Статистическая обработка результатов поиска в дифференциальных корпусах

Ю. Куратов, М. Б. Лагутин
ГИКРЯ, МФТИ, МГУ

Задачи

- Реализация статистического модуля в ГИКРЯ
  - Определение доверительного интервала для результатов поиска
  - Отображение найденных смещений в поисковой выдаче по некоторым дифференциальным признакам
- Оптимизация поиска в корпусе
  - Определение минимально необходимого размера подкорпуса для заданного запроса

Зачем это все?

- Нет известных инструментов/корпусов предоставляющих информацию о точности полученных результатов поиска
- Поиск в больших корпусах затратен по вычислительным ресурсам
- Данные могут быть неоднородны по различным признакам

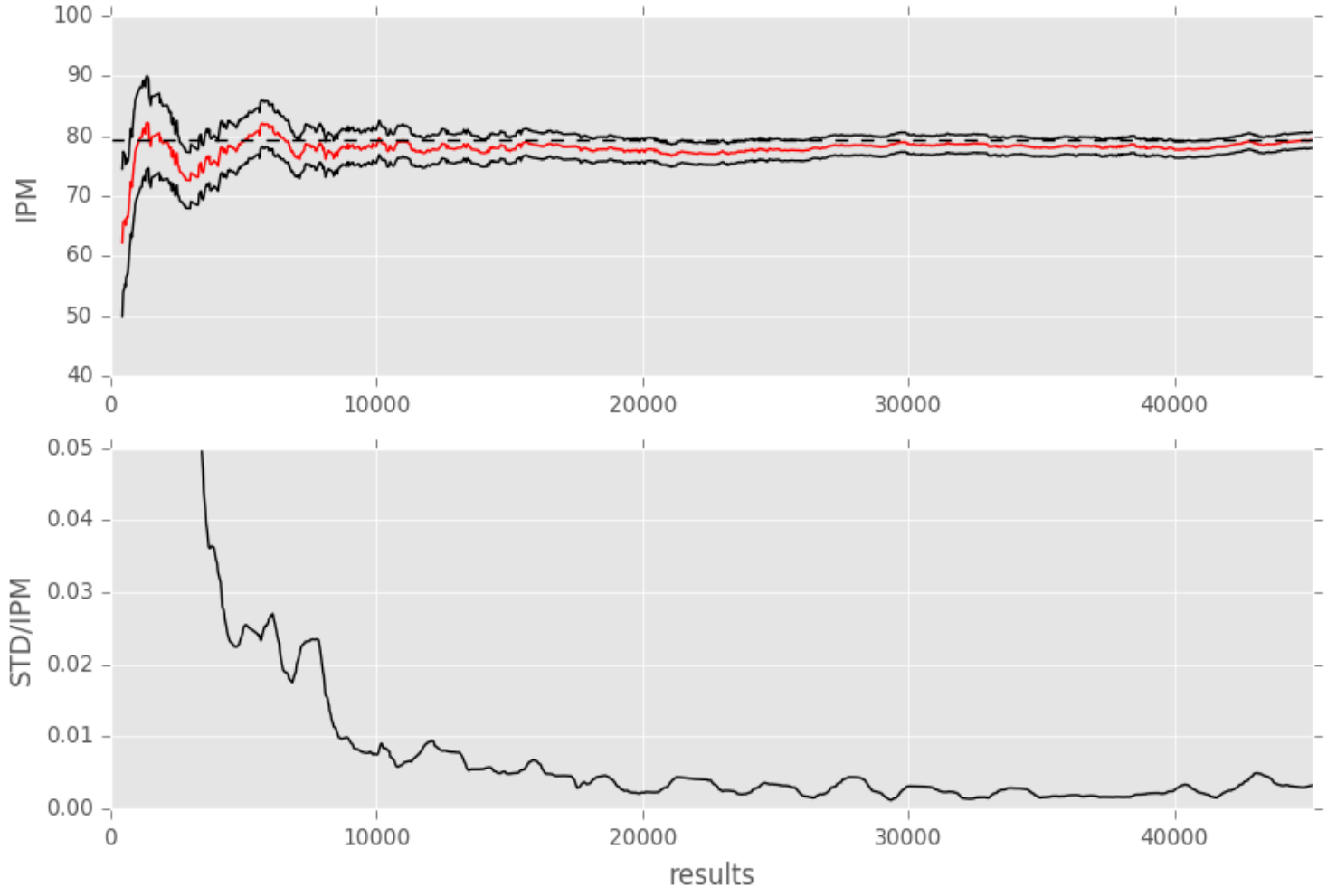
Проблемы

- Вхождения некоторой единицы в корпус не являются независимыми случайными событиями
- Нельзя просто взять число вхождений *K* среди *N* слов и за частоту принять величину $\frac{K}{N}$
- Моделирование подбрасыванием монетки

Подходы для решения

- Базовая методика
  - Основана на подокументных частотах $p_i = \frac{v_i}{l_i}$
  - Доверительный интервал: $\bar{p} \pm Q_\alpha \frac{S}{\sqrt{n}}$
- Методика со стратификацией
  - Разбиваем тексты в корпусе на T страт
  - Внутри каждой страты оцениваем $\bar{p_j}$, S_j
  - Зная доли страт q_j во всем корпусе оцениваем $\bar{p}$ и S и строим доверительный интервал

[lemma=фильм]. Ж3.



Результаты

- Тестирование: 36 различных запросов на 5 корпусах (180 значений частот)
- Фиксированное значение порога = 0.01:
  - 111 из 128: 86.71%
- Уменьшение размера выборки:
  - Экономия в среднем в 6 раз (по сравнению с поиском на всем корпусе)
  - Уменьшение достигало 20 раз для запросов с большим числом результатов

Спасибо за внимание!