



Яндекс на РОМИП-2005

И. Сегалович

Версия 2005.10.06_01

Дорожки

1. Поиск: небольшие изменения параметров, связанные со спецификой веб-поиска (при этом перед отправкой результатов из-за спешки НЕ удалось воспроизвести результаты прошлого года по прогонам РОМИП-2004, AvrPres меньше на 7-10%)
2. Классификация страниц (из-за дефицита времени не проверялись результаты перед отправкой, не анализировались оценки РОМИП, нет комментариев)
3. Новости (нет оценок)
4. 2006 ? Надеемся исправиться ☹️😊

Новостная дорожка

- Суммарно 24 дня
 - "отставка Шеварднадзе с 18.11.2003 по 24.11.2003 включительно (8 дней) – проблема очень большого кластера, повышенный интерес к качеству из-за взрывного роста аудитории
 - "Взрыв в Ессентуках, Выборы в Госдуму" с 01.12.2003 по 09.12.2003 включительно (8 дней)
 - "обычная неделя" с 31.03.2004 по 7.04.2004 (8 дней)
- Объем
 - 17 источников, ~24 тыс. из ~163 тыс. сообщений, что составляет ок. 15%
- Описание кластеров и документов
- Договор
 - Переписан договор с источниками, подписывается отдельное соглашение, приглашение к участию на news.yandex.ru (специально для «генерального директора»)

Дубли в коллекции РОМИП

- 530 тыс. документов, входящих в коллекцию РОМИП, и состоящих не менее, чем из 10 слов
- 10 млн. 200 тыс. пар документов, входящих в коллекцию РОМИП, состоящих не менее, чем из 20 слов и являющихся «нечеткими» дубликатами с коэффициентом сходства не менее 85%
- Пулинг 12 методов: лексические и «синтаксические» из них из них 4 оригинальных (Маслов 1998, Ильинский 2002, Сегалович 2002, Зеленков 2005)
- Коэффициент сходства подсчитывался для страниц с удаленными тэгами (Perl String::Similarity)