Опыт создания системы автоматического **извлечения синонимов** для корпоративного информационного поиска

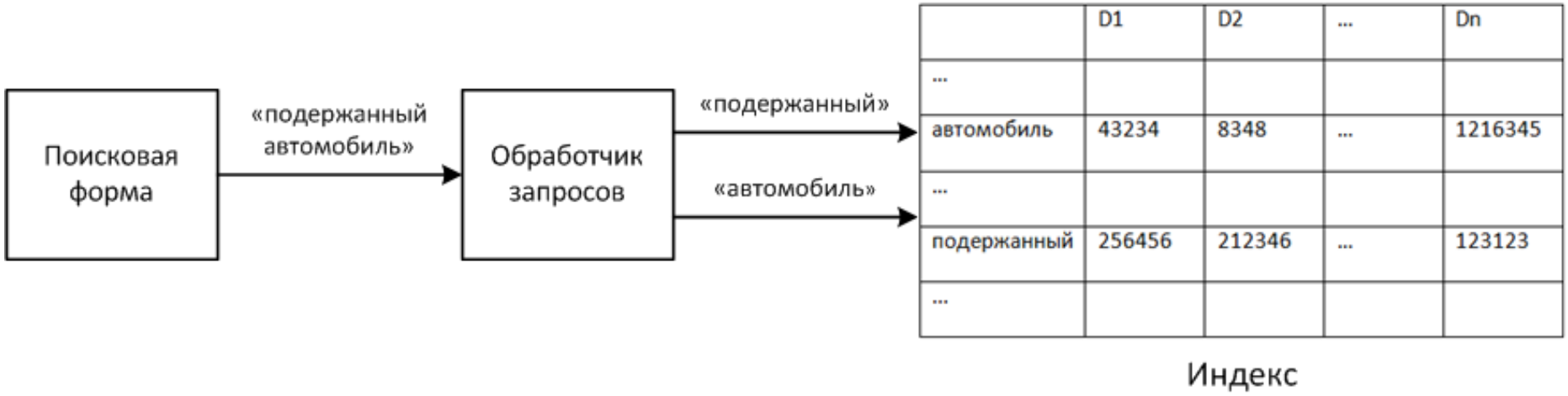
Донцов Дмитрий, аспирант СПИИРАН

Как работает поисковая машина

«Подержанный автомобиль»

↓

«подержанный + автомобиль»



	D1	D2	…	Dn
…				
автомобиль	43234	8348	…	1216345
…				
подержанный	256456	212346	…	123123
…				

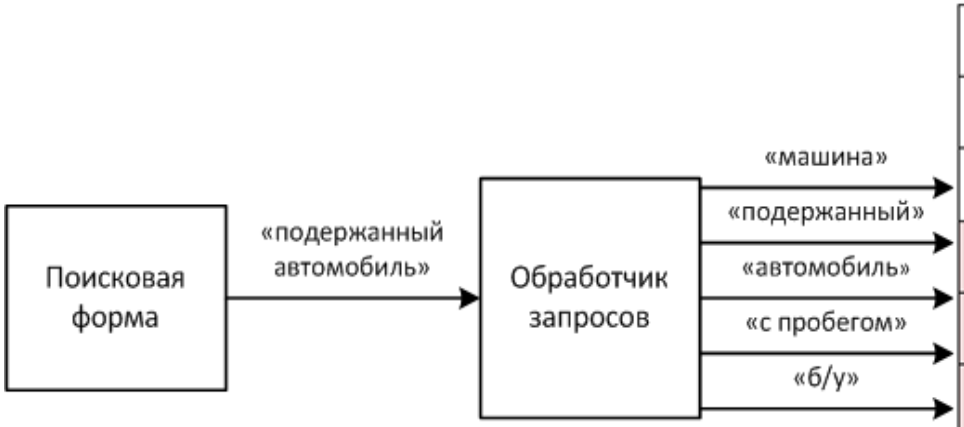
Индекс

Как могут помочь тезаурусные расширения

«Подержанный автомобиль»

↓

«подержанный + автомобиль + машина + пробег + б/у»



	D1	D2	…	Dn
…				
автомобиль	43234	8348	…	1216345
б/у	74457	367744	…	7345222
машина	846513	21488	…	328584
пробег	218877	3689813	…	6478519
подержанный	256456	212346	…	123123
…				

Индекс

Что ищут на сайтах производителей оборудования

- Документация
- Драйвера
- Информацию о сопровождении
- Цены
- Возможностях приобретения оборудования и т.д.

Постановка задачи

Используя содержимое веб-страниц сайта коммерческой организации составить словарь синонимов для её продуктов следующего вида:

…

[сущность$_1$, сущность$_2$]

[сущность$_1$, сущность$_2$]

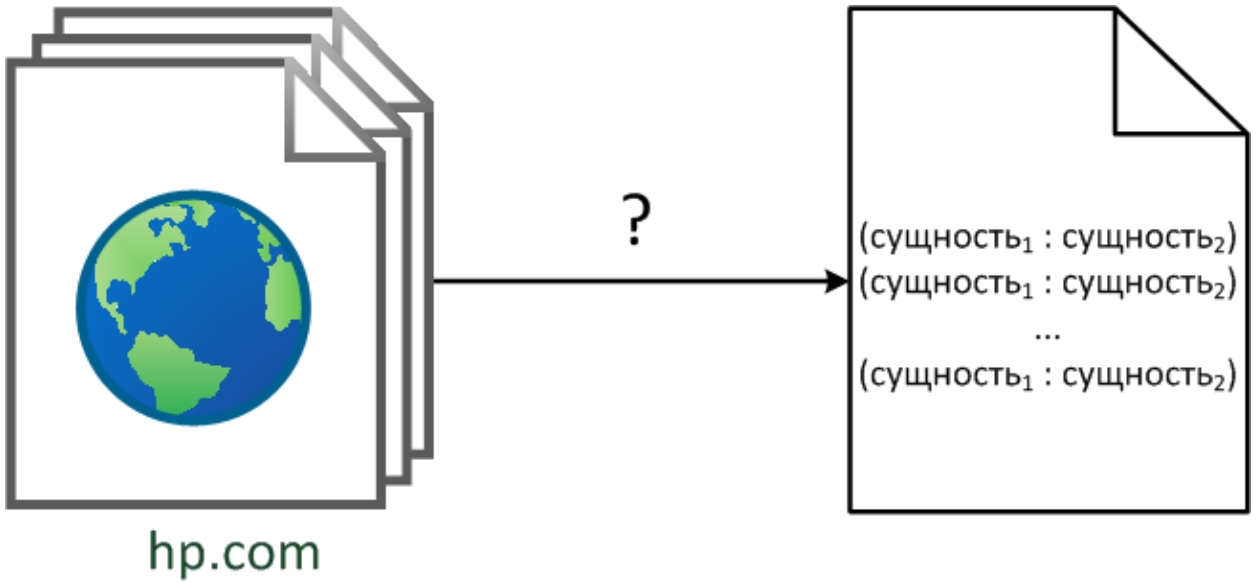
…

[HP TA5300, HP StorageWorks Tape Array 5300]

[HP-UX 11.31, HP-UX 11iv3]

Как хотим решать задачу

- Извлекать синонимы из содержимого страниц сайта hp.com



Как найти синонимы в тексте: паттерны синонимов

- Marti Hearst, *Automatic Acquisition of Hyponyms from Large Text Corpora*
- “also known as” , “also referred to as”, и т.д.
- Смартфон T-Mobile myTouch 4G Slide, **также известный как** HTC Doubleshot…
- U.S. Route 66 **also known as** the Will Rogers Highway…

Подготовка исходных данных

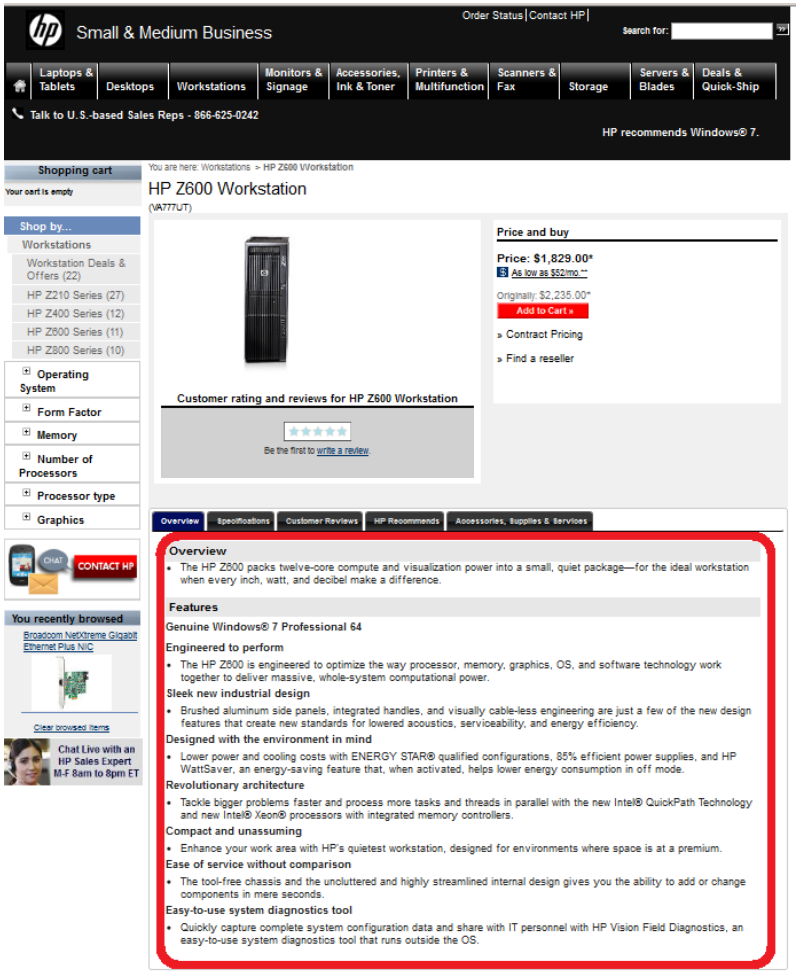
- Сохранение html-страниц
- Очистка html-страниц от шумов
- Разбивка на предложения
- Генерация словаря продуктов
- Разметка данных

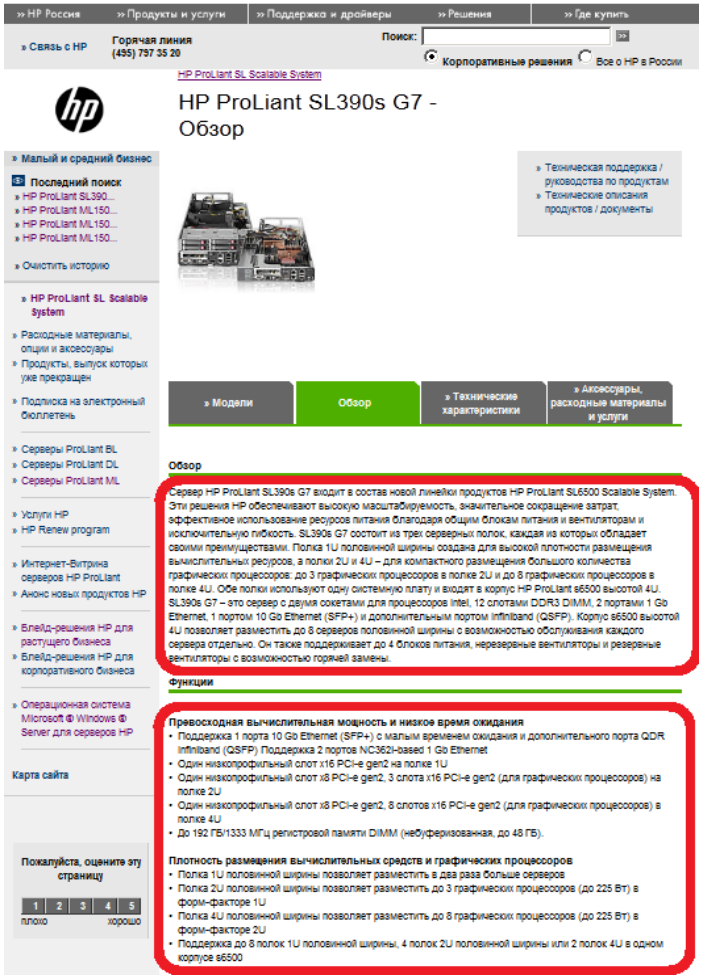
Кроулинг: статистика

15,500
> 600 МБ

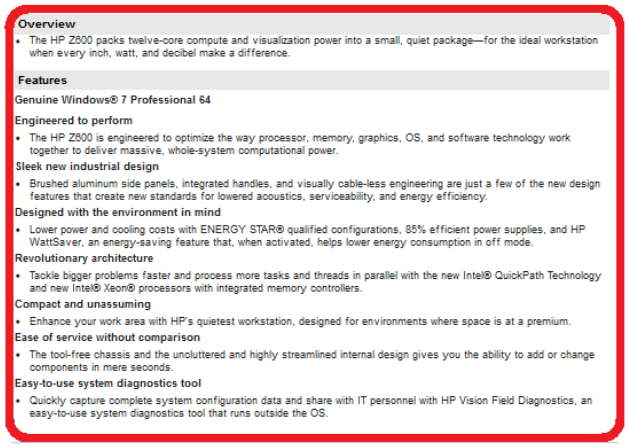
46,893
> 8,2 ГБ

Удаление шумов





Удаление шумов



Overview

- The HP Z800 packs twelve-core compute and visualization power into a small, quiet package—for the ideal workstation when every inch, watt, and decibel make a difference.

Features

Genuine Windows® 7 Professional 64

Engineered to perform

- The HP Z800 is engineered to optimize the way processor, memory, graphics, OS, and software technology work together to deliver massive, whole-system computational power.

Sleek new industrial design

- Brushed aluminum side panels, integrated handles, and visually cable-less engineering are just a few of the new design features that create new standards for lowered acoustics, serviceability, and energy efficiency.

Designed with the environment in mind

- Lower power and cooling costs with ENERGY STAR® qualified configurations, 85% efficient power supplies, and HP WattSaver, an energy-saving feature that, when activated, helps lower energy consumption in off mode.

Revolutionary architecture

- Tackle bigger problems faster and process more tasks and threads in parallel with the new Intel® QuickPath Technology and new Intel® Xeon® processors with integrated memory controllers.

Compact and unassuming

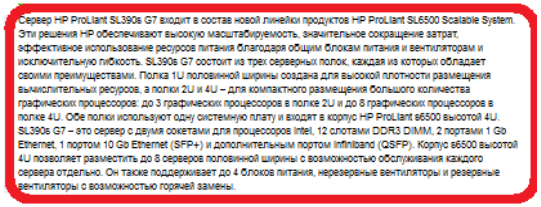
- Enhance your work area with HP's quietest workstation, designed for environments where space is at a premium.

Ease of service without comparison

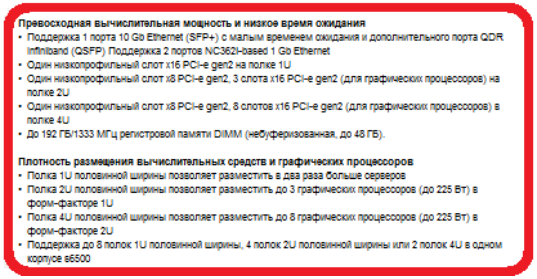
- The tool-free chassis and the uncluttered and highly streamlined internal design gives you the ability to add or change components in mere seconds.

Easy-to-use system diagnostics tool

- Quickly capture complete system configuration data and share with IT personnel with HP Vision Field Diagnostics, an easy-to-use system diagnostics tool that runs outside the OS.



Сервер HP ProLiant SL390s G7 входит в состав новой линейки продуктов HP ProLiant SL6500 Scalable System. Эти решения HP обеспечивают высокую масштабируемость, значительное сокращение затрат, эффективное использование ресурсов питания благодаря общим блокам питания и вентиляторам и исключительную гибкость. SL390s G7 состоит из трех серверных полок, каждая из которых обладает своими преимуществами. Полка 1U половинной ширины создана для высокой плотности размещения вычислительных ресурсов, а полки 2U и 4U – для компактного размещения большого количества графических процессоров: до 3 графических процессоров в полке 2U и до 8 графических процессоров в полке 4U. Обе полки используют одну системную плату и входят в корпус HP ProLiant s6500 высотой 4U. SL390s G7 – это сервер с двумя сокетами для процессоров Intel, 12 слотами DDR3 DIMM, 2 портами 1 Gb Ethernet, 1 портом 10 Gb Ethernet (SFP+) и дополнительным портом Infiniband (QSFP). Корпус s6500 высотой 4U позволяет разместить до 8 серверов половинной ширины с возможностью обслуживания каждого сервера отдельно. Он также поддерживает до 4 блоков питания, нерезервные вентиляторы и резервные вентиляторы с возможностью горячей замены.



Превосходная вычислительная мощность и низкое время ожидания

- Поддержка 1 порта 10 Gb Ethernet (SFP+) с малым временем ожидания и дополнительного порта QDR Infiniband (QSFP) Поддержка 2 портов NC362I-based 1 Gb Ethernet
- Один низкопрофильный слот x16 PCI-e gen2 на полке 1U
- Один низкопрофильный слот x8 PCI-e gen2, 3 слота x16 PCI-e gen2 (для графических процессоров) на полке 2U
- Один низкопрофильный слот x8 PCI-e gen2, 8 слотов x16 PCI-e gen2 (для графических процессоров) в полке 4U
- До 192 ГБ/1333 МГц регистровой памяти DIMM (небуферизованная, до 48 ГБ).

Плотность размещения вычислительных средств и графических процессоров

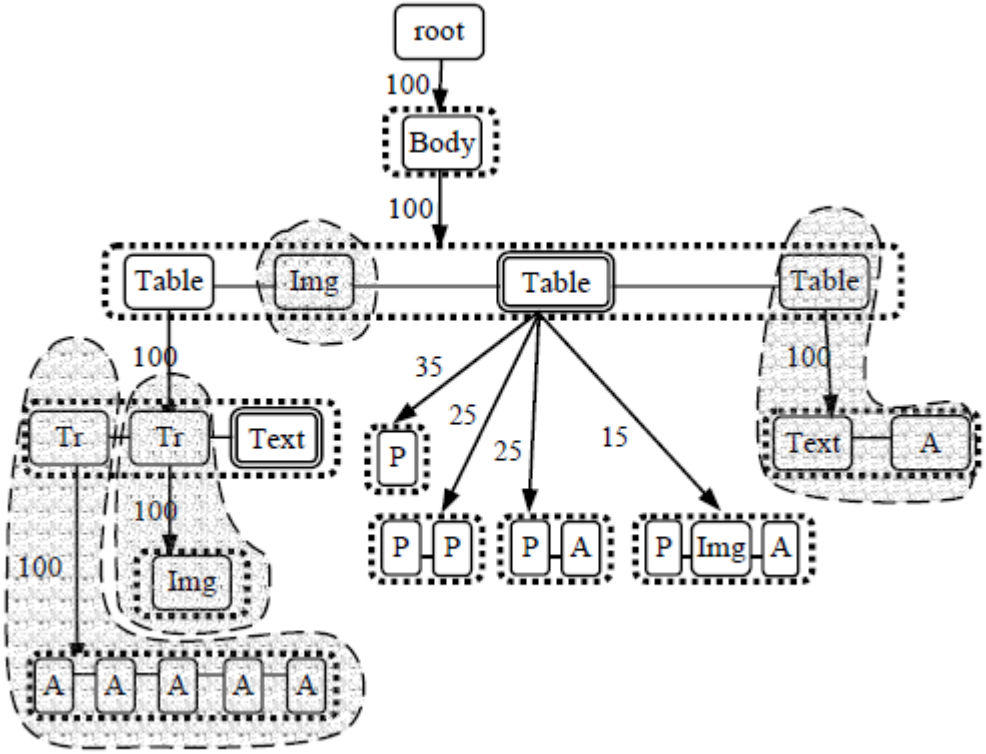
- Полка 1U половинной ширины позволяет разместить в два раза больше серверов
- Полка 2U половинной ширины позволяет разместить до 3 графических процессоров (до 225 Вт) в форм-факторе 1U
- Полка 4U половинной ширины позволяет разместить до 8 графических процессоров (до 225 Вт) в форм-факторе 2U
- Поддержка до 8 полок 1U половинной ширины, 4 полок 2U половинной ширины или 2 полок 4U в одном корпусе s6500

Подходы к очистке данных

- Site Style Tree
- Densitometric approach
- N-grams
- CSS-based (эвристика, использующая таблицы стилей)
- Сложности с оценкой (оценивается вклад в улучшение классификации/кластеризации)

Site Style Tree

- Представление DOM-модели в виде дерева



Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating noisy information in Web pages for data mining

Densitometric approach

В основном используются «поверхностные» признаки (shallow features):

- Средняя длина токена, предложения
- Общее количество слов
- Встречаемость токенов, связанных со временем/датой
- Встречаемость разделителей («|»)
- «Плотность» текста: $p(b) = \frac{\#\text{ токенов в блоке}}{\#\text{ строк в блоке}}$
- «Плотность» ссылок: $p(b) = \frac{\#\text{ токенов в тэге } <a>}{\#\text{ токенов в блоке}}$

Исходный словарь

- Нужно чем-то разметить корпус
- Pdf-каталоги продукции HP
- Извлечено около 15,000 сущностей

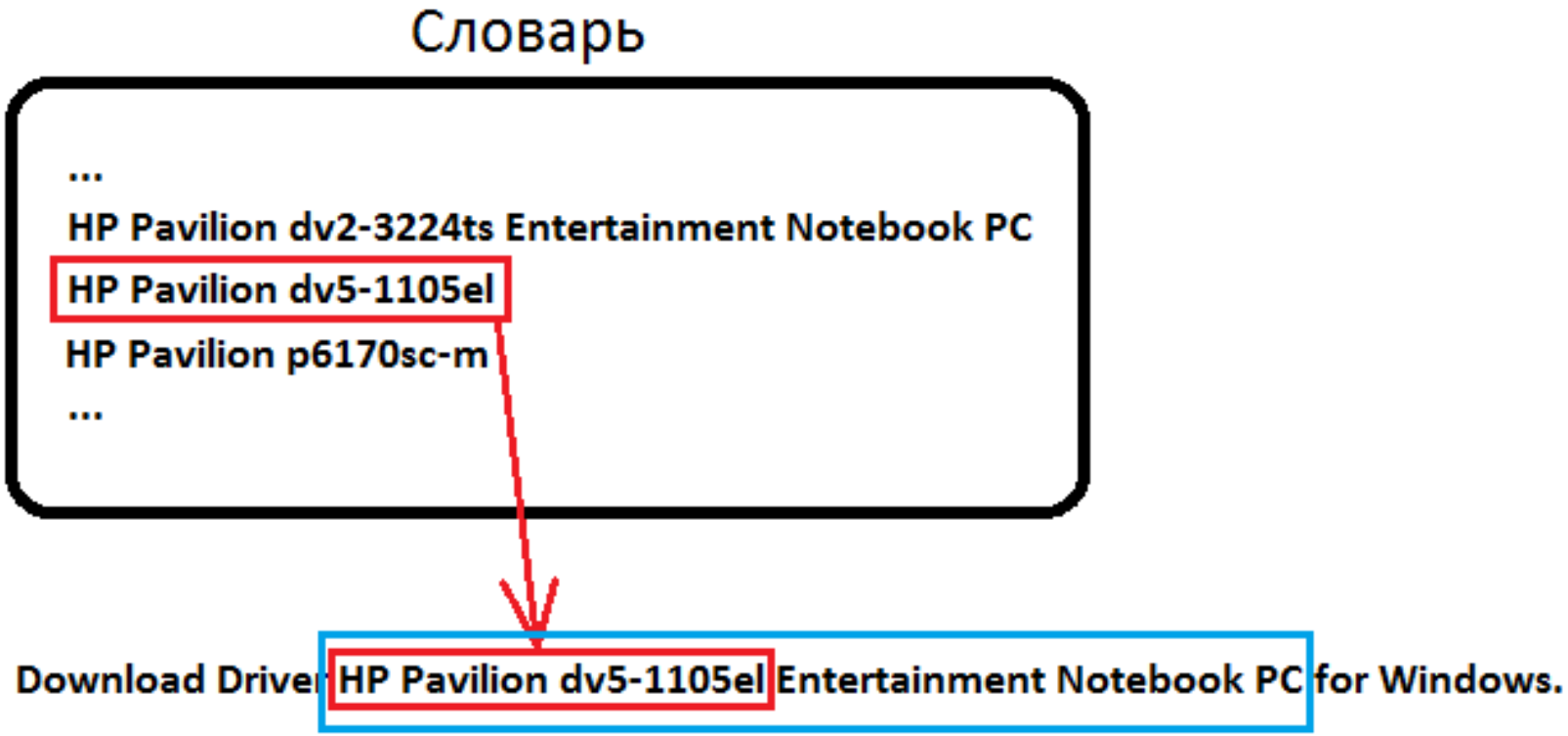
+ (120 тысяч сущностей предоставила HP)

Проблемы разметки

- Размечаются не все сущности
- Размечаются части сущностей

Проблемы разметки

- Размечаются не все строки
- Размечаются части строк



BIO-разметка

- B – начало сущности
- I – продолжение сущности
- O – не относится к сущности

Например, для сущностей разного типа используются разные метки:

Location: B-LOC, I-LOC
Organization: B-ORG, I-ORG

BIO-разметка: пример

The HP Media Center PC is available in store at Best Buy.

Токен	Метка
The	O
HP	B-HP
Media	I-HP
Center	I-HP
PC	I-HP
is	O
available	O
in	O
store	O
at	O
Best	O
Buy	O
.	O

Тренировка распознавателя: CRF

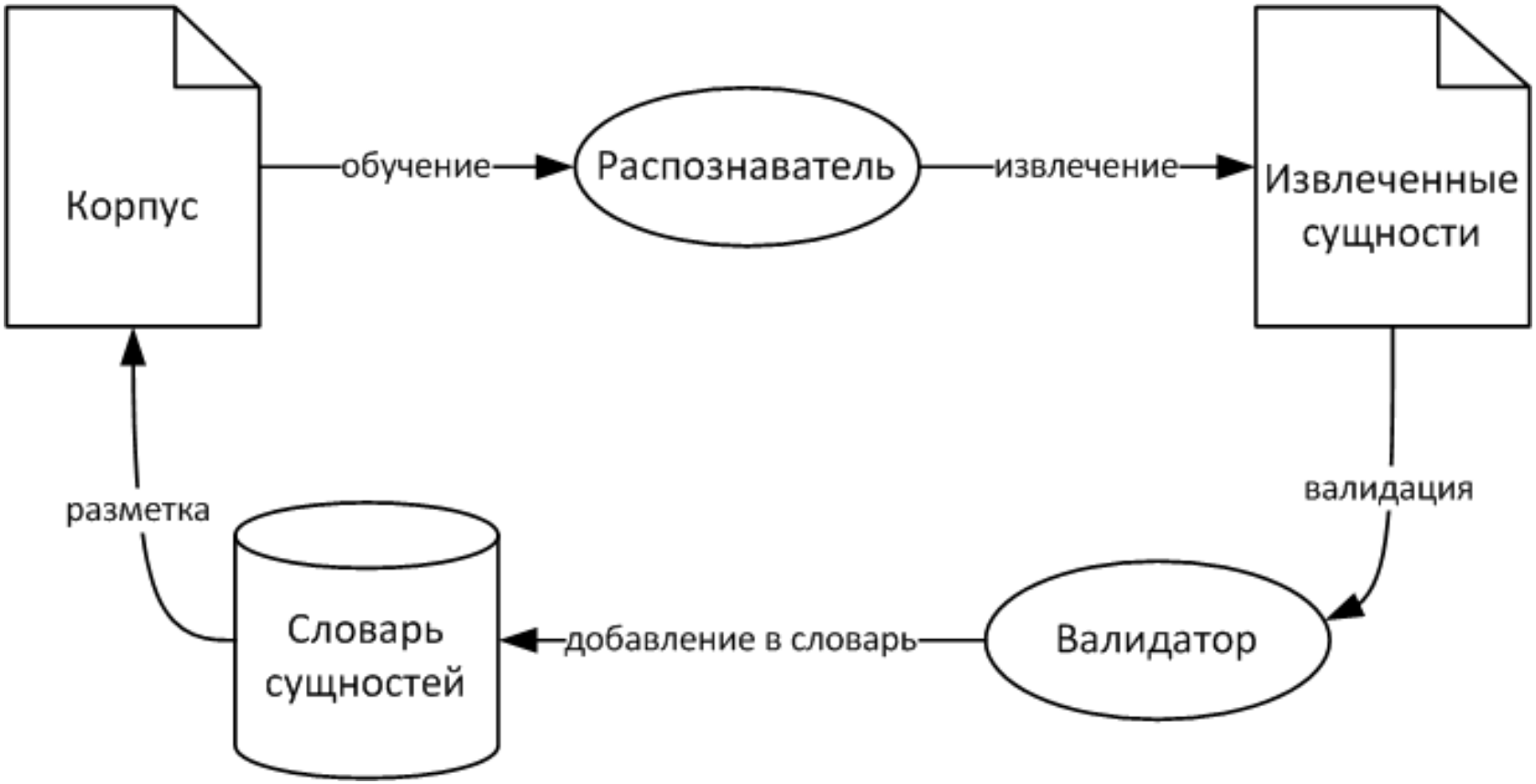
- Алгоритм Conditional Random Fields (CRF)
- Лучшие показатели для распознавания именованных сущностей (NER)
- Существующие обученные распознаватели не подходят из-за особенности сущностей
- Использовались фреймворки LingPipe и Mallet

HP ProLiant Essentials Performance Management Pack 10 Server License

Итеративный подход

- После каждого обучения и извлечения нового набора сущностей распознаватель обучается на новой, более полной разметке
- Качество распознавания улучшается
- Распознаватель правильно выделяет многие некорректно размеченные сущности (False Positives)
- Полученные сущности нужно проверить и при прохождении валидации поместить в словарь и на его основе сгенерировать новую разметку корпуса

Итеративный подход



Валидация извлеченных данных

- Каждая извлеченная сущность сравнивается со строкой из словаря

- Если в словаре находится похожая строка (расстояние между ними меньше допустимого), сущность добавляется в словарь

Кластеризация строк

Плоская кластеризация:

- Алгоритмы связей (одной связи, средней связи $O(n^2)$
- Алгоритм минимального остовного дерева – $O(n^2)$
- При квадратичной сложности и $n > 1*10^5$ для подсчета матрицы смежностей потребуется $(1*10^5)^2 = 1*10^{10}$ вычислений
- **Алгоритм k-средних O(n*m*k)**

n – количество объектов

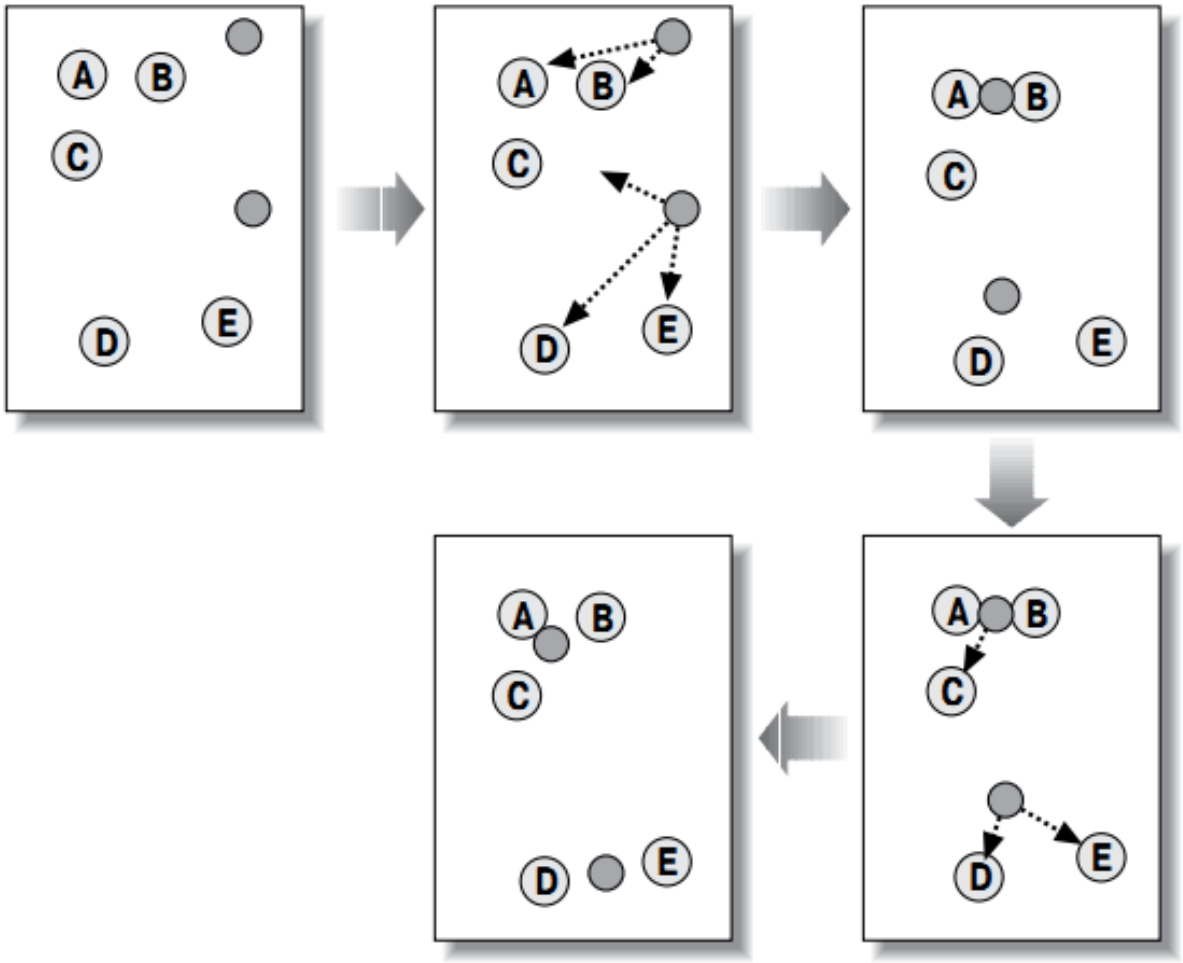
m – количество итераций,

k – предполагаемое количество кластеров

Алгоритм k-средних

- Каждая строка представляется в виде n-мерного вектора (набор n признаков)
- Выбирается k центроидов (центры кластеров), случайно разбросанных в n-мерном пространстве
- Каждому элементу назначается ближайший центроид
- После того, как назначение выполнено, каждый центроид перемещается в точку, рассчитываемую как среднее по всем приписанным к нему элементам
- Затем назначение выполняется снова
- Так происходит до тех пор, пока назначения не прекратят изменяться

Алгоритм k-средних



Segaran, Toby: *Programming Collective Intelligence* O'Reilly , August (2007) .

Преобразование строки в вектор

- Алгоритм SimHash (Charikar’s hash)
- Строка делится на n-граммы символов
- Для каждой n-граммы считается свой двоичный хэш-код
- Хэш-коды всех n-грамм побитово складываются следующим образом:

  если val[i] == 1, то simhash[i] += 1

  если val[i] == 0, то simhash[i] -= 1
- Для каждого значения simhash:

  если simhash[i] > 0, то simhash[i] = 1

  если simhash[i] < 0, то simhash[i] = 0

Алгоритм подсчета SimHash

$V = bit[32]$
$F = getFeatures(inputString)$
for $i = 1 \rightarrow size(F)$ **do**
  $H = hash(F[i])$
  for $n = 1$ **to** 32 **do**
    if $H[n] = 1$ **then**
      $V[n] = V[n] + 1$
    end if
    if $H[n] = 0$ **then**
      $V[n] = V[n] - 1$
    end if
  end for
end for
for $i = 1$ **to** 32 **do**
  if $V[i] > 0$ **then**
    $V[i] = 1$
  end if
  if $V[i] <= 0$ **then**
    $V[i] = 0$
  end if
end for

Гибридный алгоритм

- На первом шаге получаем «грубые» кластеры методом k-средних (k берется заведомо большим, чем требуется)

- Улучшаем кластеры с помощью алгоритмов связей: пытаемся найти «лишние» строки в кластерах и добавить к кластерам ошибочно отброшенные строки

- Метрики для расстояний между строками:
- Расстояние Левенштейна
- Расстояние Левенштейна (взвешенное)
- Расстояние Хэмминга
- Расстояние Жаро-Винклера
- Расстояние Жаккарда
- Коэффициент Танимото

Лучшая кластеризация (эвристика)

- В каждой строке с помощью регулярных выражений удаляем модель и другие числовые характеристики устройства

*HP Photosmart **C200** Camera AC Adapter*

- Для каждого токена считаем обычный хэш-код
- Для строки считаем сумму хэшей всех её токенов
- Группируем строки по сумме хэшей

Общая оценка

- Перекрестная проверка:

Точность	Полнота	F-мера
0,901	0,880	0,890

- Точность (P): $\frac{tp}{fp+tp}$
- Полнота (R): $\frac{tp}{tp+fn}$
- F-мера (F_1): $\frac{(\beta^2+1)PR}{\beta^2 P+R} = \frac{2PR}{P+R}, \beta = 1$

Примеры извлеченных сущностей

- (HP Server Automation software) & (Opsware Server Automation System)
- (HP Server rp7400) & (N-class server)
- (HP Indigo Press 3000) & (Indigo UltraStream 2000)
- (HP ProLiant DL760 server) & (ProLiant 8500)
- (HP ProBook s-series) & (HP Minis)

Извлеченные акронимы

- HP Message Passing Interface (MPI)
- HP OpenView Internet Usage Manager (IUM)
- HP Project Portfolio Management (PPM)
- HP Remote Graphics Software (RGS)
- HP Small Business Center (SBC)

Планы на будущее

- Разметка на уровне токенов на основе префиксных деревьев
- Расширение корпуса
- Извлечение других видов семантических отношений

Ссылки

- John D. Laferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference 12 on Machine Learning, ICML '01, pages 282{289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

- John McCrae and Nigel Collier. Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinformatics, 9(1):159, 2008.

- David Milne, Olena Medelyan, and Ian H. Witten. Mining domain-specific thesauri from wikipedia: A case study. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI '06, pages 442{448, Washington, DC, USA, 2006. IEEE Computer Society.

- Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Wikipedia mining for an association web thesaurus construction. In Proceedings of the 8th international conference on Web information systems engineering, WISE'07, pages 322{334, Berlin, Heidelberg, 2007. Springer-Verlag.

- Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL '09, pages 147{155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Ссылки

- Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, STOC '02, pages 380{388, New York, NY, USA, 2002.ACM.
- Carolyn J. Crouch and Bokyung Yang. Experiments in automatic statistical thesaurus construction. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '92, pages 77{88, New York, NY, USA, 1992. ACM.
- Edward A. Fox, J. Terry Nutter, Thomas Ahlswede, Martha Evens, and Judith Markowitz. Building a large thesaurus for information retrieval. In Proceedings of the second conference on Applied natural language processing, ANLC '88, pages 101{108, Stroudsburg, PA, USA, 1988. Association for Computational Linguistics.
- J. Gonzalo, F. Verdejo, I. Chugur, and Cigarran J. Indexing with wordnet synsets can improve text retrieval. In Proceedings of the COLING/ACL'98.
- Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics.
- Christian Kohlsch□utter, Peter Fankhauser, andWolfgang Nejdl. Boilerplate detection using shallow text features. In Proceedings of the third ACM international conference on Web search and data mining, WSDM '10, pages 441{450, New York, NY, USA, 2010. ACM.
- Christian Kohlsch□utter and Wolfgang Nejdl. A densitometric approach to web page segmentation. In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM '08, pages 1173{1182, New York, NY, USA, 2008. ACM.

Спасибо за внимание

Вопросы