Статистические распределения слов в русскоязычной текстовой коллекции

Антонов А.В.
Баглей С.Г.
Мешков В.С.
Суханов А.В.

КОРПОРАЦИЯ Галактика®

{ alexa, baglei, meshkov, sukhanov } @galaktika.ru

Представление о вероятностном порождении текста

- Алгоритмы обработки естественно-языкового текста

  Слова языка, последовательности слов в тексте.

- Криптография

  Последовательности кодовых слов

- Молекулярная биология

  Последовательности-триплеты в цепочках ДНК аминокислот (кодоны)

- Теория игр

  Последовательности действий участников

Модели вероятностного порождения текста

- Модель Бернулли

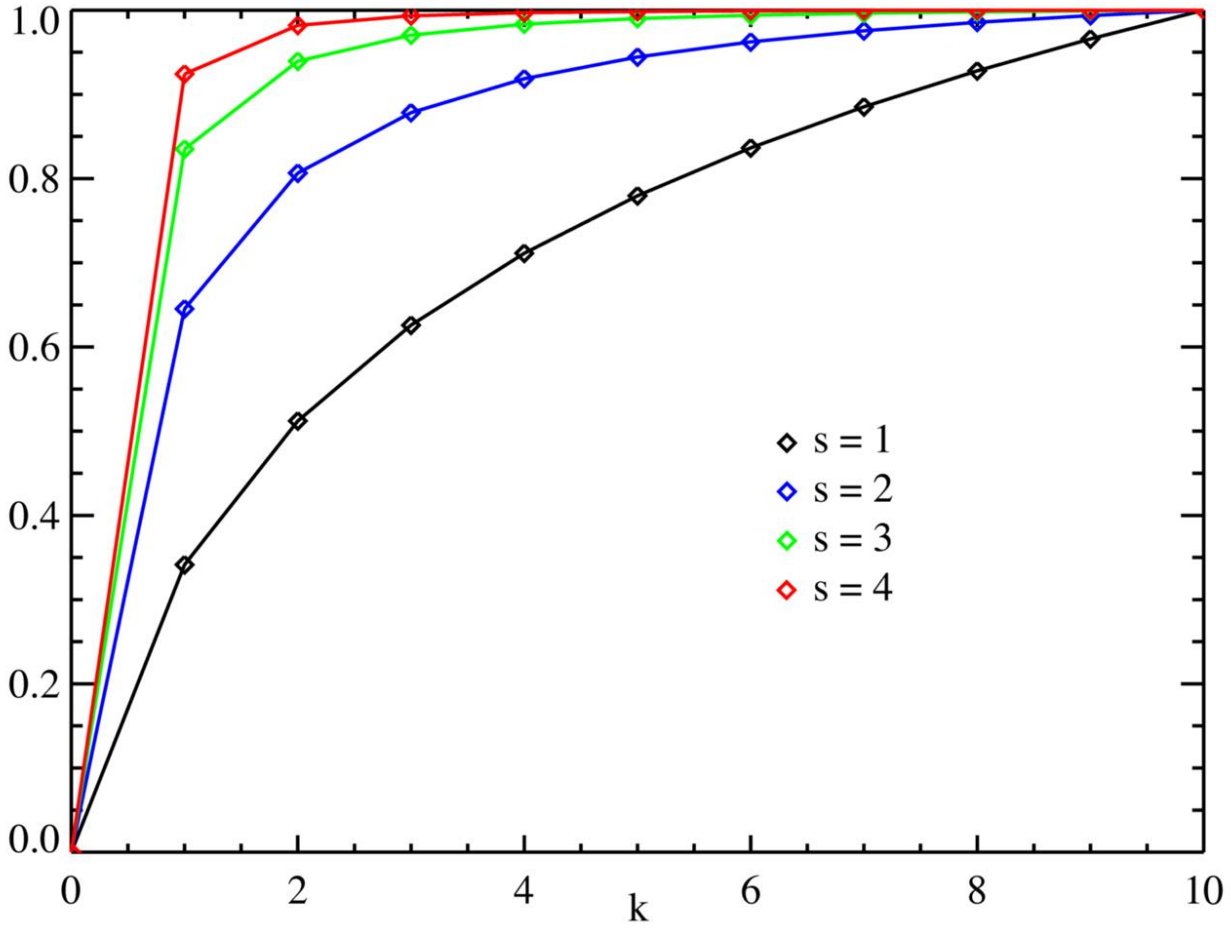
Текст формируется случайным образом. Вероятности появления словарных слов равны в любой момент времени.

- Модель Маркова

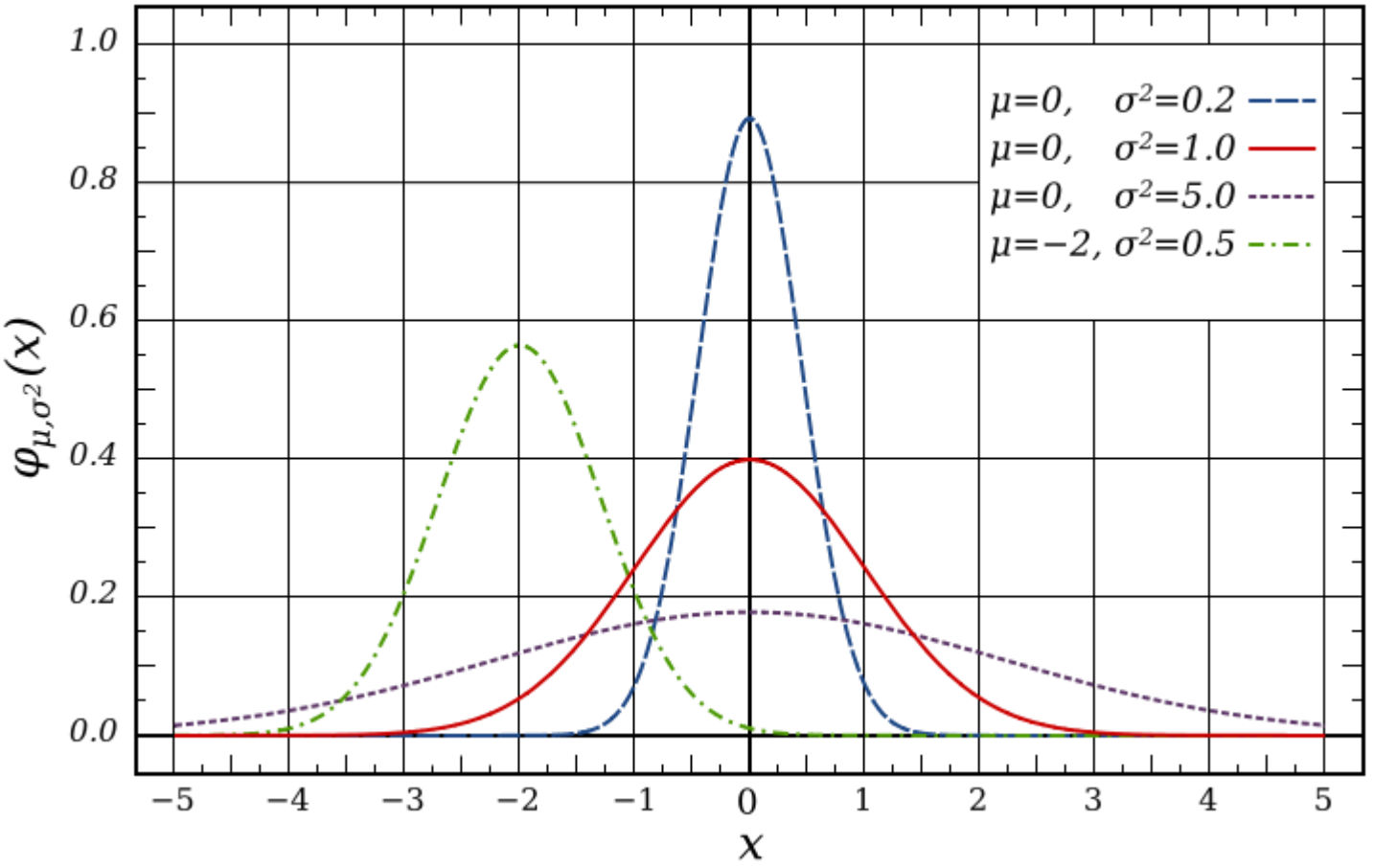
Однородная марковская цепь порядка n. Вероятность появления слова в заданной позиции зависит только от n предшествующих слов и не зависит от позиции, в которой оно появляется.

Марковская цепь порядка 0 → модель Бернулли.

Зависимость между рангом слова и его частотой: закон Ципфа

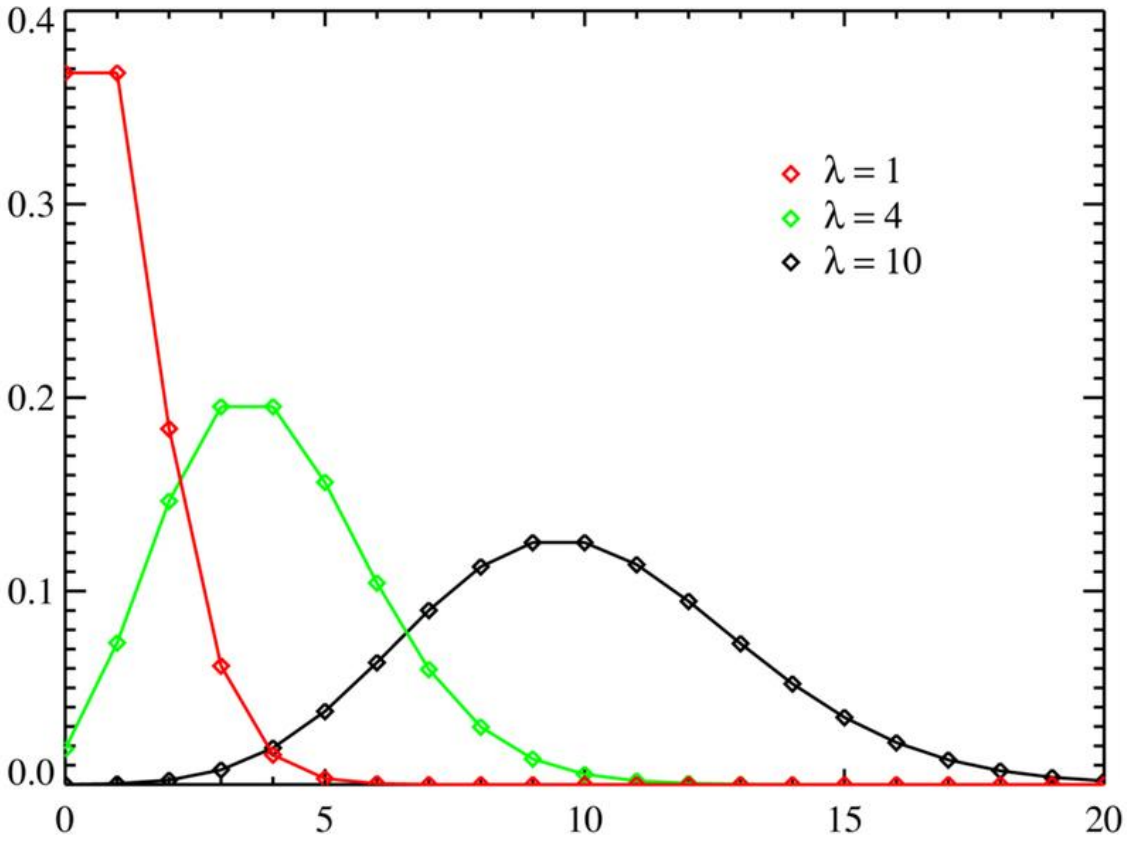


Гауссово распределение статистик частых слов в коллекции



Пуассоновское распределение статистик редких слов в коллекции

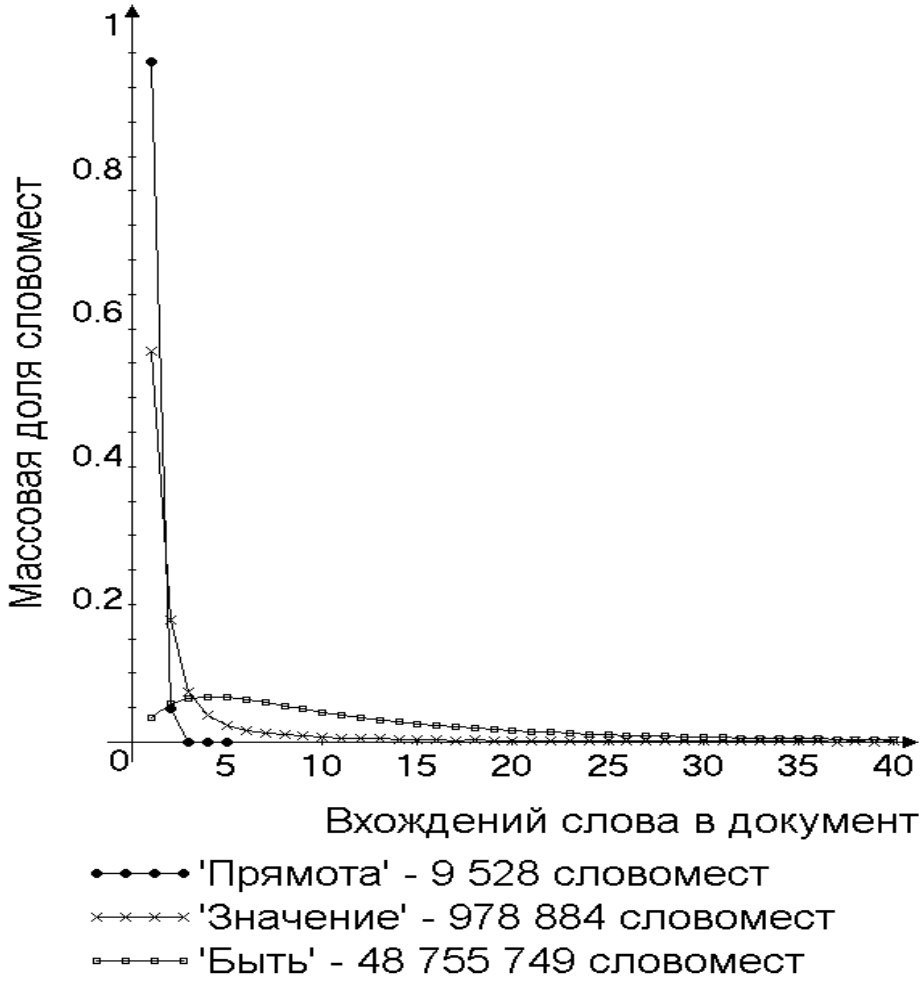
$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$



Экспериментальный массив

- Коллекция из 14,5 миллионов новостных сообщений на русском языке за 14-летний период, с 1995 по 2008 год;
- средняя длина документа в базе – 503 слова;
- общее количество словомест – 7,3 миллиарда;
- объем словаря базы – 15,6 миллиона слов;
- количество источников-СМИ – около 2 тысяч.

Взвешенные распределения некоторых слов по отдельным текстам коллекции



Представление статистических данных

- Выбраны слова с достаточной статистической базой (122 тысячи слов с не менее 500 вхождениями)
- Проведено упорядочение слов по частоте
- Сформирован вектор вхождений для каждого слова (по 1, 2, 3, …, 255 и свыше вхождений в документ)
- Элементы вектора взвешены по количеству вхождений в тексты, полученные значения нормированы

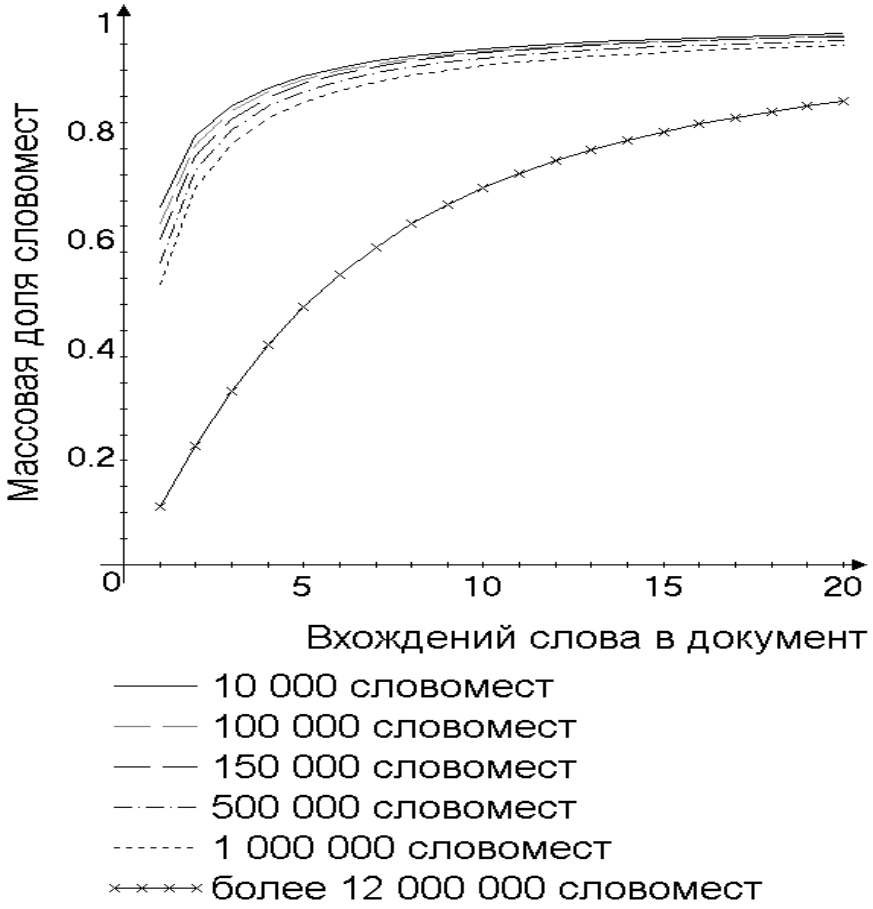
Построение распределений

- Выбраны уровни частот слов для построения распределений:

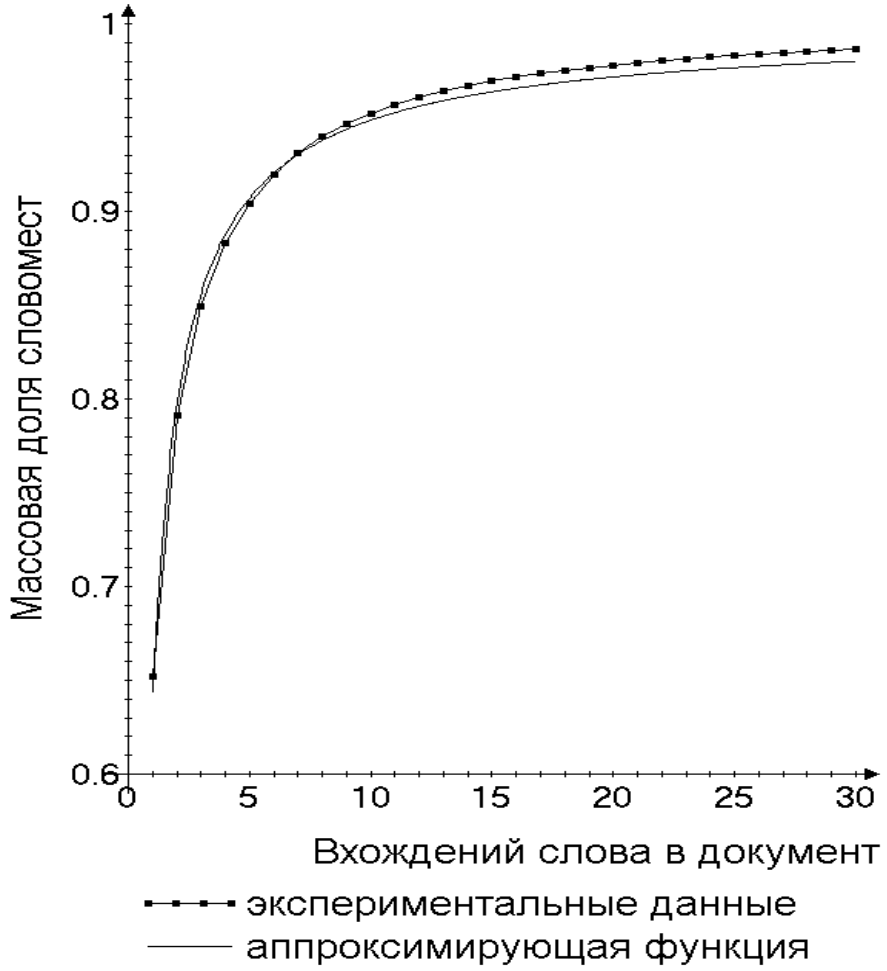
- 10 000;
- 100 000;
- 150 000;
- 500 000;
- 1 000 000

- Выделены по 1 000 слов в каждом частотном диапазоне
- В каждой группе слов вычислены средние арифметические значения элементов частотных векторов

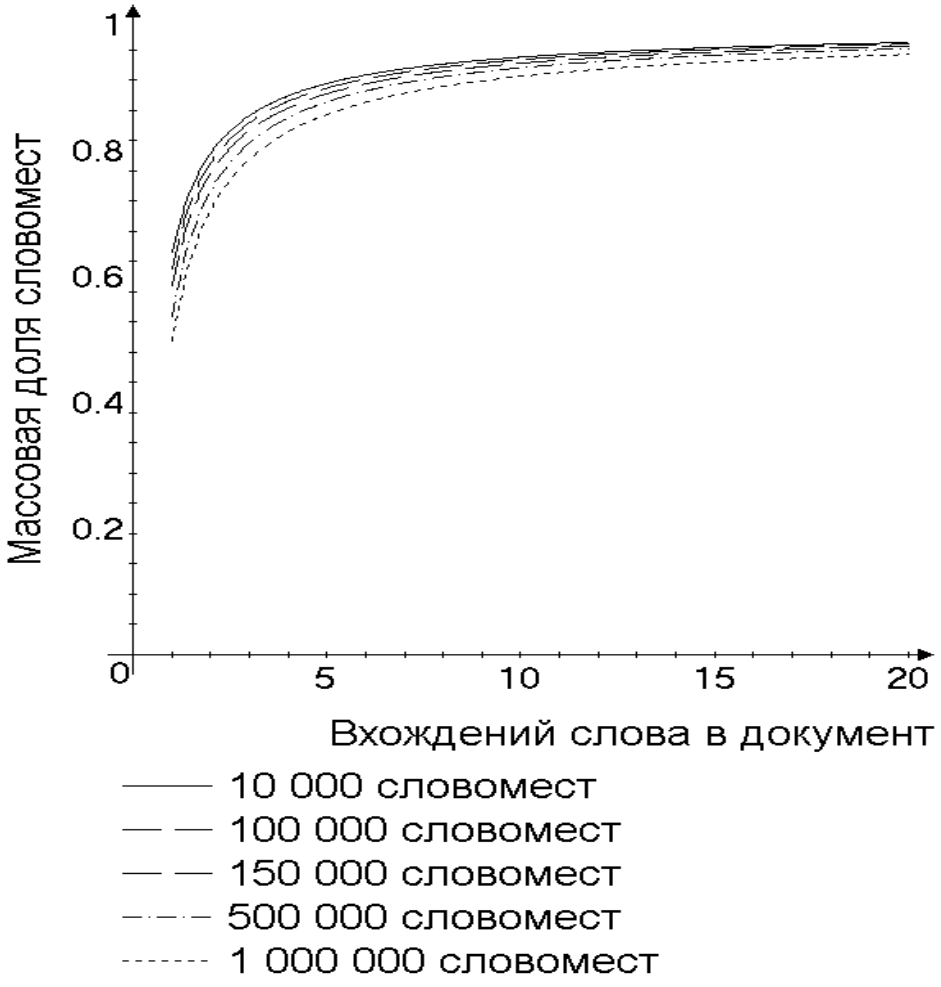
Накапливающиеся распределения в некоторых частотных диапазонах



Аппроксимация распределений



Аппроксимирующие функции некоторых частотных диапазонов

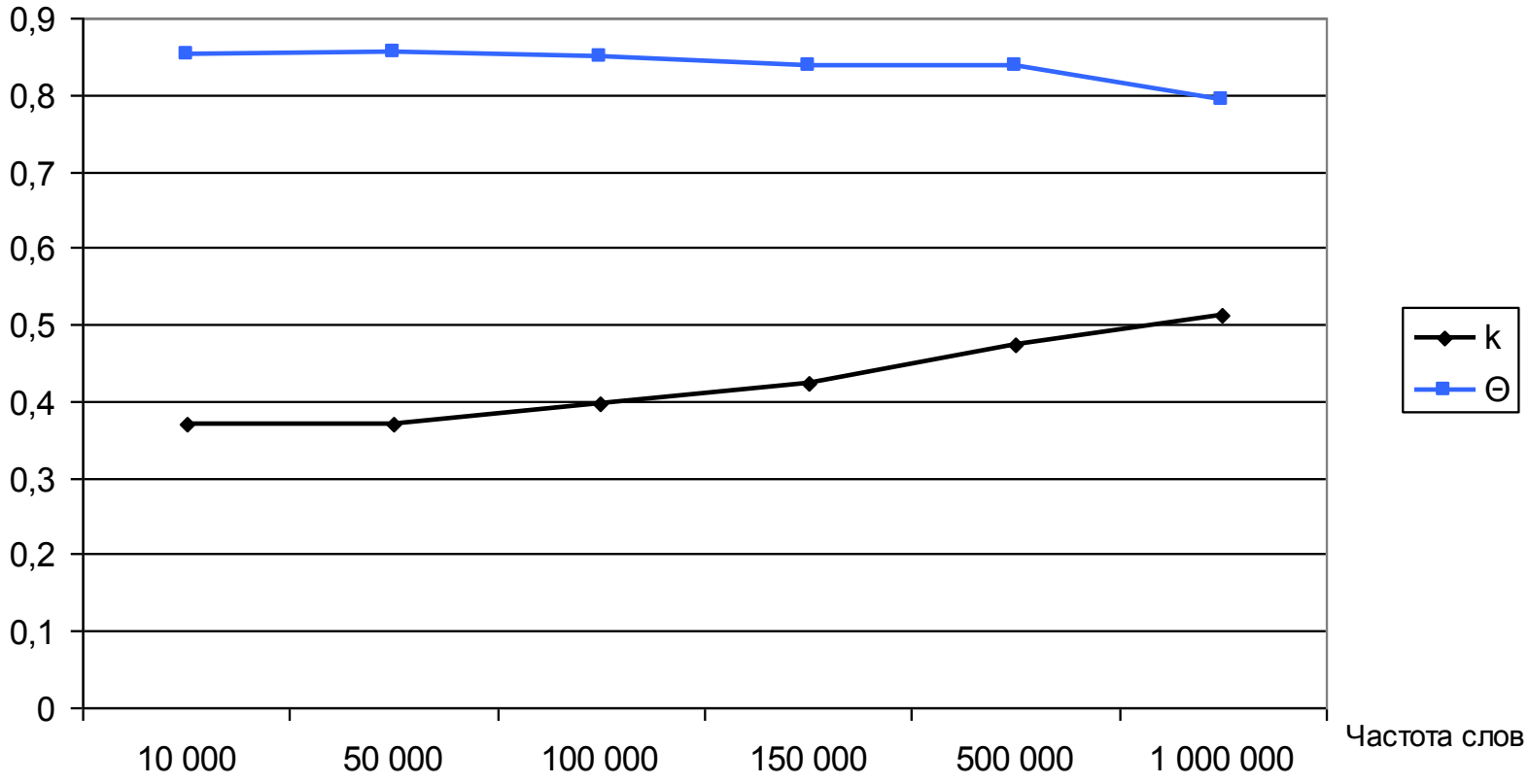


Коэффициенты аппроксимирующей функции

$$y = 1 - \frac{k}{j^{\theta}}$$

Уровень частот слов, *словомест*	*k*	*Θ*	*s*
10 000	0.370	0.854	0.0109
50 000	0.369	0.857	0.0084
100 000	0.397	0.851	0.0096
150 000	0.424	0.839	0.0106
500 000	0.473	0.838	0.0112
1 000 000	0.512	0.794	0.0101

Коэффициенты аппроксимирующей функции



Основные результаты

- Получены усредненные параметры распределения некоторых частотных диапазонов слов в большой новостной русскоязычной коллекции текстов;

- На основе анализа данных выделен частотный диапазон, свойственный словам, несущим информацию о предметных областях;

- Для некоторых частотных уровней проведена аппроксимация значений, найдены коэффициенты соответствующих функций распределения.

Перспективы развития

- получение более реальной оценки вероятности появления слова в произвольном тексте на русском языке;

- использовать функции распределения, полученные в данной работе, в сочетании с оценками вероятностей появления слов для задачи расчета оценки соответствия этих слов некоторой произвольной коллекции текстов.

Спасибо за внимание