Тематическая классификация сайтов на основе контента

Разработчик отдела Machine Learning SkyDNS

Аксёнов Александр

SkyDNS

1. Чем занимается отдел ML в нашей компании
2. Сложности мультиязычной классификации
3. Данные – нефть ML и DS
4. Особенности контентной классификации
5. Подведем итоги

SkyDNS

Чем мы занимаемся

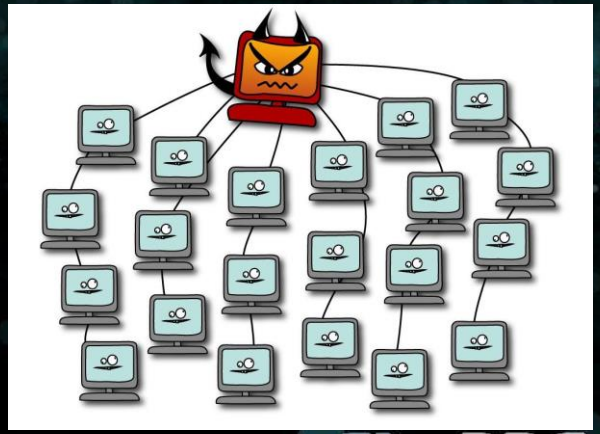
- Классификация интернет сайтов по всей интернет сети.

- На данный момент мы выделяем ***63 категории***.

- Можем предложить клиентам многомиллионную базу доменов по необходимым им категориям, для дальнейшей бизнес-аналитики (более ***109 млн уникальных сайтов***)

- Проводим анализ вредоносной активности в сети интернет: fishing сайты,
сайты распространяющие вирусы, botnets...

- Стараемся сделать интернет более ***структурированным, понятным и безопасным***



SkyDNS

Задачи и цели отдела

Цель проста - классифицировать интернет

Проблематика:

- Сотни миллионов сайтов
- Сотни новых сайтов каждый день
- Смена владельцев и контента сайтов

SkyDNS

Задачи и цели отдела

Решения:

- Открытые источники
- Контентные классификаторы
- Не контентные классификаторы
- Эвристики

SkyDNS

1. Чем занимается отдел ML в нашей компании
2. Сложности мультиязычной классификации
3. Данные – нефть ML и DS
4. Особенности контентной классификации
5. Подведем итоги

SkyDNS

Мультиязычность

1. Почему это проблема?

SkyDNS

Мультиязычность

Распределение языков по версии «Википедия»

Язык	21.07.2019, %	1.01.2019, %	2011, %
Английский	54,0	54,0	57,6
Русский	6,1	6,0	4,8
Немецкий	5,7	6.0	4,7
Испанский	5,0	4.9	6,5
Французский	3,9	4.0	4,6
Японский	3,5	3.4	3,9
Португальский	2,9	2.9	2,0
Итальянский	2,4	2,3	2,1
Персидский	2,0	2,0	4,5
Польский	1,7	1.7	1,4
Китайский	1,6	1,7	1,1

SkyDNS

Мультиязычность
Может быть нам помогут большие братья?

SkyDNS

Мультиязычность

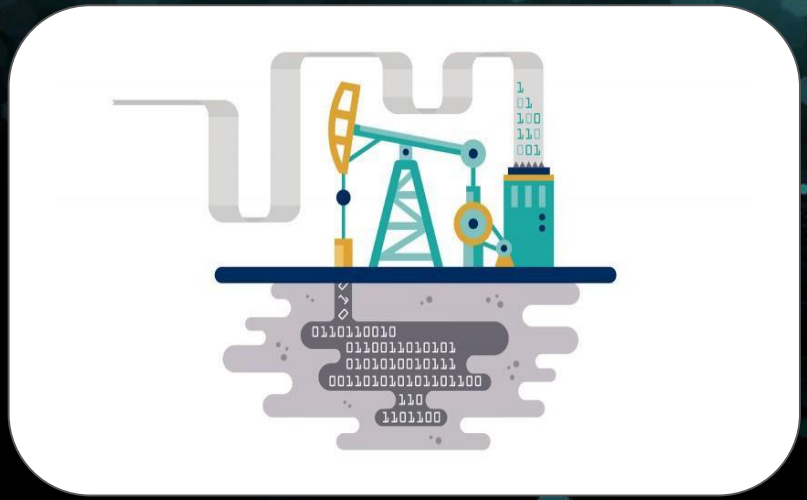
Распространенные библиотеки				
Langdetect	Langid	spacy_cld	**pycld2**	NLTK
медленная, точная	быстрая, плохо работает с романской группой	обертка pycld2	быстрая, приемлемая точность	средняя по скорости, точная
https://pypi.org/project/langdetect/	https://pypi.org/project/langid/	https://github.com/nickdavidhaynes/spacy-cld	https://pypi.org/project/pycld2/	https://pypi.org/project/nltk/

1. Чем занимается отдел ML в нашей компании
2. Сложности мультиязычной классификации
3. Данные – нефть ML и DS
4. Особенности контентной классификации
5. Подведем итоги

SkyDNS

Обучающие данные

Новая нефть и методы ее добычи



Яндекс.Толока

Легкий заработок

YouDoPartner

amazon

SkyDNS

Обучающие данные
Ручная разметка

1. Преимущества
2. Недостатки
3. Сухая статистика

2-3 недели на классификатор

Восстановительный период

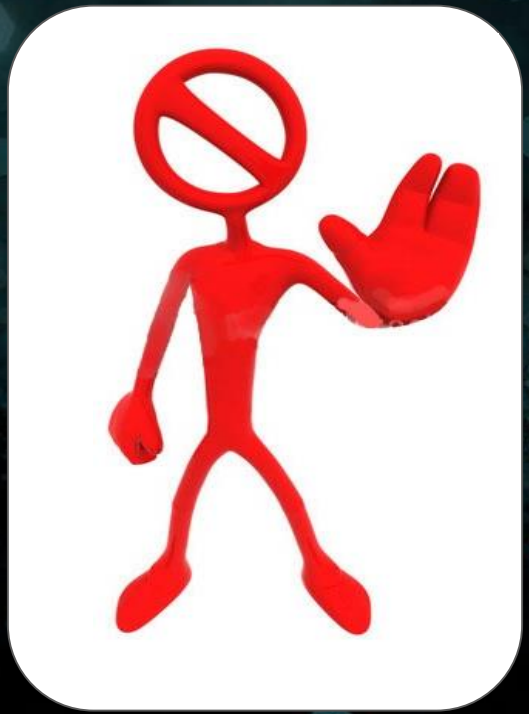
11 проектов за **9** месяцев

Анализ графов

Анализ логов запросов

Анализ взаимодей ствия

Исследова ние структур сайтов



SkyDNS

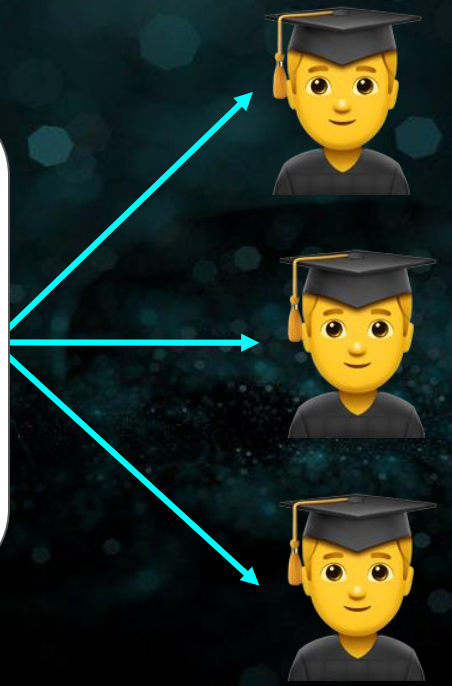
Обучающие данные
Всегда есть лучшее решение!

1. Совместим преимущества FreeLance и собственного отдела
2. Как это работает?
3. Результаты

3-4 дня на классификатор

Время на аналитику и новые проекты

19 проектов за **6** месяца



SkyDNS

1. Чем занимается отдел ML в нашей компании
2. Сложности мульти язычной классификации
3. Данные – нефть ML и DS
4. Особенности контентной классификации
5. Подведем итоги

SkyDNS

Контентная классификация
Обучающие выборки

1. Сбалансированность - это еще не все...
2. Когда высокое качество на тесте может сильно подвести в проде

Политические движения + Рестораны −

Сайты предвыборных кампаний + Фильмы −

Сайты партий и депутатов + Музыка −

НО!

?

SkyDNS

Контентная классификация
Алгоритмы не работают со словами

1. Machine Learning - это о числовых данных
2. Как сделать из слов цифру?
3. Как можно подготовить текст к превращению?

SkyDNS

Контентная классификация
Предварительная обработка

1. Приведение данных к единому регистру

   ТеЛеКом ≠ телеком

1. Очистка от знаков пунктуации и других символов

   嗨, مرحبا,ہیلو

1. Создаем вектор слов. Токенизация (NLTK)

   “на улице был чудесный день” ⟶ [“на”, “улице”, “был”, “чудесный”, “день”]

Контентная классификация
Предварительная обработка

1. Приведение к нормальной форме

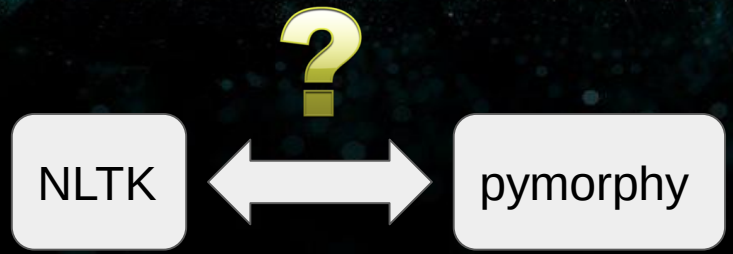
[“на”, “улице”, “был”, “чудесный”, “день”] - было

[“на”, “улиц”, “был”, “чудесн”, “ден”] - NLTK SnowballStemmer()

[“на”, “улица”, “быть”, “чудесный”, “день”] - pymorphy2

1. Удаление stop-слов

[“улица”, “чудесный”, “день”]



SkyDNS

Контентная классификация
Векторное представление и векторные модели

1. Все сделано до нас.
2. Чье решение выбрали мы.

scikit-learn	gensim	facebook	microsoft
			

Контентная классификация
Векторное представление

1. Dictionary

[“на”, “улица”, “быть”, “чудесный”, “день”, “быть”, “день”]

{0: быть, 1: день, 2: на, 3: улица, 4: чудесный} - словарь

1. bow2vec

[(0, 2), (1, 2), (2, 1), (3, 1), (4, 1)]

1. Tf-Idf (Term Frequency — Inverse Document Frequency)

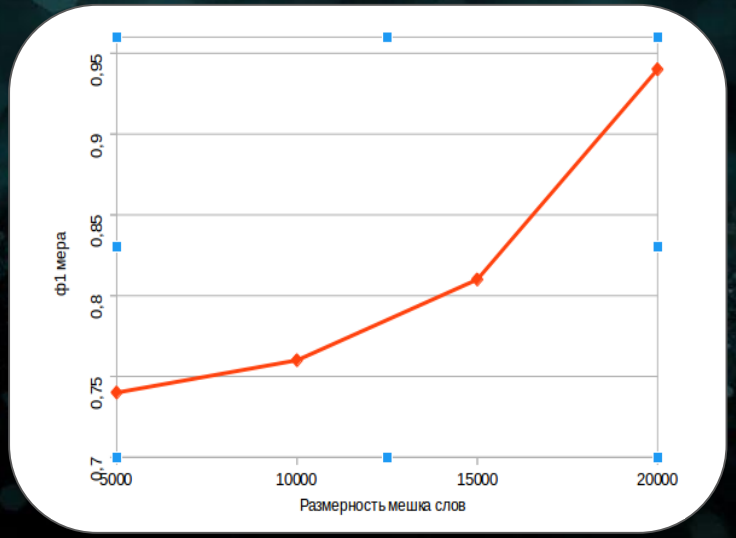
[(1, 0.582), (2, 0.291), (3, 0.701), (4, 0.291)]

Контентная классификация
Используем мешок слов!

При тренировке качество f1=0.94

НО!

- громадные вектора
- долго считаем
- Снижение размерности (PCA) не помогло. Возникли проблемы с точностью



SkyDNS

Контентная классификация
Векторные модели

1. LSI (SVD)
2. LDA
3. Word2Vec
4. Doc2Vec
5. FastText

Пример LDA вектора

'0.006*"iphone" + 0.006*"визитка" + 0.006*"полиграфия" + 0.004*"типография" + 0.004*"листовка" + 0.004*"макет" + 0.004*"широкоформатный" + 0.003*"буклет" + 0.003*"офсетный" + 0.003*"тираж"'

SkyDNS

Контентная классификация
Наш выбор и некоторые особенности работы

1. Одна модель хорошо, но три лучше
2. Какой алгоритм способен извлечь максимум результата?

	Tfidf	LDA	LSI	Doc2Vec	All
RandomForest	0.909	0.900	0.915	0.886	***0.931***
SVM	0.902	0.887	0.907	0.858	0.899
DecisionTree	0.889	0.871	0.818	0.645	0.856
LogisticRegression	0.896	0.886	0.892	0.862	0.881
NearestNeighbours	0.88	0.882	0.881	0.729	0.724
NaiveBayes	0.86	0.825	0.710	0.634	0.876

Контентная классификация
Немного о больших данных

1. На чем учить модели?
2. Преимущества более сложных моделей

TfIdf - длина вектора 574667
Время работы - 8-10 часов

vs

LDA + LSI + FastText - длина вектора 800
Время работы - 3-4 часа

	Tfidf	LDA	LSI	Doc2Vec	All
RandomForest	0.909	0.900	0.915	0.886	***0.931***

1. Чем занимается отдел ML в нашей компании
2. Сложности мультиязычной классификации
3. Данные – нефть ML и DS
4. Особенности контентной классификации
5. Подведем итоги

SkyDNS

Подведение итогов
Методы валидации и системы весов

1. Ступенчатый подход к валидации
2. Выбор доверительной отсечки
3. Несколько градаций доверия
4. Обратная связь
5. Production!!!

SkyDNS

Подведение итогов
Немного проблематики

1. Рост количества классификаторов - рост времени выполнения
2. Вопрос работы с контентом из менее распространенных языковых зон
3. Своевременный анализ актуальности доменов в базе
4. Больше метрик качества алгоритмов классификации и самой базы

SkyDNS

SkyDNS