Векторная семантика Толстого

Борис Орехов (НИУ ВШЭ, ИРЛИ РАН)

Выдающемуся цифровому исследователю Толстого Даниилу Андреевичу Скоринкину ко дню рождения

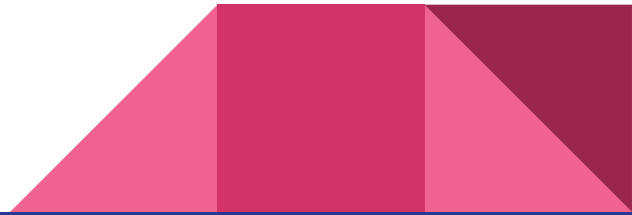
Что такое векторная семантика?

Слово и контекст

- Слова приобретают значение только в контексте.
- Слова, встречающиеся в похожих контекстах, значат похожее.
- «You shall know a word by the company it keeps» (J.R. Firth, Papers in Linguistics, 1957)

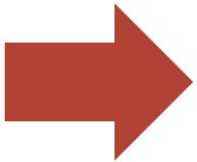
Ср.: «Я приду туда через несколько часов»

vs. «Я приду туда через несколько минут»



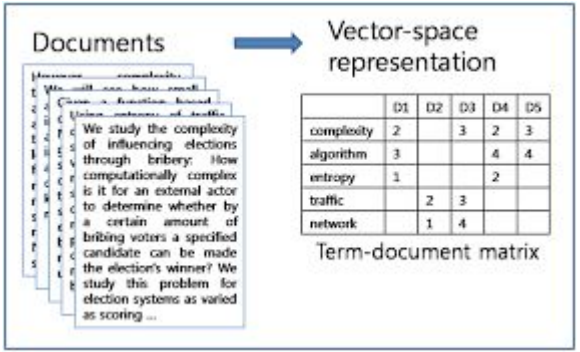
Как считать контексты?

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

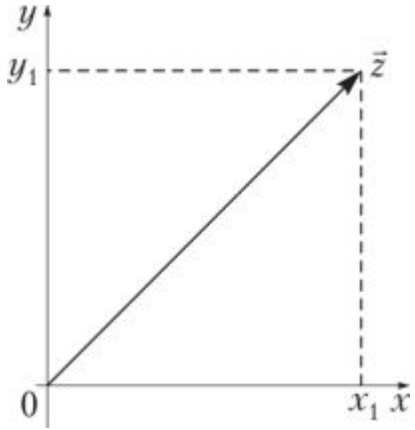


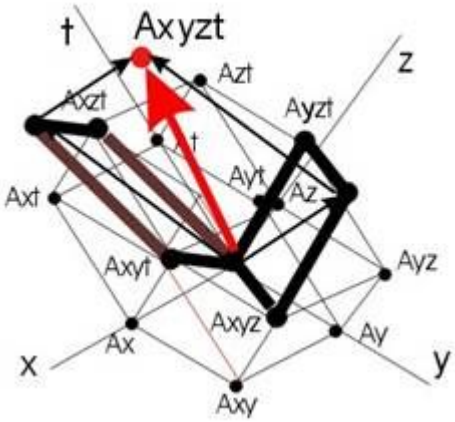
	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

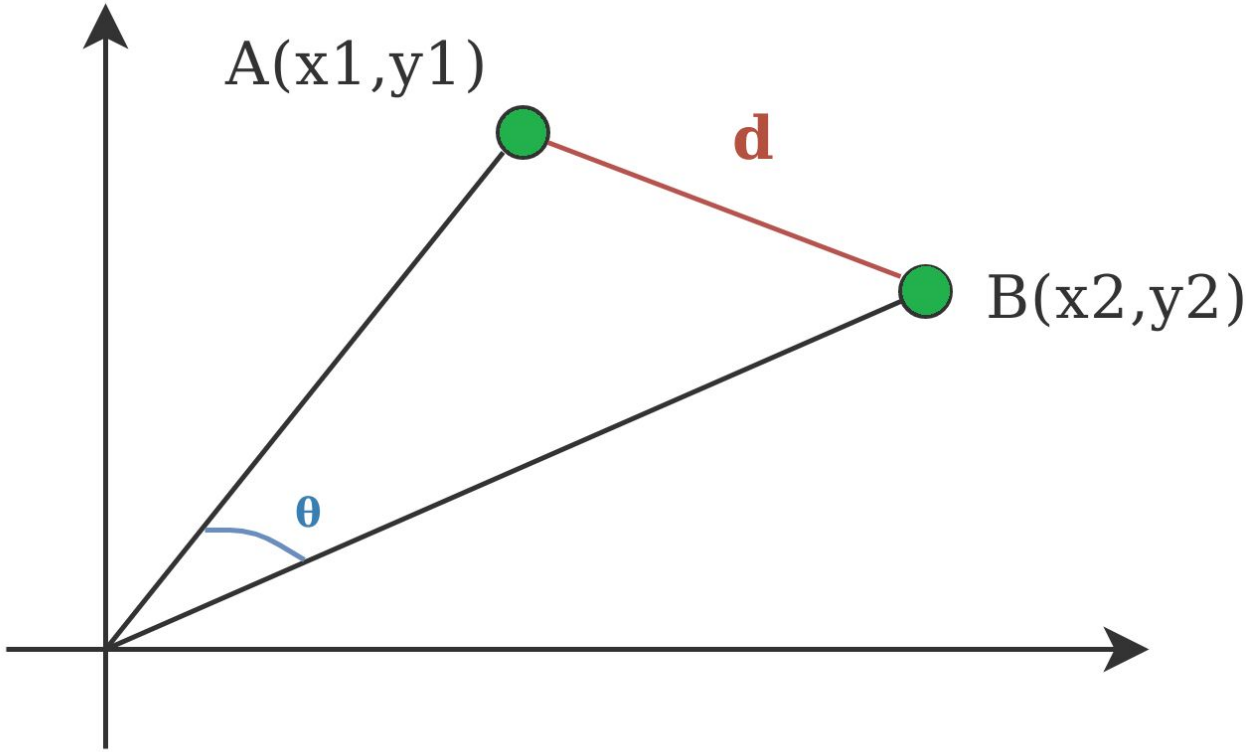
При чем тут вектора?



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		







Расстояние между векторами = между словами

Векторные романы

https://nevmenandr.github.io/novel2vec/

«Однажды весною, в час небывало жаркого заката, в Москве, на Патриарших прудах, появились два гражданина».

«Случайно весною, в полдень невиданно жаркого восхода, в Казани, на Митрополичьих ручьях, появились два сограждани́на».

Векторная модель семантики Толстого

Характеристики моделей

- Корпус: около 8 млн лемматизированных (mystem) текстов, около 100 тыс. уникальных слов, но в модели только с частотностью 2+
- Алгоритм CBOW
- Длина вектора: 300, размер окна: 10, объем: 45 тыс. слов

Модель 2:

- Тот же корпус
- Алгоритм skipgram
- Длина вектора: 500
- Размер окна: 2

Такие настройки взялись из уже существующей модели.

Хочется их сравнить

Тест модели: *любить* для Толстого (м1)

Толстой:

полюбить VERB
уважать VERB
ненавидеть VERB
страстно ADV
любящий ADJ
ценить VERB
презирать VERB
дорожить VERB
ближний ADJ
жалеть VERB

НКРЯ:

обожать VERB 0.74
полюбить VERB 0.70
любить NOUN 0.68
уважать VERB 0.66
нравиться VERB 0.66
ненавидеть VERB 0.64
любимый VERB 0.62
любить ADJ 0.60
боготворить VERB 0.59
презирать VERB 0.57

Обожать, боготворить: *ненастоящая любовь*

— Ничего нет ужасного, — сказал Новодворов, прислушивавшийся к разговору. — Массы всегда **обожают** только власть, — сказал он своим трещащим голосом. — Правительство властвует — они **обожают** его и ненавидят нас; завтра мы будем во власти — они будут **обожать** нас…

(Воскресение)

Девушка эта рассказала Николаю, как она с детства еще, по портретам, влюбилась в него, **боготворила** его и решила во что бы то ни стало добиться его внимания.

(Хаджи-Мурат)

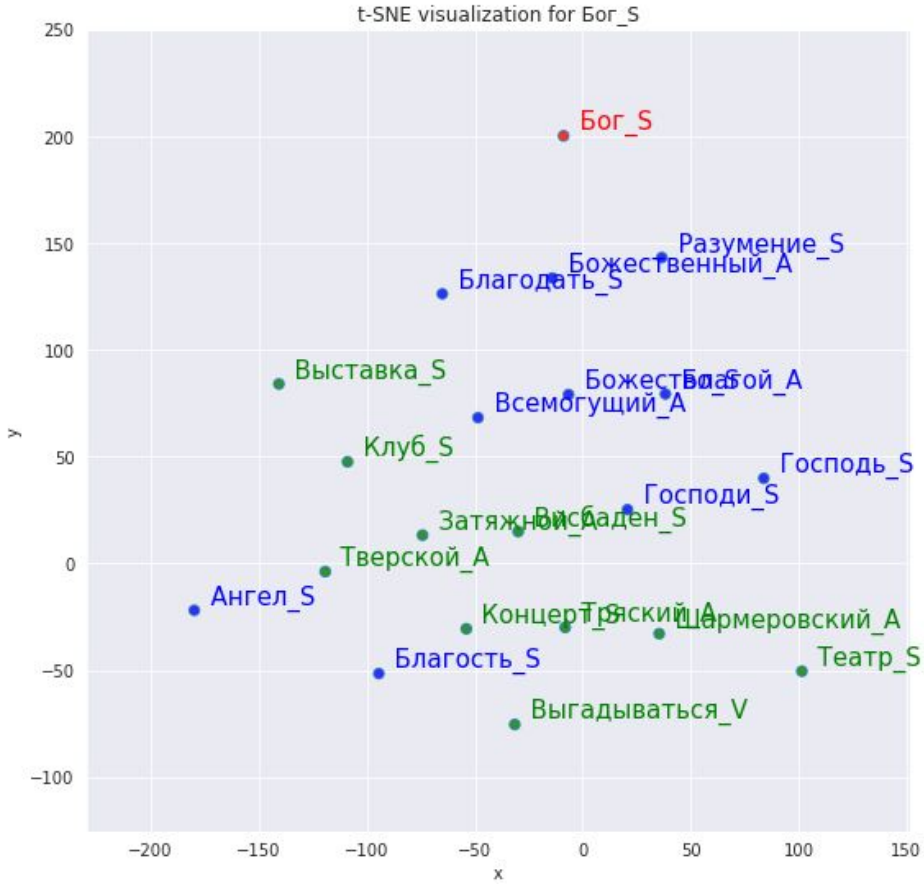
Самые близкие и самые далекие слова (1000)

красный_A, белый_A 0.904
армия_S, войско_S 0.890
сотня_S, тысяча_S 0.888
надевать_V, снимать_V 0.880
адъютант_S, генерал_S 0.878
спина_S, грудь_S 0.876
менее_ADV, более_ADV 0.875
рано_ADV, поздно_ADV 0.873
продавать_V, купить_V 0.863
неделя_S, месяц_S 0.863

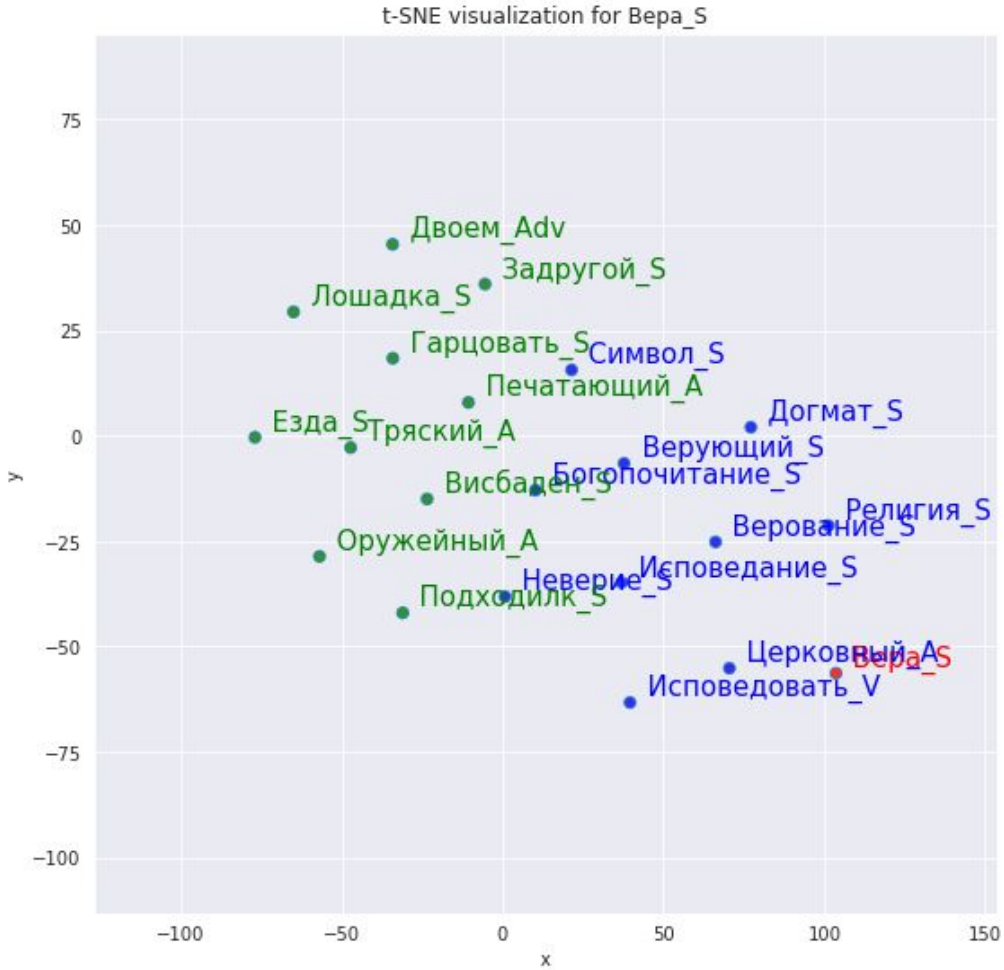
достоинство_S, пускать_V -0.579
особенность_S, велеть_V -0.579
пускать_V, характер_S -0.581
девка_S, изменение_S -0.587
исторический_A, старуха_S -0.601
связывать_V, обедать_V -0.612
изба_S, высказывать_V -0.621
купить_V, невольно_ADV -0.621
предполагать_V, старуха_S -0.622
барин_S, соединение_S -0.626
bonus: **орудие_S, маша_S** -0.569

case studies

Модель 1

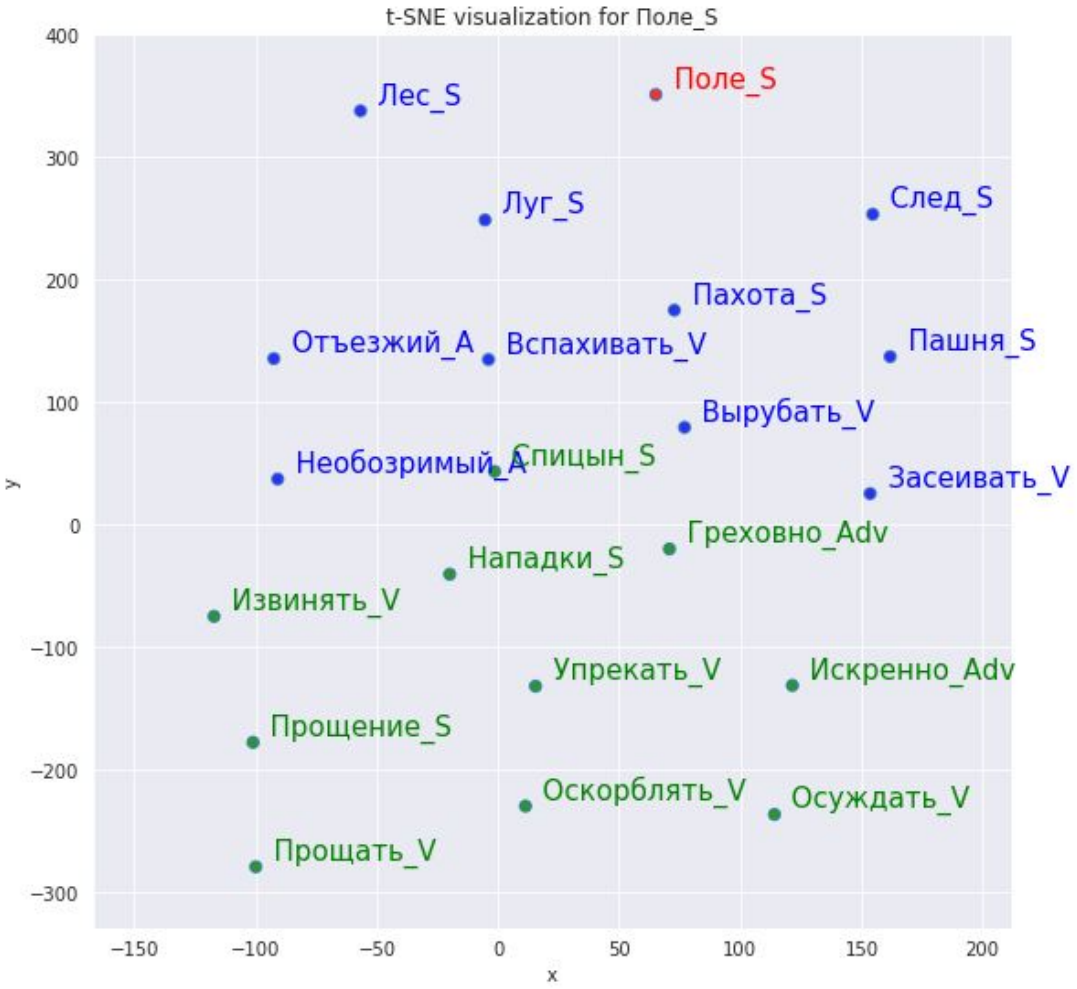


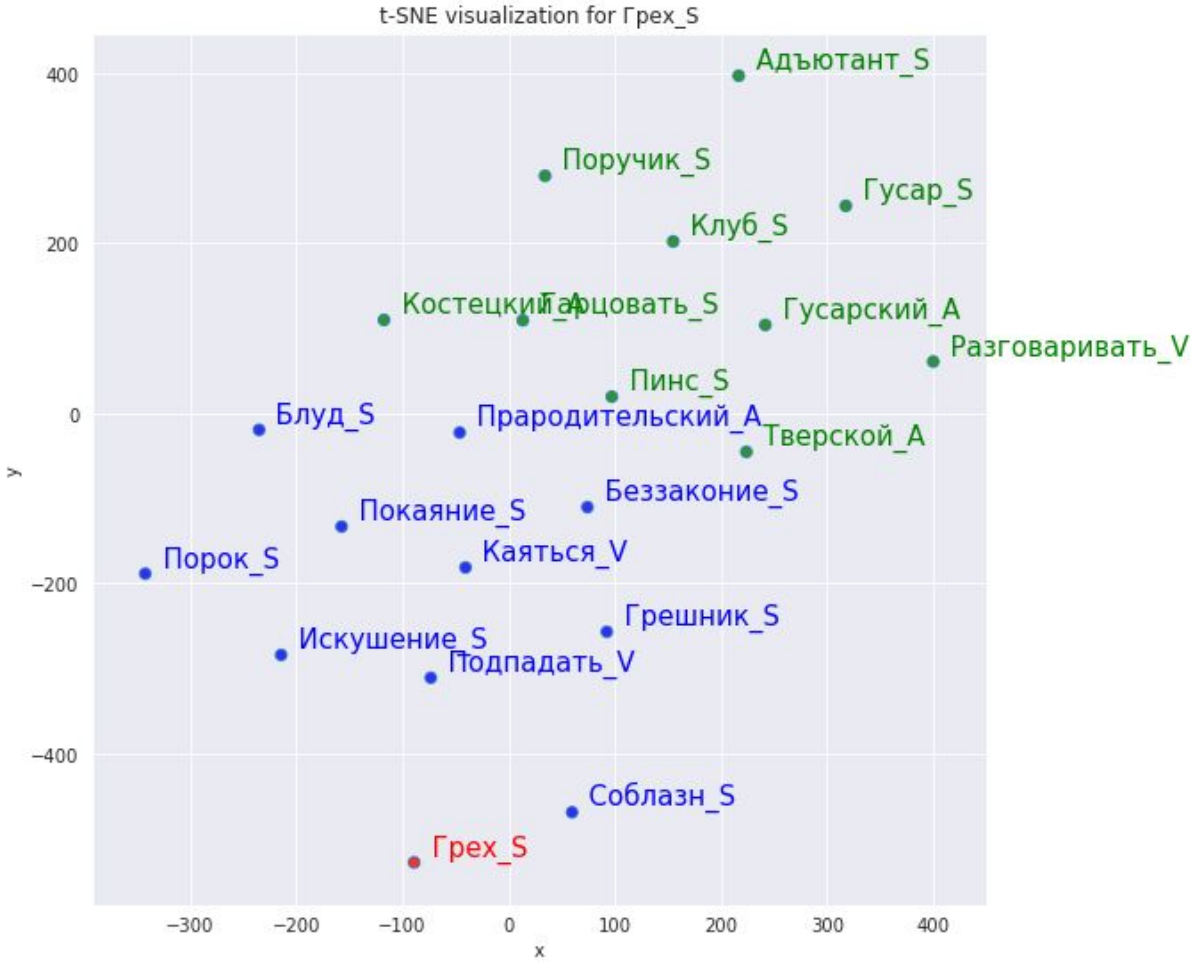
Ближайшие и наиболее далекие слова: *Бог*


t-SNE visualization for Вера_S
Двоем_Adv
Задругой_S
Лошадка_S
Гарцовать_S
Символ_S
Печатающий_A
Догмат_S
Езда_S
Тряский_A
Верующий_S
Богопочитание_S
Висбаден_S
Религия_S
Верование_S
Оружейный_A
Исповедание_S
Неверие_S
Подходилк_S
Церковный_A
Вера_S
Исповедовать_V
x
y

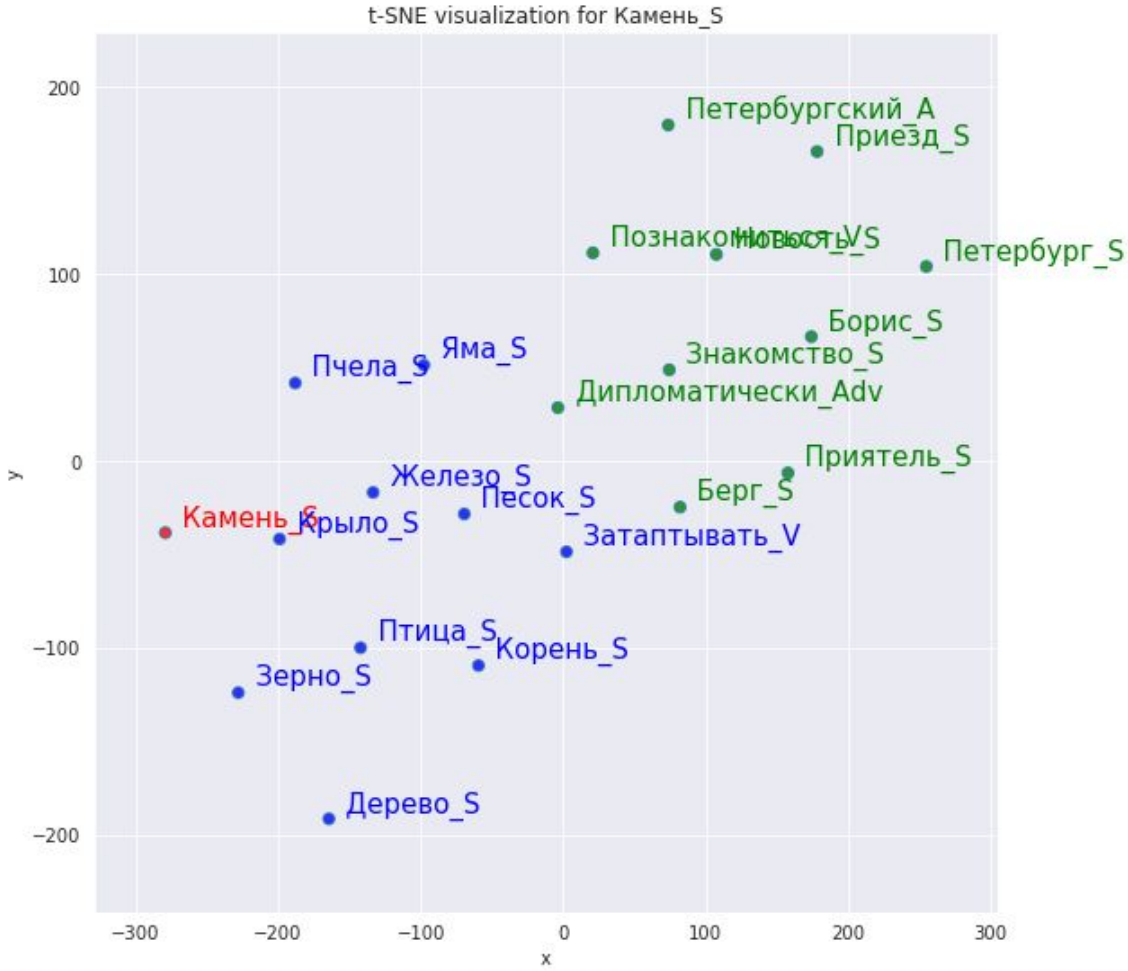
В чем *вера*?

поле: битва или пахота?





грех vs. гусар


t-SNE visualization for Камень_S
Петербургский_A
Приезд_S
Познакомиться_VS
Новость_S
Петербург_S
Борис_S
Яма_S
Пчела_S
Знакомство_S
Дипломатически_Adv
Приятель_S
Железо_S
Песок_S
Берг_S
Камень_S
Крыло_S
Затаптывать_V
Птица_S
Корень_S
Зерно_S
Дерево_S
200
100
0
−100
−200
y
−300
−200
−100
0
100
200
300
x

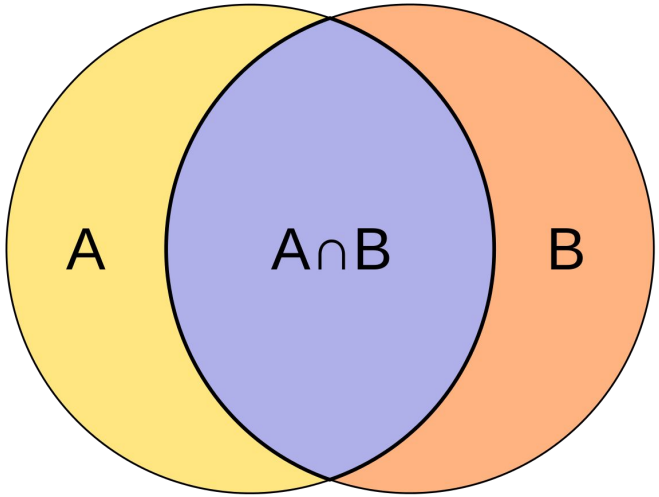
камень

Толстой и русский язык

Модель 2

Коэффициент Жаккара

Насколько совпадают списки ближайших соседей (список из 20)



Где язык побеждает идиолект

Только 8400 слов из 34000 имеют пересечения по соседям (при 40 — 12000).

Самые похожие: названия месяцев.

Глаголы: просыпаться_V (0.428), ехать_V (0.379), выпивать_V (0.333), поехать_V (0.333), грабить_V (0.290), молиться_V (0.290), подвигать_V (0.290), сидеть_V (0.25).

Существительные: таинство_S (0.428), радость_S (0.379), беспокойство_S (0.333), овца_S (0.333), ненависть_S (0.290), италия_S (0.290), социалист_S (0.290), священник_S (0.290)

Другие важные для меня тексты и проекты: http://nevmenandr.net/bo.php