ИППИ РАН. Анализ данных

Обнаружение семантических сдвигов

Муталуп Эдем

О чем это я? (План доклада)

1. Разбор статьи
2. Текстовый корпус OpenCorpora
3. Выбор алгоритма векторного представления
4. Библиотека Gensim (Word2Vec)
5. Обучение модели
6. Результат

1. Разбор статьи

- **Постановка задачи**: Оценить способы выявления и количественного анализа семантических изменений слов на исторических промежутках, а затем – установить, существуют ли универсальные статистические законы, описывающие зависимость скорости семантической эволюции от частоты и полисемии слова.

Способы векторных представлений

- Positive Point-wise Mutual Information (PPMI)
- Singular Value Decomposition (SVD)
- Skip-Gram with Negative Sampling (SGNS)

PPMI

- **Positive Point-wise Mutual Information** (PPMI) — метод, в котором каждая компонента вектора слова отражает взаимную информацию с конкретным словом-контекстом.
- Формула расчёта:

$$M_{i,j}^{PPMI} = \max\left\{\log\left(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w_i)\hat{p}(c_j)}\right) - \alpha,\ 0\right\}$$

- Матрица PPMI вычисляется на основе частот совместной встречаемости слов, при этом отрицательные значения обнуляются.
- Преимущество метода: понятная интерпретация (каждое измерение соответствует конкретному слову-соседу).
- Недостаток: высокая чувствительность к редким событиям и шуму в корпусе.

SVD

- **Singular Value Decomposition** (SVD) — метод факторизации матрицы, используемый для снижения размерности.
- Исходная матрица (PPMI) раскладывается на компоненты: $U\Sigma V^\top$.
- Сокращённая размерность позволяет устранить шум и выделить основные семантические особенности.
- Плюс: хорошо фиксирует даже слабые семантические сдвиги.
- Минус: требует значительных вычислительных ресурсов для факторизации.

SGNS

- **Skip-Gram with Negative Sampling** (SGNS) — алгоритм, реализующий модель word2vec, которая обучается предсказывать появление слова в определённом контексте.
- Для повышения точности используются «негативные примеры» — случайно выбранные слова, которые не должны встречаться вместе с данным словом.
- В результате обучения получаются векторные представления слова (и контекстные вектора), объединяемые для анализа.
- Преимущество: высокая эффективность и масштабируемость на больших текстовых корпусах.

Результаты исследования

- Исследование показало, что скорость семантических сдвигов описывается зависимостью:

$$\Delta(w) \propto f(w)^{\beta_f} \times d(w)^{\beta_d},$$

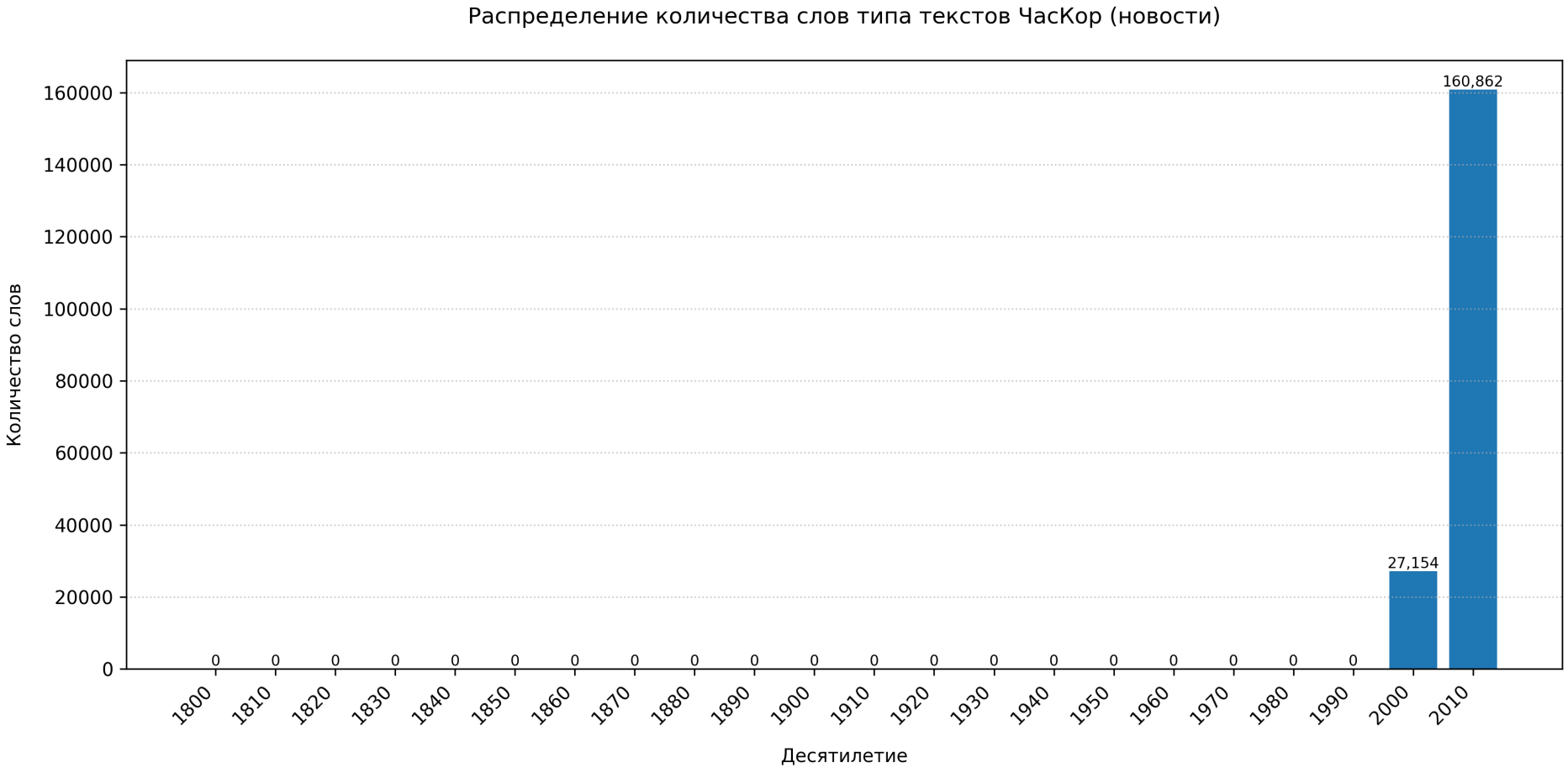
где $\beta_f < 0$ и $\beta_d > 0$.

- **Закон конформизма**: слова, употребляемые чаще, меняются медленнее.
- **Закон инновации**: слова с большим количеством значений изменяются быстрее даже при равной частоте.
- Частота и полисемия совместно объясняют от 48% до 88% дисперсии темпов семантических изменений.

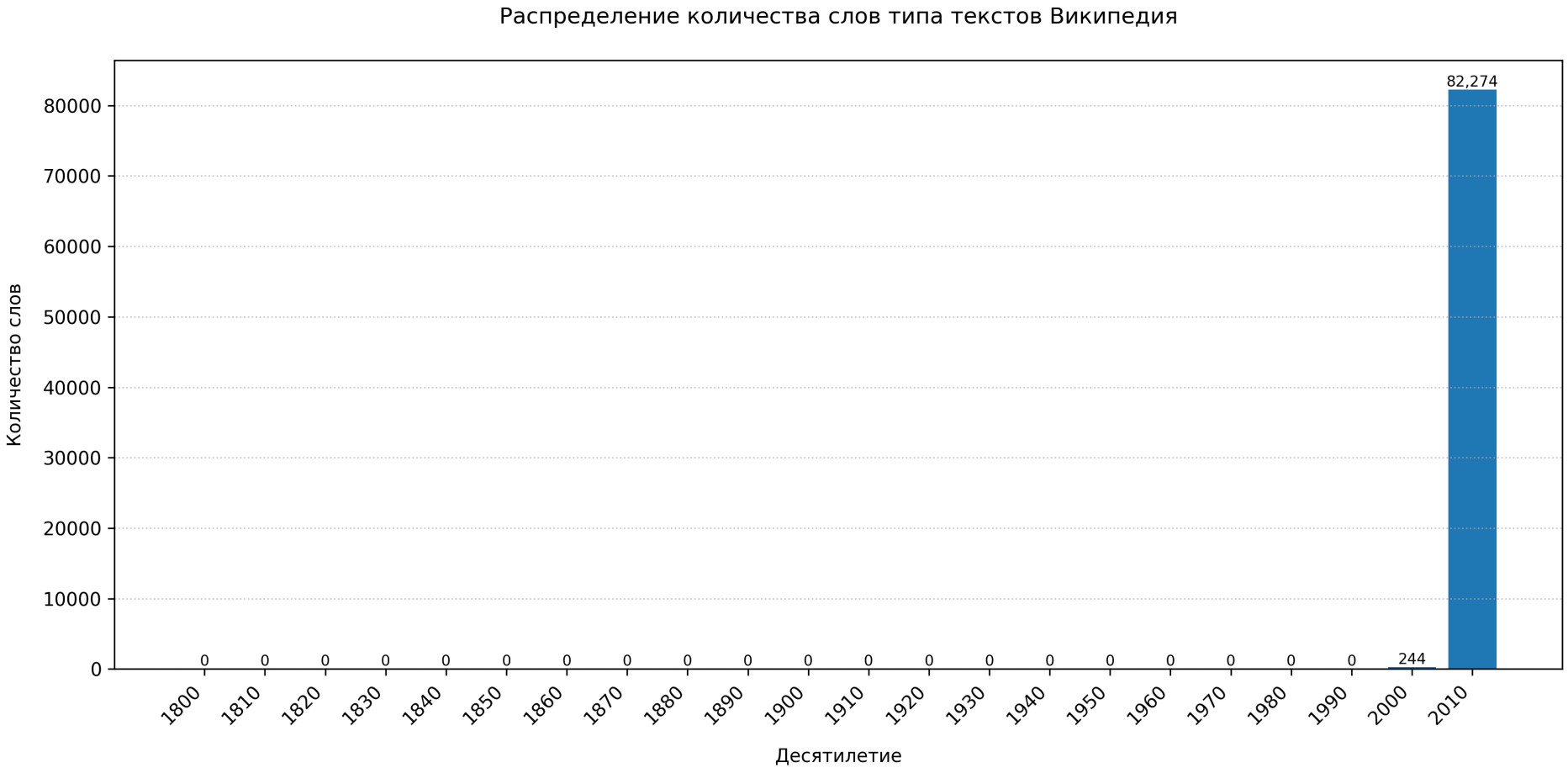
2. Текстовый корпус OpenCorpora

- **OpenCorpora** — Открытй корпус текстов русского языка, созданный в 2009 году.
- Рассмотрим несколько типов текстов:
  - ЧасКор (новости)
  - Википедия
  - Блоги
  - Худож. литература
  - Нон-фикшн
- Среди этих типо выберем пару ЧасКор(новости) и Худож. литература. Почему?

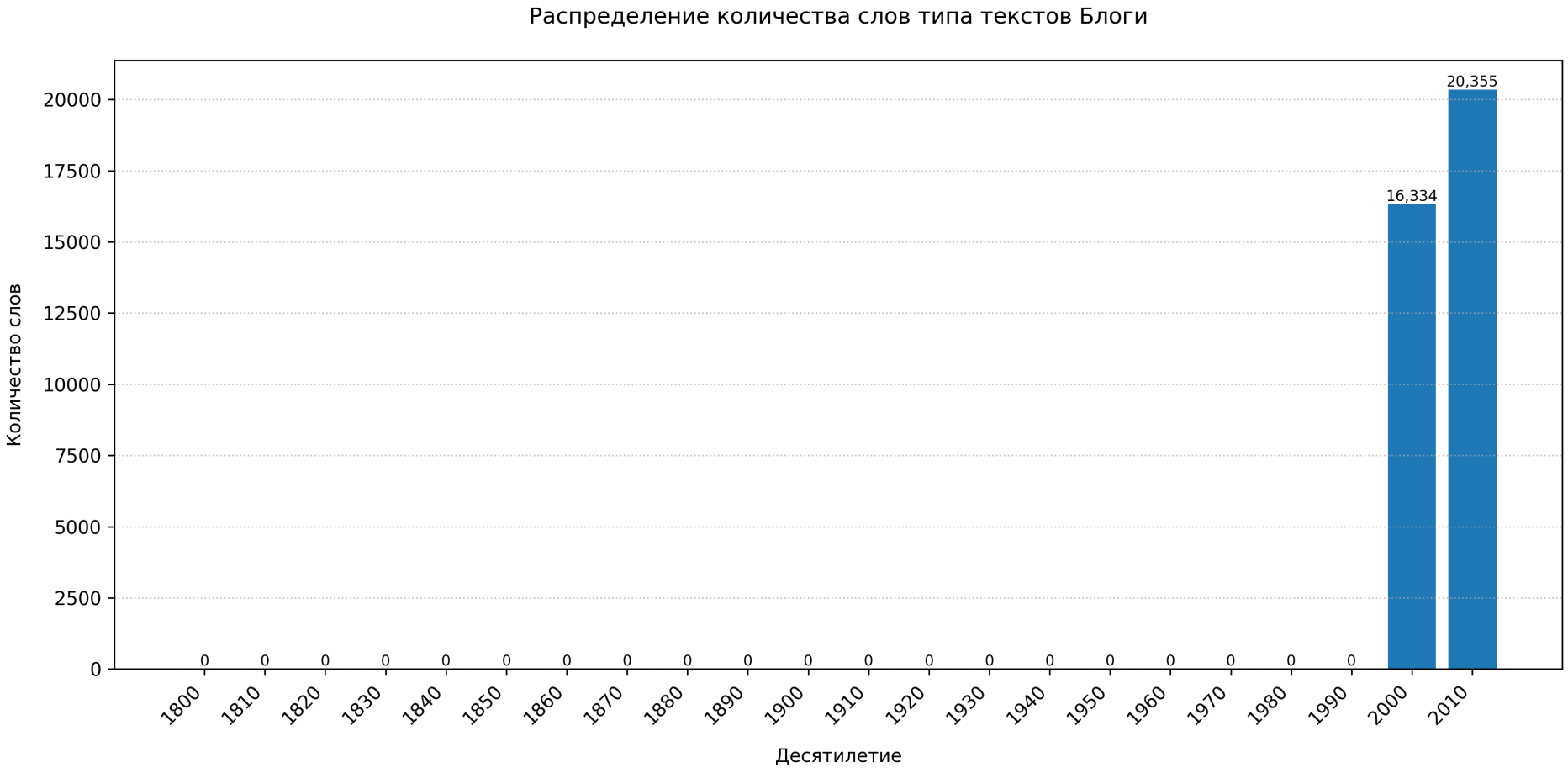
Выбор типов текстов



Выбор типов текстов

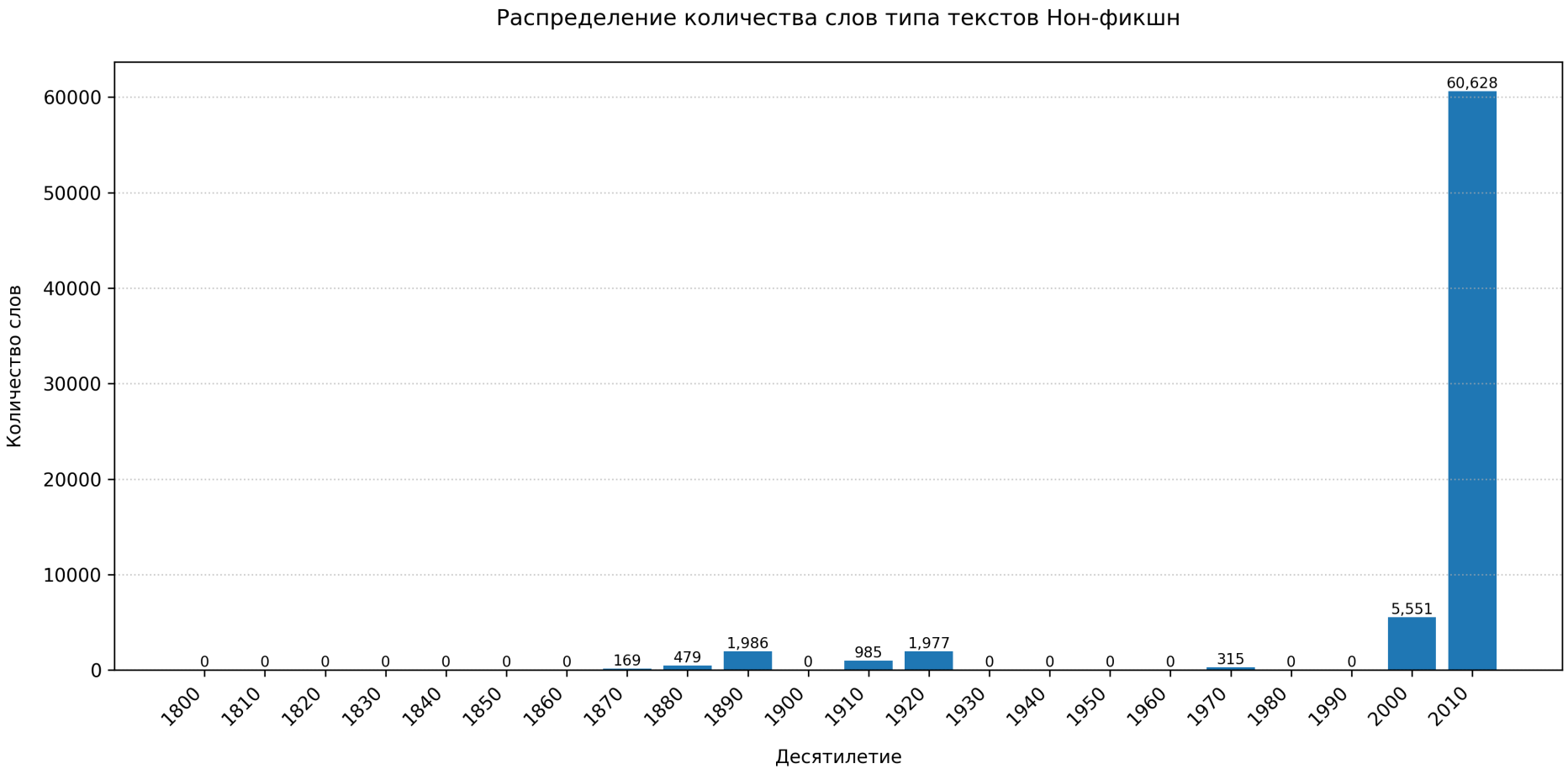


Выбор типов текстов

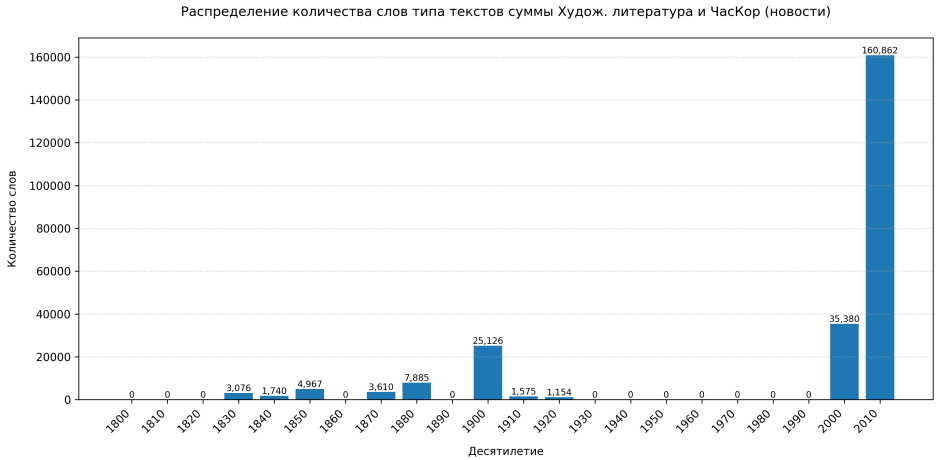


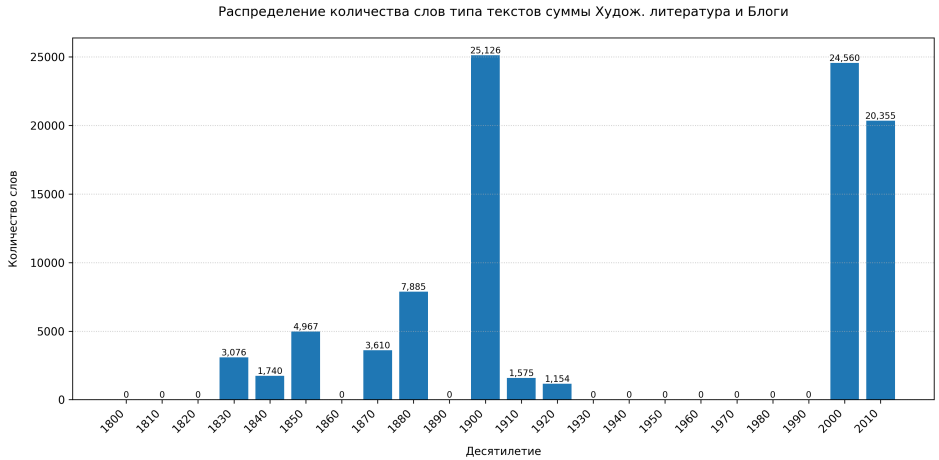
Выбор типов текстов

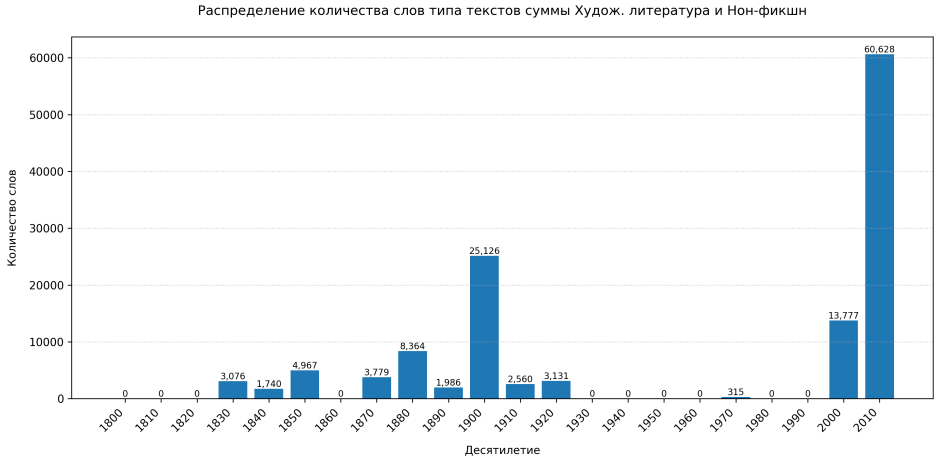
Выбор типов текстов



Совместно с Худож. литература







3. Выбор алгоритма

- В исследовании были протестированы три алгоритма построения векторных представлений: PPMI, SVD и SGNS.
- По результатам статьи, **SGNS** продемонстрировал большую чувствительность к семантическим изменениям.
- Это делает SGNS оптимальным выбором для обнаружения семантических сдвигов.

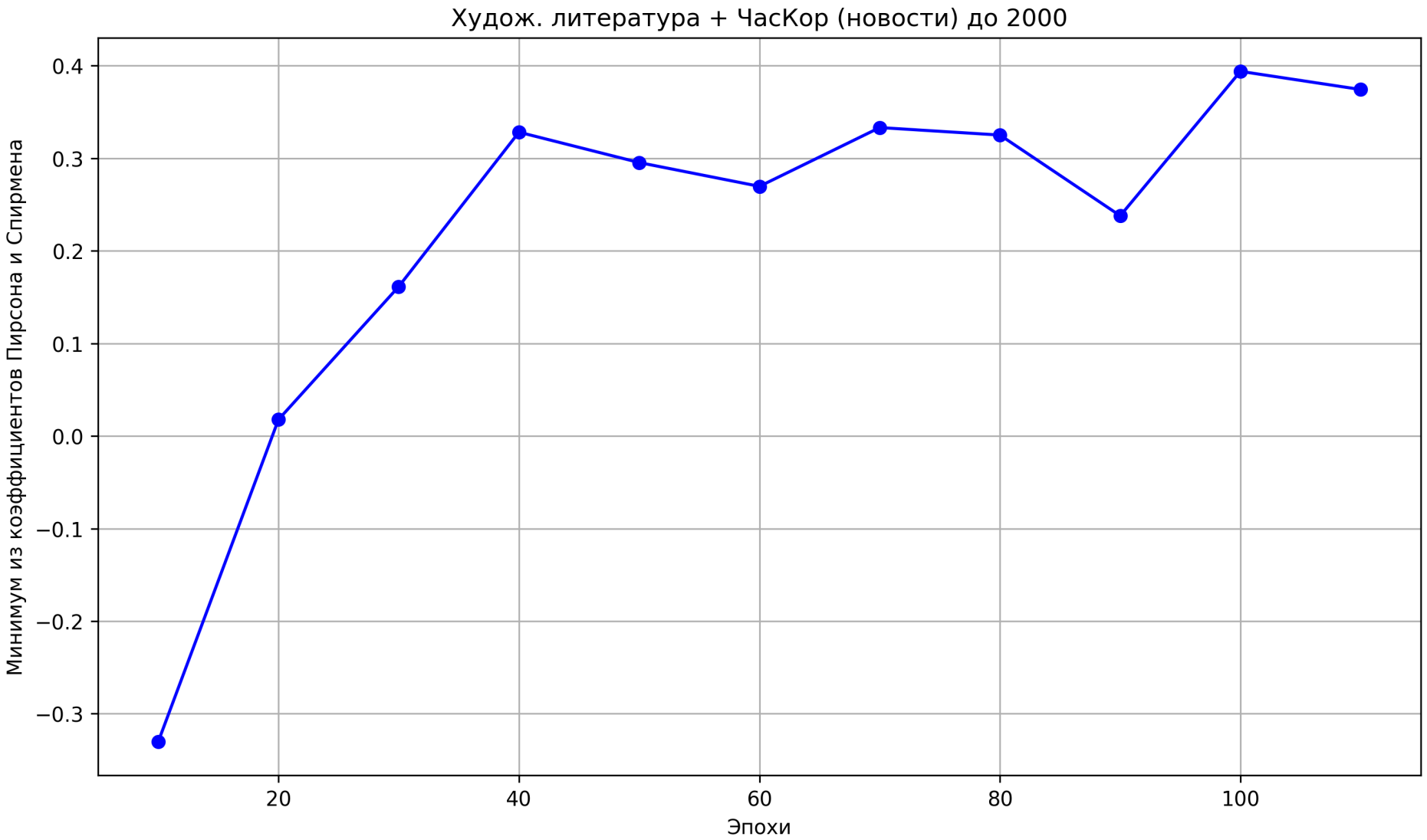
4. Библиотека Gensim и модель Word2Vec

- **Gensim** — это популярная библиотека на Python для тематического моделирования и обработки больших текстовых корпусов.
- Библиотека предоставляет реализацию модели **Word2Vec**, которая позволяет обучать векторные представления слов (таких как SGNS).

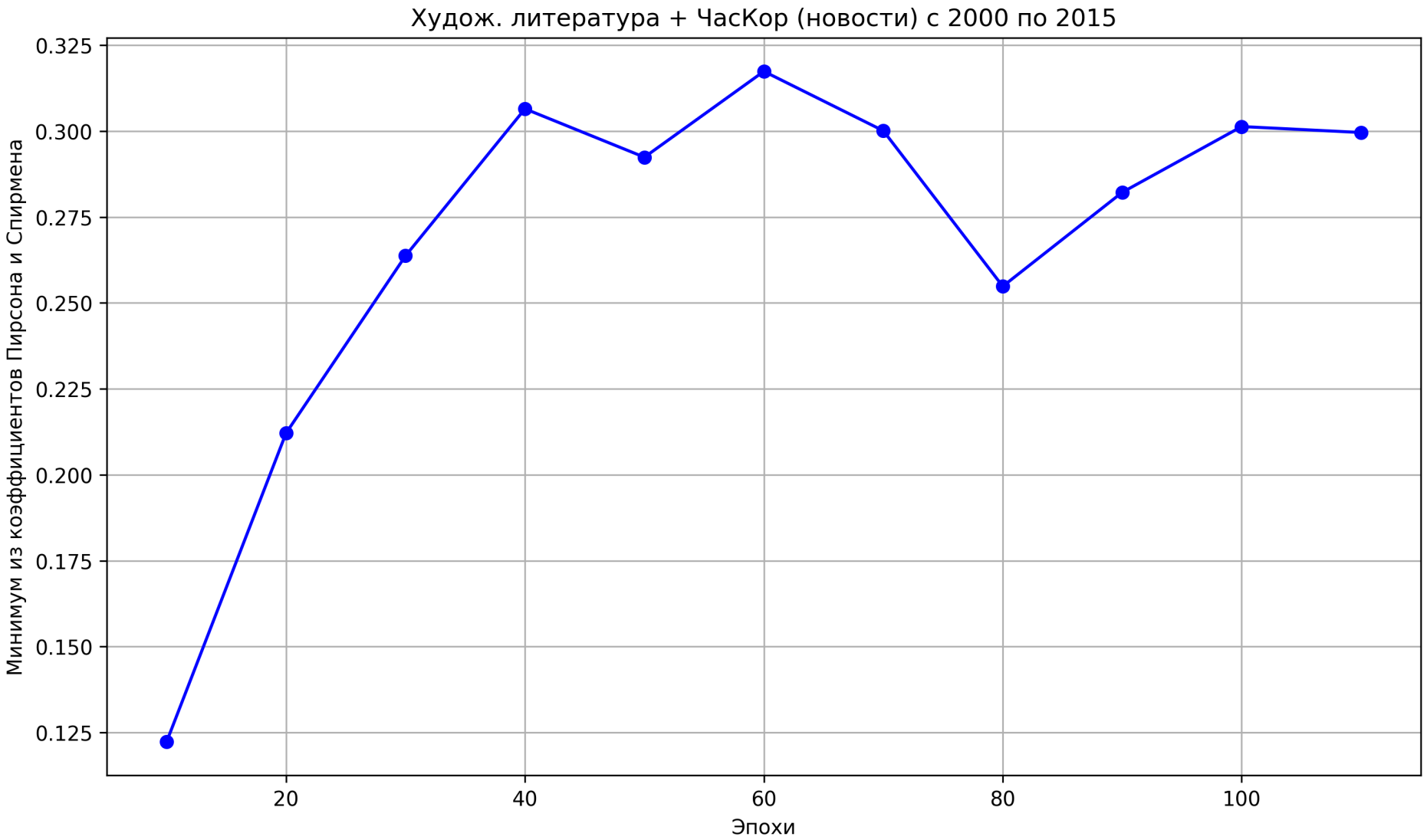
5. Обучени модели

- Для обучения модели, использовался корпус слов (Худож. литература + ЧасКор (новости)), разделенный на две части: до 2000 года и с 2000 по 2015 год
- Согласно статье в качестве пареметров машинного обучения использовались значения:
  - vector-size=300
  - window=4
  - min-count=3
- Далее описан подбор эпох

Подбор параметра эпох



Подбор параметра эпох



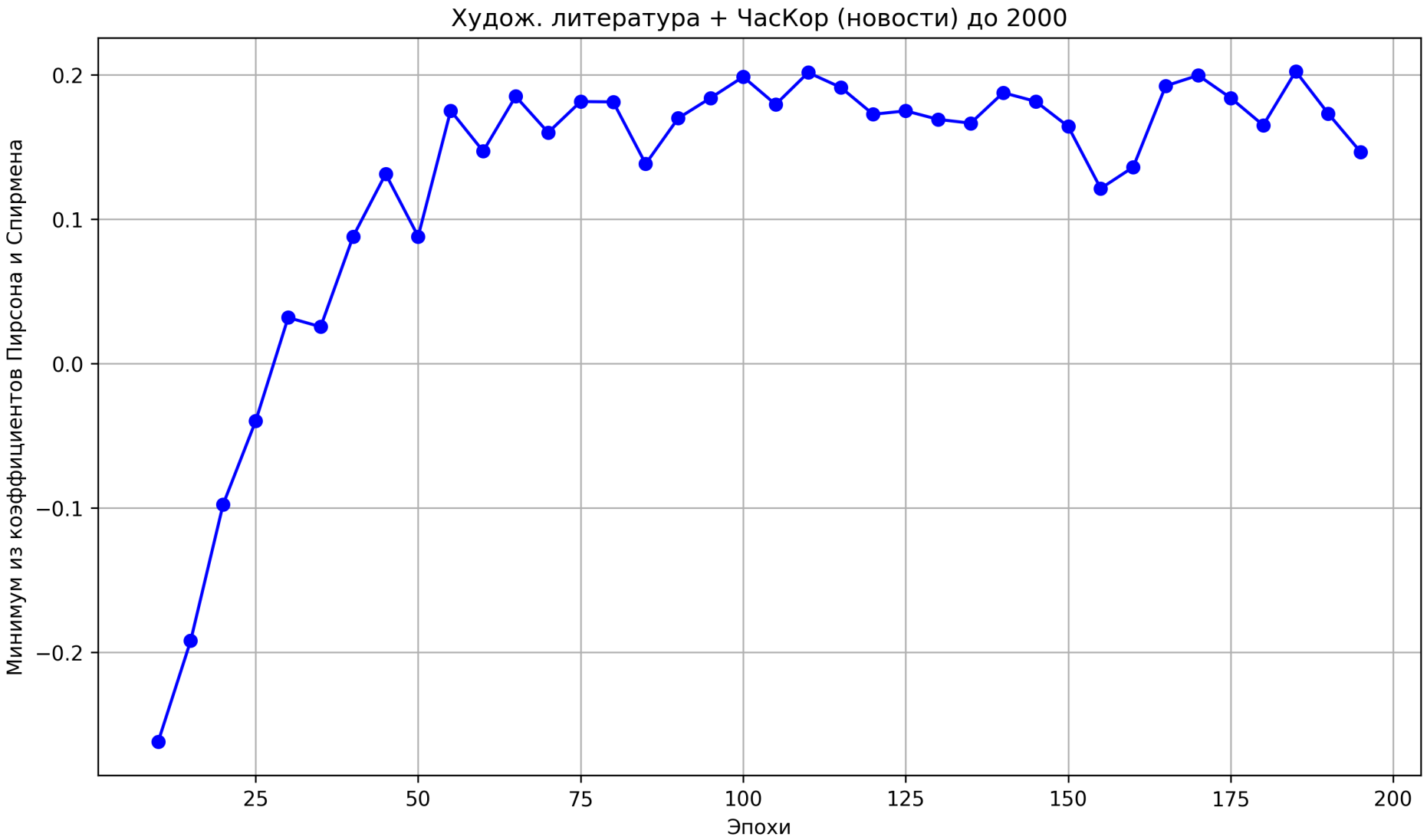
6. Результат

- **Все очень плохо**
- Качество обученной модели низкое, коэффициенты Пирсона и Спирмена, ниже 0.4.(начиная с этого значения, модель считается репрезентативной)
- Общий словарь моделей $\approx$ 1200 слов .(крайне низкое значение)
- **В чем проблема?** — Предложенный корпус слов очень мал для обучения модели, размер корпуса, используемого для обучения модели чуть больше «Преступления и Наказания», и более чем в 2 раза меньше «Войны и Мира», в то время как, действительно репрезентативыне модели обучается сотнях миллионов слов.

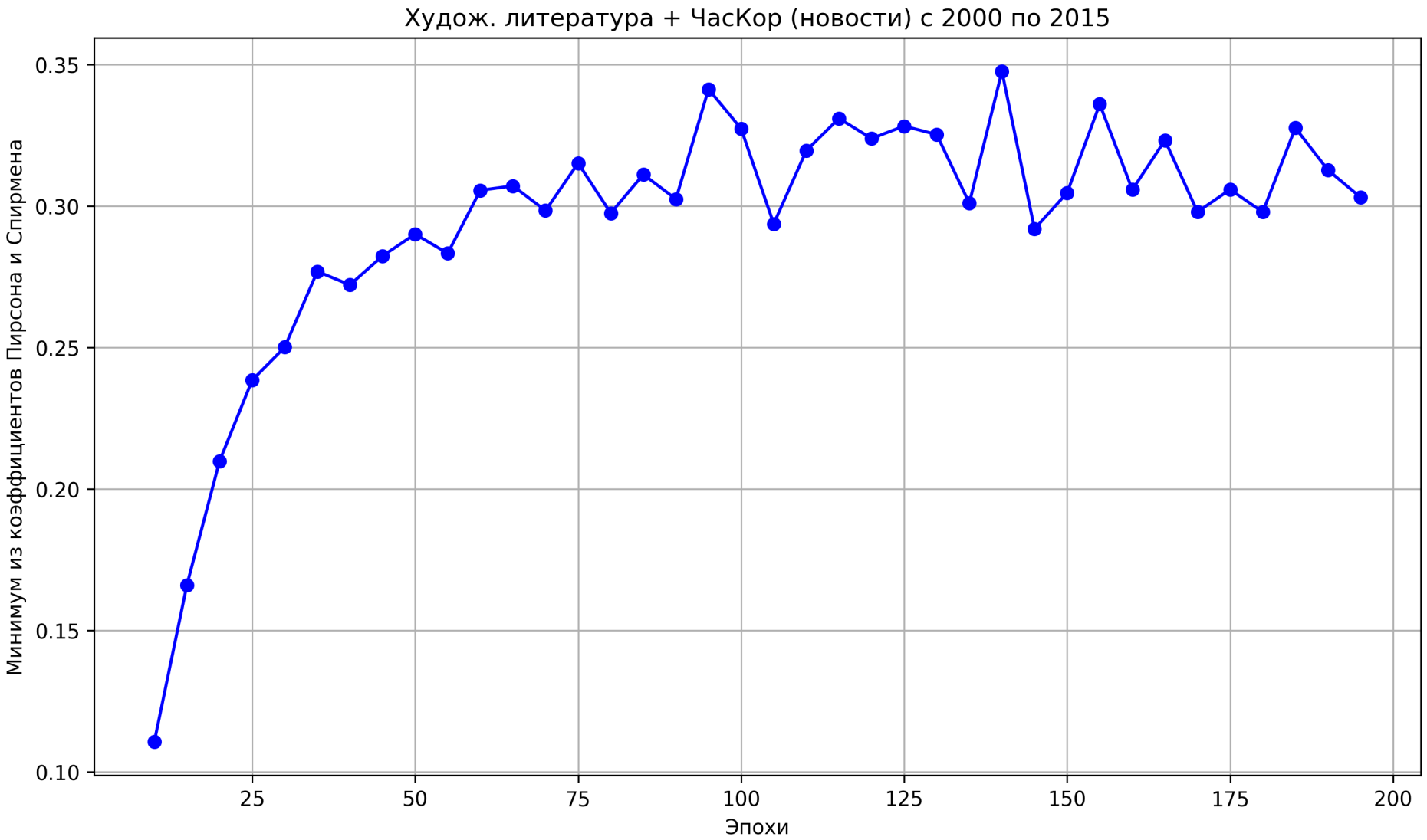
Попытка спасти

- Было замечено, что треть текстов типа «Худож. литература» не маркированы годом. (Отчего и не попали в данные для обучения)
- Вручную были добавлены метки в общий xml файл, согласно дате написания произведения или дате перевода (в случае зарубежного автора)
- Это увеличело данные для обучения приблизительно на 30 000 слова
- Но и это не спасло ситуацию.

Качество моделей на новых данных



Качество моделей на новых данных



Конец

- Спасибо за внимание!