Про поисковую систему и машинное обучение

Цели и вопросы

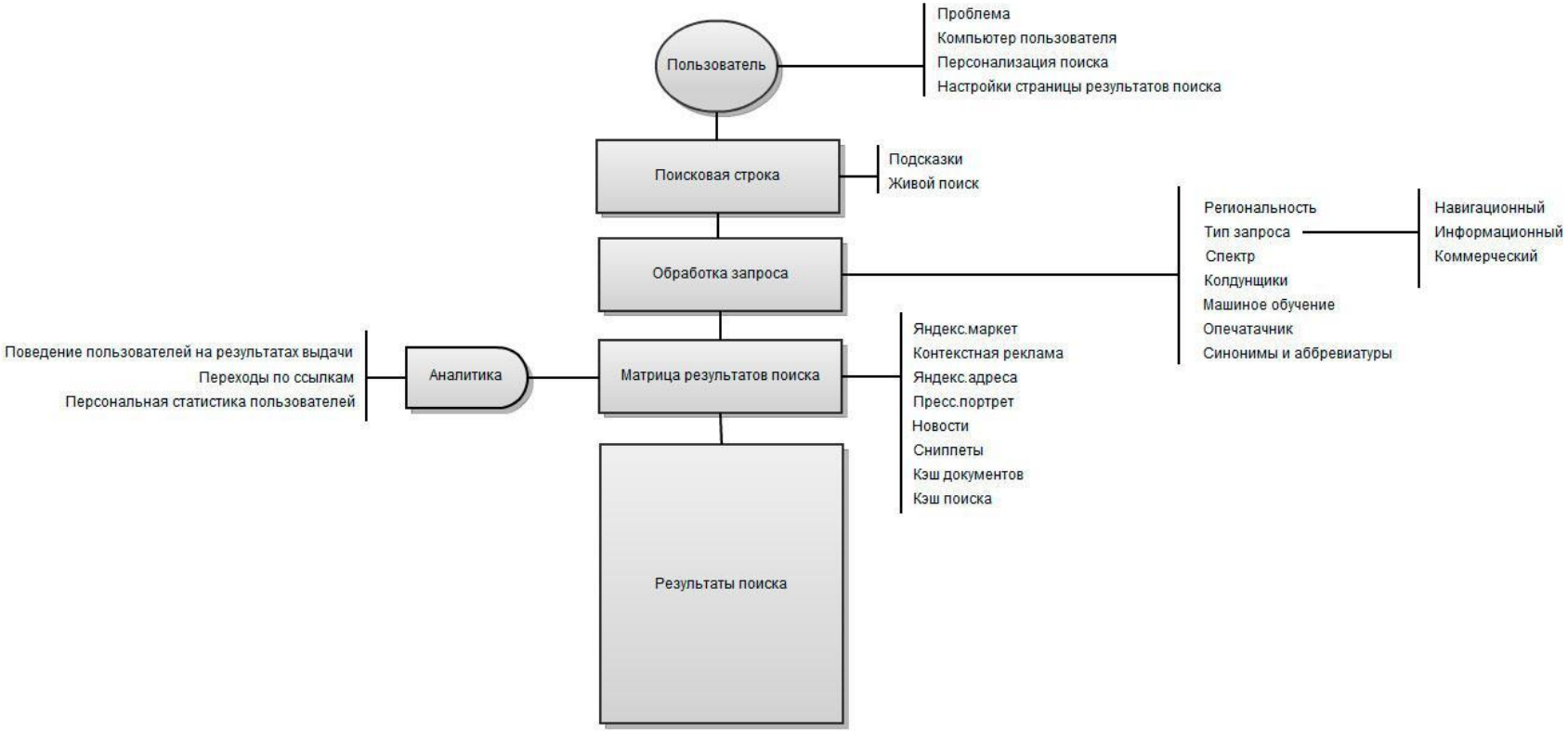
1. Мои цели
2. Переход из мира магии в мир науки
3. Станет понятней как работает алгоритмы поиска, мо и в какую сторону смотреть
4. Наконец-то узнать что LSA - это не частотный словарь)))
5. Накопление опыта

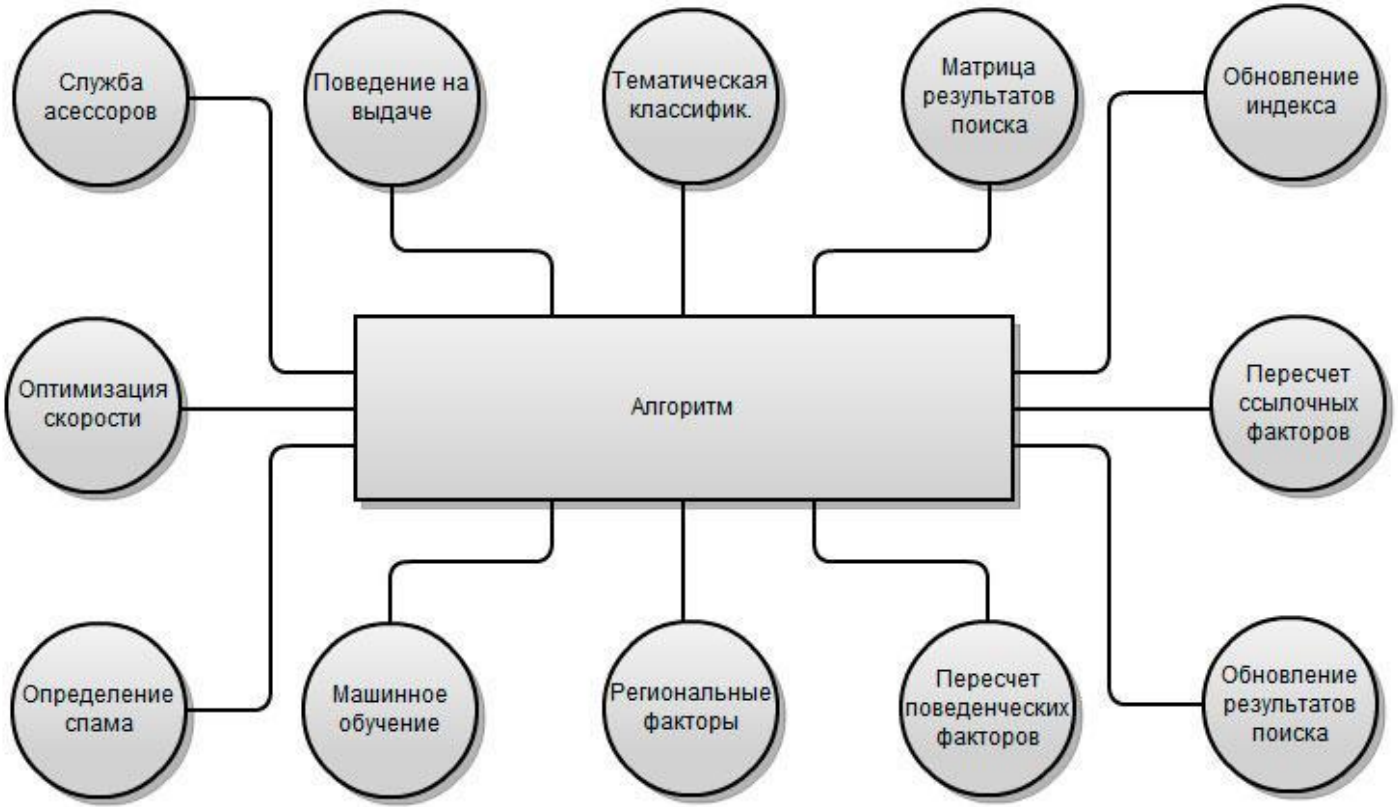
Чего не будет

1. Я не могу за 2 часа сделать вас Data Science спецами
2. Четких рецептов из мира магии
3. Ответов на вопросы: "Какое нужно количество вхождений", "Сколько нужно купить ссылок..." и т.д.
4. Мы сделали какую-то фигню, смотрите какой красивый гр

Как устроен поиск?


MAGIC
HISTORY.COM
risovach.ru


Пользователь
Проблема
Компьютер пользователя
Персонализация поиска
Настройки страницы результатов поиска
Поисковая строка
Подсказки
Живой поиск
Обработка запроса
Региональность
Тип запроса
Спектр
Колдунщики
Машиное обучение
Опечатчник
Синонимы и аббревиатуры
Навигационный
Информационный
Коммерческий
Аналитика
Поведение пользователей на результатах выдачи
Переходы по ссылкам
Персональная статистика пользователей
Матрица результатов поиска
Яндекс.маркет
Контекстная реклама
Яндекс.адреса
Пресс.портрет
Новости
Сниппеты
Кэш документов
Кэш поиска
Результаты поиска


Служба асессоров
Поведение на выдаче
Тематическая классифик.
Матрица результатов поиска
Обновление индекса
Оптимизация скорости
Алгоритм
Пересчет ссылочных факторов
Определение спама
Машинное обучение
Региональные факторы
Пересчет поведенческих факторов
Обновление результатов поиска

Ранжирование

Ранжирование - процесс упорядочивания документов в соответствии со степенью их соответствия поисковому запросу.

Ранжирующие признаки

Запросо-независимые или статические признаки — зависящие только от документа, но не от запроса. Например, PageRank или длина документа.

Признаки, зависящие только от запроса. Например, «запрос про порно или нет».

Запросо-зависимые или динамические признаки — зависящие и от документа, и от запроса. Например, TF-IDF.

Как происходит поиск

- Запрос токенизируется, к словам запроса применяется морфологический анализ, ищутся синонимы
- Из документов индекса отбираются те, которые с большой вероятностью отвечают на запрос
- Для отфильтрованных документов рассчитываются признаки (фичи)
- К признакам применяется формула, дающая конечную оценку релевантности

Общая функция релевантности

Функция релевантности документа d относительно запроса q

$$fr(q,d)=a_1h_1(q,d)+a_2h_2(q,d)+...+a_nh_n(q,d)$$

количество функций $h_k(q,d)$ достаточно большое, десятки тысяч.
Коэффициенты a_k – малые величины.

$h_k(q,d)$ - мономы факторов

Качество поиска

Ассесоры нужны не для ручного управления, а для оценки качества алгоритма

Определяют фичи

Постоянное переобучение

Типы запросов и регионы

Выводы

Среднее по топу не всегда покажет порог релевантности

Никто не знает какие факторы как влияю на конкретный топ

Все факторы важны, один может помочь вытянуть другой

Нет смысла считать все факторы, если релевантных документов мало

Факторы дают + или - в ранжирование

Ссылочный антиспам

Вероятность что текст анкора коммерческий

Вероятность что сайт продает ссылки

Вероятность что сайт покупает ссылки

Тематика дорона-акцептора

Длинна текста в блоке

и т.д.

Текстовый антиспам

ПФ на странице с текстом

Вероятность встретить слово в тексте

Тематический вектор текста документа и сайта

Статические признаки спама(сжимаемость текста, количество знаков препинания и т.д)

Перечисление запросов и пр. фичи.

Поведенческие факторы и антиспам

Поведение в топе

Поведение на странице и хосте

Ласт-клик.

Тематические фичи

Количество и частота новыйх объявлений - для класифайдов

Наличие интентов на странице(купить, скачать и т.д.)

Общая релевантность сайта запросу

Для авто-сайтов не нужна кнопка "купить"

Про поиск и МО

Антиспам построены на МО

У яндекса алгоритм ранжированя работает на МО

Все задачи по кластеризации и классификации текстов - МО

Типы задач

Классификация
Кластеризация
Регрессия
Понижение размерности данных
Восстановление плотности распределения вероятности по набору данных
Одноклассовая классификация и выявление новизны
Построение ранговых зависимостей

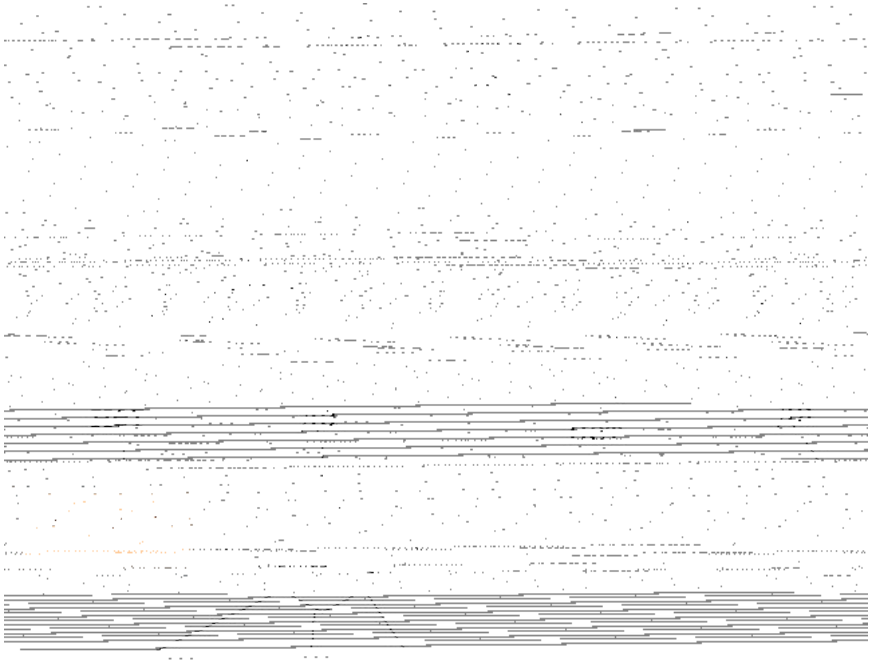
Класическая задача: Кредитный скоринг

Объект - человек

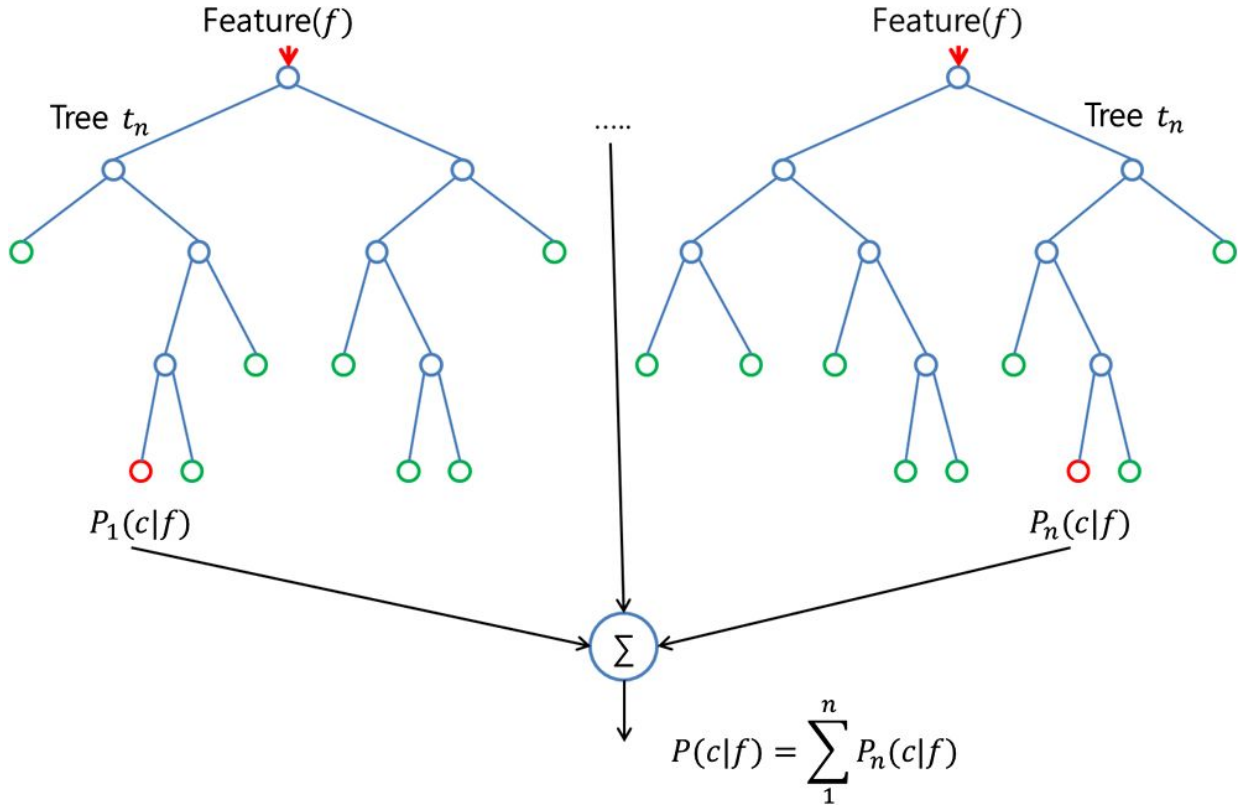
доход, есть квартира, есть жена у которой есть машина и т.д. - признаки

Задача: Найти вероятность того что клиент вернет кредит.

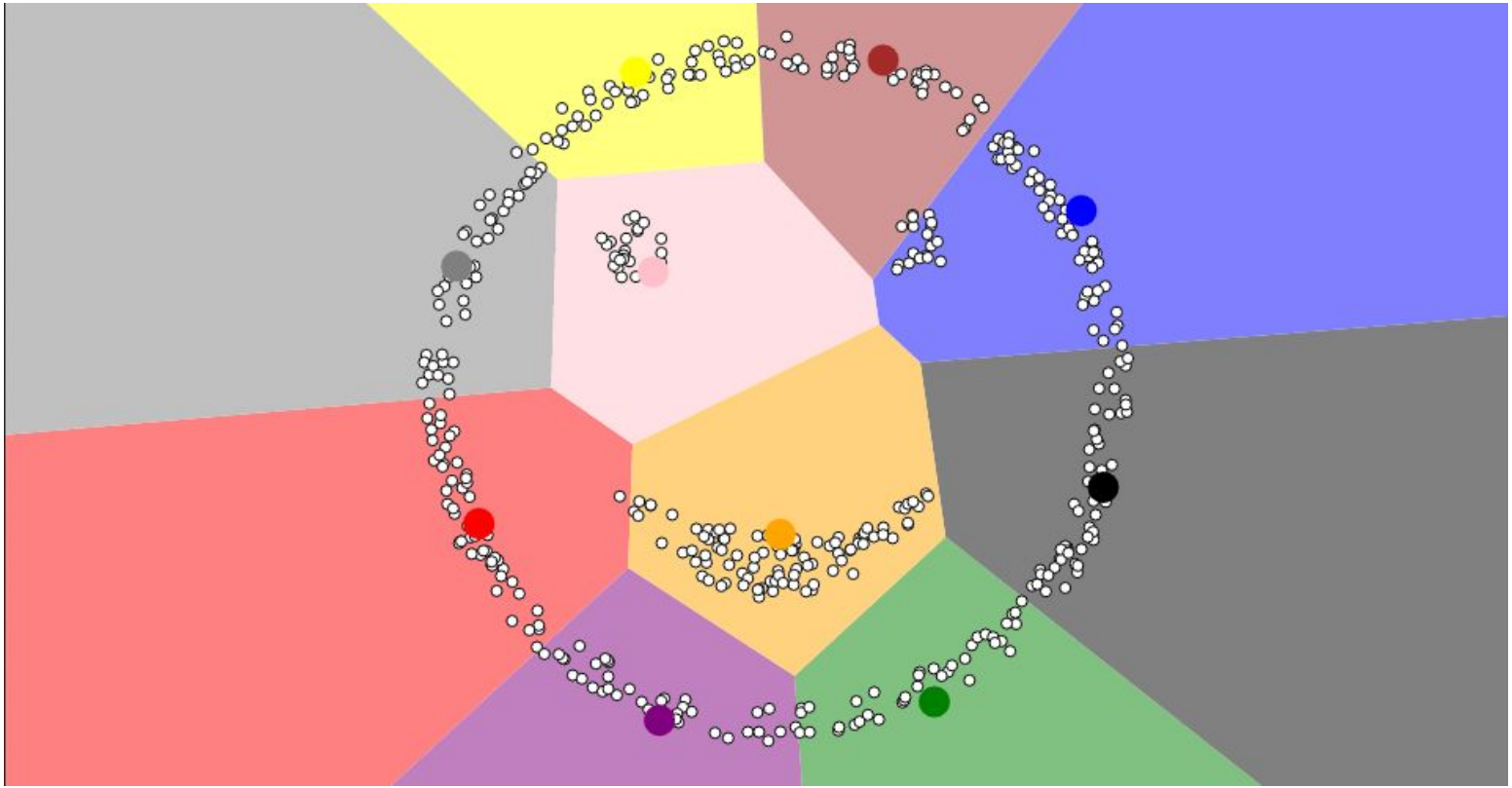
Алгоритмы. Деревья решений



Random Forest



k-means



Коллаборативная фильтрация

МО для текстов

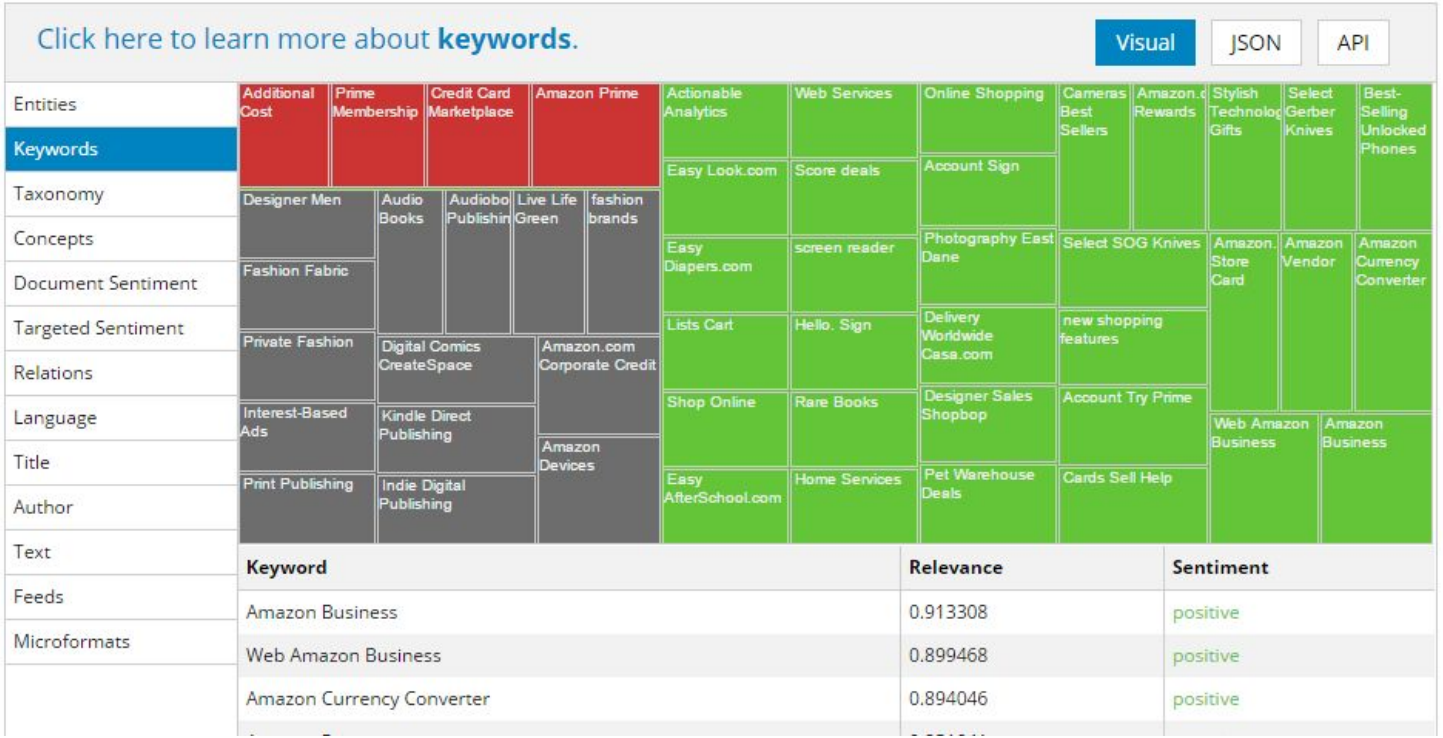
Катеригоризация

Кластеризация

Таксономия

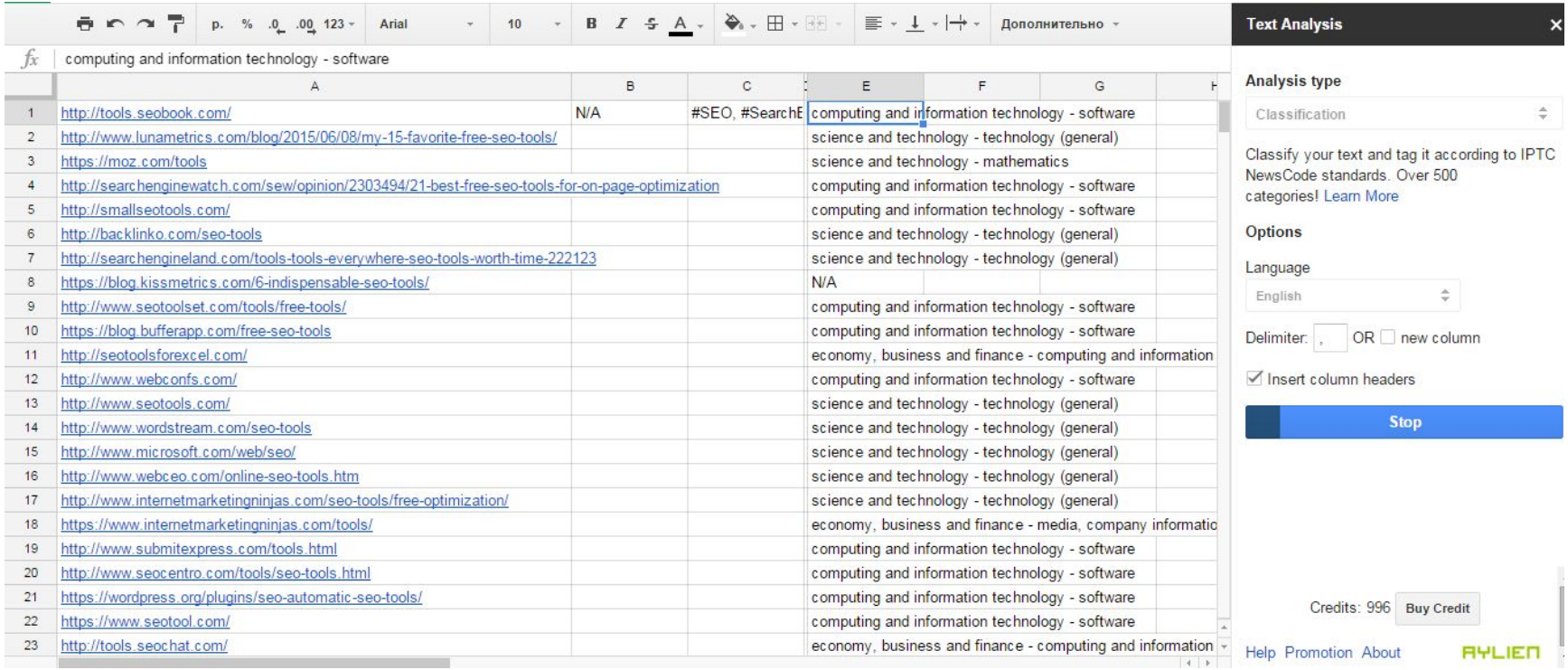
Классификация

AlchemyLanguage API



http://www.alchemyapi.com/products/demo/alchemylanguage

Классификация текстов



	A	B	C	E	F	G
1	http://tools.seobook.com/	N/A	#SEO, #SearchE	computing and information technology - software		
2	http://www.lunametrics.com/blog/2015/06/08/my-15-favorite-free-seo-tools/			science and technology - technology (general)		
3	https://moz.com/tools			science and technology - mathematics		
4	http://searchenginewatch.com/sew/opinion/2303494/21-best-free-seo-tools-for-on-page-optimization			computing and information technology - software		
5	http://smallseotools.com/			computing and information technology - software		
6	http://backlinko.com/seo-tools			science and technology - technology (general)		
7	http://searchengineland.com/tools-tools-everywhere-seo-tools-worth-time-222123			science and technology - technology (general)		
8	https://blog.kissmetrics.com/6-indispensable-seo-tools/			N/A		
9	http://www.seotoolset.com/tools/free-tools/			computing and information technology - software		
10	https://blog.bufferapp.com/free-seo-tools			computing and information technology - software		
11	http://seotoolsforexcel.com/			economy, business and finance - computing and information		
12	http://www.webconfs.com/			computing and information technology - software		
13	http://www.seotools.com/			science and technology - technology (general)		
14	http://www.wordstream.com/seo-tools			science and technology - technology (general)		
15	http://www.microsoft.com/web/seo/			science and technology - technology (general)		
16	http://www.webceo.com/online-seo-tools.htm			science and technology - technology (general)		
17	http://www.internetmarketingninjas.com/seo-tools/free-optimization/			science and technology - technology (general)		
18	https://www.internetmarketingninjas.com/tools/			economy, business and finance - media, company informatic		
19	http://www.submitexpress.com/tools.html			computing and information technology - software		
20	http://www.seocentro.com/tools/seo-tools.html			computing and information technology - software		
21	https://wordpress.org/plugins/seo-automatic-seo-tools/			computing and information technology - software		
22	https://www.seotool.com/			computing and information technology - software		
23	http://tools.seochat.com/			economy, business and finance - computing and information		

http://docs.aylien.com/docs/addon-introduction

Пример. Как найти похожие документы

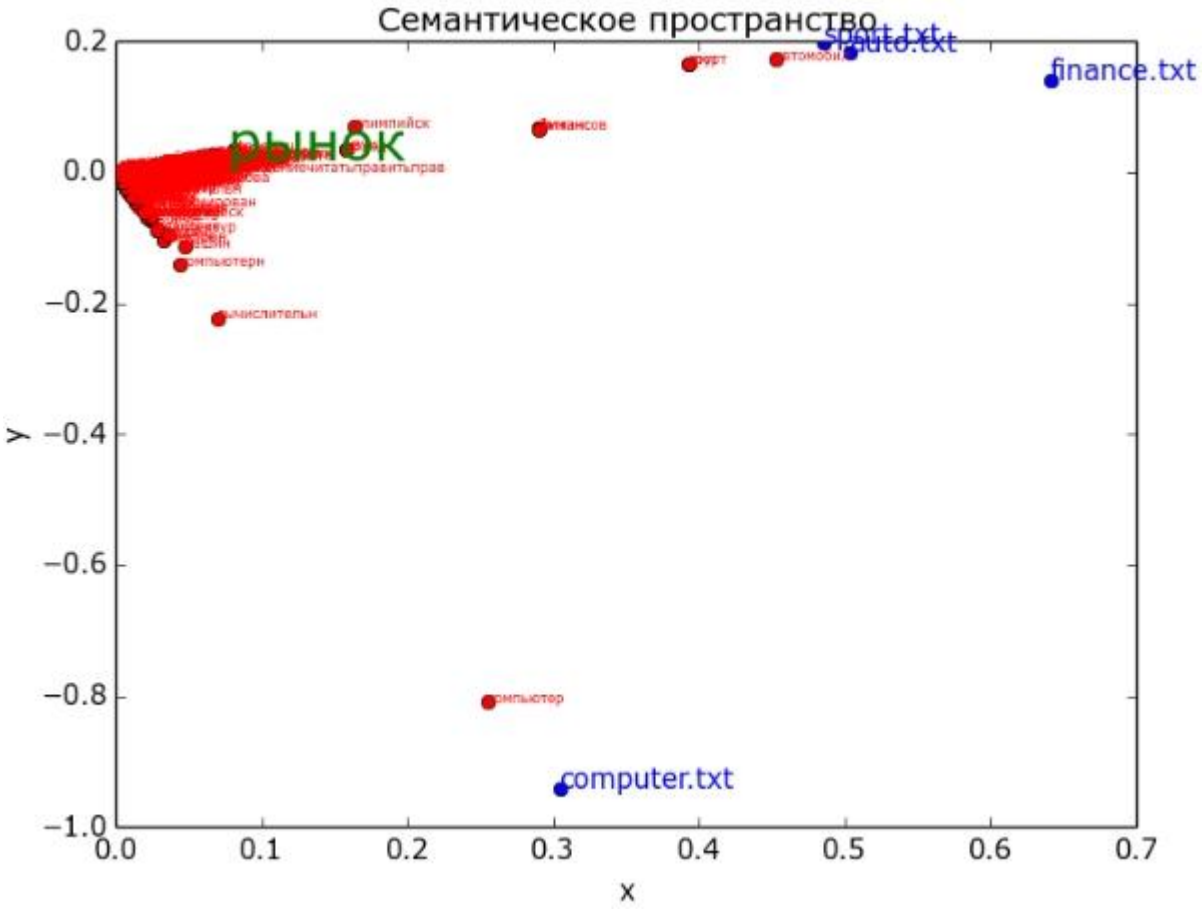
Пацаны сказали что LSA - это круто.

(на самом деле нет, Дирихле - рулит)

LSA?

1. Как работает:
    a. удаление стоп-слов, стемминг или лемматизация слов в документах;
    b. исключение слов, встречающихся в единственном экземпляре;
    c. построение матрицы слово-документ (бинарную есть/нет слова, число вхождений или tf-idf);
    d. разложение матрицы методом SVD (A = U * V * WT);
    e. выделение строк матрицы U и столбцов W, которые соответствуют наибольшим сингуляр ным числам (их может быть от 2-х до минимума из числа терминов и документов). Конкретное количество учитываемых собственных чисел определяется предполагаемым

LSA


Семантическое пространство
рынок
sport.txt
auto.txt
finance.txt
computer.txt
x
y
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Пример. Как найти похожие документы

LSA:

На выходе получаем координаты в тематическом пространстве

LDA:

На выходе получаем вероятность принадлежности слова к тематике, и тематики пренадлежащие документам

Как найти схожие документы. Обучение. LDA

```
77 76 0.059*азербайджан + 0.054*армения + 0.047*армянский + 0.038*азербайджанский + 0.029*divot + 0.019*армянин + 0.014*азербайджанец + 0.014*ереван +
   0.011*албания + 0.009*алиев
78 77 0.029*здание + 0.027*памятник + 0.014*строительство + 0.013*построить + 0.010*архитектор + 0.009*жила + 0.009*архитектура + 0.008*дворец + 0.007*комплекс +
   0.007*городской
79 78 0.073*серия + 0.055*сезон + 0.041*персонаж + 0.031*эпизод + 0.030*сериал + 0.022*flagicon + 0.022*аниме + 0.018*comics + 0.014*комикс + 0.014*anime
80 79 0.084*сельсовет + 0.039*республика + 0.039*башкортостан + 0.016*башкирский + 0.015*герб + 0.015*красный + 0.014*деревня + 0.013*посёлок + 0.012*белгородский
   + 0.012*калининградский
81 80 0.020*войско + 0.016*армия + 0.013*противник + 0.012*битва + 0.011*сражение + 0.010*отряд + 0.007*военный + 0.007*солдат + 0.007*захватить + 0.007*потеря
82 81 0.021*религия + 0.019*религиозный + 0.011*христианство + 0.010*христианский + 0.009*праздник + 0.009*церковь + 0.009*христос + 0.008*библия + 0.007*учение +
   0.007*иисус
83 82 0.088*волость + 0.048*губерния + 0.024*житель + 0.024*уезд + 0.022*деревня + 0.019*files + 0.016*крымский + 0.015*двор + 0.012*узбекистан + 0.011*крым
84 83 0.036*государство + 0.027*международный + 0.019*республика + 0.014*организация + 0.014*отношение + 0.013*европейский + 0.012*правительство + 0.010*политика
   + 0.008*договор + 0.008*независимость
```

Пример обучения:

http://pastebin.com/PMrCAQpz

Мера схожести

1. Косинусная мера
2. Коэффициент корреляции Пирсона
3. Евклидово расстояние
4. Коэффициент Танимото
5. Манхэттенское расстояние и т.д.

Пример работы

Таблица схожести	
URL	LDA (≥ 0,01)
http://rozetka.com.ua/forte_ew1230_37850/p294041/	0,9999999835
http://rozetka.com.ua/intertool_dt_9140/p229434/	0,6353030427
http://rozetka.com.ua/poctex_bpt_1402_38403/p303691/	0,4734146187
http://kiev.all.biz/betonomeshalka-betoniar-hcm-65-g8889466	0,1130037189
http://kishinev.all.biz/betonosmesitel-bro-350-500-g149523	0,0905475162
http://dergachi.all.biz/smesi-betonnye-betonosmesiteli-kupit-zakazat-g2502279	0,0880606089
http://www.nl.ua/ru/otdelochnye_materialy/smesi/plastifikatory/dobavka_dlya_teplogo_pola_Coral_Mas	0,0779477065
http://www.nl.ua/ru/stroimaterialy/stroitelnye_smesi/tsement/tsement_shpts_m_400_25_kg_otsz.html	0,0776410032
http://chelyabinsk.all.biz/betonosmesitel-dizelnyj-sbr-1200a-so-skipom-g1541229	0,0767527349
http://www.nl.ua/ru/instrumenty/eilektroinstrument/perforatory/perforator_zenit_zpp_1000_2.html	0,0739039488
http://www.nl.ua/ru/stroimaterialy/stroitelnye_smesi/tsement/tsement_kredo_m_400_5_kg.html	0,0729614846
http://www.nl.ua/ru/instrumenty/eilektroinstrument/perforatory/20111410.html	0,0717495587
http://www.nl.ua/ru/otdelochnye_materialy/germetiki/germetiki/germetik_Baumit_Baumasol_silikonovyi	0,0715833345
http://www.nl.ua/ru/otdelochnye_materialy/klei/klei_dlya_plitki/90314012.html	0,0704553144
http://intertool.ua/catalog/instrument-dlya-otdelochnih-rabot/shpateli/shpateli-iz-nerzhaveyushchey-stali	0,0630298777
http://proforma.com.ua/root/betonosmes.php	0,0388188289

Gensim


gensim
topic modelling for humans
Download
latest version from the Python Package Index
Direct install with:
easy_install -U gensim
Home
Tutorials
Install
Support
API
About
>>> from gensim import corpora, models, similarities
>>>
>>> # Load corpus iterator from a Matrix Market file on disk.
>>> corpus = corpora.MmCorpus('/path/to/corpus.mm')
>>>
>>> # Initialize Latent Semantic Indexing with 200 dimensions.
>>> lsi = models.LsiModel(corpus, num_topics=200)
>>>
>>> # Convert another corpus to the latent space and index it.
>>> index = similarities.MatrixSimilarity(lsi[another_corpus])
>>>
>>> # Compute similarity of a query vs. indexed documents
>>> sims = index[query]
Gensim is a FREE Python library
Scalable statistical semantics
Analyze plain-text documents for semantic structure
Retrieve semantically similar documents

Сложность фраз. Задача и ограничения

Ограничения:

1. Нужно посчитать за вменяемое время "сложность" для ~ 100 млн фраз.
2. Нельзя заходить на страницы
3. Нет ссылочного профиля

Задача:
Найти переменную которая выражает сложность продвижения фразы.

Сложность фразы. Параметры.

SEO-score = вхождение фразы в разные участки снипета.

Вспомагательные параметры:

- число главных страниц
- число результатов в выдаче
- длина фразы + число подсказок и запросов в базе
- сила домена
- частота запроса
- стоимость клика в контексте
- конкуренция в контексте

Сложность фраз

1. Поиск признаков
2. Разметили обучающую выборку
3. Отдельно обучили SEO-score
4. Нормализовали другие параметры и обучили
5. Проверили результат на 2 выборках НК и ВК
6. Обучали пока небыло заметной разницы между НК и ВК
7. Еще раз проверили на другой группе

Сложность фразы

type	kw	se	tscore	sigma	lin	top_si	seo_s	main_	result	ph_len	doma	sugg	query	cost	conc	keyw	norm_	norm_	norm_	norm_	norm_	norm_	norm_	conc
mix	arthritis	g_us	97,90	98,51	96,49	10,00	845,0	2,00	7390	1,00	84,57	2431	2010	2,69	59,00	5490.	84,50	54,00	99,93	98,83	84,57	97,45	83,64	59,00
mix	cars	g_us	97,30	98,37	95,51	10,00	860,0	4,00	1930	1,00	79,13	1890	1220	1,24	39,00	1947	86,00	90,00	100,0	100,0	79,13	99,57	70,20	39,00
hard	форекс	g_ru	96,41	98,15	94,05	10,00	760,0	5,00	9300	1,00	65,59	1113	1100	4,28	84,00	4590.	76,00	100,00	99,44	97,83	65,59	95,43	89,05	84,00
mix	divorce	g_us	95,30	97,83	92,25	10,00	780,0	1,00	1790	1,00	74,49	4107	6050	4,38	90,00	1305	78,00	30,00	99,97	99,74	74,49	92,00	89,27	90,00
hard	forex	g_ru	95,03	97,75	91,81	10,00	755,0	8,00	1580	1,00	61,11	7024.	4950	4,58	90,00	3329.	75,50	100,00	99,97	95,84	61,11	90,39	89,69	90,00
mix	anxiety	g_us	94,63	97,62	91,17	10,00	845,0	1,00	1620	1,00	78,68	3487	3010	2,20	23,00	6945.	84,50	30,00	99,97	99,27	78,68	98,28	80,69	23,00
mix	news	g_us	90,01	89,75	89,39	10,00	985,0	4,00	1018	1,00	81,44	3596	1110	2,09	4,00	4244	98,50	90,00	100,0	100,0	81,44	99,95	79,88	4,00
mix	pizza delivery	g_us	92,47	96,81	87,75	10,00	752,0	2,00	1480	2,00	66,91	1014	9050	2,91	56,00	1823.	75,20	54,00	99,96	89,16	66,91	94,50	84,68	56,00
mix	хостинг	g_ua	92,23	96,70	87,38	10,00	955,0	7,00	2250	1,00	55,03	1052	9900,	1,55	95,00	2606.	95,50	100,00	99,77	95,46	55,03	65,29	74,68	95,00
mix	ray ban	g_us	90,68	95,95	84,98	10,00	760,0	1,00	6290	2,00	74,27	1012	3680	0,29	80,00	2616.	76,00	30,00	99,92	91,66	74,27	98,59	35,53	80,00
mix	royal caribbean	g_us	90,63	95,92	84,89	10,00	700,0	3,00	2200	2,00	61,90	6828,	1000	1,57	36,00	1620.	70,00	73,64	99,76	86,91	61,90	99,48	74,89	36,00
mix	mortgage calculator	g_us	89,94	95,54	83,84	10,00	794,0	1,00	3550	2,00	57,37	7502,	2240	1,19	53,00	1578.	79,40	30,00	99,85	86,97	57,37	99,77	69,33	53,00
hard	брокер	g_ru	88,64	94,72	81,86	10,00	865,0	4,00	2800	1,00	55,48	2395,	8100,	3,93	81,00	1119.	86,50	90,00	98,15	77,88	55,48	60,61	88,19	81,00
mix	spinning	g_us	88,16	94,39	81,14	10,00	705,0	2,00	1340	1,00	78,73	9099,	1480	1,40	15,00	3040.	70,50	54,00	99,96	95,99	78,73	73,77	72,68	15,00
mix	dolls	g_us	84,40	88,15	80,31	10,00	815,0	0,00	1550	1,00	79,58	2332	3310	0,87	100,0	1612	81,50	0,00	99,97	99,71	79,58	86,28	62,31	100,0

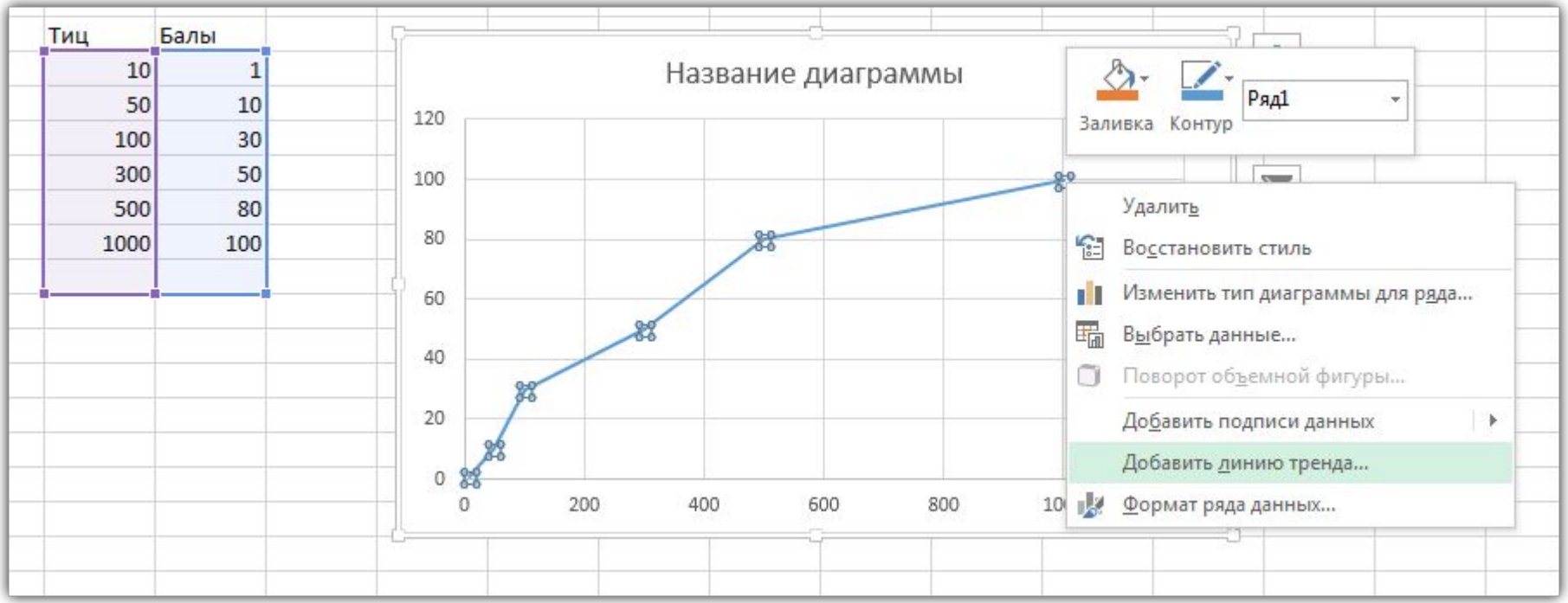
Полином

$$fr(q,d)=a_1h_1(q,d)+a_2h_2(q,d)+...+a_n$$

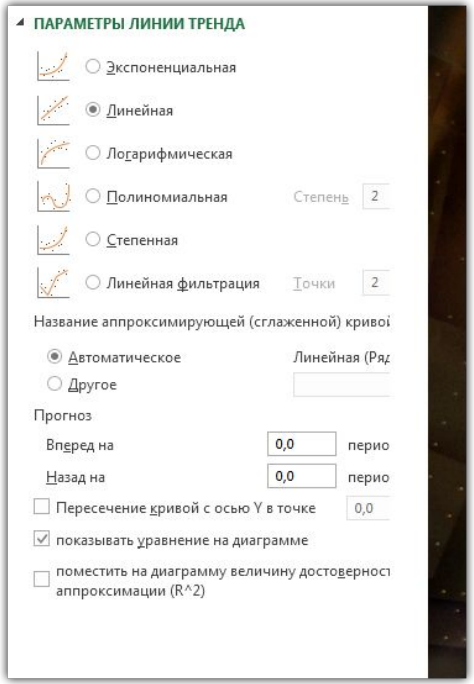
Нормировка линейная

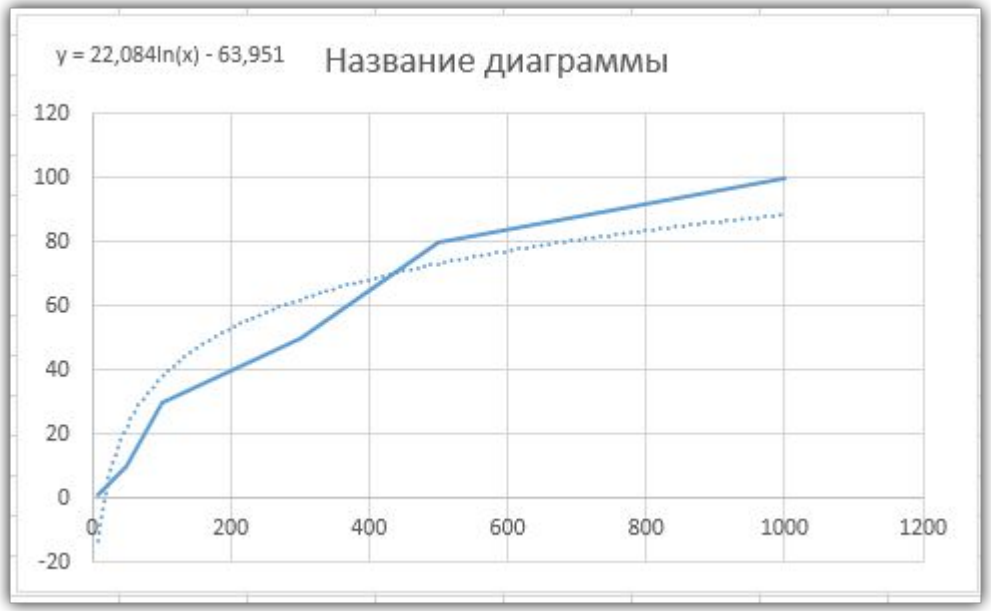
$$\tilde{x}_i = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}}$$

Апроксимация



Виды апроксимации





Итоговая формула

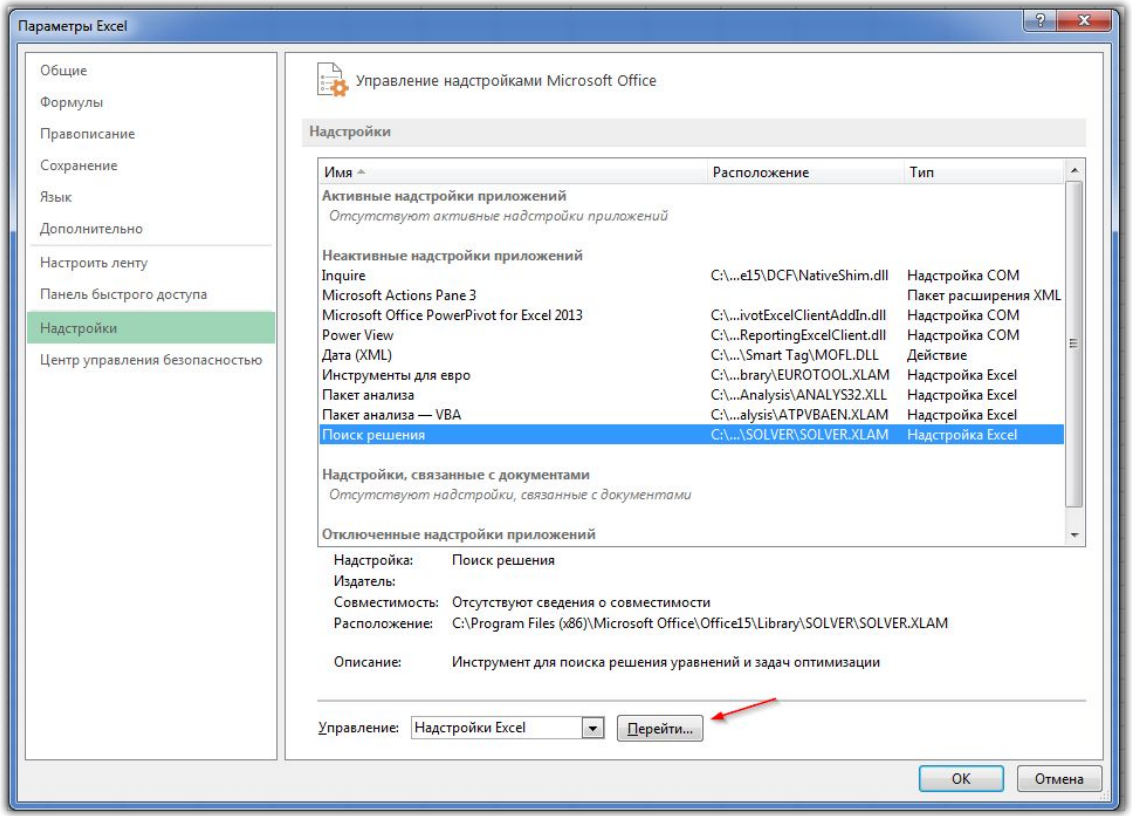
$$score = af(тиц) + bf(pr) + cf(BC) + d$$

Как найти коэффициенты a,b,c

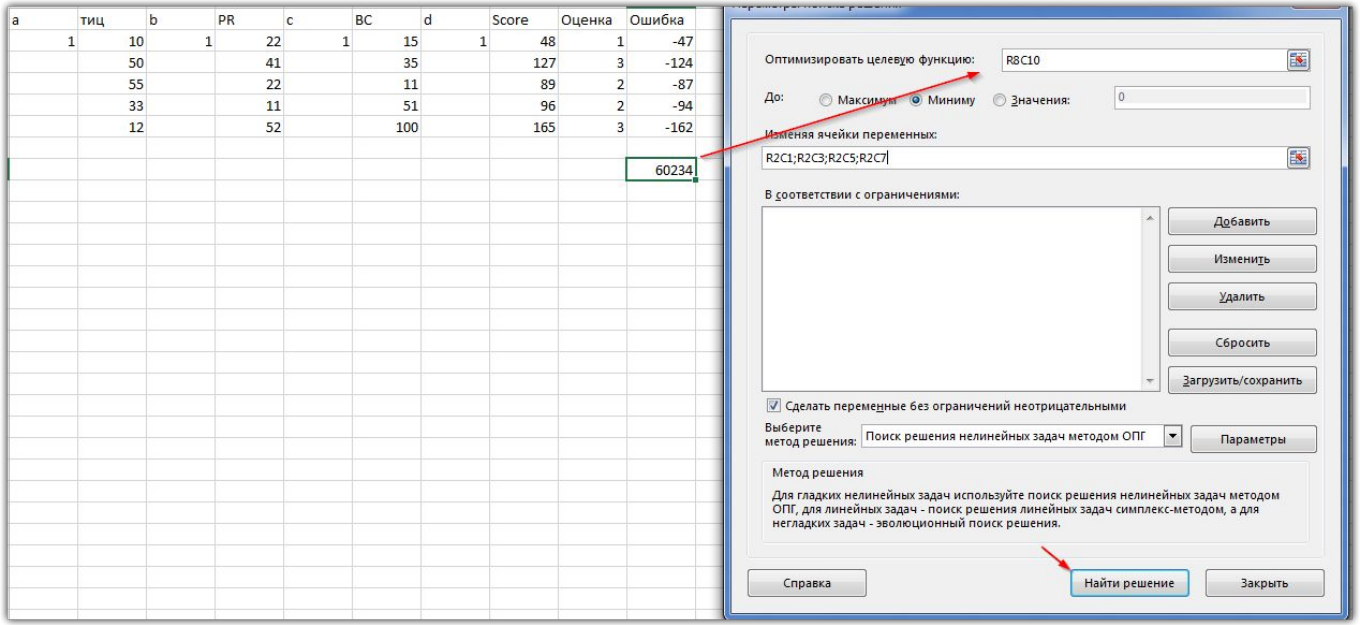
Таблица для поиска коэффициентов

a	тиц	b	PR	c	BC	d	Score	Оценка	Ошибка
1	10	1	22	1	15	1	48	1	-47
	50		41		35		127	3	-124
	55		22		11		89	2	-87
	33		11		51		96	2	-94
	12		52		100		165	3	-162
									60234

Поиск решения



Поиск решения



a	тиц	b	PR	c	BC	d	Score	Оценка	Ошибка
1	10	1	22	1	15	1	48	1	-47
	50		41		35		127	3	-124
	55		22		11		89	2	-87
	33		11		51		96	2	-94
	12		52		100		165	3	-162
									60234

Конец)

a	тиц	b	PR	c	BC	d	Score	Оценка	Ошибка
0,014486	10	0,012706	22	0,010979	15	0,926209	1,515287	1	-0,51529
	50		41		35		2,555725	3	0,444275
	55		22		11		2,123244	2	-0,12324
	33		11		51		2,103948	2	-0,10395
	12		52		100		2,858661	3	0,141339
									0,508872

Усовершенствованный алгоритм

1. Выбираем параметры
2. Нормируем
3. Находим корреляцию с правильными результатами
4. Строим формулу
5. Помним про эффект переобучения

Реальный пример

https://docs.google.com/spreadsheets/d/1KSXignNr7SvNGhUU0W_uWCaxp5Ka3ea1jHRiWQKOFrM/edit#gid=573531330

Рекомендации

kime, rapidminer - комбайны

Gensim - библиотека python

SciPy - библиотека python

Национальный корпус русского языка - http://www.ruscorpora.ru/

Обработка текста
http://www.alchemyapi.com/products/demo/alchemylanguage

http://www.wordfrequency.info/ - ENG корпус

https://github.com/buriy/python-readability - очистка текстов

Вопросы?

CEO of Prodvigator
Олег Саламаха
Facebook

www.prodvigator.ru