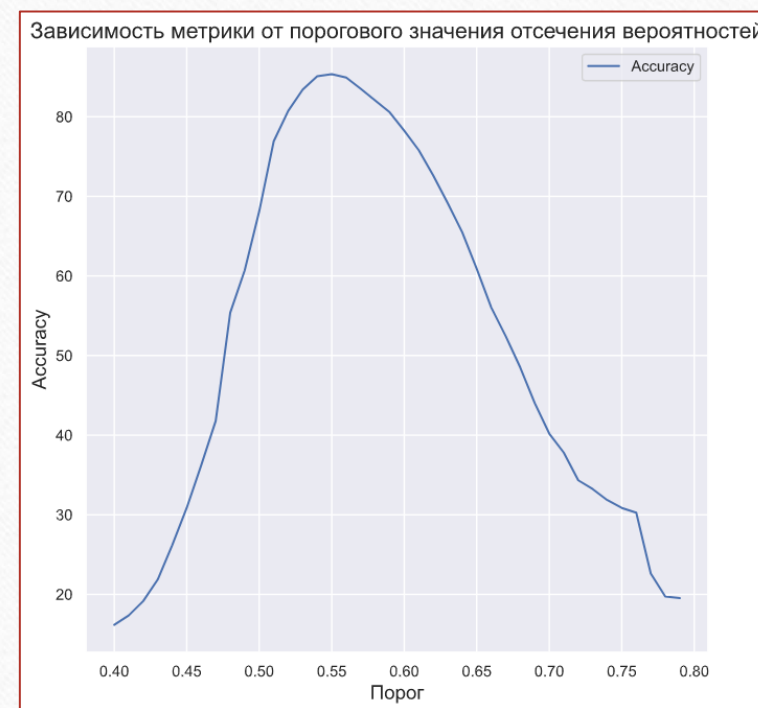
62%

65%

58%

**Коэффициент важности**

- Офисный стол - 11
- Столы со шкафом - 10
- Столы для двоих (Тандем) - 9
- Угловые столы - **8**
- Столы с надставкой - 7
- Белые столы - 6
- Столы для детей и школьников - **5**
- Столы с ящиками - 4
- Маленькие столы - 3
- Компьютерные столы - 2
- Письменные столы - **1**



Accuracy: 85.35%

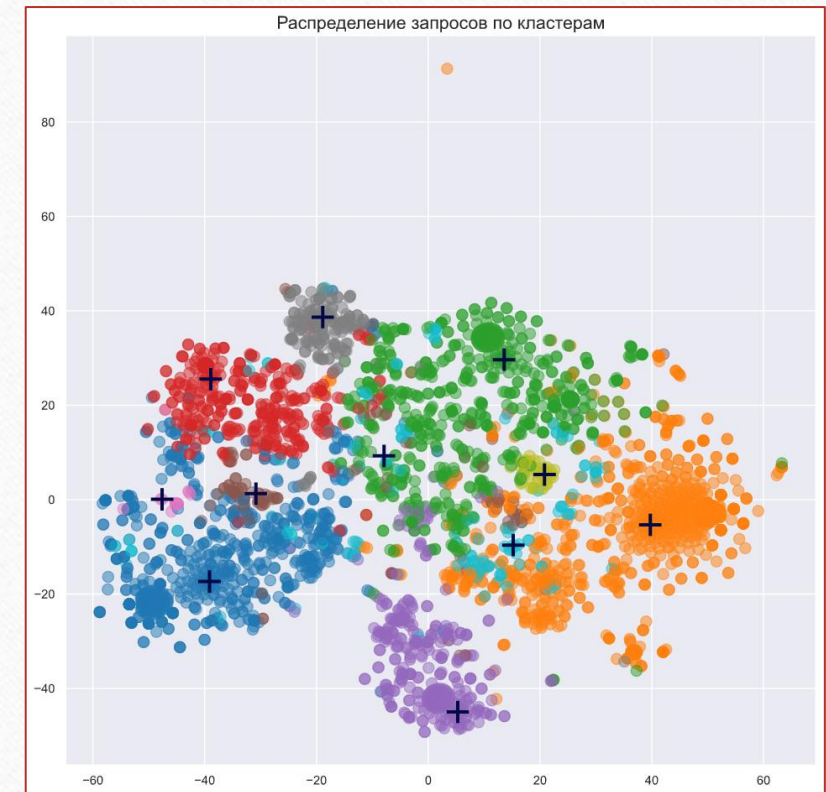
# Процесс работы



## Уточнение центров кластеров

Если после распределения поисковых запросов по группам дать пользователю подтвердить правильность распределения некоторых запросов, а потом на основе подтвержденных запросов скорректировать центры кластеров, результат распределения можно будет улучшить.

<b>письменный</b> Точность: 92.09%	<b>компьютерный</b> Точность: 92.04%	<b>угловой</b> Точность: 90.89%	<b>школьник, детский</b> Точность: 71.12%
<input checked="" type="checkbox"/> письменный стол 122559	<input checked="" type="checkbox"/> компьютерный стол 107513	<input checked="" type="checkbox"/> угловой стол 28134	<input checked="" type="checkbox"/> стол + для школьника 18430
<input checked="" type="checkbox"/> купить письменный стол 24047	<input checked="" type="checkbox"/> купить компьютерный стол 23058	<input checked="" type="checkbox"/> угловой компьютерный стол 10909	<input checked="" type="checkbox"/> письменный стол + для школьника 12305
<input checked="" type="checkbox"/> письменный стол москва 7463	<input checked="" type="checkbox"/> компьютерный стол москва 7357	<input checked="" type="checkbox"/> угловой письменный стол 5586	<input checked="" type="checkbox"/> купить стол + для школьника 3180
<input type="checkbox"/> купить письменный стол + в москве 5550	<input type="checkbox"/> купить компьютерный стол + в москве 5887	<input type="checkbox"/> купить угловой стол 4367	<input type="checkbox"/> купить письменный стол + для школьника 2397
<input type="checkbox"/> детский письменный стол 3406	<input type="checkbox"/> компьютерный стол недорого 3569	<input type="checkbox"/> купить угловой компьютерный стол 2215	<input type="checkbox"/> стол + для школьника + с полками 2168
<input type="checkbox"/> письменный стол недорого 3404	<input type="checkbox"/> купить компьютерный стол недорого 2888	<input type="checkbox"/> угловой стол + с надстройкой 1538	



Accuracy: 86.6%

Accuracy: 93%



# Процесс работы

## Пользовательское приложение

<div><div>письменный</div><div>Точность: 92.09%</div><div><div><input checked="" type="checkbox"/>122559</div><div>письменный стол</div></div><div><div><input checked="" type="checkbox"/>24047</div><div>купить письменный стол</div></div><div><div><input checked="" type="checkbox"/>7463</div><div>письменный стол москва</div></div><div><div><input type="checkbox"/>5550</div><div>купить письменный стол +в москве</div></div><div><div><input type="checkbox"/>3406</div><div>детский письменный стол</div></div><div><div><input type="checkbox"/>3404</div><div>письменный стол недорого</div></div><div><div><input type="checkbox"/>2936</div><div>письменный стол +с полками</div></div><div><div><input type="checkbox"/>2311</div><div>купить стол письменный недорого</div></div><div><div><input type="checkbox"/>1696</div><div>купить компьютерный стол +в москве недорого</div></div><div><div><input type="checkbox"/>1464</div><div></div></div></div>
---

## Выводы по проекту

---

- Проведена многоэтапная аналитическая работа по **выбору алгоритмов** подготовки, фильтрации и распределения поисковых запросов по необходимым группам.
- Удалось достичь **высокой скорости** и **точности** работы алгоритмов, что позволит пользователям обрабатывать более широкий спектр целевых запросов, выявляя тенденции движения рынка мебели.
- Достигнута общая **точность** работы алгоритма на уровне **93%**, что превосходит точность ручного распределения поисковых запросов.
- Сформированы **основные принципы** работы **пользовательского приложения** и создан прототип для тестирования.





**Спасибо за внимание!**

---

Skill Factory. DST-10.  
Никишин Андрей