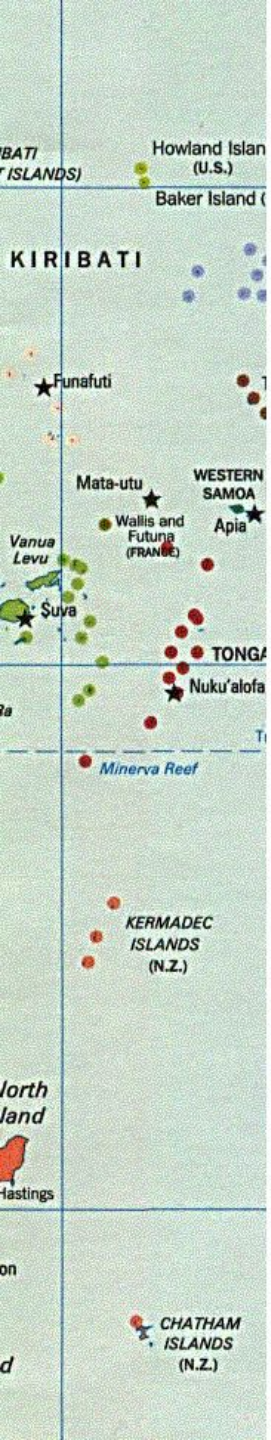
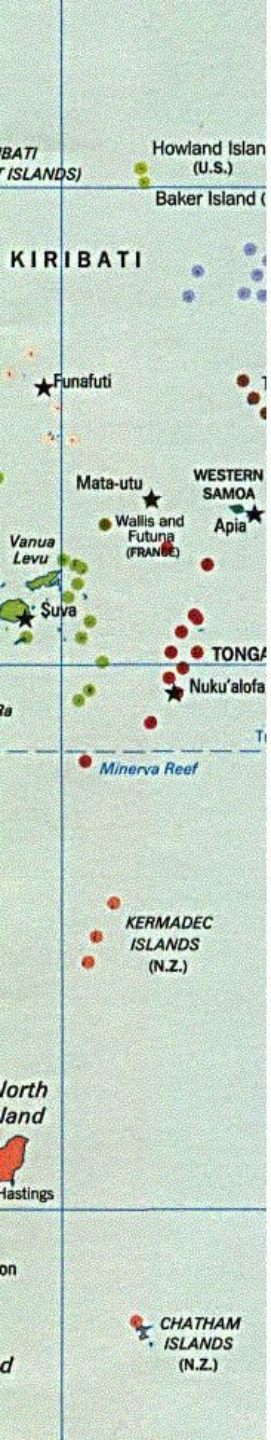


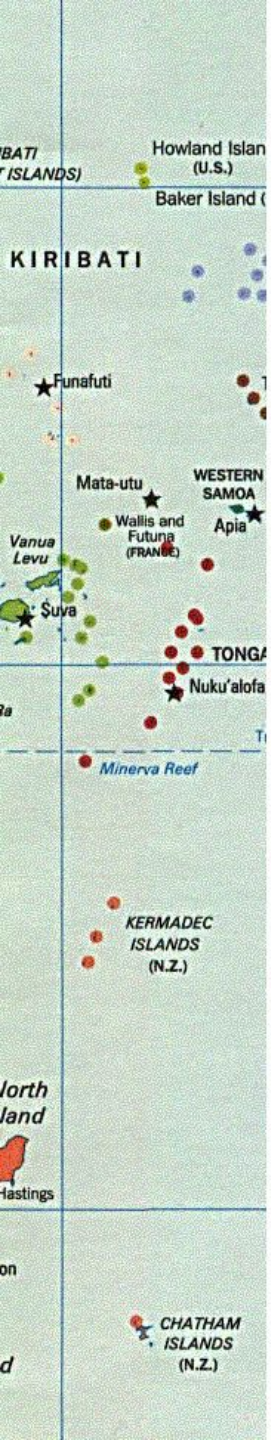
АВТОМАТИЧЕСКОЕ РАЗБИЕНИЕ ТЕКСТА НА ПРЕДЛОЖЕНИЯ ДЛЯ РУССКОГО ЯЗЫКА

Ольга Урюпина (uryupina@gmail.com)
Институт Языкознания РАН,
Ашманов и Партнеры
06.06.08



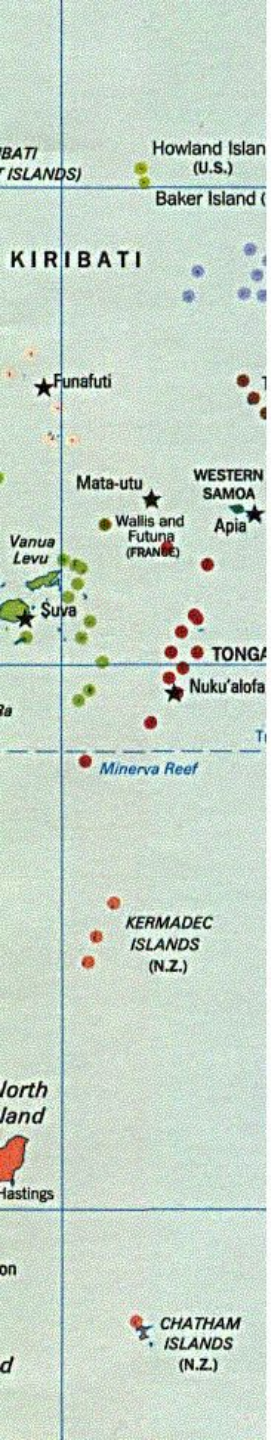
Вкратце

- Зачем и почему
- Примеры
- Признаки
- Эксперименты



Вкратце

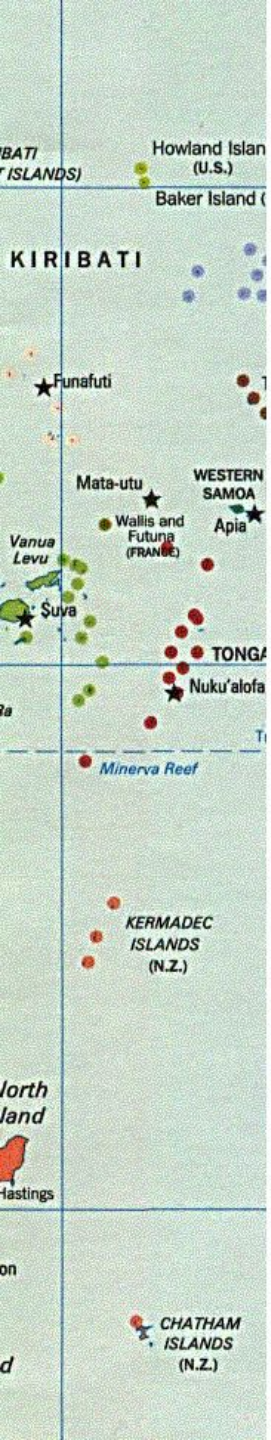
- Зачем и почему
- Примеры
- Признаки
- Эксперименты



Автоматическая обработка текста

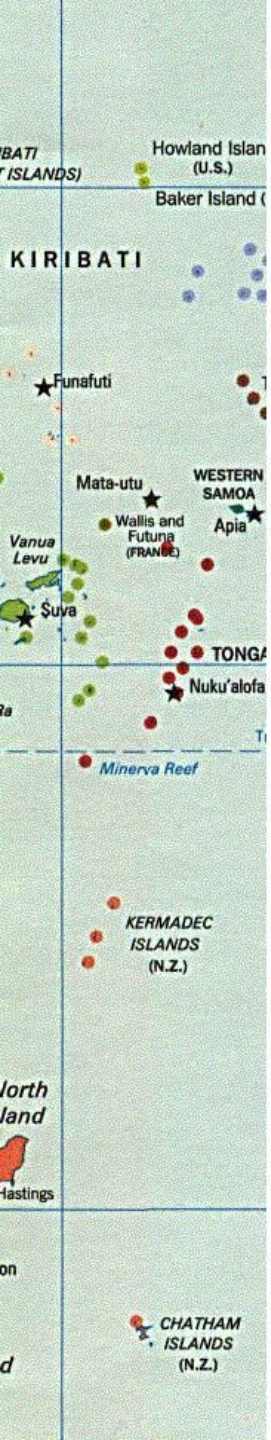
- синтаксический анализ (парсеры)
- системы автоматического реферирования
- машинный перевод
- экспертные системы
- …

Текст, разбитый на предложения



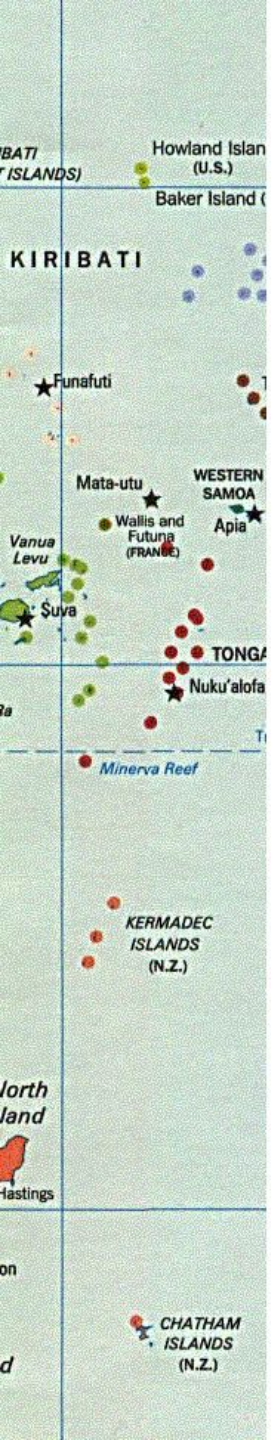
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



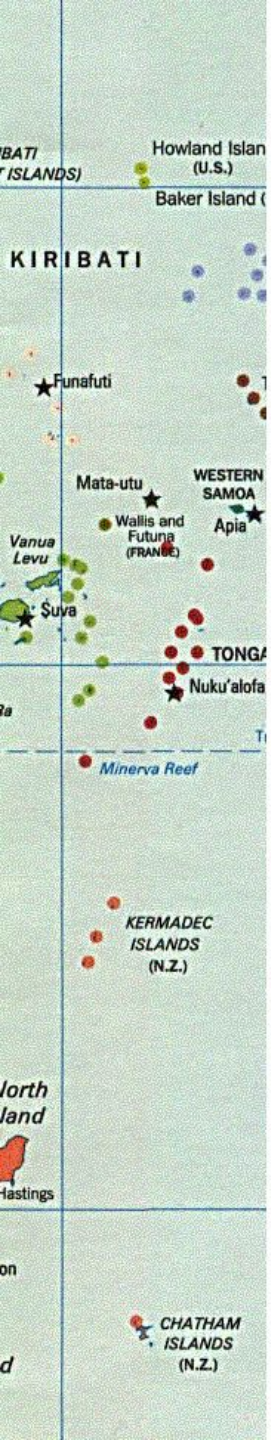
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



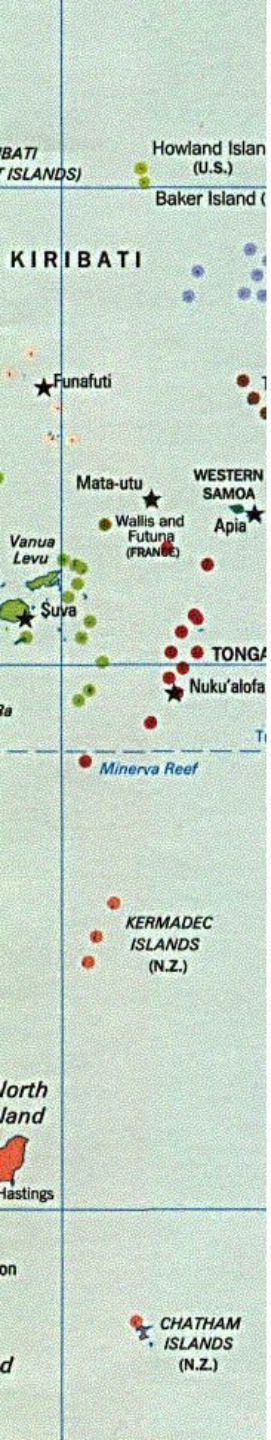
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



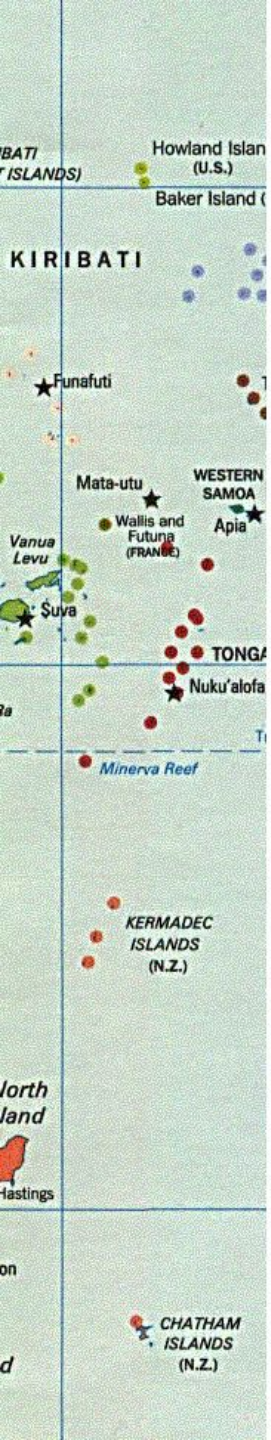
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



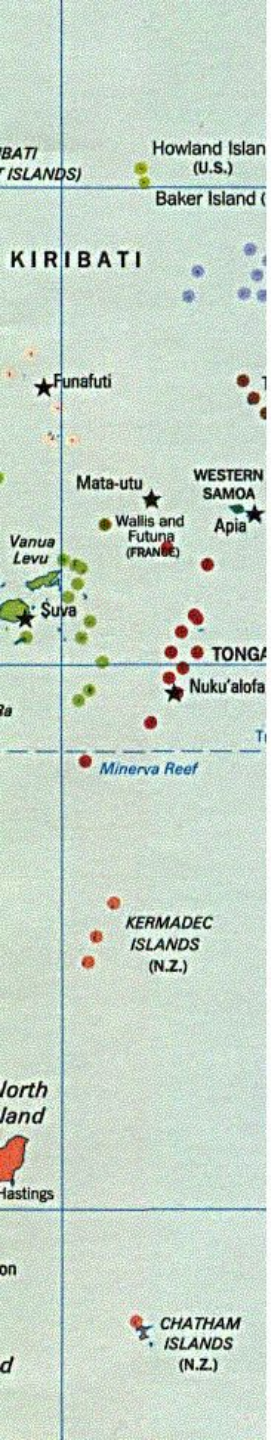
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



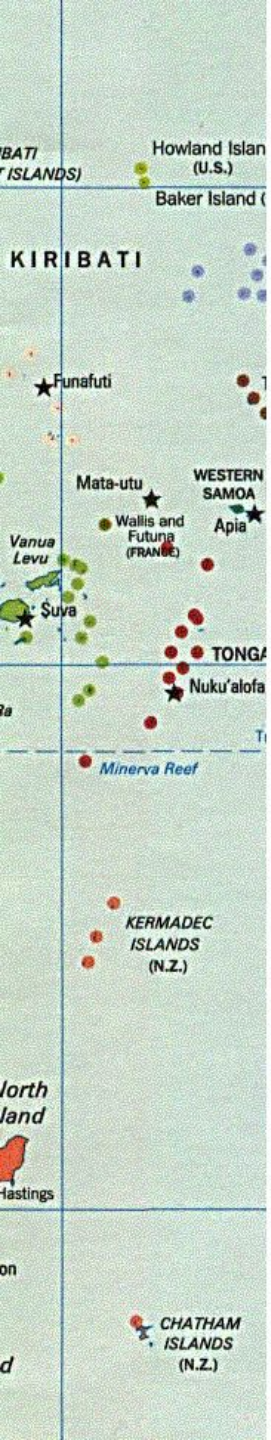
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



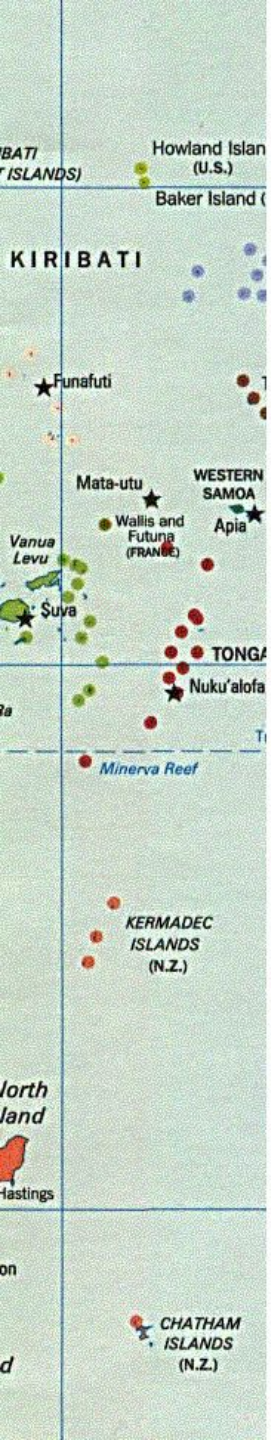
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



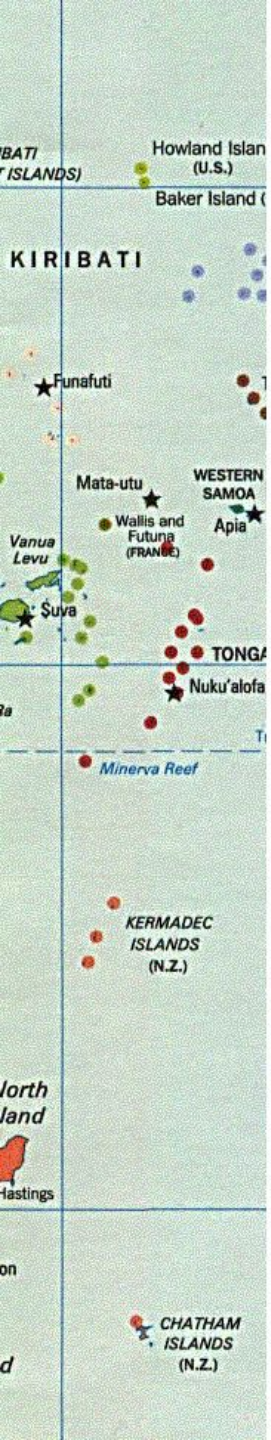
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



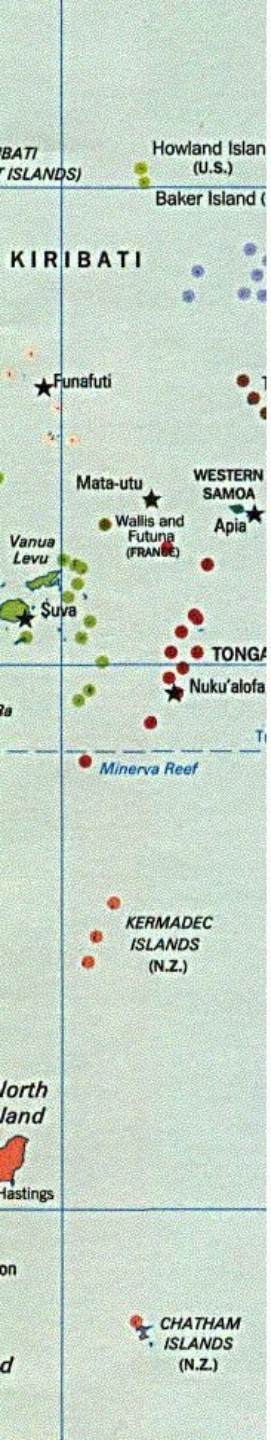
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



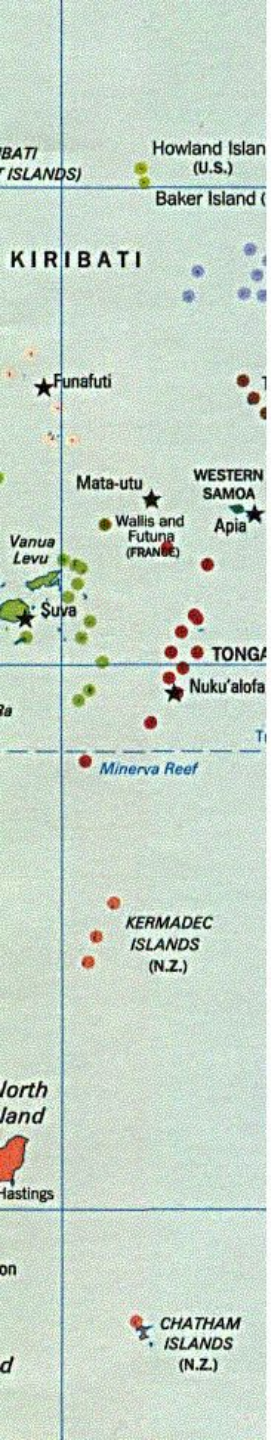
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



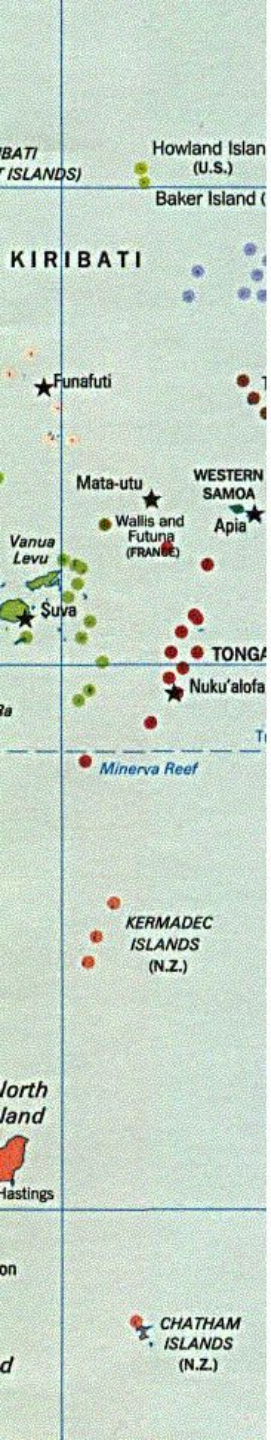
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



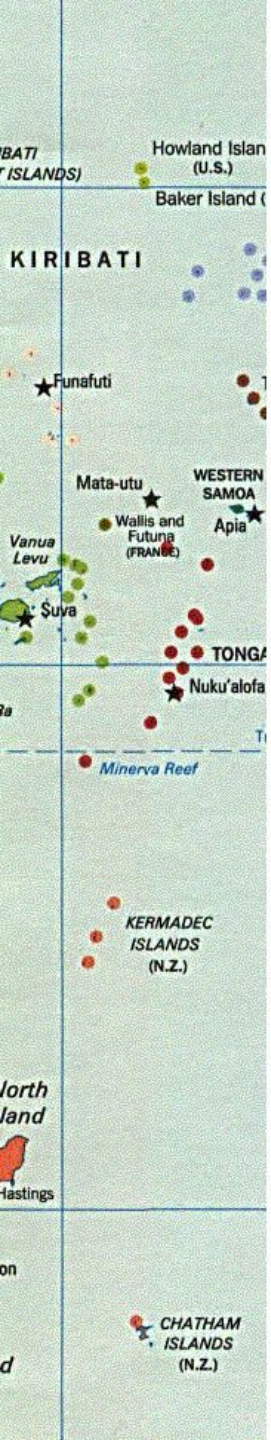
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



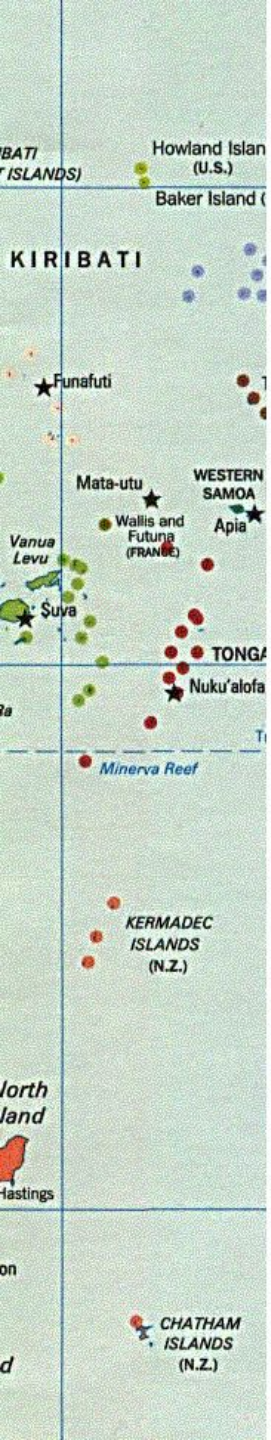
Наивная сегментация

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



Наивная сегментация

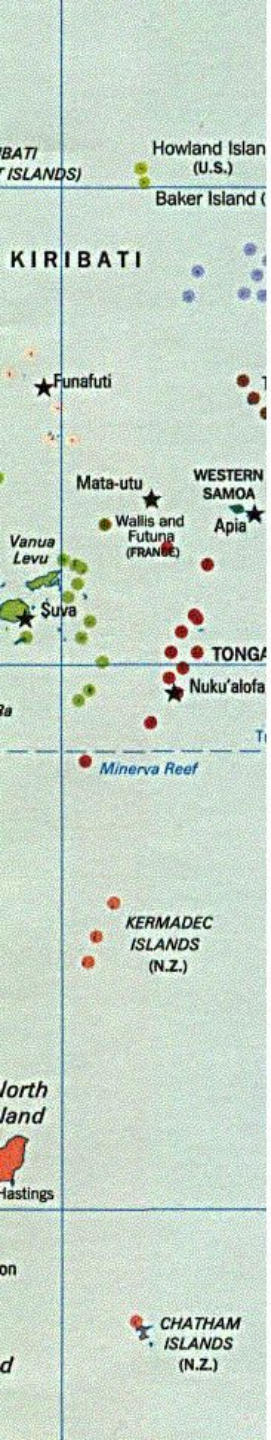
В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



Наивная сегментация

км), второй интервал -- 700.

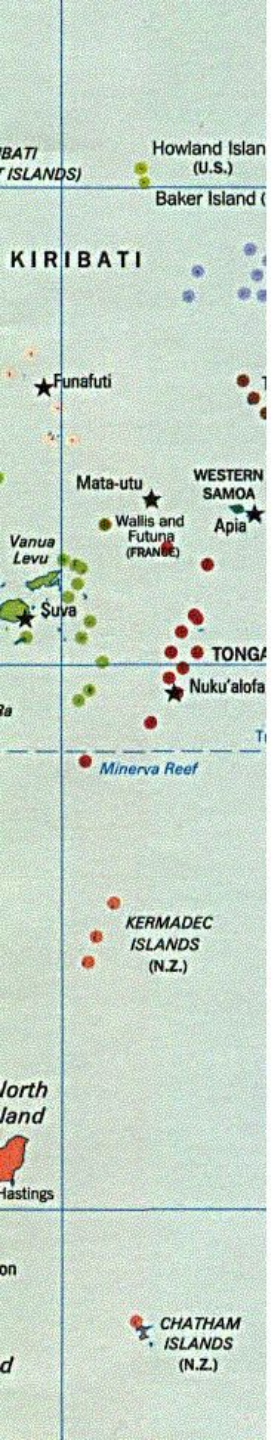
- синтаксический анализ (парсеры)
- системы автоматического реферирования
- машинный перевод
- экспертные системы
- ...



Практические решения

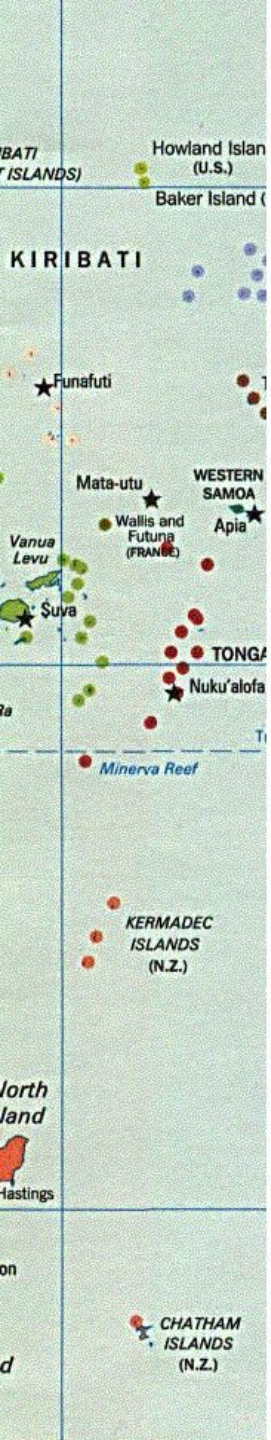
Эвристики:

- Предложение должно содержать буквы
- Предложение должно начинаться с заглавной буквы
- Сокращения (из списка) требуют «особого внимания»
- …



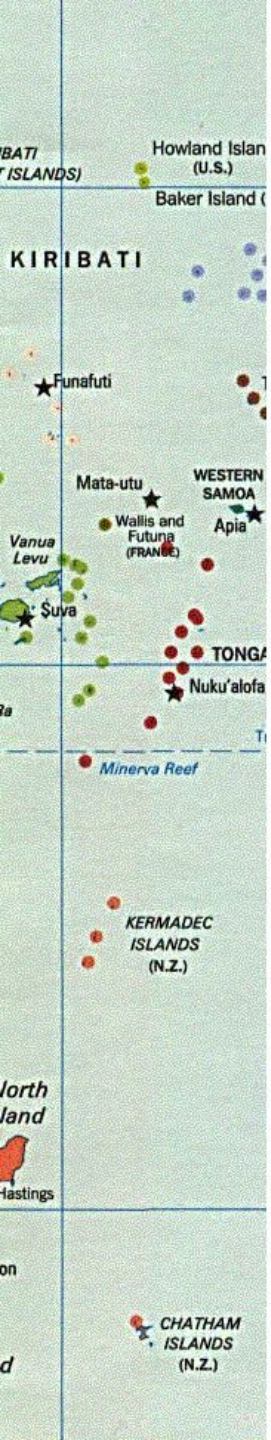
Проблемы

- Сложно адаптировать к новым данным
- Сложно адаптировать к новым задачам
- Сложно оценить роль отдельных факторов



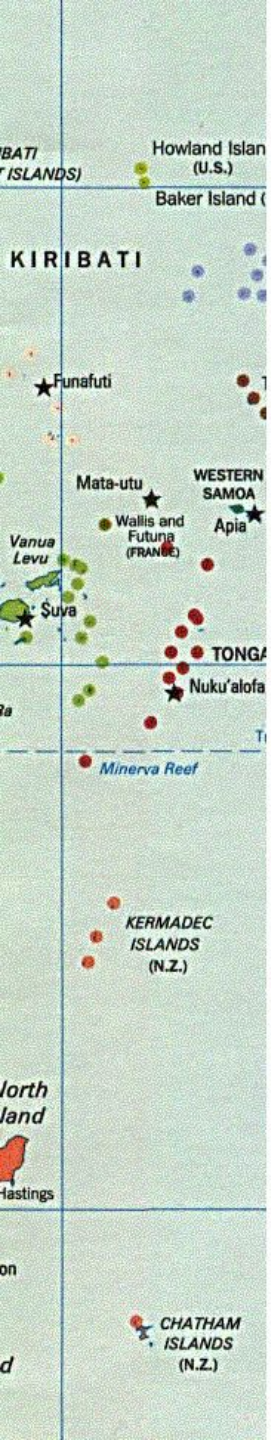
Вкратце

- Зачем и почему
- Примеры
- Признаки
- Эксперименты



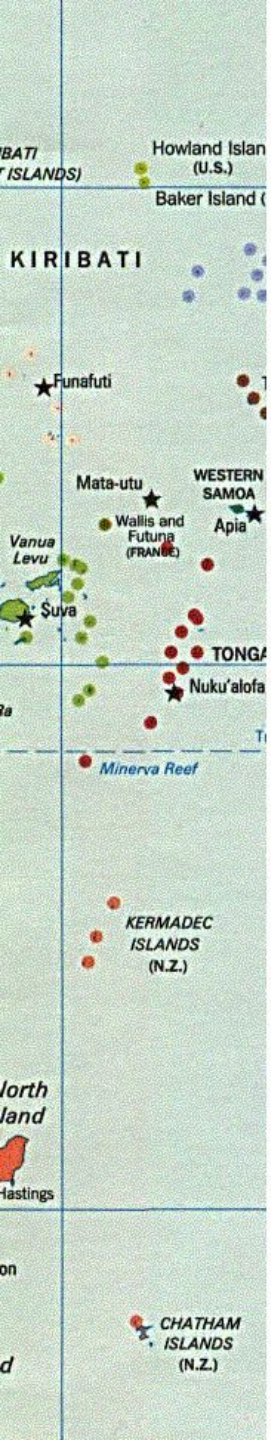
Точка

- URL: www.dialog-21.ru
- даты, время: 06.06.08
- сокращения: тыс. руб.
- сокращения в конце предложения
- опечатки: Михаил. Бычков
- многоточия: эээ...
  100...200
- форматирование: Введение.....1
  Данные.....5



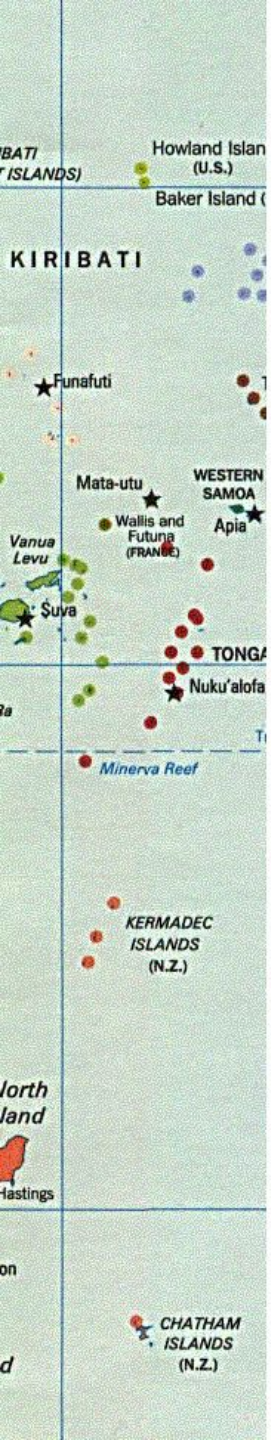
Вопросительный и восклицательный знаки

- комментарии: (правда?)
  - о ужас! –
- комбинации знаков: да ну?!
  - xxx: ??????
- URL: http://maps.google.com/maps?f=q&hl=de&geocode=&q=bekasovo&sll=37.0625,-95.677068&sspn=49.310476,76.640625&ie=UTF8&z=15&iwloc=addr
- кодировка: ?Локомотив?



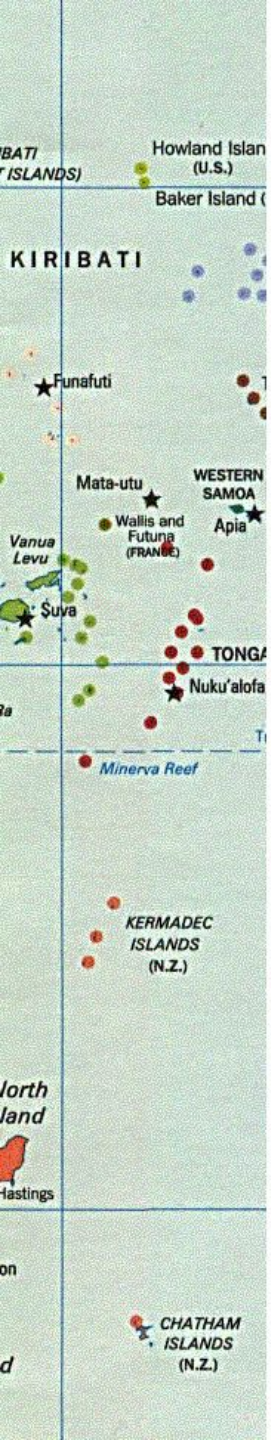
Скобки и кавычки

- .) и). не по правилам
- «», „“, “”, ‘’, ‚‘



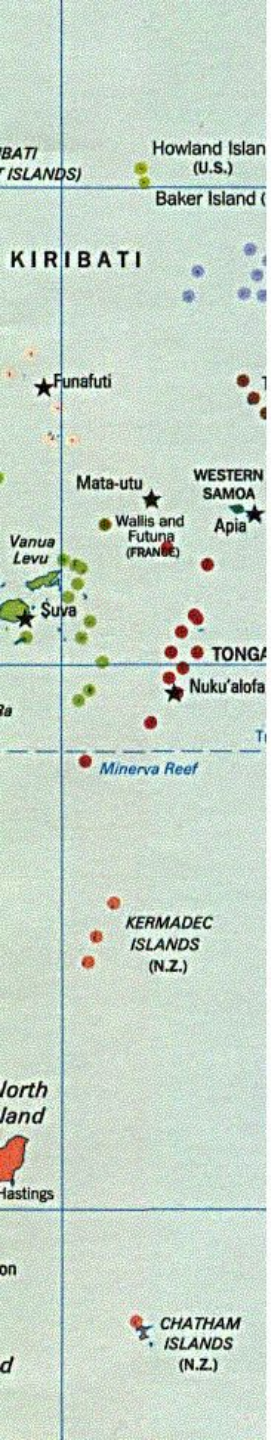
Вкратце

- Зачем и почему
- Примеры
- Признаки
- Эксперименты

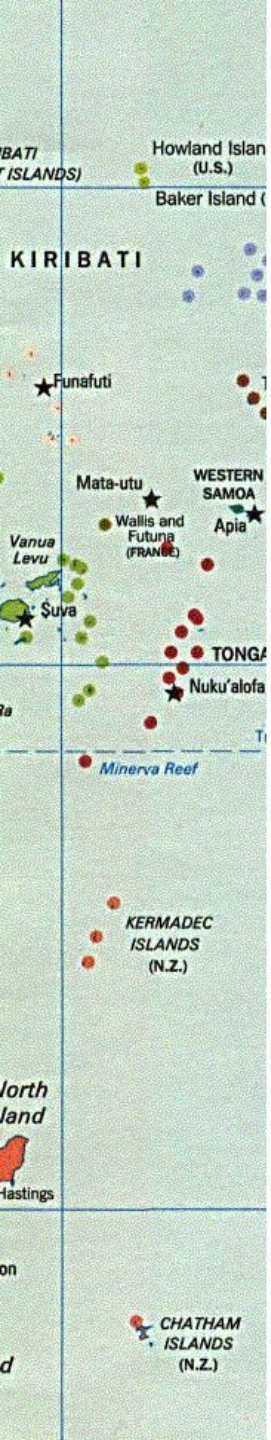


Контексты

- знак препинания
- слово слева
- слово справа
- «настоящее» слово справа

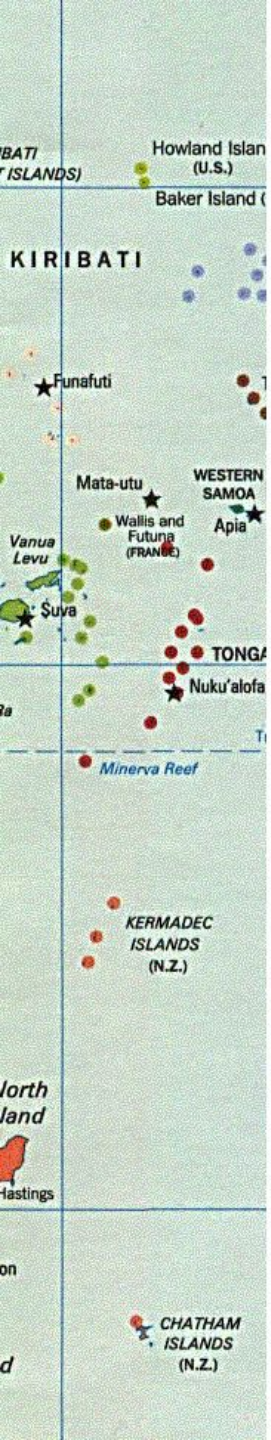


В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).



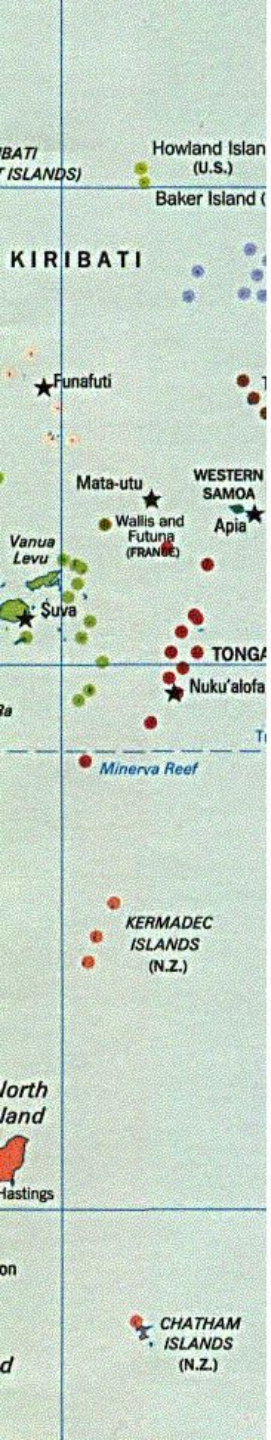
Контексты

- знак препинания .
- слово слева 700
- слово справа .
- «настоящее» слово справа 1050



Признаки

- сокращения
- «тип» слова
- начало и конец абзаца
- расстояния до потенциальных границ

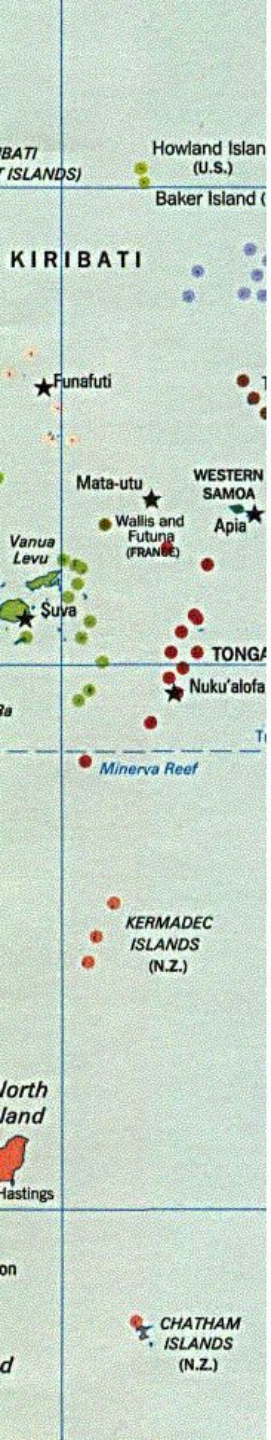


Сокращения

Извлечены автоматически из НКРЯ:

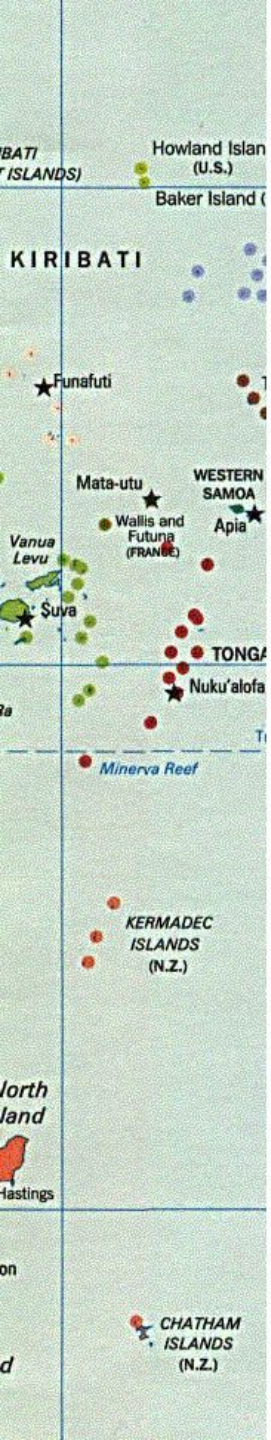
слово . слово_со_строчной

(дополнительно: по разметке)



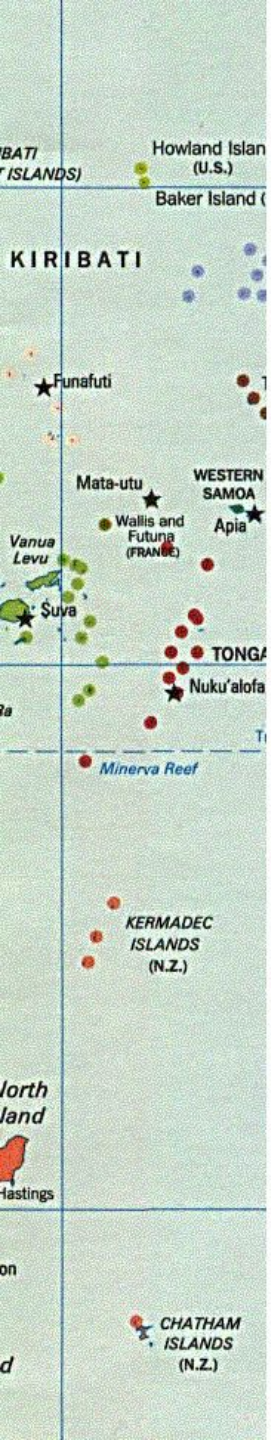
Вектор признаков

- знак препинания .
- слово слева 700
- слово справа .
- «настоящее» слово справа 1050



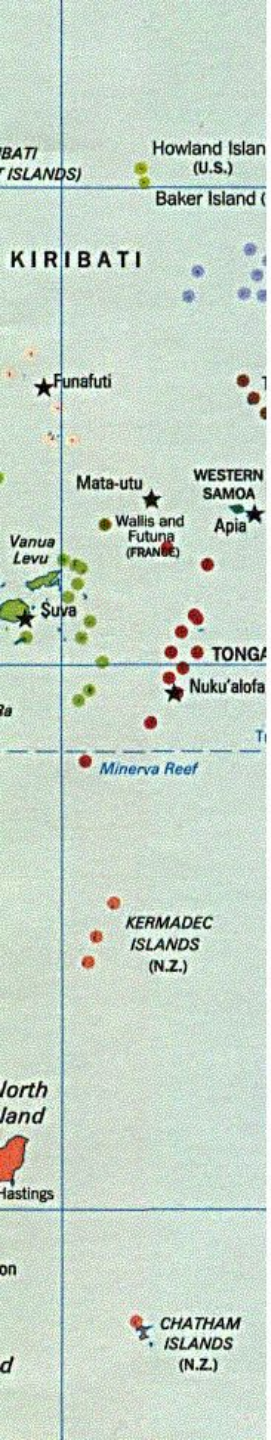
Вектор признаков

- знак препинания .
- слово слева 700
- слово справа .
- «настоящее» слово справа 1050
- расстояние1 6
- расстояние2 1
- сокращение справа нет
- сокращение слева нет
- тип слова слева цифры
- тип слова справа пунктуация
- тип «настоящего» слова справа цифры
- начало абзаца нет
- конец абзаца нет



Вкратце

- Зачем и почему
- Примеры
- Признаки
- Эксперименты

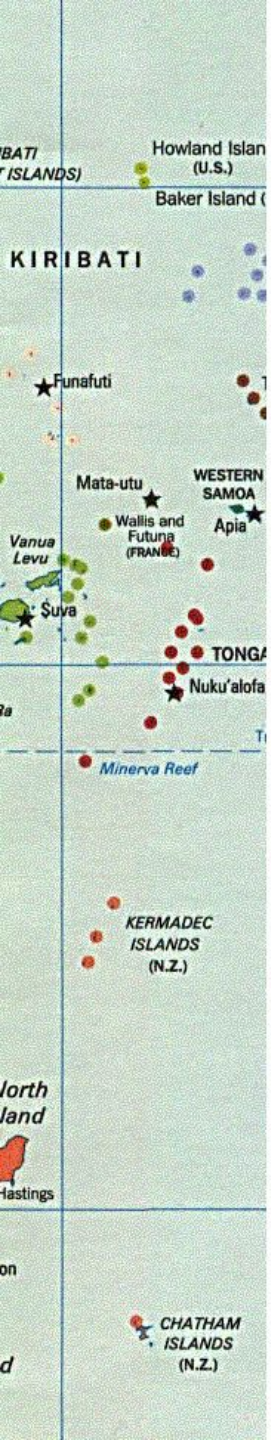


Данные

НКРЯ, 33 документа:

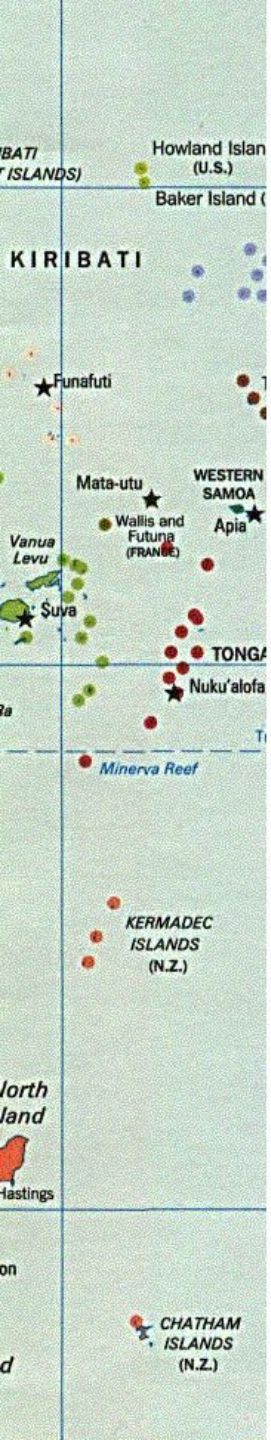
- политика, культура
- ремонт локомотивов

Ручная разметка



Данные - статистика

предложений	1639
предложений с .?!	1414
контекстов	5230(=4230+1000)
контекстов с .?!	2048



Контрольные эвристики

termpunct:

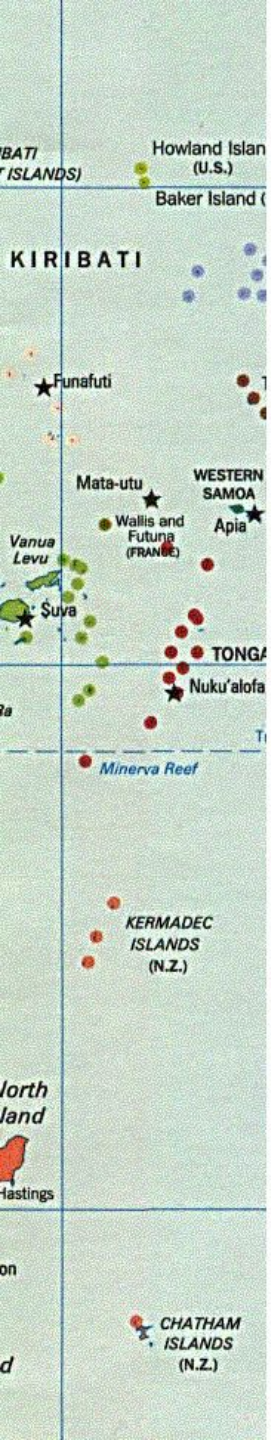
Предложение должно заканчиваться «.», «?», или «!».

termpunct_cap:

+Предложение должно начинаться с заглавной буквы.

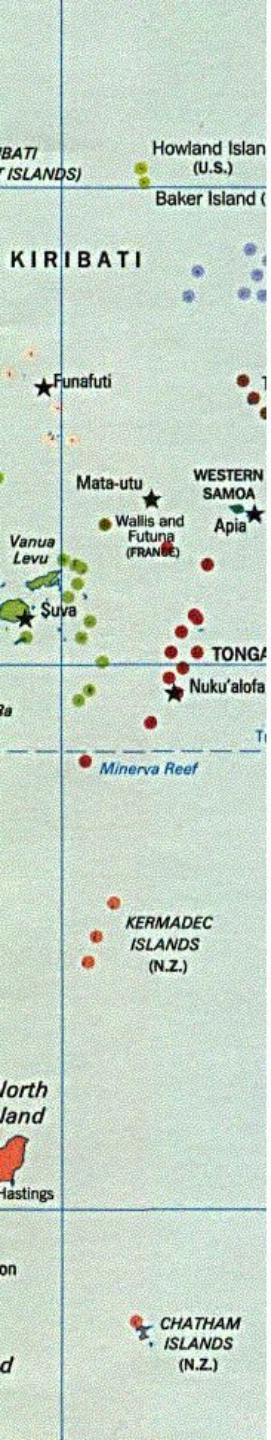
advanced:

+ Предложение не должно заканчиваться сокращением и «.».



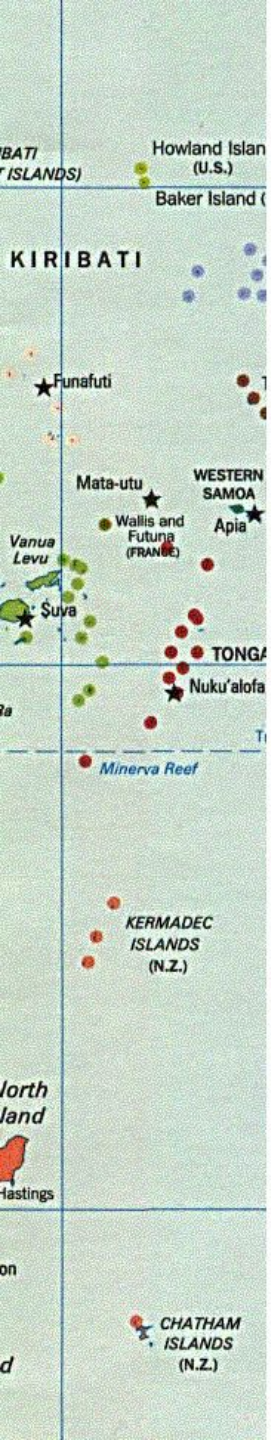
Результаты-1

	точность	полнота
termpunct	67.2	**100
termpunct_cap	90.7	**97.0
advanced	96.4	90.4
C4.5	97.8	**98.5
Ripper	98.5	**98.5
SVM	**99.6	**98.5



Результаты-2

	точность	полнота
termpunct	66.9	**98.9
termpunct_cap	89.6	**96.0
Advanced	95.0	89.6
C4.5	*98.5	**97.5
Ripper	**98.9	**96.0
SVM	**99.6	**97.5



Пример

<s> Был на церемонии момент , когда прозвучала пронзительно высокая и чистая нота . " Ника " за " Честь и Достоинство "-- вот так , всё с заглавной буквы -- вручалась Петру Ефимовичу Тодоровскому .</s>

<s> Петру Тодоровскому -- оператору и режиссёру , композитору и музыканту , солдату и просто замечательному человеку .</s>

<s> Он молодой , ошалевший от победной весны 45-го , смотрел на нас с экрана в хуциевском фильме " Был месяц май ".</s>

<s> Он вышел на сцену под гром аплодисментов и " Рио-риту ".</s>

<s> Для своих ровесников и друзей так и оставшийся в его - то годы Петей Тодоровским .</s>

<s> Он прошёл через зал ," по главной улице с оркестром ", держа в руках гитару .</s>

<s> Спасибо вам , дорогой Петр Ефимович !</s>

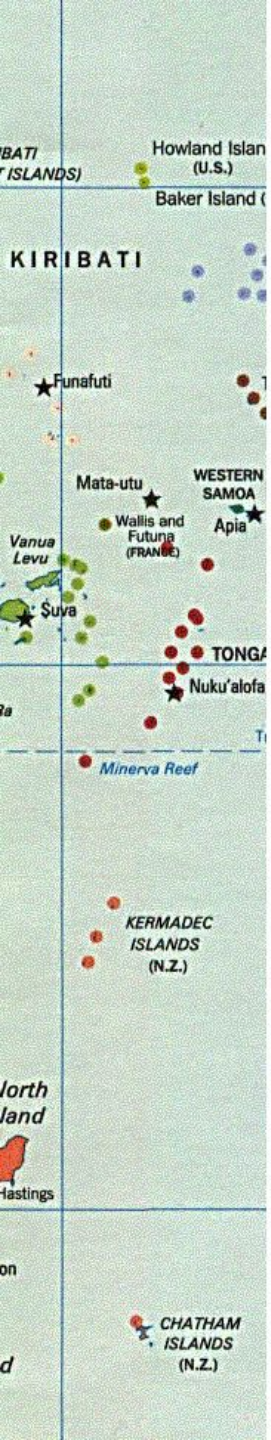
<s> За веру , верность и " Верность ", за всё ваше кино , за то , что вы сделали для нас , за вашу нескончаемую любовь , за то , что вы есть .</s>

<s> За то , что " и всё-таки , и всё-таки , и всё-таки мы победили "!</s>

<s> Той весной .</s>

<s> За то , что у нас есть эта весна .</s>

<s> И это ее семнадцатое мгновение .</s>



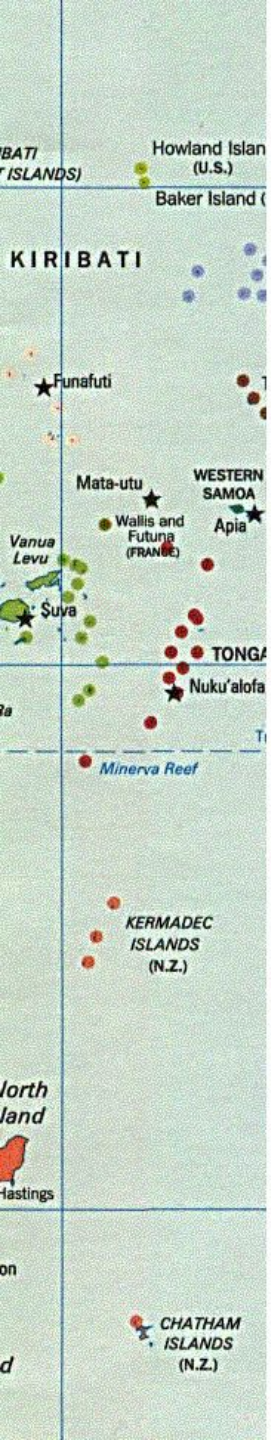
Заключение

статистический подход к задаче определения границ предложений в произвольном тексте на русском языке:

- легко адаптировать к новым данным и задачам
- высокая скорость
- высокая полнота и точность

В будущем:

- лингвистическая экспертиза (сокращения)
- новые данные (кавычки)


Howland Islan
(U.S.)
Baker Island
KIRIBATI
Funafuti
Mata-utu
WESTERN
SAMOA
Wallis and
Futuna
(FRANCE)
Apia
Vanua
Levu
Suva
TONGA
Nuku'alofa
Minerva Reef
KERMADEC
ISLANDS
(N.Z.)
Hastings
CHATHAM
ISLANDS
(N.Z.)

Спасибо!