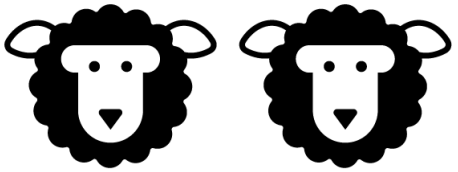
Avito Duplicate Ads Detection



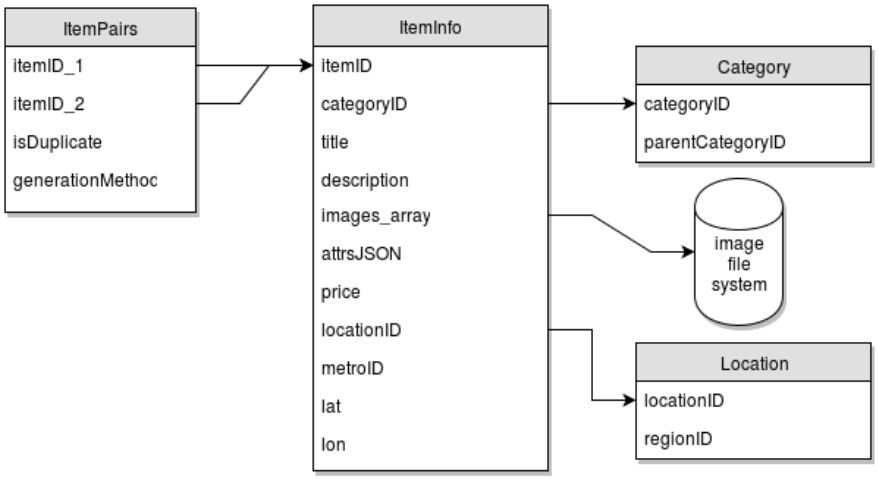
AVITO DUPLICATE
ADS DETECTION

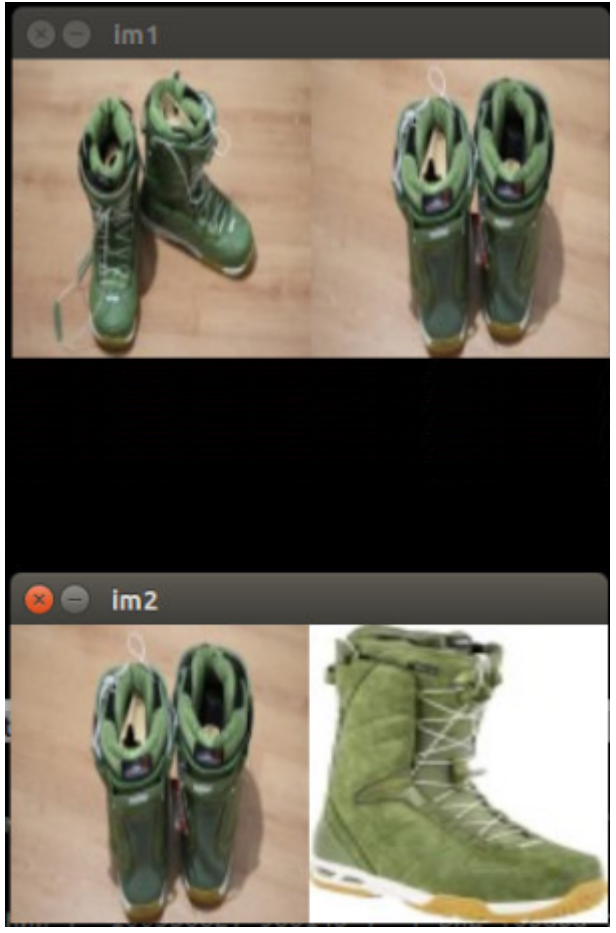
- Семенов Станислав , Цыбулевский Дмитрий, 2016

#	Δrank	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Devil Team * • Stanislav Semenov • u1234x1234	0.95829	162	Mon, 11 Jul 2016 23:09:03 (-0.7h)
2	—	TheQuants * • Sonny Laskar • NoName • anokas • Μαριος Μιχαηλιδης KazAnova	0.95294	197	Mon, 11 Jul 2016 19:53:15 (-46.7h)
3	—	ADAD * • Gerard Toonstra • Kele Xu • Praveen Adepu • Gilberto Titericz Junior • Mario Filho	0.94971	226	Mon, 11 Jul 2016 23:57:54
4	—	8 + 9 = 11 • ZFTurbo • Alexander Vikulin	0.94694	193	Mon, 11 Jul 2016 13:01:32 (-6.3h)
5	—	ololobhi • Abhishek • ololo	0.94587	133	Mon, 11 Jul 2016 22:18:01
6	—	otivA	0.94560	117	Mon, 11 Jul 2016 13:09:48 (-0.1h)
7	—	Native Russian Speakers :P • Evgeny Eltyshev • Georgiy Danshchin	0.94449	43	Mon, 11 Jul 2016 22:49:32 (-1.3h)
8	↑1	frist	0.94438	158	Mon, 11 Jul 2016 21:28:56 (-1h)

Данные

- 2991396 – тренировочная выборка
- 1044197 – тестовая выборка
- 10824317 Изображений





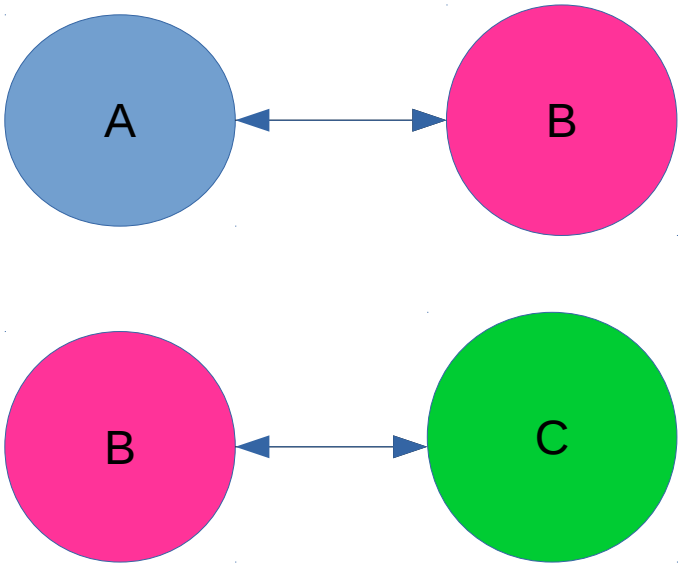
Сноуборд ботинки Nitro Team 10 us

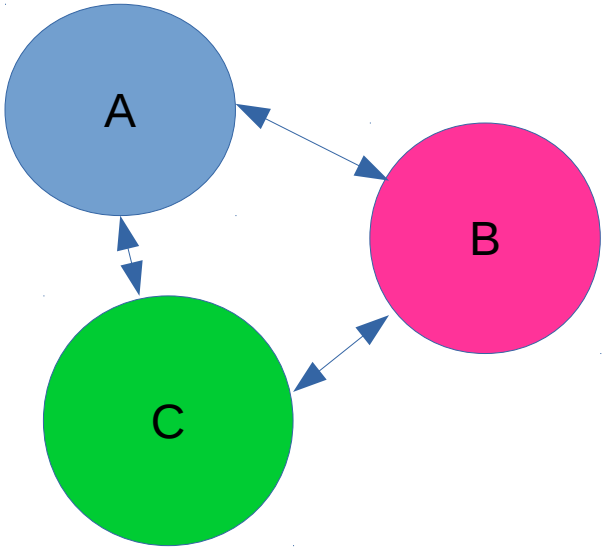
сноубордические ботинки Nitro Team
Размер 42,5
28см, 10 us
новые!!!

Сноубордические ботинки

ботинки Nitro Team
Размер 42,5
28см, 10 us
новые!

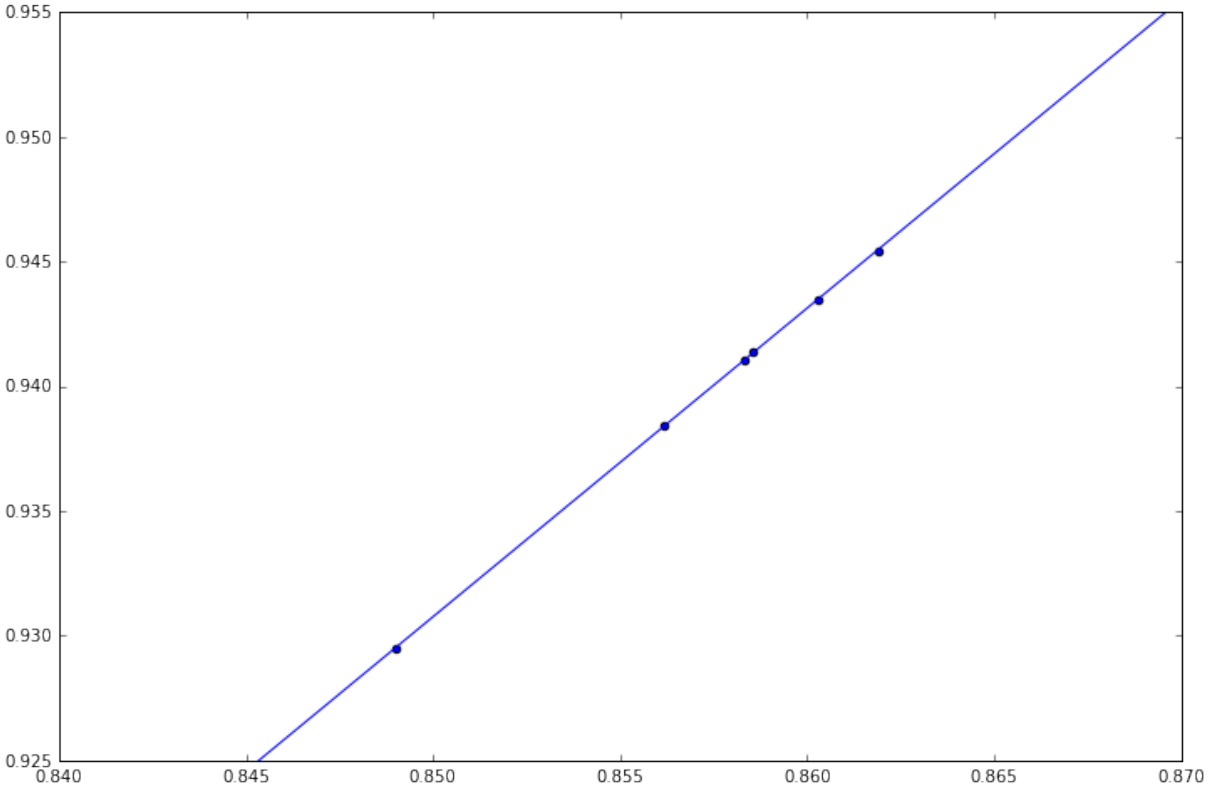
Валидация





Сайлент Режим

Заменяем каждое 10е значение на 0 (минимальное из всех).



По оси x – сайлент режим, по оси y – нормал.

Нужно заметить, что у нас объект – не просто объявление, а пара. Поэтому, не очень хорошо фичи по одному объявлению совать как просто feature_ad1, feature_ad2. Я всегда в одну колонку записывал минимум из этих двух, в другую – максимум из них. Т.е. например, получается минимум(цена двух объяв), максимум(цена двух объяв). В дальнейшем, если в качестве фичи указывается фича именно по одному объявлению, значит используется логика с мин+макс.

Общие фичи

Цена, абсолютная разность цен, относительная разность цен, есть ли цены вообще, широта, долгота, декартово расстояние между городами, есть ли картинки вообще, совпадение широты, совпадение долготы, совпадение метро, совпадение региона, совпадение полностью всех json-аттрибутов, совпадение полностью тайтла, совпадение полностью дескрипшна и т.д.

Числовые фичи

Вычленение всех чисел из тайтла, дескрипшена, тайтла+дескрипшена

Количество чисел в объявлении, количество повторяющихся чисел, относительное количество повторяющихся чисел, количество уникальных чисел, относительное количество уникальных чисел, сравнение медиан наборов чисел двух объявлений и т.д.

Json фичи

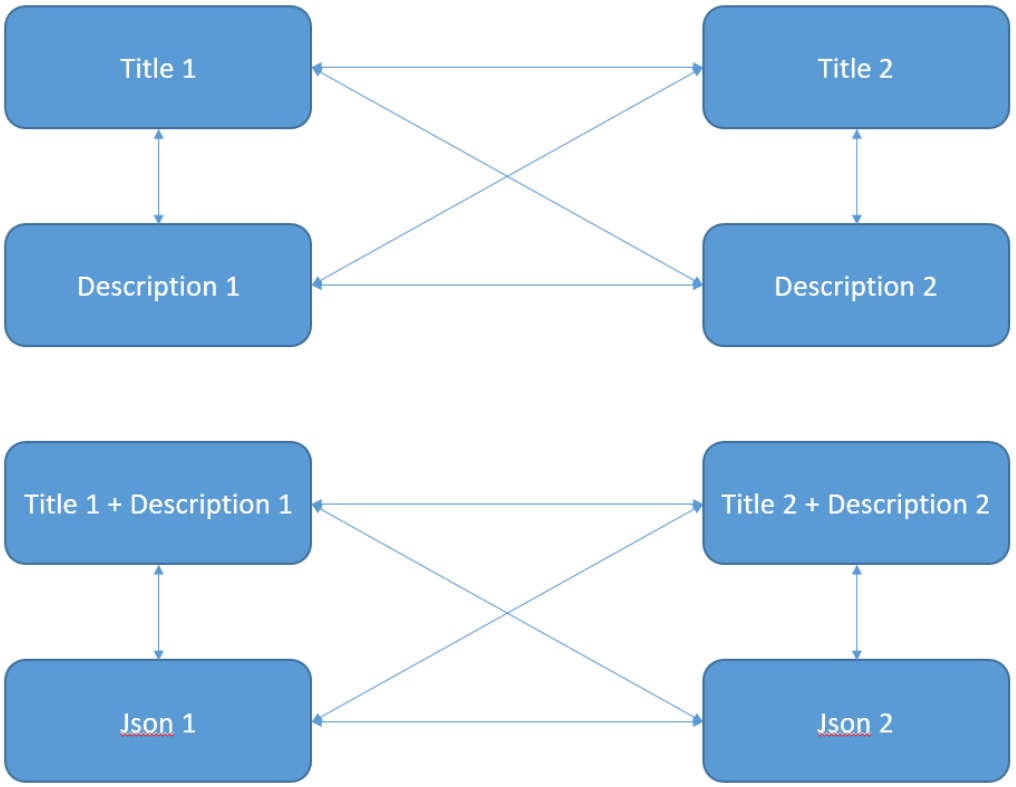
Общее количество json-аттрибутов объявлений, разность этого количества, относительная разность количества, число уникальных json-аттрибутов, число совпадающих json-аттрибутов, относительное число совпадающих json-аттрибутов и т.д.

Текстовые фичи

Наборы: тайтл vs дескрипшн, тайтл + дескрипшн vs json

Сущность: слова, символы, триграммы символов

Косинусная мера между сущностями: тайтл-тайтл, дескрипшн-дескрипшн, тайтл-дескрипшн одного объявления, тайтл-дескрипшн разных объявлений, тайтл+дескрипшн-тайтл+дескрипшн, json-json, тайтл+дескрипшн-json одного объявления, тайтл+дескрипшн-json разных объявлений

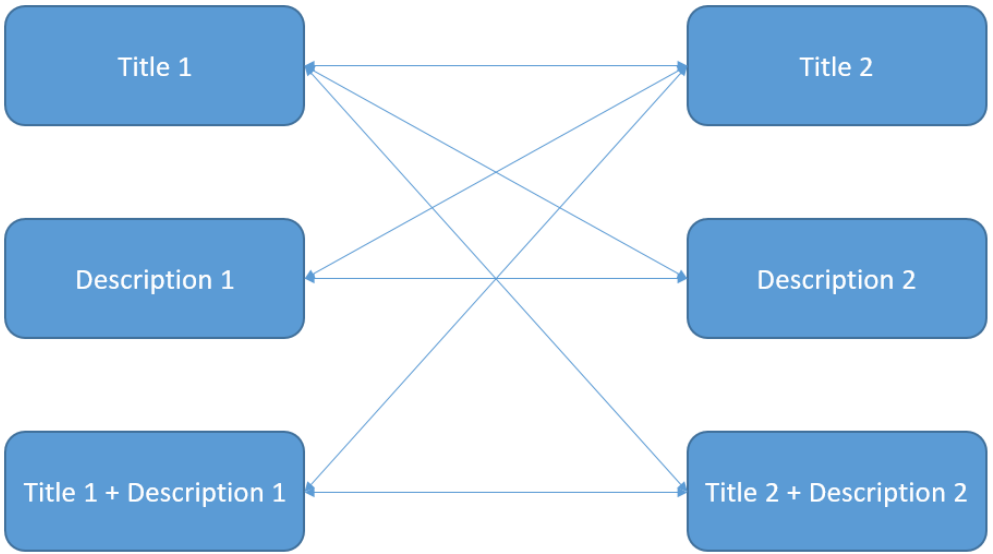


Текстовые фичи

Наборы: тайтл, дескрипшн, тайтл + дескрипшн

Сущность: слова, символы

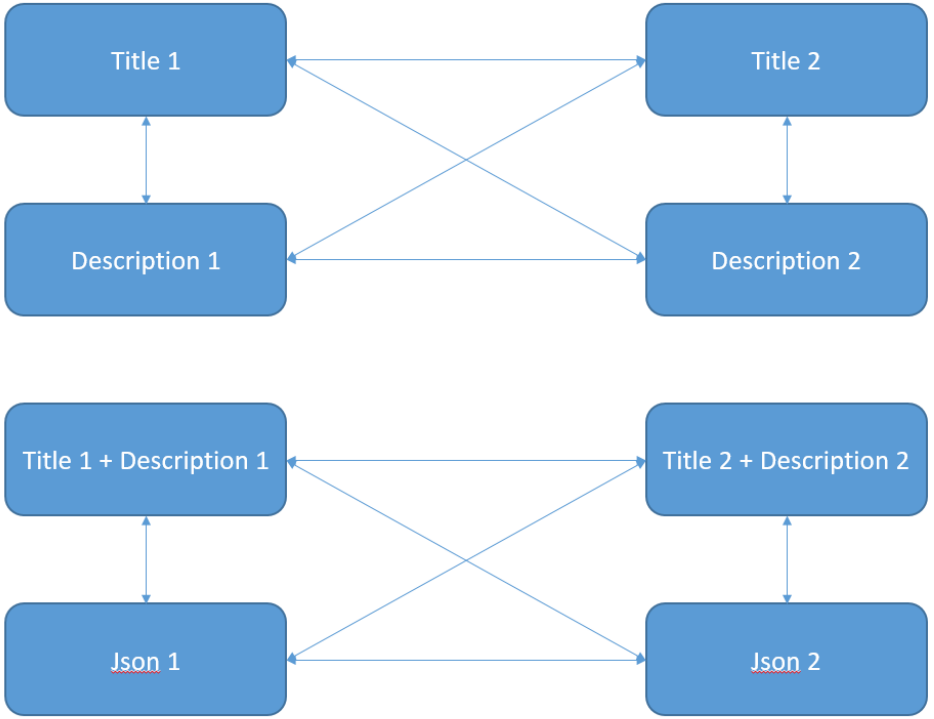
Число сущностей в каждом из наборов, относительное покрытие сущностей тайтла к сущностям тайтл+дескрипшн, абсолютная разность сущностей в каждом из наборов, относительная разность сущностей в каждом из наборов, матричное произведение сущностей каждого из наборов, матричное произведение между тайтлом и дескрипшном одного объявление, матричное произведение между тайтлом и дескрипшном разных объявлений, матричное произведение между тайтлом и тайтлом + дескрипшном разных объявлений, то же, но всё относительное и т.д.



Word2Vec

Наборы: тайтл vs дескрипшн, тайтл + дескрипшн vs json

Обучаем ворд2век на тайтлах+дескрипшнах. Строим n_similarity между: тайтл-тайтл, дескрипшн-дескрипшн, тайтл-дескрипшн одного объявления, тайтл-дескрипшн разных объявлений, тайтл+дескрипшн-тайтл+дескрипшн, json-json, тайтл+дескрипшн-json одного объявления, тайтл+дескрипшн-json разных объявлений



LSI (LSA)

- На объединении объявлений
- На разности объявлений

Текстовые фичи

- Различные расстояния title-title, title-description

  Jaccard, Levenshtein, Cosine, NCD

- Ручной парсинг единиц измерения (Gb)

Текстовые фичи

- Stemming
- Lemmatization
- Transliteration
- Дешевый стемминг: token = token[:-7]

Картиночные фичи

Совпадение хешей картинок при разных степенях сжатий

Картиночные фичи

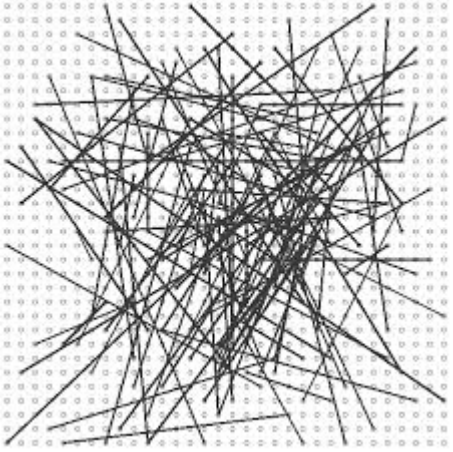
0.9, 0.5, 0, 0.8

Признаки - перцентили

Картиночные фичи

- BRIEF: Binary Robust Independent Elementary Features

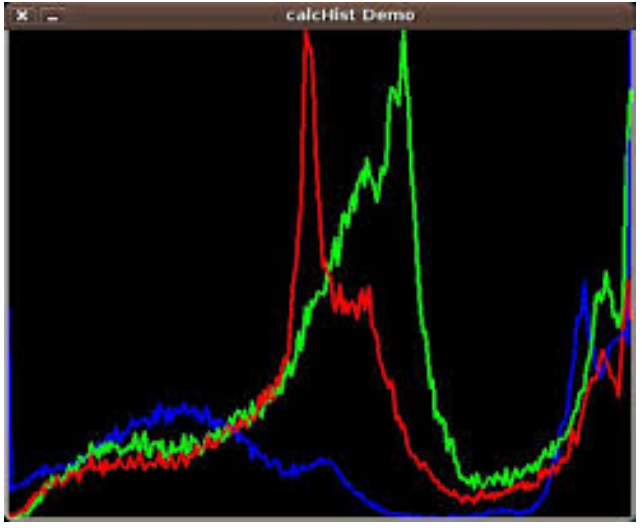
Hamming distance



Картиночные фичи

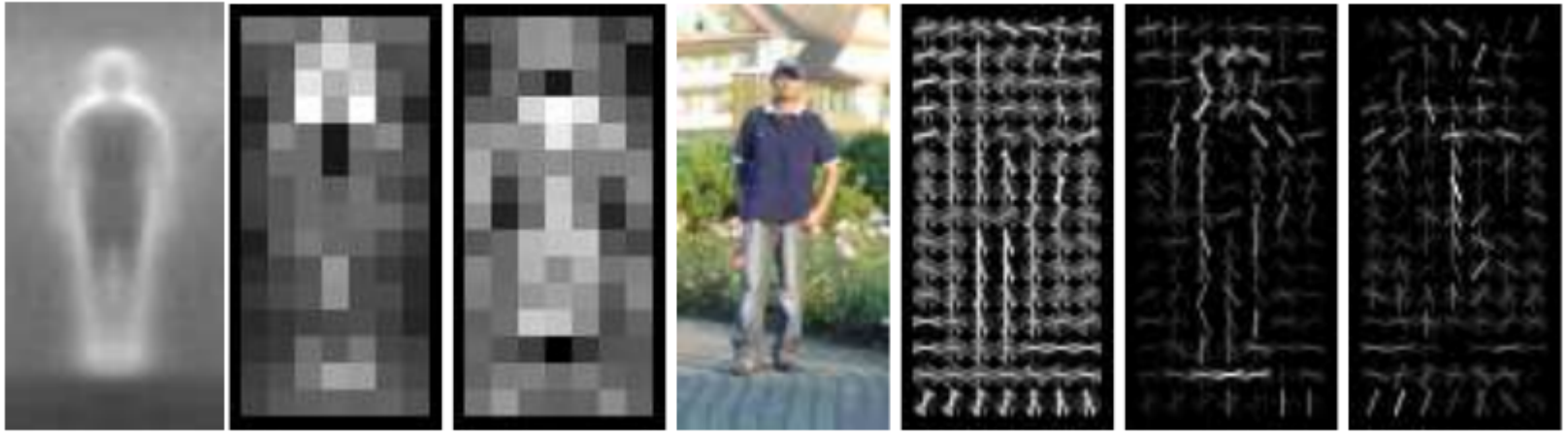
- Цветовые гистограммы

Пространство LAB, HSL
Bhattacharyya distance



Картиночные фичи

- Histogram of Oriented Gradients



Картиночные фичи

- Нейронная Сеть обученная на ImageNet
- Full ImageNet Network
- Batch normalization: Accelerating deep network training by reducing internal covariate shift

Картиночные фичи

- AKAZE

Построенные модели:

1) Различного вида NN
2) RF
3) ET
4) XGBoost
5) XGBoost без DevilFeas
6) XGBoost обученный только на generation_method = 1
7) XGBoost обученный только на generation_method = 3
8) XGBoost обученный только на generation_method = 1 или 3
9) XGBoost обученный на транзитивной выборке (создание всех связей 1го уровня)
10) Xgboost – мультикласс, предсказывающий тип generation_method (3 класса)
11) Xgboost – мультикласс, предсказывающий generation_method-isDuplicate (5 классов)

Что дало основной вклад?
XGBoost, XGBoost обученный только на generation_method = 1 или 3

Модель поверх них - XGBoost

Обучение

- 7 серверов 40 / 160