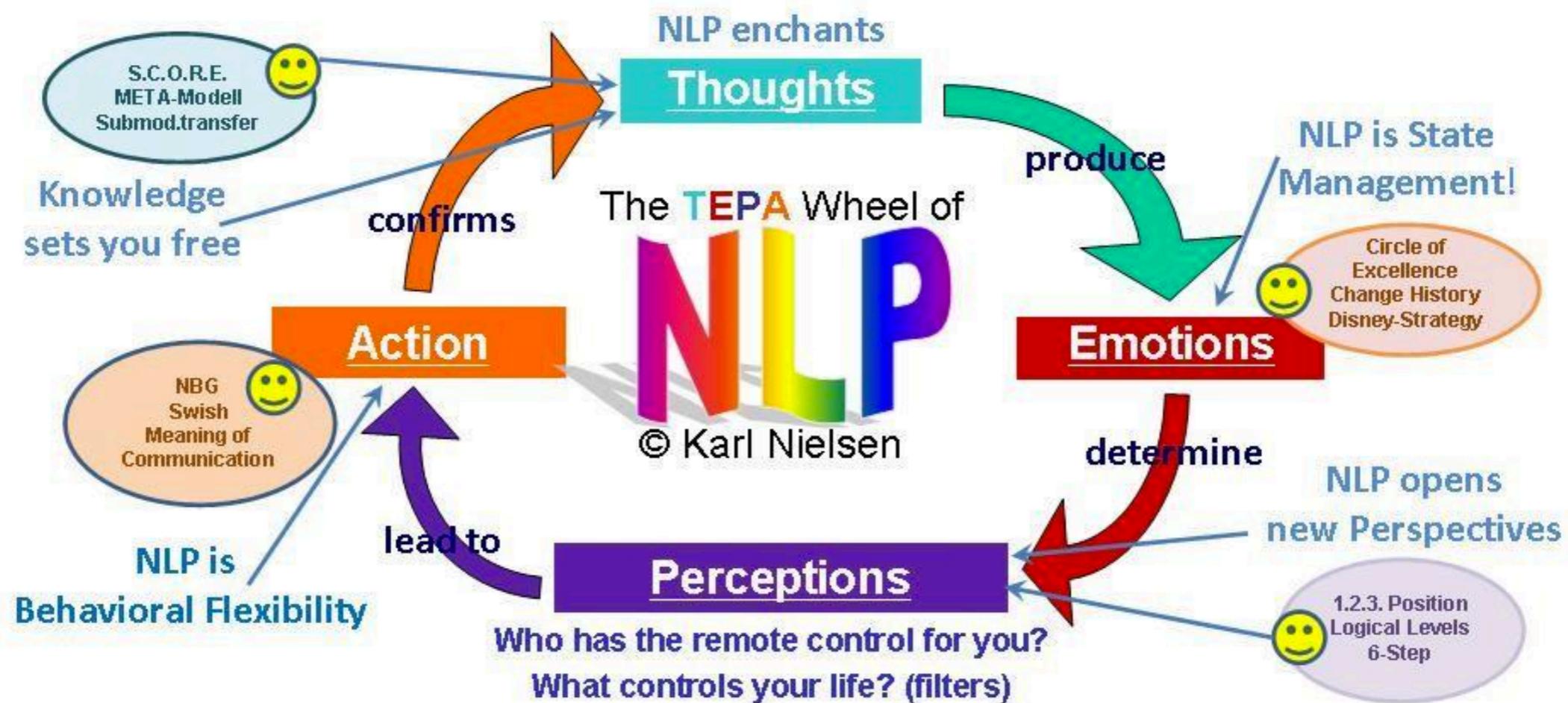


Text Mining and NLP

Ульянкин Филипп

NLP, the freedom in Thinking, Feeling, Perceiving and Behavior



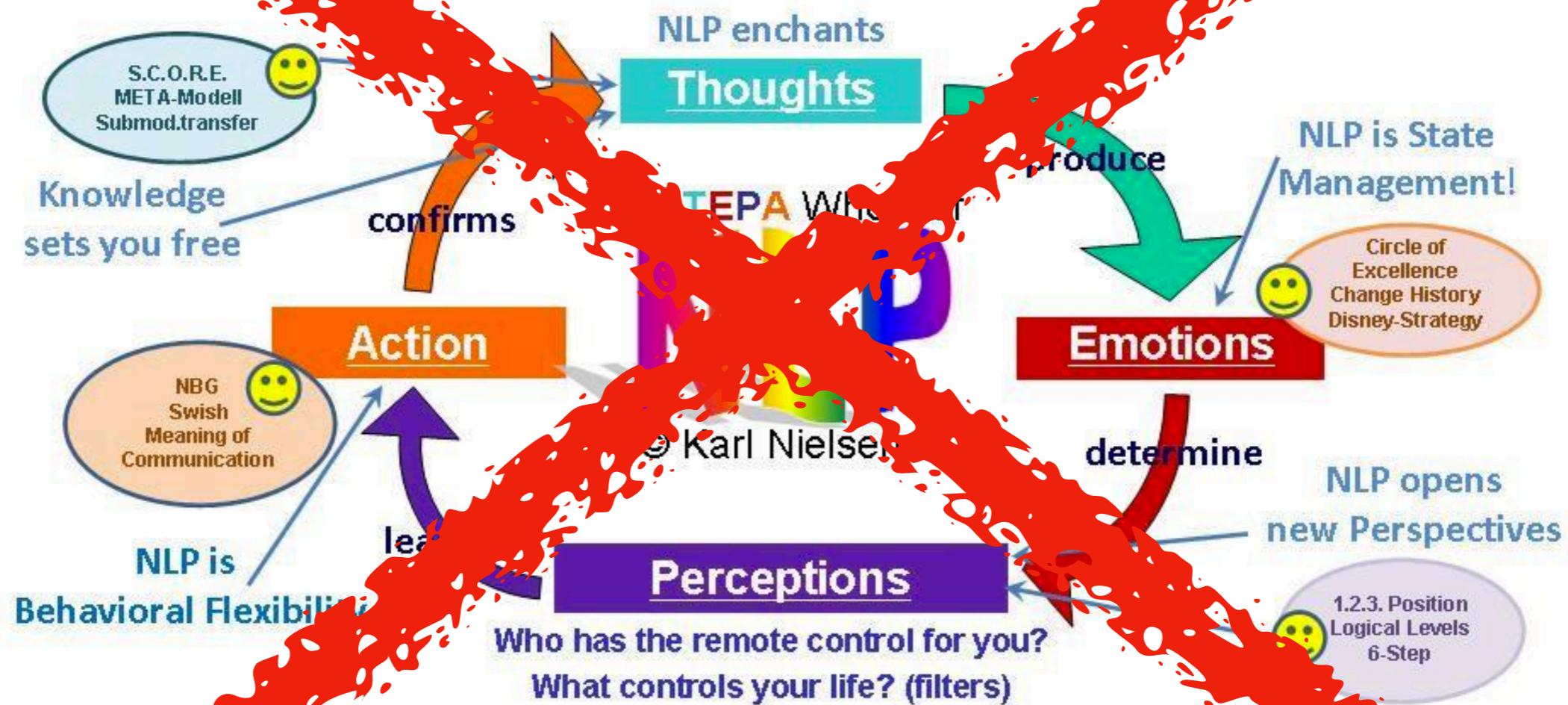
Neuro-linguistic programming



Karl Nielsen

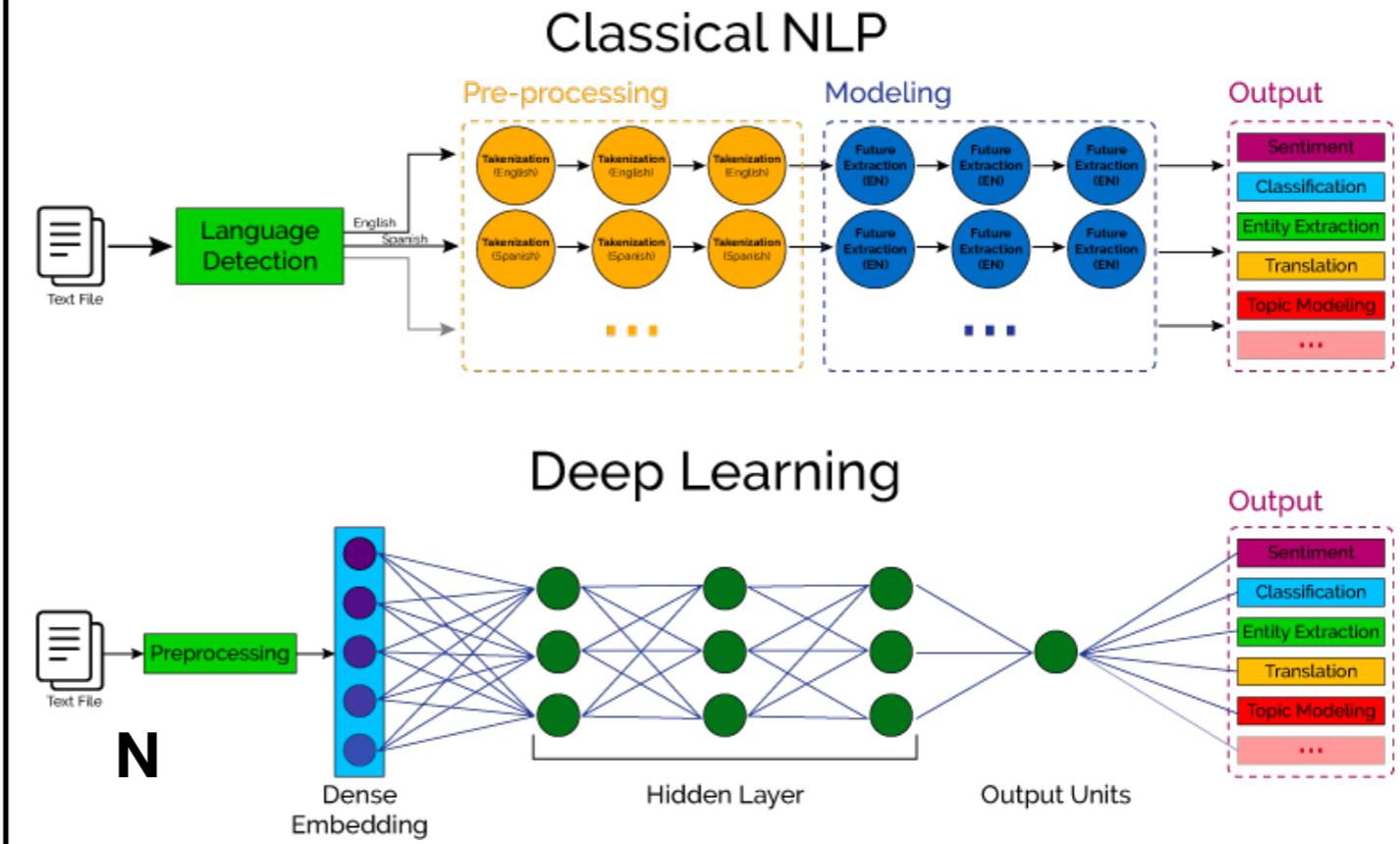


NLP, the freedom in Thinking, Feeling, Perceiving and Behavior





Neuro-linguistic programming



Natural Language Processing

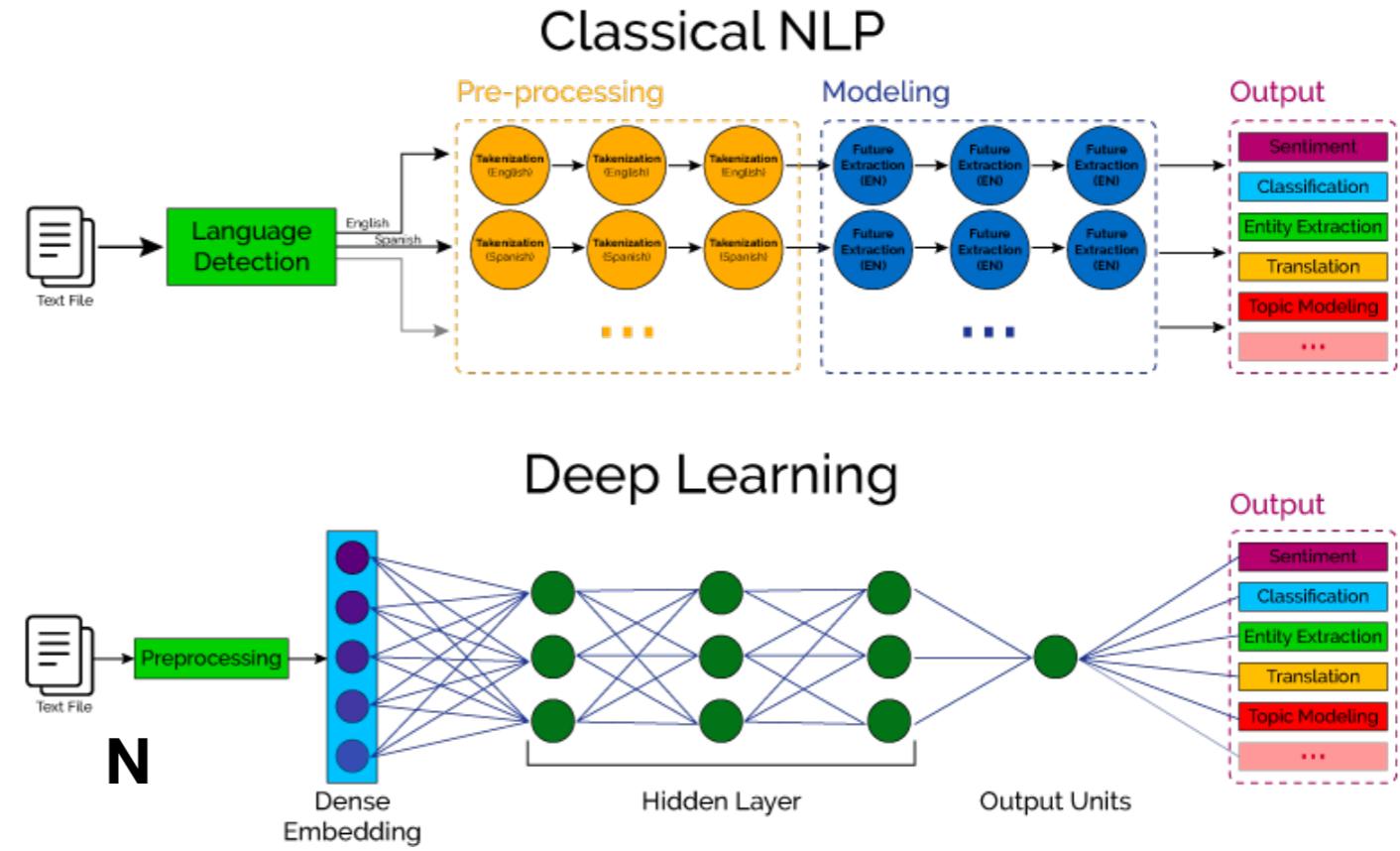
Neuro-linguistic programming

NLP курильщика



Natural Language Processing

NLP здорового человека



Задачи на текстах

- Классификация текстов (спам, порно, расчленёнка, токсичность)
- Регрессия на текстах (рейтинг статьи, лайки, просмотры)
- Кластеризация текстов, выделение тематик в текстах
- Извлечение информации (фактов и событий, именованных сущностей)



Задачи на текстах

- Контентные рекомендательные системы
- Поиск слов, похожих по смыслу на данное
- Диалоговые системы
- Автопереводы
- Генерация текстов



Эти задачи новее
и сложнее

Наши данные:

Боевики открыли огонь в христианской церкви в Нигерии; 19 человек погибли

ПРЕТОРИЯ, 7 августа. /ИТАР-ТАСС/. Боевики открыли в минувший понедельник огонь по прихожанам в христианской церкви в городе Отите /центр Нигерии/. В результате ЧП 19 человек погибли, сообщил сегодня представитель правоохранительных органов. Ни одна группировка пока не взяла на себя ответственность за теракт. В июне в Нигерии прокатилась волна насилия против христиан. Тогда жертвами стали более 130 человек. За терактами в церквях часто стоят экстремисты из террористической организации "Боко харам" /"Западное образование - грех" в переводе с языка хауса/, которая добивается введения шариата на всей территории страны.

Международная панорама

Россия лучше готова к возможному кризису, чем в 2008 году - Путин

МОСКВА, 17 октября. /ИТАР-ТАСС/. Россия лучше готова к возможному кризису, чем в 2008 году, несмотря на меньшие объемы резервов. Об этом заявил сегодня премьер-министр РФ Владимир Путин на заседании консультативного совета по иностранным инвестициям. "Я считаю, что если, не дай Бог, какие-то турбулентные процессы разгорятся слишком сильно, то все-таки Россия в целом к этим сложностям сегодня готова", - сказал он. По его словам, "сам факт прохождения сложностей кризисных явлений в мировой экономике в предыдущие два года позволил наработать нам определенные инструменты и опыт борьбы с трудностями". "Если некоторые уменьшения резервных фондов - это минус, то наличие опыта и отработанных инструментов борьбы с кризисом - это, безусловно, плюс", - добавил Путин.

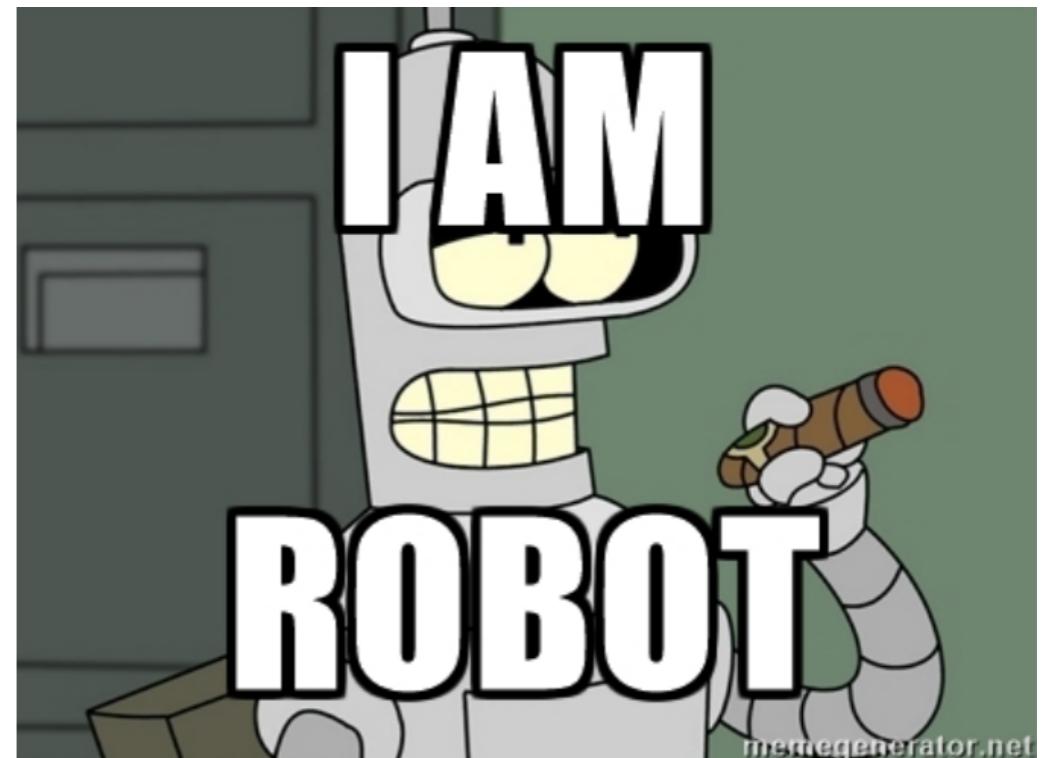
Экономика и бизнес

Наши данные:

text → features → ML model → $P(y|x)$

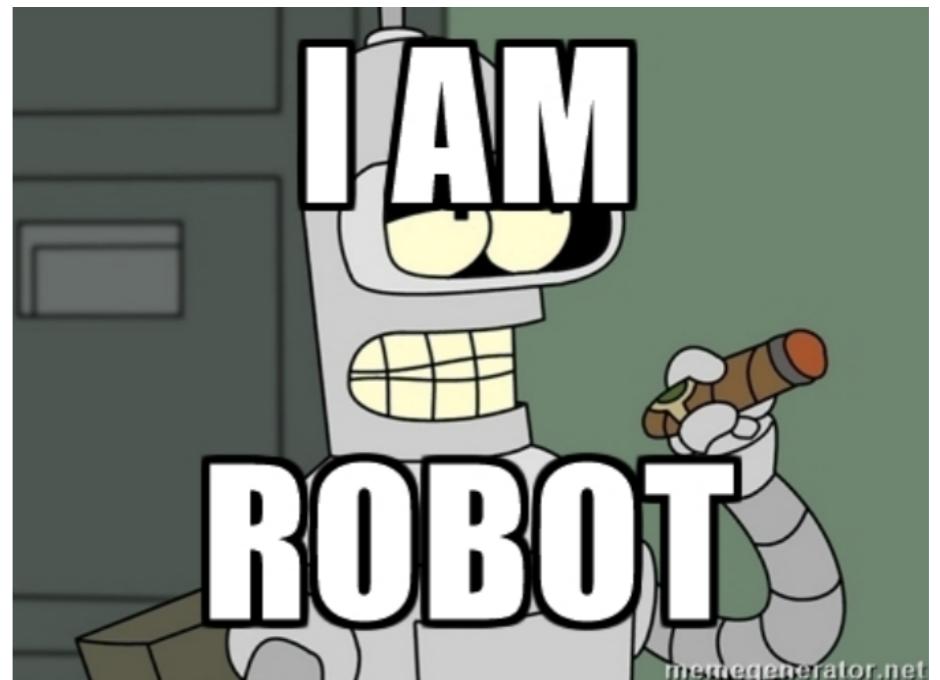


Как представить текст
в виде, который могла
бы понять модель?



Основная идея

- Компьютер не понимает голый текст, он понимает только цифры
- Значит нам нужно превратить текст в цифры!
- Объект для работы - **документ**, то из чего документ состоит - **токены**

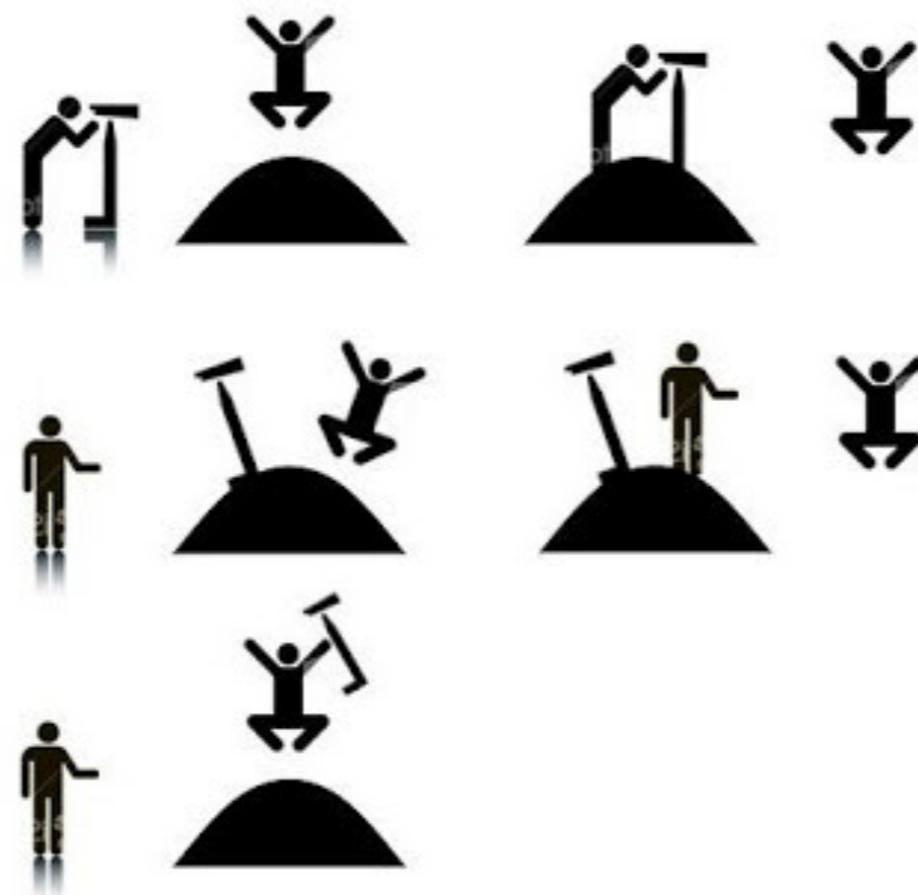


Основные проблемы

- Лексическая неоднозначность: орган, парить, рожки
- Морфологическая неоднозначность: «хранить деньги в банке», «что делают белки в клетке»
- Синтаксическая неоднозначность: «Эти типы стали есть на складе»
- Неологизмы: печеньки, репостнуть, расшарить, зарисёрчить, рашка
- Разное написание: Россия, РФ, Российская Федерация
- Нестандартное написание, орфографические ошибки и очепятки

Синтаксическая неоднозначность

I saw a man on the hill with a telescope



**Проблемы созданы
для того, чтобы с ними
бороться!**

Предположения

- Документ это множество из слов
- Порядок слов неважен
- Порядок неважен слов
- Неважен слов порядок
- Неважен порядок слов
- Слов неважен порядок
- Слов порядок неважен
- Мешок слов (bag of words)



Term-document matrix

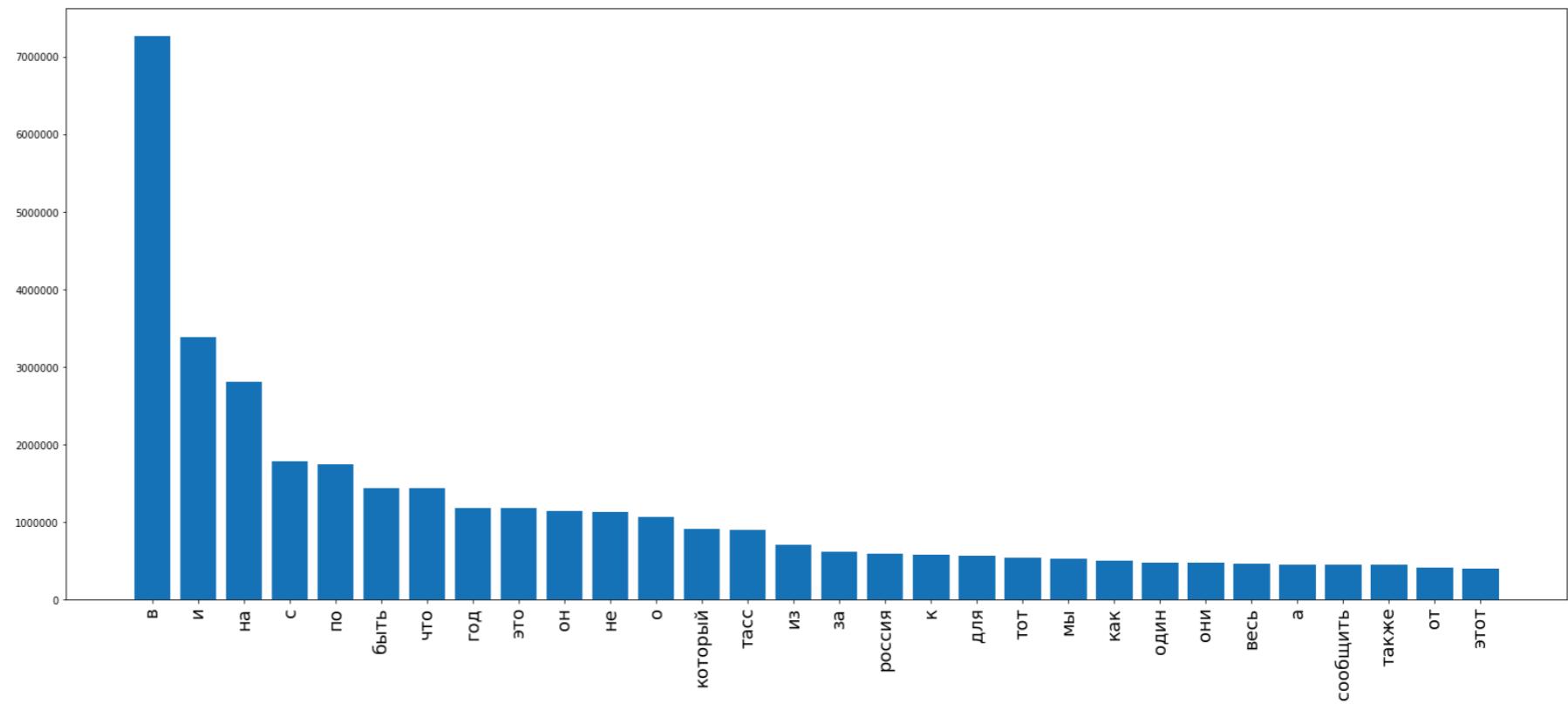
- Посчитаем частоту каждого слова в каждой статье!
- Готово, можно строить модели!

	я	просмотреть	выборы	вместе
Статья 1	4	1	1	1
Статья 2	6	2	0	0
Статья 3	7	0	1	0
....
Статья 150000	2	0	0	0

HET!!

Стоп-слова

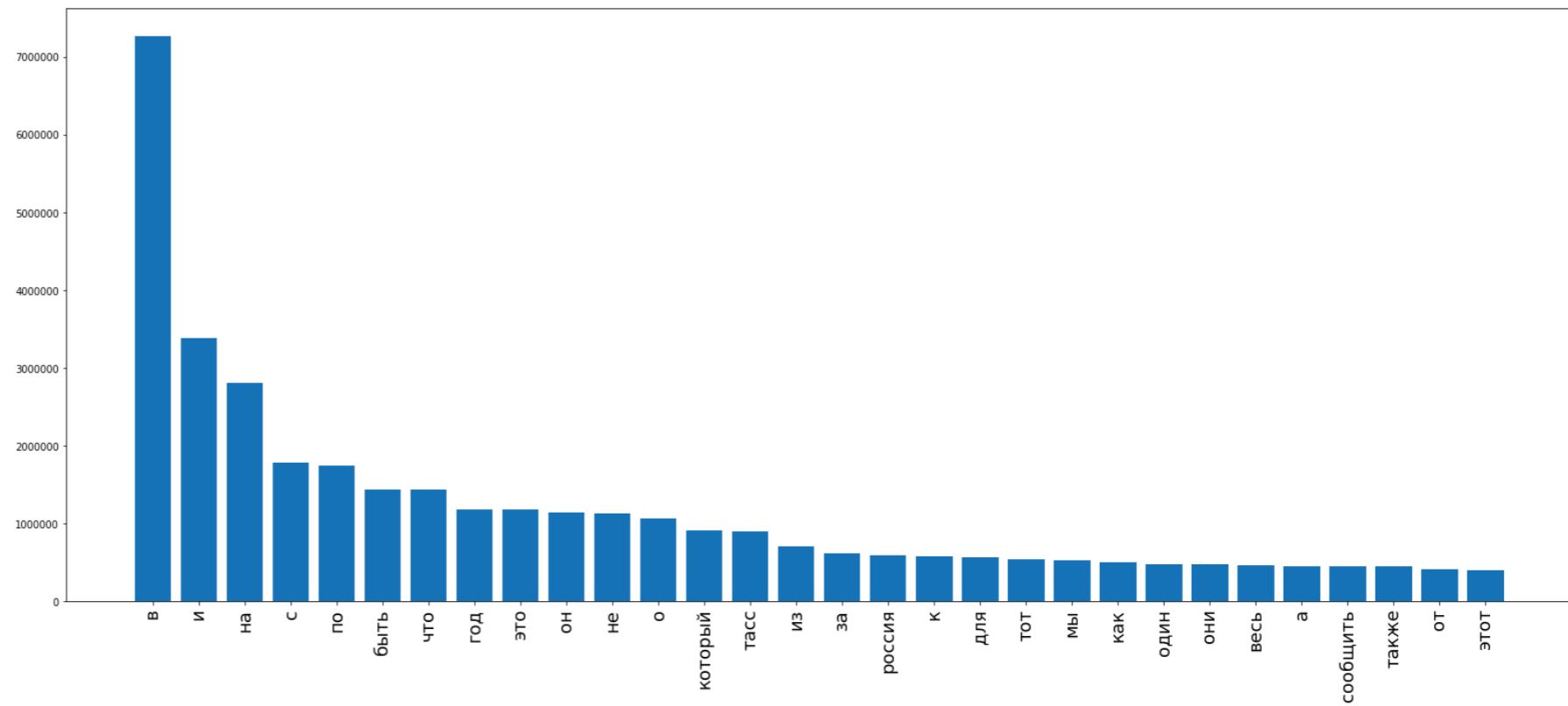
- На 40 000 наблюдений есть 2 058 294 уникальных слова в текстах



- Слова а, но, в, за можно найти в каждом тексте
- Выбрасываем стоп-слова

Стоп-слова

- На 40 000 наблюдений есть 2 058 294 уникальных слова в текстах



- Слова а, но, в, за можно найти в каждом тексте
- Выбрасываем стоп-слова

2 058 294 → 2 031 302

Нормализация

красивый, красива, красивые ...

банк, банков, банками ...

человек, люди ...



Стеминг

Обрезаем приставки
и окончания

Лемматизация

Меняем словоформу
по словарю

Нормализация

красивый, красива, красивые ...

банк, банков, банками ...

человек, люди ...



Стеминг

Обрезаем приставки
и окончания

Лемматизация

Меняем словоформу
по словарю

2 031 302 → 408 979

Редкие слова

- Частые слова: встречаются везде, неинформативны. Оставим те, что встречаются менее чем в 80% текстов.
- Редкие слова: нельзя получить несмешённую оценку по малому числу наблюдений, слишком много параметров.
- Оставим те, что встречаются больше, чем в 30%

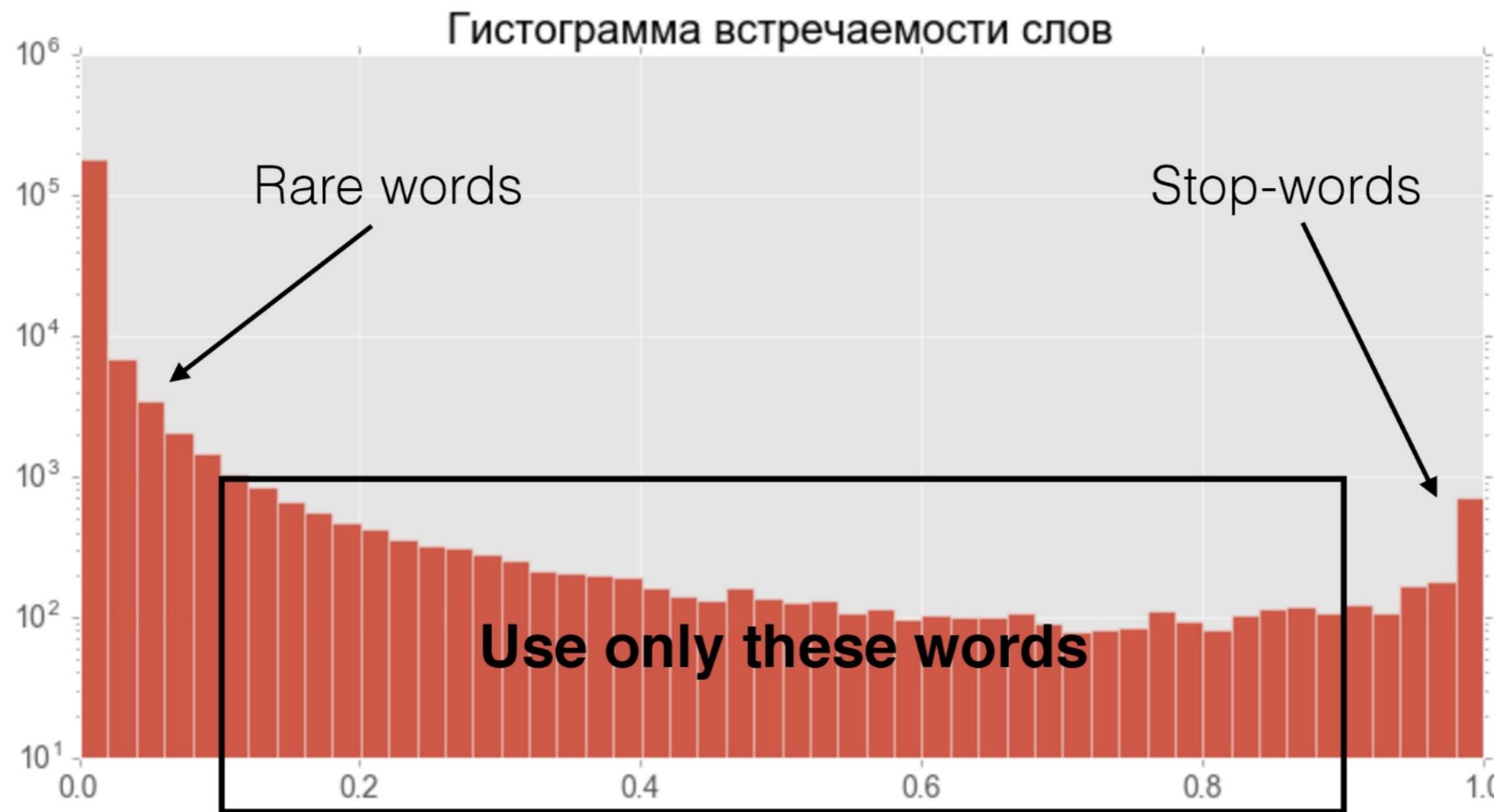
Редкие слова

- Частые слова: встречаются везде, неинформативны. Оставим те, что встречаются менее чем в 80% текстов.
- Редкие слова: нельзя получить несмешённую оценку по малому числу наблюдений, слишком много параметров.
- Оставим те, что встречаются больше, чем в 30%

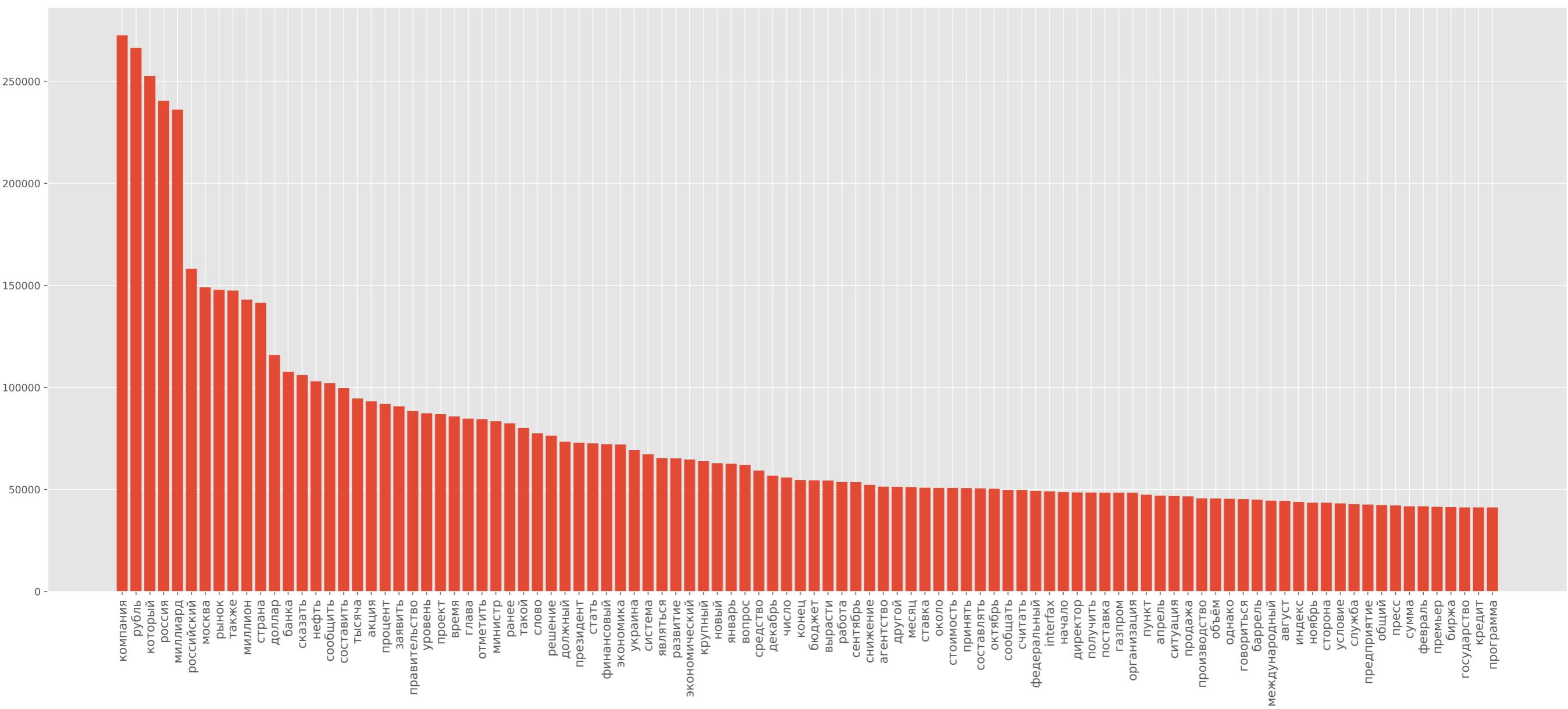
408 979 → 6773

Отсев по частоте

- Для разных корпусов текстов стоп-слова могут быть специфическими (Например, ЦБ для экономических текстов)



Отсев по частоте



Матрица термы-на-документы

- После всех этих предобработок, наконец, можно строить матрицу термы-на-документы!

	я	просмотреть	выборы	вместе
Статья 1	4	1	1	1
Статья 2	2	0	0	0
Статья 3	0	1	1	0
....
Статья 150000	2	0	0	0

Что мы сделали

- Выкинули стоп-слова
- Лемматизация текста
- Отфильтровали по частоте

Это недостаточно круто!

Tf-idf

Очень важная идея о том, как строить
матрицу термы-на-документы

tf: term frequency,
нормализация матрицы по строкам

idf: inverse document frequency,
нормализация матрицы по столбцам

TF

- Если слово встречается в документе часто, но оно не стоп-слово => оно важное (tf)
- О том, как правильно подсчитать слова в документе
 - Количество (уже делали это)
 - Частота (tf)
 - Булева частота (1 или 0)
 - $\log(\text{частота})$

Tf-idf

n	x	z
1	Испания	Нежился на пляже
2	Крым	Копали яму на пляже
3	Дача	Копал картошку
4	Крым	Ел картошки и картошку

Tf

	нежиться	пляж	копать	яма	картошка	есть
1	1/6	1/6	0	0	0	0
2	0	1/6	1/6	1/6	0	0
3	0	0	1/6	0	1/6	0
4	0	0	0	0	2/6	1/6

IDF

- Слово встретилось только в этом документе, а в других нет => оно важное и описывает природу этого документа
- Idf пытается увеличить вес редких слов в матрице термы-на-документы
 - $\text{idf} = 1$
 - $\text{idf} = \ln\left(\frac{N}{n_t}\right)$ (canonical idf)
 - $\text{idf} = \ln\left(1 + \frac{N}{n_t}\right)$
 - любой свой вариант!

Tf-idf

n	x	z
1	Испания	Нежился на пляже
2	Крым	Копали яму на пляже
3	Дача	Копал картошку
4	Крым	Ел картошки и картошку

Tf

	нежиться	пляж	копать	яма	картошка	есть
1	1/6	1/6	0	0	0	0
2	0	1/6	1/6	1/6	0	0
3	0	0	1/6	0	1/6	0
4	0	0	0	0	2/6	1/6

X

Idf

	нежиться	пляж	копать	яма	картошка	есть
1	$\ln 4$	$\ln 2$	0	0	0	0
2	0	$\ln 2$	$\ln 2$	$\ln 4$	0	0
3	0	0	$\ln 2$	0	$\ln 2$	0
4	0	0	0	0	$\ln 2$	$\ln 4$

=

Tf-idf

=

	нежиться	пляж	копать	яма	картошка	есть
1	0.23	0.11	0	0	0	0
2	0	0.11	0.11	0.23	0	0
3	0	0	0.11	0	0.11	0
4	0	0	0	0	0.23	0.23

Tf-idf

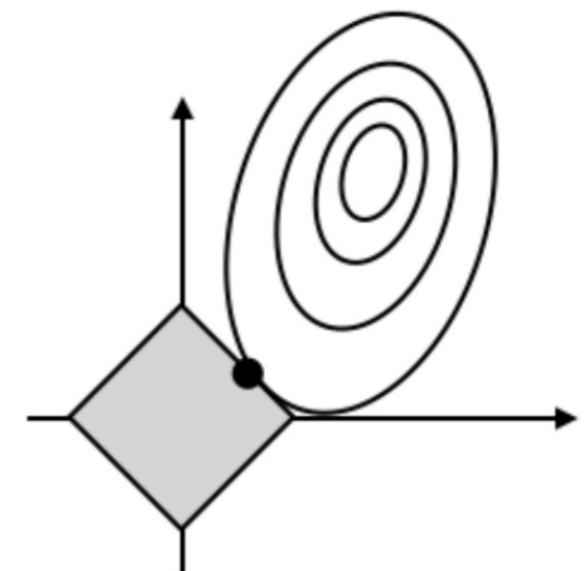
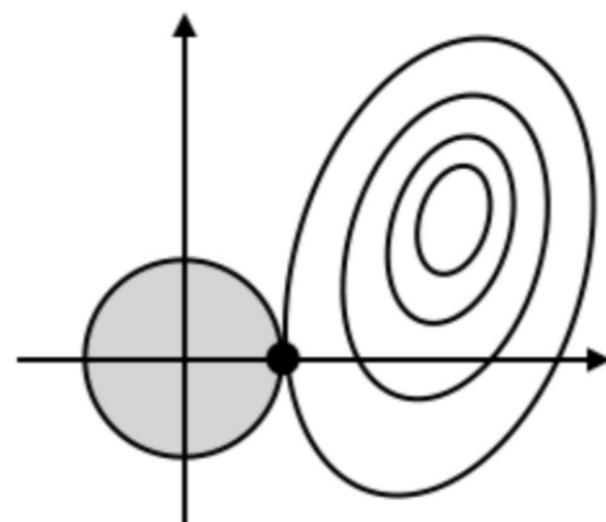
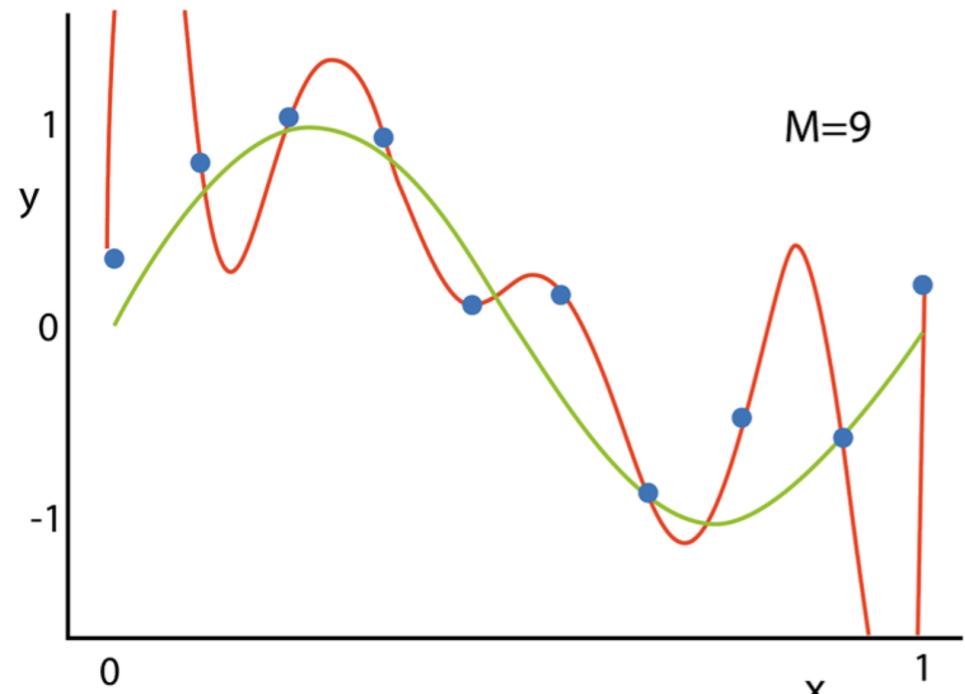
- Выбирая пороговые значения tf и idf можем решать какое число фичей взять в модель
- Idf убивает стоп-слова из-за того, что они встречаются почти в каждом документе
- Tf убивает редкие слова

Модель

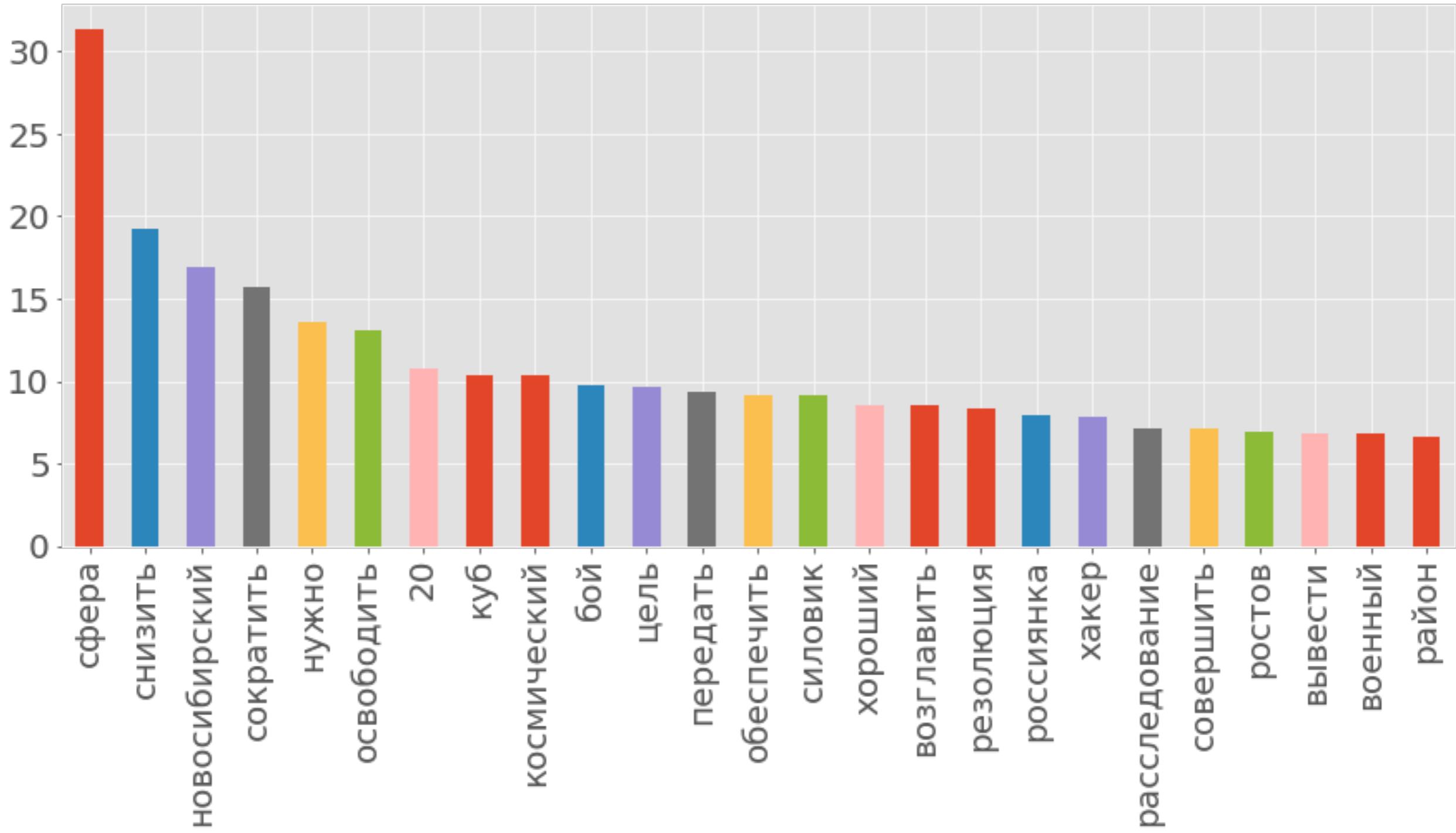
- Текстовая специфика сказывается на выборе модели для обучения
- Случайный лес жадно строит глубокие деревья минимизируя локальную ошибку вместо глобальной до тех пор, пока в листе не будет мало объектов. В нашей ситуации деревья будут очень глубокими => долгий перебор
- Текст хорошо описывается только с помощью совместного использования множества регрессоров. Лес требует очень большого числа деревьев для поиска закономерностей между разнесёнными по тексту словами. Чем больше деревьев, тем выше будет шум и сигнал затеряется.
- Бустинг строит маленькие деревья. Каждое учит небольшое подмножество признаков. Целевая переменная объясняется комбинацией большого набора признаков. Приходится использовать много деревьев, сигнал рассеивается.

Модель

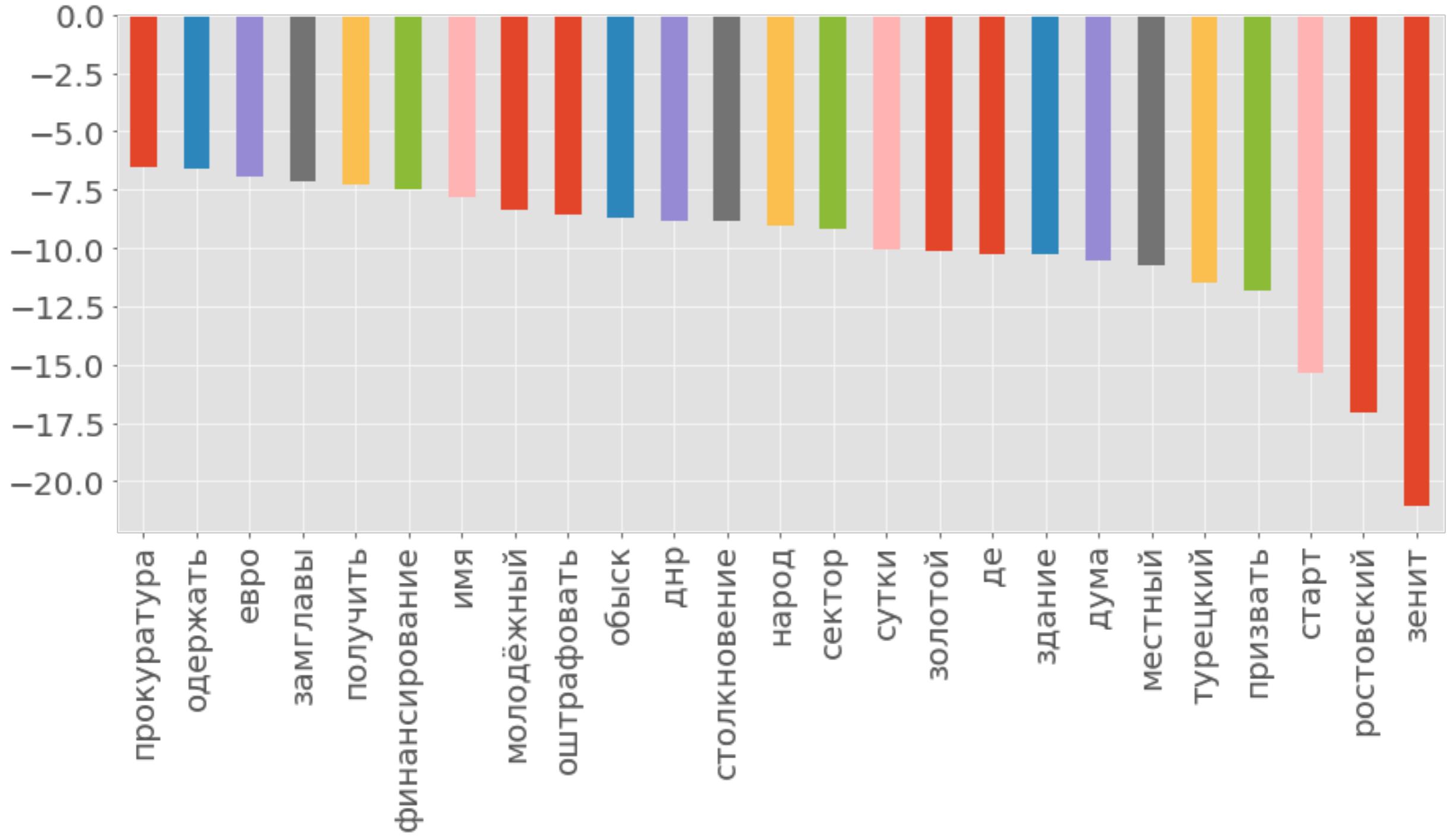
- Логистическая регрессия
- Lasso-регуляризация
- Загуляет плохие регрессоры



Слова слова



Слова слова опять слова



Как сделать модель лучше?

- Сейчас мы используем слова как фичи
- Почему бы не использовать пары из слов?
- В общем виде n-граммы - последовательности из n слов

МОЖЕТ ХВАТИТЬ ПРИМЕРОВ С ХУРМОЙ

УНИГРАММЫ:

1. МОЖЕТ
2. ХВАТИТЬ
3. ПРИМЕРОВ
4. С
5. ХУРМОЙ

МОЖЕТ ХВАТИТЬ ПРИМЕРОВ С ХУРМОЙ

БИГРАММЫ:

1. МОЖЕТ ХВАТИТЬ
2. ХВАТИТЬ ПРИМЕРОВ
3. ПРИМЕРОВ С
4. С ХУРМОЙ

МОЖЕТ ХВАТИТЬ ПРИМЕРОВ С ХУРМОЙ

ТРИГРАММЫ:

1. МОЖЕТ ХВАТИТЬ ПРИМЕРОВ
2. ХВАТИТЬ ПРИМЕРОВ С
3. ПРИМЕРОВ С ХУРМОЙ

Как сделать модель лучше?

- Это улучшает результат, но словарь разрастается
- Можно редуцировать пространство признаков с помощью РСА или другого метода понижения размерности

Что мы сделали

- Вычислили tf-idf
- Отфильтровали по tf-idf
- Мы остались в концепции мешка слов

Всё ещё недостаточно круто!

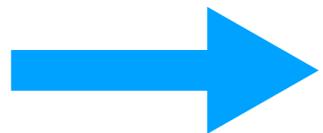
Word2Vec

- Мы хотим понимать смысл слов
- Давайте опишем каждое слово вектором длины d так, чтобы похожие слова обладали близкими векторами.
- Идея: схожие слова имеют схожие контексты.
- Будем оценивать вероятность встретить слово i в контексте слова j

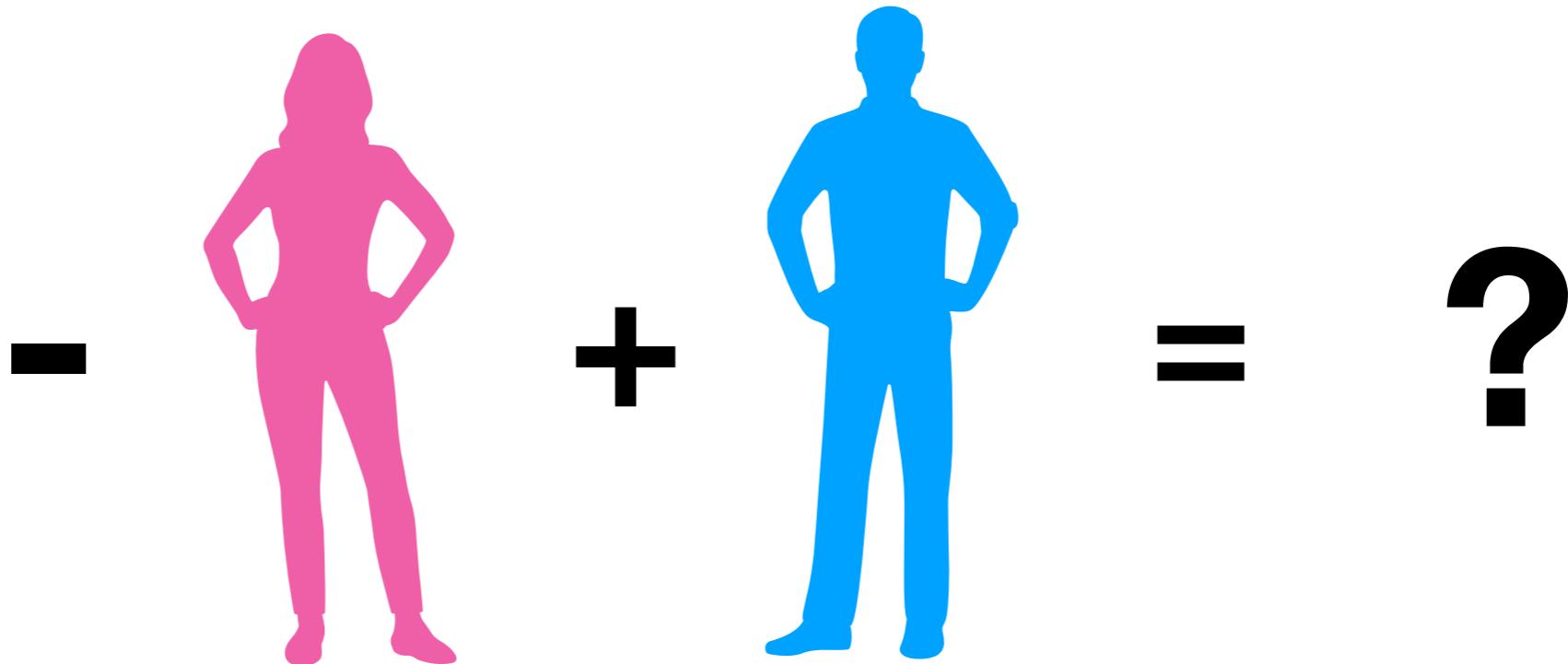
$$p(w_i | w_j) = \frac{\exp(\langle \bar{w}_i, \bar{w}_j \rangle)}{\sum_w \exp(\langle \bar{w}_i, \bar{w} \rangle)}$$

$$\sum_{i=1}^n \sum_{j=-k}^k \log p(w_{i+j} | w_i) \rightarrow \max$$

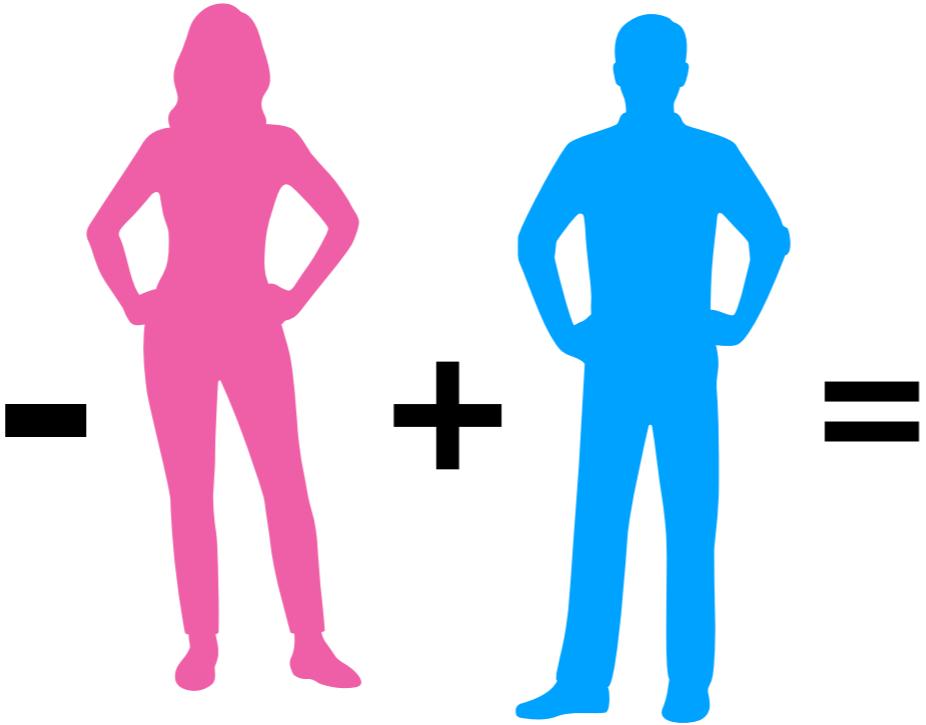
Слово



Вектор



Enter word or sentence (EXIT to break): гугл
— гугол 0.850174
гугле 0.809912
гогл 0.786360
гугль 0.760508
гоогл 0.734248
гуг 0.731465
гугла 0.726011
гуугл 0.725497
гкгл 0.724901
гугул 0.722874
гогле 0.719596
гугд 0.719277
гугел 0.715329
гугал 0.713950
яндекс 0.695366
google 0.690433
googl 0.669867



«Бог продолжал ещё грешить».

Марковская цепь, обученная на библии