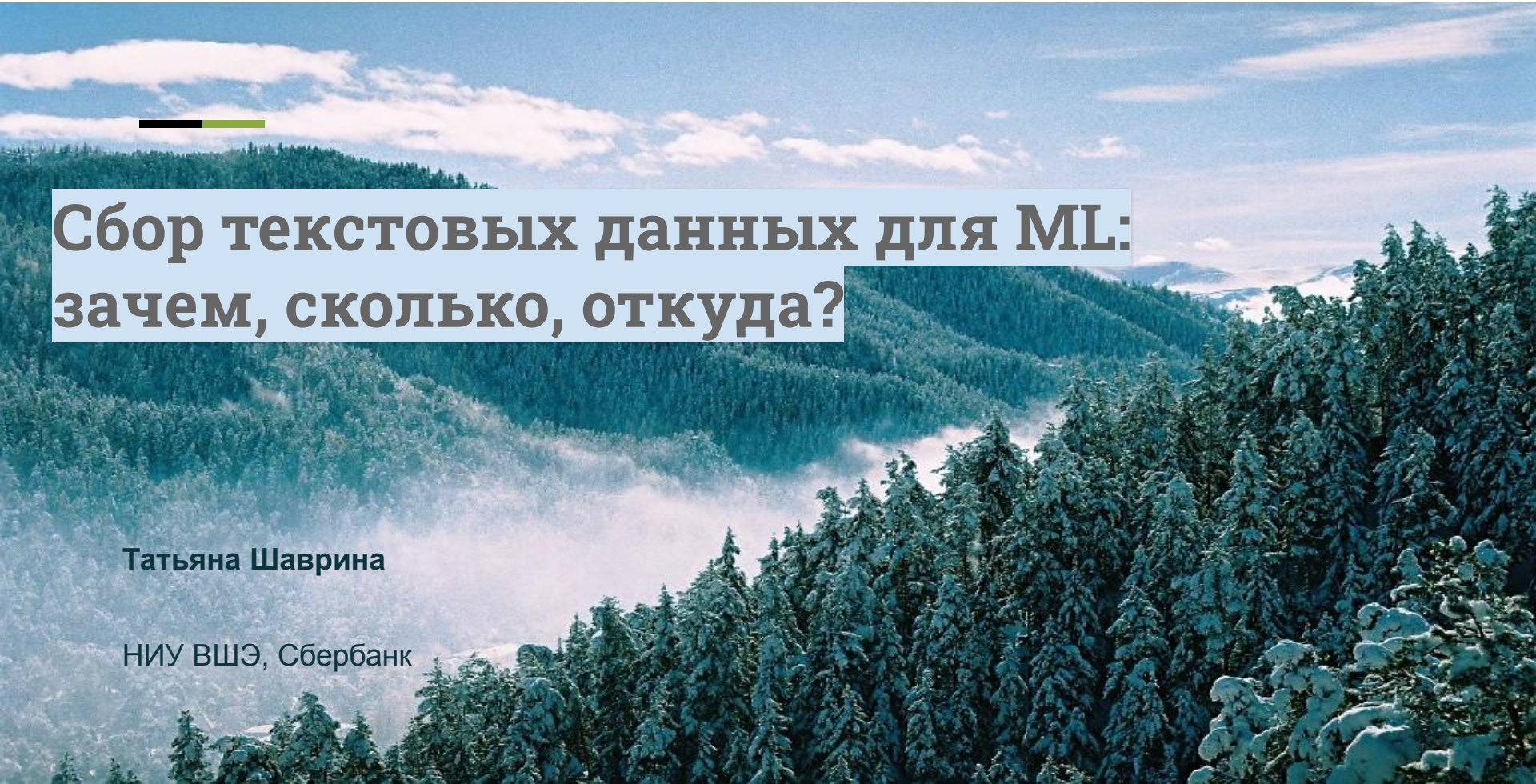


Сбор текстовых данных для ML: зачем, сколько, откуда?

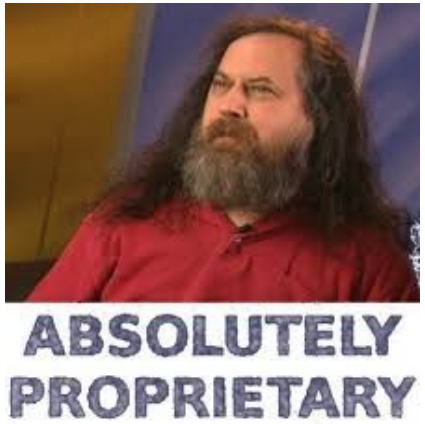
Татьяна Шаврина

НИУ ВШЭ, Сбербанк

Big Text Data

Миллиардные выборки их литературы, веба

- и все закрыто?



Смена корпусной парадигмы

- 1961 г - Brown Corpus, 1 млн слов
- 1990-00е гг Национальные корпуса. 100-500 млн слов
- 2000-2010е гг - Web as corpus, миллиардные корпуса
- 2010-2020е гг - индексы поисковых компаний, суррогаты корпусов

Сегодня мы пользуемся не только самими корпусами, но их суррогатами:

- fasttext, word2vec, GloVe
- BERT, ELMo, ULMFit

При этом:

- получить оригинальные текстовые данные из суррогатов нельзя,
- проверить соотношения жанров, источников - тоже
- скачать собственноручно данные подобного объема стоит огромного труда
- доступ к большим текстовым данным есть только у компаний-монополий

Что делать и кто виноват?

С точки зрения науки наличие оригинальных данных всегда лучше.

“Черный ящик” моделей не позволяет узнать оригинальное распределение слов, словосочетаний, жанров, авторов, гендерного и возрастного баланса в оригинальных данных.

Новые корпусные проекты должны облегчать лингвистам и разработчикам сбор оригинальных текстовых данных.

“Fair use” - многие источники разрешают сбор текстовых данных для научных целей, некоммерческих целей

закрыты юридически

Либрусек, Флибуста

THE QUESTION

стихи ру, проза ру

КиноПоиск

Common Crawl

ответы@mail.ru

Open Subtitles

YouTube

TED

Два.ч

Новостные сайты

закрыты технически

Как рассчитать примерный объем данных?

Как рассчитать примерный объем данных?

Источник данных должен быть масштабируемый + с необходимыми метаданными

1) максимально конкретизировать задачу, выделить конкретные примеры явлений под задачу
   - word2vec модель: хотим, чтобы похожие по смыслу слова считались близкими
   - примеры: известные синонимы оценочных слов
2) оценить частоту изучаемых явлений

- “плохой” IPM ~34.087471
- “хреновый” IPM ~0.715949
- “лажовый” IPM ~0.006211

выборка, содержащая по 100 контекстов на каждое слово - 16,6 миллиардов слов

Перплексия vs частота слов

Слова языка сильно различаются по перплексии контекстов вокруг себя

Самые однородные контексты у:

- числительных
- оценочной лексики
- названий профессий, национальностей, геолокаций
- тематических групп, терминов
- стилистически окрашенных слов

Самые разнообразные контексты:

- у имен собственных

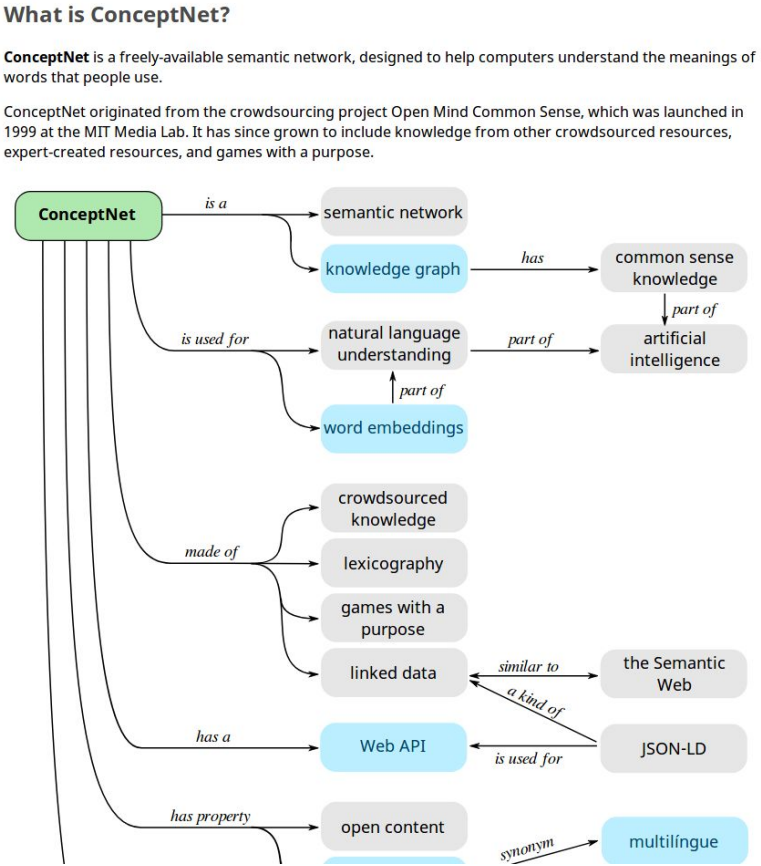
Где взять примеры для оценки?

Сами примеры:

- Ваша собственная интуиция
- Любые словари синонимов
- Онтологии - WordNet, YARN, Conceptnet...

Частоты слов

- корпуса, дающие статистику по источникам, которые вам нужны:
  - ГИКРЯ - социальные сети, новости, худлит
  - Sketch Engine - средневзвешенный рунет
  - НКРЯ - худлит



Настройка поиска

Введите ваш запрос

лажовый

Запрос количества слов

Результатов: 61

IPM: ~0.006211

Добавить запрос | Построить график

Искать среди слов: | **или** найти столько результатов:

Имя запроса: *** | Сохранить | Удалить | Поделиться с

Стат. запрос | Тестовый режим | Базовый запрос

Включить?	Сегмент	Слов:	Баланс корпусов Нормировать
☐	Живой Журнал - Кассандра	8720 млн.	
☑	ВКонтакте - Кассандра	9820 млн.	
☐	Новости - Кассандра	851 млн.	
☐	Журнальный Зал Кассандра	313 млн.	

Запустить | Остановить

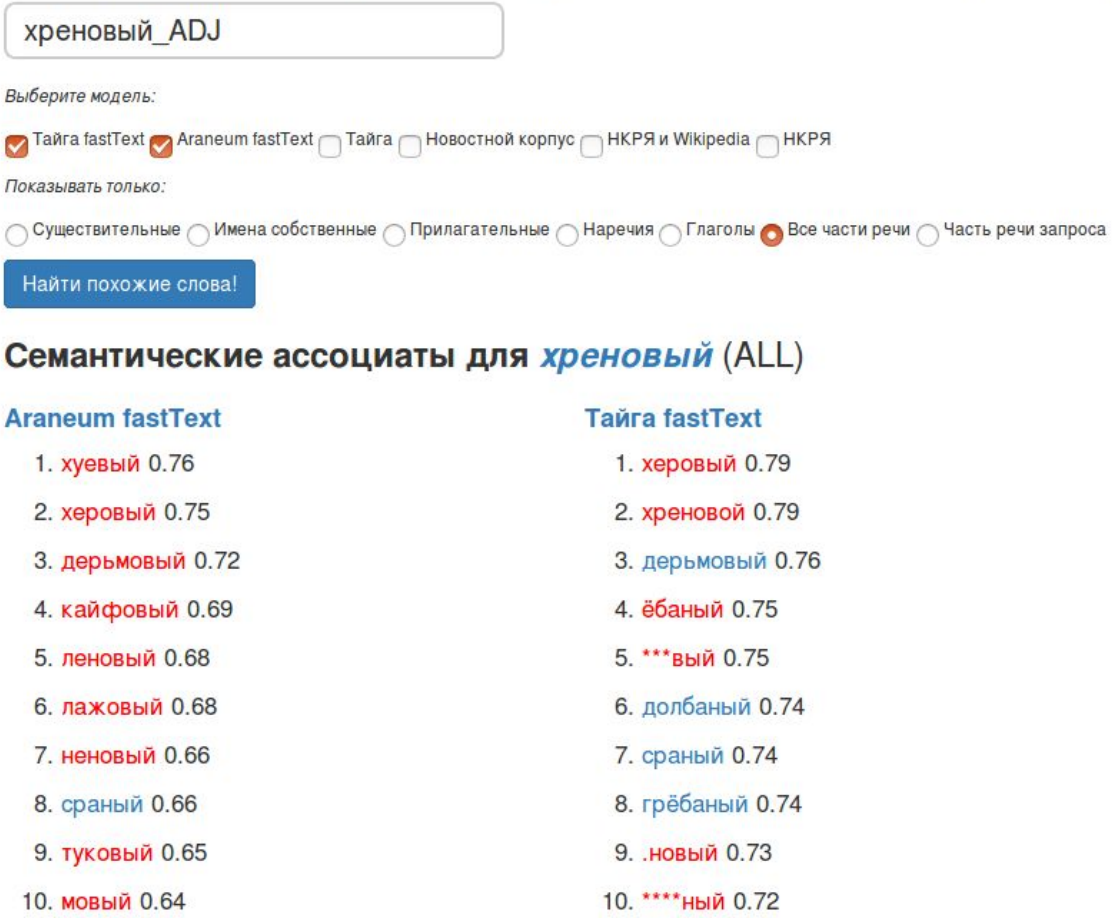
Где взять примеры для оценки?

Готовые модели и визуализации - **RusVectores**

https://rusvectores.org/ru/

Все модели обучены на корпусах до 10 млрд слов

Будет ли это вообще работать?



Самостоятельная ценность корпусов

Корпус - выборка из редких событий языка.

Репрезентативный корпус отличают:

- полнота примеров изучаемых явлений
- правильное соотношение явлений в корпусе vs “in the wild”
- контролируемое соотношение источников текстов, жанров в корпусе, авторского баланса и т.д.

Такой корпус может служить материалом для моделирования

Тем не менее, современные корпуса лучше не использовать именно как “обобщенное знание” напрямую

Векторные модели vs знание

Популярные word2vec модели и доля отношений RuWordNet, которые они сохраняют

relation	value	taiga	rnc	news	aranea
antonymy	FALSE	73.052	57.468	75.649	45.563
antonymy	TRUE	25.108	37.771	16.775	**48.701**
antonymy	not in vocabulary	**1.84**	4.762	7.576	5.736
cause	FALSE	68.439	55.15	78.239	31.229
cause	TRUE	19.435	**41.03**	10.133	15.282
cause	not in vocabulary	12.126	3.821	**11.628**	53.488
entailment	FALSE	91.922	80.13	89.322	64.438
entailment	TRUE	4.457	**17.734**	6.778	12.813
entailment	not in vocabulary	3.621	**2.136**	3.9	22.748
hypernymy	FALSE	88.951	82.181	87.644	61.34
hypernymy	TRUE	6.806	14.331	7.731	**19.714**
hypernymy	not in vocabulary	4.243	**3.488**	4.625	18.946
hyponymy	FALSE	81.565	76.485	79.872	63.195
hyponymy	TRUE	7.87	13.403	7.183	**17.751**
hyponymy	not in vocabulary	10.565	**10.112**	12.945	19.055

Как мы собираем ресурсы в Тайге

Подходят ресурсы:

- Собственнописные краулеры
- Не используем АПИ ресурсов, если они есть - Selenium + CSSSelector
- Azure
- Очистка - Beautiful Soup + скрипты унификации символов + Onion дедупликация
- Разметка - UDPipe

Демо: как собрать комментарии youtube без API

Omnia Russica: Taiga + Aranea + Common Crawl

Мы готовим самый большой когда-либо существовавший открытый корпус русского языка

- 50 млрд слов в едином пайплайне, с дедубликацией, в единой разметке
- возможность скачать текстовые данные + метаданные
- интерфейс для онлайн-поиска примеров

Araneum станет открытым корпусом русского языка

Формат разметки - UD 2.0, UDPipe

Формат xml для хранения как требование для поискового движка

Парсинг Common Crawl - можно ли сделать “грязные данные” чистыми?

Материалы

блог Александра Вейсова - как скачивать common crawl
https://spark-in.me/post/parsing-common-crawl-in-four-simple-commands

open-source инструменты для сбора корпусов http://corpus.tools/

мой блог - есть список всех ресурсов https://tatianashavrina.github.io/

Спасибо за внимание!