

Визуализация статистики вхождения слов

Ландэ Дмитрий Владимирович,
д.т.н., зам. директора ИЦ «ЭЛВИСТИ»

Киев-2009

При подборе ключевых слов для поиска важно учитывать такое их свойство, как «различительная» или дискриминантная сила.

Ведь если слово равномерно распределено по тексту (очень часто или даже редко), то вряд ли оно может использоваться для эффективного содержательного поиска.

Данная мысль была «материализована» Солтоном в его знаменитой векторно-пространственной модели поиска, где именно для учета дискриминантной силы слов он ввел понятие инверсной частоты появления слова в отдельных документах массива (IDF).

В работе испанских исследователей [*] для этой же цели была предложена технология спектограмм слов, которые внешне напоминали штрих-коды товаров.

Вместе с тем не позволяли рассматривать вхождения слов в разных масштабах измерений, как это делается например в средствах вейвлет-анализа .

[*] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado, J. L. Oliver.

Level statistics of words: Finding keywords in literary texts and symbolic sequences //
PHYSICAL REVIEW E 79, 035102, 2009. –P. 035102-1-035102-4

Нами реализованы инструментальные средства позволяющие визуализировать плотность встречаемости слова в тексте в зависимости от ширины окна наблюдения. Через веб-интерфейс вводится текст и слово для анализа (<http://edu.infostream.ua/down/jag1.html>).

Введите слово и текст, нажмите "Start", получите визуализацию плотности встречаемости слова в зависимости от ширины окна наблюдения

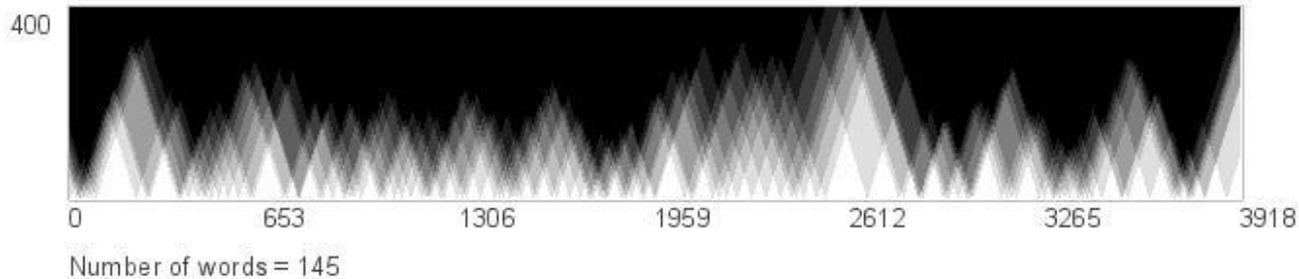
Если Вы хотите ввести полное слово, завершите его символом], например, поэмы]
Без этого знака по умолчанию запрос воспринимается как начало слова, можете ввести, например, поэм

В результирующей таблице ось OX - номер слова в тексте, OY - ширина окна наблюдения.
(Пока есть ограничение - обрабатывается до 5000 слов текста)

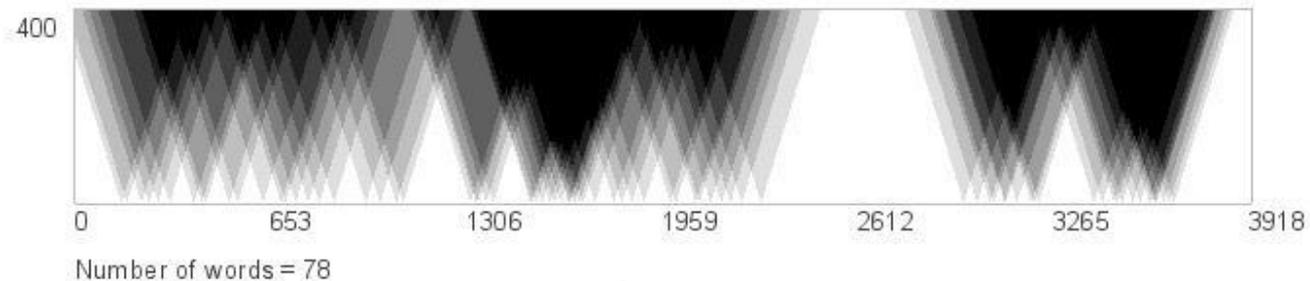
Слово: [Clear](#)

Текст:

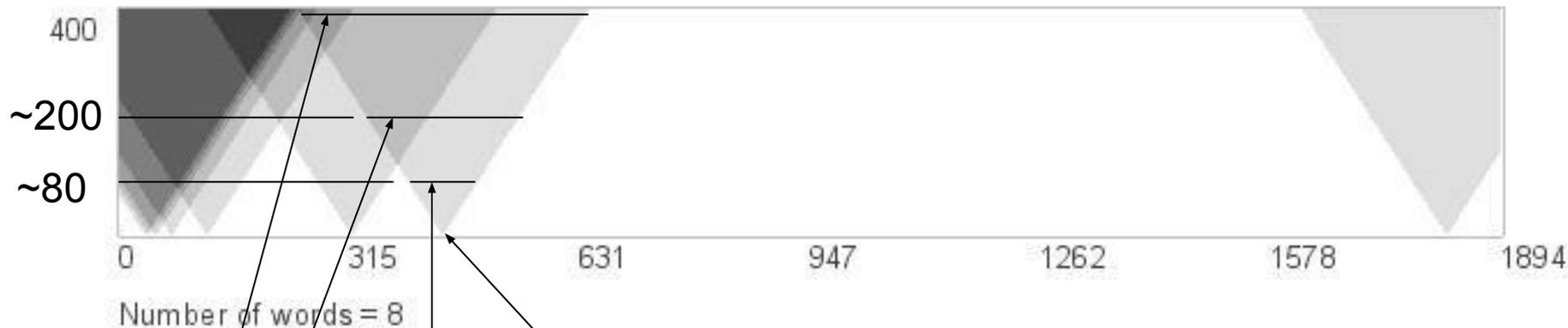
В результирующей спектограмме по горизонтали откладываются номера вхождения слова в тексте, а по вертикали - ширина окон наблюдения (начиная со значения 1 в самом низу, вхождения слова в данном случае выделяется светло-серым цветом). Если в соответствующее окно наблюдения попадает несколько целевых слов, то оно закрашивается более интенсивным оттенком темного. Всего предусмотрено 16 оттенков.



*Спектограмма вхождения слова «и» в рассказе
Стругацких
«Ночь на Марсе»*



*Спектограмма вхождения слова «сказал» в рассказе
Стругацких
«Ночь на Марсе»*

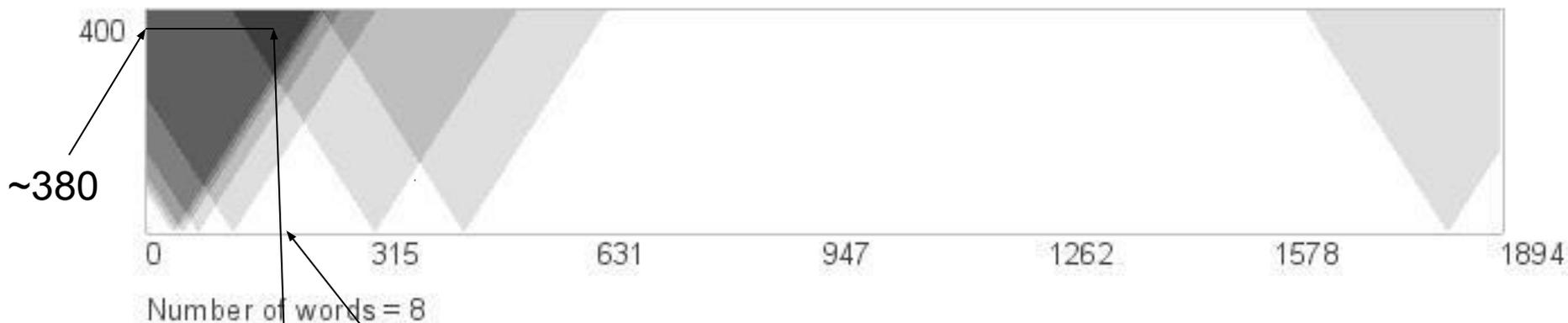


В этой позиции найдено слово
Окно наблюдения примерно 80 слов –
в нем пока слово только одно

Окно наблюдения примерно 200 слов –
в нем найдено 2 слова

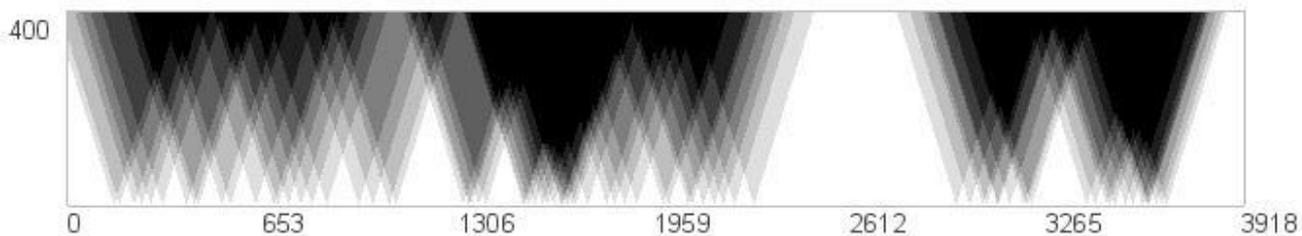
Окно наблюдения примерно 400 слов –
в нем найдено 4 слова – это видно по расцветке
наиболее темного участка

Читать в этой последовательности



В этой позиции нет вхождения искомого слова

Зато при окне наблюдения примерно в 380
(190 слов до данного слова и 190 после) – целых 5 слов –
это место самого плотного вхождение слова на диаграмме.



Number of words = 78

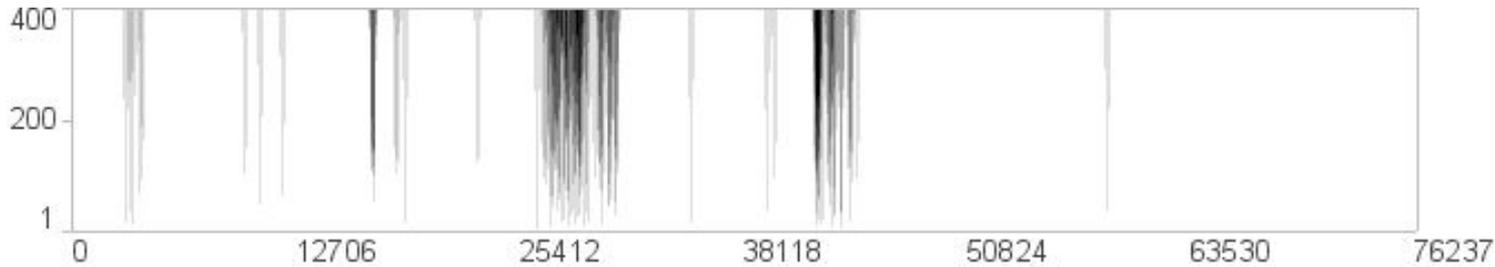
*Спектограмма вхождения слова «сказал» в рассказе
Стругацких
«Ночь на Марсе»*



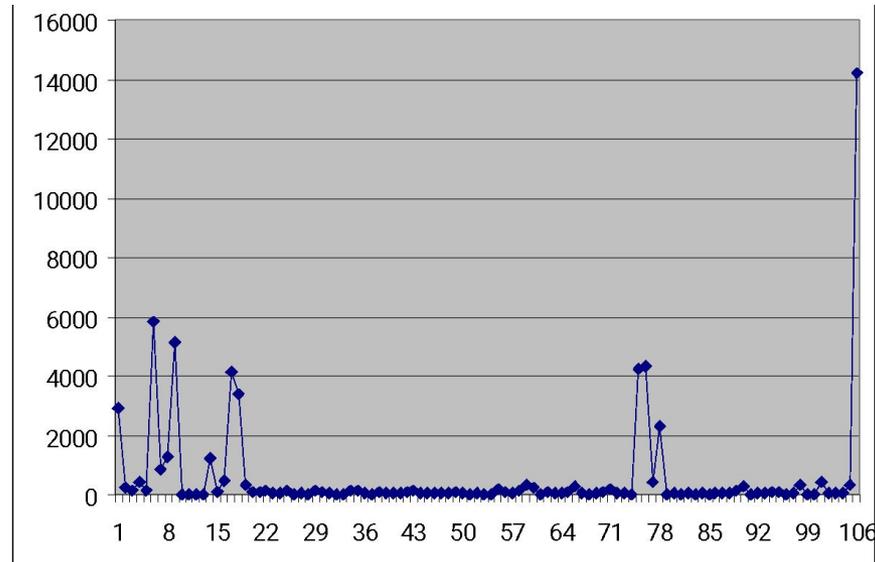
Number of words = 10

*Спектограмма вхождения слова «подумал» в рассказе
Стругацких
«Ночь на Марсе»*

Для исследований распределения слов представляет интерес числовая последовательность, составленная из расстояний между появлениями слов в тексте. Пример: Гоголь, Мертвые души, том первый.
Слово: Собакевич

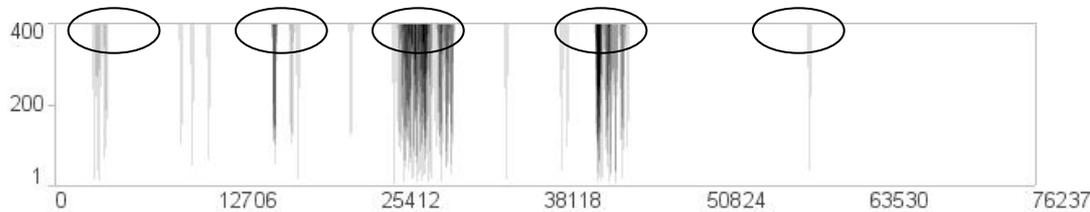


Number of words = 106



Такие последовательности позволяют ответить на вопросы, актуальные при автоматическом поиске и реферировании текстовых массивов/документов. Например, представляется, что автоматический реферат текста по аспекту, выраженному словом будет тем лучше, чем:

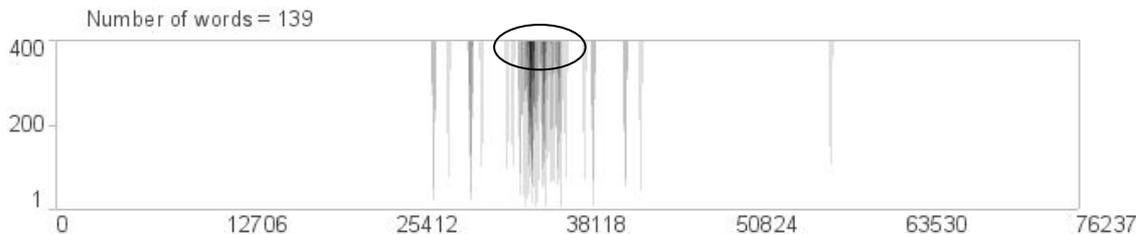
- более явно выражено скопление этих слов в текстах;
- таких «явно выраженных» скоплений больше.



Собакевич



Ноздрев



Плюшкин

Number of words = 46

В естественных науках как величина меры «изрезанности» числовых последовательностей используется показатель Херста, который вычисляется на основании R/S-анализа.

Нам показалась естественной аналогия с приведенными выше свойствами. Параметр Херста был рассчитан для рассмотренных выше персонажей «Мертвых душ».

Собакевич – 0.71

Ноздрев – 0.57

Плюшкин – 0.44

СПАСИБО ЗА ВНИМАНИЕ!

Ландэ Дмитрий Владимирович,
dwl@visti.net

<http://www.visti.net>

<http://www.infostream.ua>

<http://www.uaport.net>

Киев-2009