**Вопросы осень 2016**

1. **Основные свойства естественного языка**

Книга “АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ И КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА”

1. Открытость: язык постоянно меняется
2. Нестандартная сочетаемость единиц: большое количество грамматически правильных сочетаний реально не употребляется
3. Большая системность: наличие многозначности, синонимии, омонимии
4. Наличие словоформ (лексем в определенной форме): значит, что есть род/число/падеж и тд
5. **Что такое графематический анализ? Что такое лемматизация?**

Слайды 2-17 лекции 02-2morf.ppt

*Графематический анализ:*

1. Разделение текста на слова, разделители
2. Выделение устойчивых оборотов, не имеющих словоизменительных вариантов
3. Выделение предложений
4. Выделение абзацев
5. Выделение дат
6. Определение языка слова (русский, нерусский)
7. Определение формата написания слова (прописные, строчные буквы)

*Лемматизация* – тип морфологического анализа – приведение к нормальной форме

Лесной, лесного, лесному->лесной

леса -> лес

Танцующая -> танцевать

1. **Как работает словарный морфологический анализ?**

Слайды 18, 22-27 лекции 02-2morf.ppt

*Cловарный:*

* со словарем словоформ

Каждой словоформе поставлена в соответствие основа или лемма

* со словарем основ

*Схема морфологического анализа со словарем:*

Для неслужебных слов:

1. Выделить возможные окончания слова длиной от 0 до 3 символов
2. Для каждого полученного окончания определить код окончания по таблице окончаний и номер флективного класса по словарю основ (лемм)
3. Если номер флективного класса и номер окончания найдены, то проверить их согласованность по морфологической таблице
4. Если согласованность подтверждается, то сохранить данный вариант

*Проблемы:*

1. Дают максимально полный анализ словоформы ?чем плохо?
2. На реальных текстах дают сбои (опечатки, уникальные слова)
3. Не существует абсолютно полных словарей – лексика языка непрерывно пополняется
4. Для примера – невозможно включить в словарь всю существующую терминологию, имена, фамилии и т.д.
5. **Как морфологические анализаторы обрабатывают слова, отсутствующие в словаре**

Слайды 28-32 лекции 02-2morf.ppt

Используется метод предсказаний: *Метод предсказания* – выявление аналогий со словоформами, распознаваемыми имеющимся словарем.

*Алгоритм:*

1. предсказание префиксального образования
* Попытка найти существующую словоформу языка, которая максимально совпадала бы справа со входным словом.
* Если левая часть (потенциальный префикс) не длиннее M символов (пяти), а правая часть (совпавшая с известной словоформой) не короче N символов (четырех), то слово разбирается по образцу известной словоформы.
[евро]технологию, [супер]коньками.
1. предсказание по концовке, взятой из известных словоформ
* Отделяются инвертированные концовки известных словоформа – длины K (пять букв),
Сопоставляются с морфологическими характеристиками:
«анием» - как «ср. род, ед. ч., тв. пад.»
* Такая строка заносится в исходный лексикон, если она встречается:
* не менее L раз (трех) и
* чаще конкурентов в пределах одной части речи
* ВСЕГДА предусматривается разбор именем существительным, хотя бы неизменяемым.
1. **Что такое постморфологический анализ. Основные методы.**

\’Нужно подробно про методы? Слайды 33-43 лекции 02-2morf.ppt

1. =предсинтаксический анализ
2. Предназначен для устранения морфологической омонимии (многозначности) слов
* Выбор правильной леммы
* Уточнение морфологических характеристик

*Основные методы*

1. Написание правил,
2. Статистические методы, прежде всего, на основе морфологически размеченного корпуса

*Примеры правил:*

* Удаление признаков служебных частей речи для однобуквенных слов, за которыми следуют точки
* Удаление омонимов слова «уже», соответствующих прилагательным, если за ним не стоит запятая или слово в родительном падеже

*Морфологическая разметка корпуса:*

1. Морфологический анализ всех словоформ текста (информация о морфологических (грамматических) характеристик в виде тегов)
2. Снятие неоднозначностей (или исправление ошибок)
3. Добавление информации о результатах в электронное представление текста

Снимаем омонимию с помощью скрытых марковских моделей:

- встречаемость каждого тега в определенном месте цепочки зависит только от предыдущего тега;

- то, какое слово находится в том или ином месте цепочки, полностью определяется тегом (а не, допустим, соседними словами).
Таким образом, порождение правильно построенной цепочки тегов уподобляется действию конечного автомата, где дуги помечены тегами с приписанными им вероятностями, а слова – это наблюдаемые реализации тегов. Состояния определяются парой «текущий тег + предыдущий тег».

1. **Основные понятия информационного поиска**

Слайды 14-18 в лекции 01-01appl.ppt

Сфера науки, которая исследует методы поиска информации, называется информационный поиск (information retrieval (IR))

Документ - материальный объект, содержащий информацию в зафиксированном виде и специально предназначенный для её передачи во времени и пространстве.

Примеры

Интернет-страницы, электронные письма, книги, новости, посты форумов, патенты и многое другое

Общие свойства

* Значительное текстовое содержание
* Некоторая структура:
* заголовок,автор, дата - для статей;
* тема, отправитель, адресат - для писем

Записи базы данных (структурированные таблицы) состоят из хорошо определенных полей и атрибутов

* e.g., банковские записи балансы, номера счетов, имена, адреса, даты рождения, номера социального обеспечения
* Relevance – релевантность
	+ Простое (и упрощающее) определение: Релевантный документ содержит информацию, которую искал пользователь, когда задавал запрос поисковой машине
	+ На релевантность оказывают влияние много различных факторов: задача, контекст, опыт пользователя, новизна, стиль
	+ Тематическая релевантность(отражение заданной темы) vs. пользовательская релевантность(все остальные факторы)
* Evaluation - оценка качества
	+ Экспериментальные процедуры и меры для сравнения результатов работы систем с ожиданиями пользователей
	+ Методы оценки качества поиска сейчас используются во многих областях
	+ Типично используются тестовые коллекции документов, запросов, и оценки релевантности
	+ *Полнота и точность – простые примеры оценки качества*
* Users and Information Needs – потребность пользователя, информационная потребность
	+ Оценка качества поиска – является “пользователецентричной”
	+ Ключевые слова – это слишком бедное описание действительных информационных потребностей
	+ Взаимодействие и контекст – важны для понимания потребности пользователя
	+ Методы уточнения запроса: расширение запроса, предложение запроса, *relevance feedback*
1. **Виды поисковых систем по охвату и направленности. Особенности разных типов поисковых систем**

Что еще?? Слайды 28-35 в лекции 02-1search\_engines.ppt

* Интернет-поиск
	+ Собирает результаты по общедоступному Интернету
	+ Проблема ранжирования результатов
	+ Большие объемы
	+ Громадная индустрия – Интернет-реклама
	+ Активные исследования: хорошее качество
* Корпоративный поиск
	+ Собирает информацию разных форматов из совокупности хранилищ
	+ Ранжирование документов разного типа
	+ Относительно малый объем исследований
	+ Хуже качество поиска – сложнее проводить сравнительные исследования
	+ Активная сфера исследований
1. **Особенности научного поиска**

Слайды 36-44 в лекции 02-1search\_engines.ppt

1. Рост публикаций
2. Устаревание
3. Проблема литературы до интернетовской эпохи
4. Цитаты, которые можно использовать как ссылки в Интернет-поиске
5. Большое количество ссылок на работу – фактор значимости работы
6. наукометрия
7. **Основные этапы обработки текстов в поисковой машине**

Нужно ли подробно рассказывать? Слайды 9-20 в лекции 02-1search\_engines.ppt

1. Извлечение текстов

Идентифицирует и сохраняет тексты для индексирования

1. Трансформация текстов

Трансформирует документы в индексные термы (parser, стоп-слова, стемминг, анализ ссылок, извлечение информации, классификация по темам)

1. Создание индексов

Берет индексные термы и создает индексы для быстрого поиска (статистика по документам, определение весов, инвертирование, распределенное хранение индекса)

1. **Основные этапы обработки запроса в поисковой машине**

Нужно ли подробно рассказывать? Слайды 21-27 в лекции 02-1search\_engines.ppt

1. Взаимодействие с пользователем
Поддерживает создание и уточнение запроса, показ результатов (интерфейс и парсер для языка запроса, трансформация запроса: спеллчек, подсказки, расширение запроса; выдача результатов: подсветка, релевантная реклама и пр)
2. Ранжирование
Использует запрос и индексы, чтобы породить ранжированный лист документов (для каждого документа есть вес соответствия запросу; оптимизация выполнения запроса; распределенное выполнение)
3. Оценка качества
Мониторит и измеряет качество поиска
4. **Алгоритм сопоставления запроса с документами (Алгоритм Merge)**

Слайды 16 в лекции 03search.ppt



Если длины списков x и y, то пересечение O(x+y) операций
Необходимо: записи в списке должны быть отсортированы по docId

1. **Булевская модель информационного поиска. Преимущества и недостатки булевской модели поиска**

Нужно ли подробно рассказывать? Слайды 3-23 в лекции 03search.ppt

1. Два возможных результата для сопоставления запроса и документа
	1. TRUE и FALSE
	2. Поиск по полному совпадению
	3. Простейшая форма ранжирования
2. Запрос может специфицироваться посредством Булевых операторов
	1. AND, OR, NOT
	2. Могут быть использоваться операторы близости (proximity)
3. Инвертированный индекс
	1. Для каждого терма храним список номеров документов, где этот терм встречается

*Преимущества*

1. Результаты предсказуемы, их легко объяснить
2. Могут быть встроены многие различные признаки
3. Эффективная обработка

*Недостатки*

1. Качество выдачи зависит исключительно от пользователя
2. Простые запросы дают слишком много документов (нет упорядочения)
3. Длинные запросы сложно составить
4. **Как измеряется качество булевского поиска**

Слайды 9-10 в лекции 04-01eval.ppt

Оценка булевского поиска

* *Специфика*: булевский поиск не имеет ранжирования (упорядочения)
1. Поисковая система разделяет коллекцию на два множества
	* Разделение по принципу выдано в ответ на запрос или нет
	* Разделение по принципу релевантен документ или нет
2. Вычисляются меры качества:
	* *Обозначения*

*Доли различных видов документов среди всех*

|  |  |  |
| --- | --- | --- |
|  | Relevant | Nonrelevant |
| Retrieved | tp | fp |
| Not Retrieved | fn | tn |

* + *Точность (Precision, P)*

Доля релевантных документов в выдаче

P(relevant | retrieved) = $P = \frac{tp}{tp + fp}$

* + *Полнота (Recall, R)*

Доля выданных документов среди релевантных

P(retrieved | relevant) = $R = \frac{tp}{tp + fn}$

* + *F-мера*

$F1 = \frac{2}{1/R + 1/P}=\frac{2PR}{P + R}$

1. **Алгоритм сопоставления запроса с документами (Алгоритм Merge)**

 см. 11.

1. **Векторная модель информационного поиска. Показатели *idf* и *tf.idf*.**

Слайды 30, 34-36, 38-39, 45-47 (в основном) в лекции 03search.ppt

*Общие сведения*

Рассмотрим v-мерное векторное пространство, v - количество слов. Слова - отдельные позиции (оси) в векторах. Каждый документ и запрос - вектор в $N^{v}$. Введем понятие близости документов. Предположение: чем меньше угол между векторами, тем больше они похожи.



*Мера сходства документов*

Для определения близости документов определяется, насколько одинаково направлены соответствующие вектора документов (то есть, насколько мал угол между ними). Используется косинусная мера:

$q , d\_{} \in N^{|v|}=> ρ(q,d) =cos(q,d) = \frac{q\*d}{|q|\*|d|}= \frac{\sum\_{i}^{}q\_{i}d\_{i}}{\sqrt{\sum\_{i}^{}q\_{i}^{2}}\sqrt{\sum\_{i}^{}d\_{i}^{2}}} $

Если векторы $q , d\_{}$уже нормализованы, то

$ ρ(q,d) =cos(q,d) = \frac{q\*d}{|q|\*|d|}=\sum\_{i}^{}q\_{i}d\_{i}$

*Построение векторного пространства*

Определим представление документа в векторном пространстве. Необходимо учесть долю пересечения слов запроса с документом. Векторное представлением строится на основе двух факторов.

* *Первый фактор:* матрица частоты употребления терма в документе. Рассмотрим число вхождений терма в документ. Если всего v термов, каждый документ - вектор частот в $N^{v}$, обозначается tf (term frequency). Совокупность векторов частот порождает матрицу частот всех термов во всех документов.
* *Второй фактор:* подокументная частотность. Частотные слова менее информативны, чем редкие, самые частотные слова в документе - служебные. Сверхчастотные слова типа: предлоги, союза, которые есть во многих документах вообще иногда рассматриваются как стоп-слова и выбрасываются из документа.Чтобы учесть эту распространенность - вводится фактор, df - количество документов, в которых употреблялось это слово - подокументная частотность.

*Применение факторов для построения представления*

* **Вес** **idf**. t - терм, $df\_{t}$ - подокументная частотность t: количество документов, содержащих t. Таким образом, df - обратная мера информативност, df<=N, N - число документов. Определяем idf (обратная подокументная частота):

$idf\_{t} = log\_{10}(N/df\_{t})$

Логарифм используется для смягчения разницы в употреблении более и менее частотных слов. Данная величина вычисляется для каждого слова во всех коллекции документов

* **tf.idf взвешивание.** Произведение tf на idf. Tf - количество входждений терма в документ, вычисляется отдельно для каждого документа и терма (первый фактор).Увеличивается при росте количества упоминаний в документе. Увеличивается для редких терминов в коллекции. Каждый документ представлен вещественным вектором размерностью число разных слов в коллекции.
* **Модификация 1 tf.idf.** $w\_{t,d} = log(1+tf\_{t,d})\*idf\_{t}$



1. **Классическая процедура оценки качества информационно поиска**

Нужны ли слайды 2,3,5? Слайды 6,8,12 в лекции 04-01eval.ppt

Классическая процедура оценки (Cranfield, Крендфилдские эксперименты) :

1. Составим список запросов и ограничим коллекцию документов
2. Для каждой пары запрос/документ выставим экспертную оценку “релевантности”
3. Будем рассматривать ответ системы как множество/последовательность оценок релевантности
4. На полученной последовательности/множестве оценок релевантности построим метрики

Оценка релевантности выдачи

* Информационная потребность выражается запросом
* Релевантность оценивается по отношению к информационной потребности, а не к словам запроса (даже, если все слова запроса присутствуют в тексте, текст может быть не релевантен)

Оценка ранжированных результатов:

* Современные системы выдают упорядоченные результаты
* Выдача может быть достаточно большой
* Релевантные документы должны выдаваться раньше нерелевантных
* Можно измерять точность на каждом уровне полноты
1. **Что такое РОМИП, какие задачи в нем решаются?**

Слайды 5 в лекции 04-01eval.ppt + википедия + слайды 2-18 в лекции 04-02eval.ppt

Российский семинар по оценке Методов Информационного Поиска (РОМИП) (2003) - это открытый семинар, проводимый ежегодно с 2003 года группой российских исследователей и разработчиков, занимающихся информационным поиском. Основная цель семинара — создание плацдарма для проведения независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией.

Структурно семинар представляет собой набор дорожек — секций, посвященных конкретным проектам с определенной задачей и правилами оценки. Оргкомитет формирует тестовые наборы данных, заданий и распространяет их участникам.

Примеры задач:

1. текстовый поиск (поиск по запросу)
2. текстовая классификация
	1. веб-сайты
	2. веб-страницы
3. контекстно-зависимое аннотирование (составление аннотации документа по запросу)
4. поиск изображения по образцу
5. выявление нечетких дубликатов изображений
6. построение текстовых меток по изображению
7. автоматический анализ тональности текста (2011г) (эмоциональная оценка, выраженная в тексте)
8. **Кривая полнота-точность. 11-точечный график TREC?**

?? Слайды 13-20 в лекции 04-01eval.ppt

Усреднение по запросам:

* Кривая полнота-точность для одного запроса не очень интересна
* Нужно построить кривую полнота-точность для совокупности запросов
	+ Пока Кривая – это совокупность точек
	+ Как интерполировать?



Кривая полнота-точность.

Интерполированная точность:

Идея: Если локально точность возросла с увеличением полноты, то засчитаем ее максимум…
Т.е. берем максимум точности справа на графике


11-точный график TREC:

* Значения полноты от 0 до 1 с шагом 0.1
* Интерполяция точности
* Если $r\_{t }> recall(q\_{j}),$ то $p(r\_{t}, q\_{j}) = 0$
* Если $r\_{t }\leq recall(q\_{j}),$ то $p(r\_{t}, q\_{j}) = max\_{n}(precision(n)), n\geq pos(r\_{t}, q\_{j})$



1. **Что такое пулинг в информационном поиске? Сложности, связанные с пулингом**

WTF?? Слайды 32-36 в лекции 04-01eval.ppt

Для каждого запроса:

1. Собрать результаты систем участников глубины A
2. Выбрать из полученных результатов B первых
3. Удалить дубликаты
4. Проставить оценки релевантности
5. Не оцененные документы считать нерелевантными
6. Оценить весь ответ системы (с глубиной А)

Сложности:

* Взаимное усиление систем
* Недооценка систем, не участвовавших в оценке
* Получаемая оценка – оценка снизу
* Но: участники относительно в равных условиях
1. **Оценка качества в поисковых машинах Интернет.**

WTF?? Слайды 39 в лекции 04-01eval.ppt

1. Полноту невозможно измерить
2. K- первых документов
3. Релевантные документы должны показываться раньше
4. NDCG (Normalized Cumulative Discounted Gain)
5. Использование кликов пользователей
	1. A/B testing
6. **Шкалы оценок. Мера NDCG**

Слайды 40-43 в лекции 04-01eval.ppt

Шкалы оценок:

* В прошлом: TReC – бинарные
* Сейчас TReC:
* Высоко релевантный
* Релевантный
* Нерелевантный
* РОМИП
* Соответствует
* Скорее соответствует
* Возможно соответствует
* Не соответствует
* Не может быть оценен

Оценка качества выдачи по небинарным оценкам:

* Предположения
* Лучше, если релевантные документы находятся в начале списка
* Если есть несколько типов релевантных документов, то лучше, чтобы документы с высокими оценками были раньше в списке
* Существует наилучшее упорядочение расположения оценок от лучших к худшим
* В суммированной оценке выдачи каждая следующая позиция в списке должна давать меньший вклад, чем предыдущая

Discounted Cumulative Gain: $DCG\_{λ}=g\_{1}+\sum\_{i = 2}^{λ}\frac{g\_{i}}{log\_{}}, g\_{i}$- значение релевантности для i-ого документа в выдаче

nDCG: Нормализация DCG по отношению к лучшему упорядочению по данному запросу:

$nDCG\_{p} = \frac{DCG\_{p}}{IDCG\_{p}}$

1. **Что такое информационно-поисковые тезаурусы? Зачем они нужны? Где применяются сейчас**

Надо ли 34-37??? Слайды 28-33 в лекции 05-01exp.ppt

*Тезаурус* – словарь, в котором

* Единицы соответствуют близким по смыслу словам (понятия, дескрипторы и др.)
* Между единицами установлены формализованные отношения

Типы тезаурусов

* Информационно-поисковые тезаурусы
* Тезаурусы типа WordNet
* И др.

*Информационно-поисковый тезаурус* – нормативный словарь терминов предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области

* Нормативные ключевые слова
* Национальные и международные стандарты
* Используются в ряде международных организаций и парламентский организаций
	+ Европейский парламент – EUROVOC
	+ ООН – UNBIS Thesaurus

Цели:

1. Перевод языка авторов на нормативный язык, используемый для индексации и поиска
2. Обеспечение последовательности в присваивании индексных терминов
3. Обозначение отношений между терминами
4. Облегчение информационного поиска

Использование

1. Особенно популярны в 60-80х годах 20 века
2. Повышение интереса в последнее время за счет роста интереса к предметно-ориентированному и корпоративному поиску

Структура ИПТ:

1. Основные понятия предметной области - дескрипторы
2. Условные синонимы - аскрипторы
	1. отношение эквивалентности
	2. аскриптор - дескриптор
3. Отношения между дескрипторами
	1. родовидовые отношения (выше-ниже): BT (broader term) - NT (narrower term)
	2. ассоциации (транспорт-путешествие и пр): RT (related term)
	3. часть - целое (органы тела, географические объекты, дисциплины, иерархические структуры)
4. **Назовите методы расширения запросов пользователей при информационном поиске.**

Подробности??? Слайды 2-3, 26, 41 в лекции 05-01exp.ppt

* Несовпадение слова запроса:

самолет – лайнер

* Методы расширения запроса:
	+ Глобальные методы
		- Ручные тезаурусы
		- Автоматически порождаемый тезаурус
	+ Локальные методы (по конкретному запросу)
		- Relevance feedback (обратная связь по релевантности)
		- Pseudo Relevance feedback (обратная связь по псевдорелевантности

*Relevance FeedBack:*

1. Пользователь оценивает документы в поисковой выдаче
* Пользователь задает относительно простой, короткий запрос
* Затем пользователь размечает часть результатов как релевантные и нерелевантные
* Система вычисляет улучшает соответствие документов запросу на основе пользовательской разметки
* Процедура может выполняться итеративно.
1. *Основная идея*: сформулировать хороший запрос трудно, если пользователь не знаком с коллекцией, поэтому – итеративное построение запроса

*Pseudo Relevance FeedBack:*

1. Pseudo-relevance feedback автоматизирует «ручную» часть реального relevance feedback.
2. Pseudo-relevance алгоритм:
* Строит поисковую выдачу по запросу
* Предполагает, что первые k документов - релевантны
* Выполняет relevance feedback
1. В среднем хорошо работает
2. Но может получить очень плохие результаты для некоторых запросов
3. Несколько итераций могут вызвать «искажение запроса»

*Ручной тезаурус:*

* Для каждого термина t в запросе происходит расширение синонимичными словами или близкими по смыслу (связанными отношениями с исходным словом) – из тезауруса
* feline → feline cat
* Как расширять:
* Можно добавлять в вектор запроса (с более низкими весами и в зависимости от типа отношения к слову запроса)
* Можно вставлять в булевское выражение
* Налог →( НАЛОГ или НАЛОГОВЫЙ)
* Используется в предметно-ориентированных системах
* Современные тезаурусы, встроенные в ПО поисковые системы, могут иметь другие формы, чем описано в стандартах, например, только список синонимов и вариантов
1. **Алгоритм Роккио для *relevance feedback***

Слайды 13-14, 17 в лекции 05-01exp.ppt

*Ключевые понятия:*

* Центроид – это центр масс совокупности точек
* Документы – это точки в многомерном пространстве
* Определение: Центроид



где C – множество документов.

*Алгоритм Роккио:*

1. Алгоритм Rocchio использует векторное пространства найти наилучший запрос на основе пользовательской разметки
2. Rocchio ищет запрос $q\_{opt}$ , который максимизирует



1. Пытается отделить релевантные и нерелевантные документы



1. Проблема: мы не знаем все релевантные документы

*Алгоритм Роккио (SMART):*

1. На практике используется:



1. $D\_{r}$ = множество известных релевантных doc векторов
2. $D\_{nr}$ = множество известных нерелевантных doc векторов
* Отличны от $C\_{r}$ и $C\_{nr}$
1. $q\_{m}$ = модифицированный вектор запроса; $q\_{0}$ = исходный вектор запроса; α,β,γ: веса
2. Новый запрос «сдвигается» по направлению к релевантным документам и «уходит» от нерелевантных документов
3. **Назовите проблемы расширения запроса при помощи обратной связи по релевантности**

 Слайды 21-24 в лекции 05-01exp.ppt

* Длинные запросы – неэффективны для типичной поисковой машины
* Большее ожидание для пользователя
* Высокая стоимость для поисковой системы
* Частичное решение:
	+ Использование только слов с наиболее высоким весом: Например, 20 первых по весу
* Пользователи часто не хотят размечать документы
* Трудно понять, почему данный документ был выдан после relevance feedback

*Предположения для relevance feedback (тоже нарушаются):*

1. A1: Пользователь имеет достаточно знаний для исходного запроса
2. A2: Прототипы релевантных/нерелевантных документов “ведут себя хорошо”
	1. Распределение слов в релевантных документах сходно
	2. Распределение слов в нерелевантных документах отлично от распределения слов в релевантных документах
		1. 1) Все релевантные документы похожи на один прототип
		2. 2) Имеется несколько прототипов, но у них значительное пересечение по составу
		3. Сходство между релевантными и нерелевантными документами относительно небольшое
3. **Вопросно-ответные системы: постановка задачи, основные компоненты, особенности тестирования.**

 Тестирование?? Постановка?? Слайды 11-12 в лекции 06-02qa.ppt + вики

Вопросно-ответная система (QA-система; от англ. QA — англ. Question-answering system) — информационная система, способная принимать вопросы и отвечать на них на естественном языке, другими словами, это система с естественно-языковым интерфейсом.

Этапы работы:

1. Анализ вопроса

* определяется тип вопроса (вопрос времени, места, количества и другие) и тип ответа
* Формируется запрос в информационно-поисковую систему

2. Поиск релевантных документов или абзацев информационно-поисковой системой

* формируется упорядоченный список наиболее релевантных документов (абзацев)
* выбирается первых n (например, n=100-1000) документов (абзацев) для дальнейшей обработки

3. Анализ полученных документов

* содержит ли документ требуемый тип ответа
* Анализ предложений: близость слов ответа и вопроса, сходство синтаксических структур и т.п.

4. Извлечение ответа заданного типа

* суммирование ответов от разных документов
1. **Классификация вопросов в вопросо-ответных системах. Типы вопросов и типы ответов**

Слайды 7, 14-19 в лекции 06-02qa.ppt

*Типы вопросов:*

1. Фактоидные вопросы - краткий ответ
2. Нефактоидные (нарративные) вопросы - ответ включает одно или более предложений

*Типы вопросов:*

* Вопрос типа почему
* по какой причине, на каком основании, с какой целью, какова причина, какова цель,
* зачем, для чего, из-за чего, благодаря чему, и т.п. все предлоги из типа фрагм.
* что + (вызвать/ вызывать/ обуславливать/ обусловить/ определить/ определять/ повлечь/ повлиять/ породить/ порождать/ привести/ приводить/ способствовать)
* Вопрос типа когда
* в какое время (\_ / случиться / происходить / быть / произойти /cовершиться / случаться / приключиться / приключаться ; с какого времени, до какого времени / к какому сроку / к какому времени)
* Вопрос типа «где»
* Вопрос типа «как»
* Вопрос типа «кто»
* Вопрос типа «что такое»
* Вопрос типа «куда»
* Вопрос типа «сколько»
* Вопрос типа «какой параметр»
* Вопрос типа «что»
* Вопрос типа «как называется»
* Вопрос на сравнение

*Типы ответов на вопрос “когда”:*

* Ответ типа ФРАГМЕНТ
	+ в начале
	+ в конце
	+ в то время как …
* Ответ типа СПИСОК
	+ день
	+ ночь
	+ Утро…
* Ответ типа ЧИСЛО
	+ Даты, точное время

*Таксономия ответов:*

* 6 классов верхнего уровня
* Сокращение, сущность, определение, человек, место, число
* 50 классов второго уровня
* Место: город, страна, гора
* Человек: группа, титул, персона,
* Сущность: животное, тело, цвет, валюта
1. **Исправление несловарных ошибок на основе применения правила Байеса**

Слайды 11, 13-31 в лекции 06-01spell.ppt

1. Распознавание:
* Любое слово, которое не найдено в словаре, - это ошибка
* Чем больше словарь, тем лучше
1. Исправление:
* Порождение кандидатов: реальных слов, которые похожи на ошибку
* Выбор наилучшего слова:
	1. Сходство
* По написанию - наикратчайшее редакционное расстояние (=расстояние Левенштейна - измеряется в количестве минимальных редакционных операций (вставка, удаление, замена, swap соседних), которые требуются для преобразования одного слова в другое)
* По звучанию
	1. Вероятность по методу зашумленного канала (noisy channel)

*Теорема Байеса:* $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

*NoisyChannel + Байес:*

* Мы видим неправильно написанное слово x
* Найдем правильное слово $\overline{w}$

$\overline{w}= argmax\_{w}P(w|x) = argmax\_{w}\frac{P(x|w)P(w)}{P(x)}=argmax\_{w}P(x|w)P(w)$
Вероятность ошибки:



* del[x,y]: количество (xy написано как x)
* ins[x,y]: количество(x написано как xy)
* sub[x,y]: количество(y написано как x)
* trans[x,y]: количество(xy написано как yx)
* можно сгладить, добавив 1 в числитель и нормализовав, А - алфавит символов



Языковая статистическая модель:

* Нужно собрать большой корпус.
* Пусть C(w) = # количество вхождений w

$P(w) = \frac{C(w)}{T}$

* Для запросов – корпусом может быть множество всех заданных пользователем запросов
1. **Исправление ошибок перехода в другое словарное слова на основе применения правила Байеса**

Слайды 12, 36-44 в лекции 06-01spell.ppt

1. Для каждого слова порождается множество кандидатов:
	1. С похожим произношением
	2. С похожим написанием (все слова с небольшим редакционным расстоянием 1-2)
	3. Текущее слово w включается в множество
2. Выбор лучшего кандидата
	1. Подход зашумленного канала
	2. Использование контекста
	3. Flying form Heathrow to LAX Flying from Heathrow to LAX

*Noisy Channel:*

* Дано предложение $w\_{1}$,$w\_{2}$,…,$w\_{n}$
* Множество кандидатов для каждого слова wi
	+ Candidate($w\_{1}$) = {$w\_{1}$, $w'\_{1}$ , $w''\_{1}$ , $w'''\_{1}$ ,…}
	+ Candidate($w\_{2}$) = {$w\_{2}$, $w'\_{2}$ , $w''\_{2}$ , $w'''\_{2}$ ,…}
	+ Candidate($w\_{n}$) = {$w\_{n}$, $w'\_{n}$ , $w''\_{n}$ , $w'''\_{n}$ ,…}
* Нужно выбрать последовательность W, которая максимизирует P(W)

Упрощение: одна ошибка на предложение.

Строим список всех возможных предложений.

*Подсчет вероятностей:*

1. Подсчет вероятностей по корпусу
	* Униграммы
	* Биграммы
	* И др.
2. Channel model
* То же самое, как для несловарной ошибки
* Плюс нужна вероятность отсутствия ошибки: P(w|w)
	+ Оценивается вероятность отсутствия ошибки в слове
* 0.9-0.99
1. **Учет контекста при исправлении ошибок написания**

Слайды 32-35, в лекции 06-01spell.ppt

Учет контекста:

1. Учет условных вероятностей появления одного слова после другого
* P($w\_{1}$…$w\_{n}$) = P($w\_{1}$)P($w\_{2}$|$w\_{1}$)…P($w\_{n}$|$w\_{n-1}$)
1. Вероятности насчитываются на некотором корпусе. Нужно насчитать:
* P($w\_{i}$) – вероятность униграмм
* P($w\_{n}$|$w\_{n-1}$) – вероятность биграмм
* Для униграмм P(w) всегда ненулевое
* Поскольку наш словарь построен на текстовой коллекции
* Но P($w\_{k}$|$w\_{k-1}$) может быть нулевым.
* Нужно сглаживание
* Можно применить сглаживание add-1

**Задачи на следующие темы:**

1. **Точность, полнота, F-мера – меры качества**

Точность: $P = \frac{tp}{tp + fp}$(доля релевантных в выдаче)

Полнота:$R = \frac{tp}{tp + fn}$(доля выданных среди релевантных)

F-мера:$F1 = \frac{2}{1/R + 1/P}=\frac{2PR}{P + R}$

1. **Мера качества упорядочения: средняя точность**

$AP = \frac{\sum\_{rel}^{}P\_{rel\_{k}}}{N\_{rel}}$, где $P\_{rel\_{k}}$ - точность в момент k, когда в выдаче релевантный документ. $N\_{rel}$ - число релевантных документов.

1. **Нахождение близости между запросом и документом по векторной модели, языковой модели**
2. Строим вектора для документов и запроса tf или tf-idf
* $tf\_{t,i}$ - количество вхождений терма t в i-ый документ
* $idf\_{t} =log\_{10}(N/df\_{t})$, N - число документов, $df\_{t}$ - количество документов, содержащих терм t
* $w\_{t,i }= tf\_{t,i}.idf\_{t} = tf\_{t,i} \* idf\_{t}$
1. Находим косинус между ними

$cos(d, q) = \frac{(d, q)}{|d|\*|q|}; (d, q) = \sum\_{i=1}^{n}d\_{i}\*q\_{i};|d| = \sqrt{d\_{1}^{2}+...+d\_{n}^{2}} $

1. **Мера упорядочения: NDCG**

$nDCG\_{p} = \frac{DCG\_{p}}{IDCG\_{p}}$

$DCG\_{p}=g\_{1}+\sum\_{i = 2}^{p}\frac{g\_{i}}{log\_{i}}, g\_{i}$- значение релевантности для i-ого документа в выдаче

$IDCG\_{p}$- для наилучшего расположения документов по релевантности

1. **Расширение запроса методов relevance feedback.**
2. Строим вектора для документов и запроса
3. Рассчитываем новый запрос $q\_{m}$ по формуле:

$q\_{m}=α\*q\_{0} + β\*\frac{1}{|D\_{r}|}\sum\_{d\_{j in D\_{r}}}^{}d\_{j}- γ\*\frac{1}{|D\_{nr}|}\sum\_{d\_{j in D\_{nr}}}^{}d\_{j}$

$q\_{0}$ - исходных запрос,$D\_{r}$ - множество релевантных документов, $D\_{nr}$ - множество нерелевантных документов,$α, β, γ$ - веса