Матчинг отзывов и организаций в Яндексе

Антон Алексеев, CSC
anton.m.alexeyev@yandex.ru

Яндекс.Отзывы

5 июня 2014 г.

План

1. Что такое Яндекс.Отзывы
2. Постановка задачи
3. Старый матчинг
4. Новый матчинг
5. Результаты и выводы

Яндекс.Отзывы

- Яндекс — это не только Яndex.ru
- Много сервисов с отзывами
- Единая платформа для хранения, обработки и предоставления доступа к

[Здесь должен быть скриншот смешного отзыва]

Какие бывают отзывы?

По типу источника

- Свои (user generated content)
- Партнёрские
  - Микроразметка
  - Фиды
  - Спецроботы

По типу объекта

- На организации
- На приложения
- На автомобили
- etc.

Партнёрские отзывы на организации

То, что получаем от партнёров

- Собственно отзыв
  - текст
  - ссылка
  - рейтинг
  - etc.
- Организация
  - наименование
  - ссылка
  - адрес
  - телефон
  - категория?
  - etc.
- Бессмертный автор
  - имя
  - etc.

Задача

- Каждому отзыву умеем сопоставлять организацию из Яндекс.Справочника
- Много, много жалоб на матчинг от владельцев организаций, партнёров и пользователей

Задача

- Каждому отзыву умеем сопоставлять организацию из Яндекс.Справочника
- Много, много жалоб на матчинг от владельцев организаций, партнёров и пользователей
- **Повысить точность сопоставления, не убивая полноту**

Отступление

$$Recall_{total} = \frac{Right + Wrong}{Right + Wrong + Skipped + Trash}$$

$$Precision_{important} = \frac{Right}{Wrong}$$

$$Recall_{important} = \frac{Right}{Right + Skipped}$$

Как всё устроено

Старая магия

”+7 (812) 2-12-85-0-6”	→	[phone:60581222187]
”Санкт-Петебург, Столярный пер., 5.”	→	[address:санкт-петербург address:столярный address:пер address:5]
”Бабушкин комод”	→	[name:бабушкин name:комод]

Фильтрующий запрос: должны встречаться хотя бы один терм из адреса и один терм из названия.
Для **ранжирования** используется всё.
Получаем список top-5 по Lucene score и берём верхний.

Что можно сделать?

- Разметить небольшую выборку (верно/неверно; если неверно, есть ли в базе)
- Подогнать параметры, добавить больше эвристик, увеличивающих точность
- ...Учиться автоматически принимать решение «берём/не берём»?

Новая, улучшенная магия [1]

- Не рассматриваем организации без адреса и телефона
- **Фильтрация**
  - Расширяем адрес городами, определёнными по номеру телефона
  - Расширяем имя компании «разрезанием» по заглавным буквам
  - Расширяем имя компании транслитерацией на латиницу
- **Ранжирование**
  - Всё как раньше

Да мы же увеличиваем $Recall_{total}$!

Новая, улучшенная магия [2]

Постфильтрация: на входе найденный кандидат (top-1) из базы и партнёрский отзыв

- *Numbers* — у кандидата и у отзыва в адресах есть числа
- *Intersection* — множества этих чисел пересекаются
- *StreetIsDefined* — у кандидата в адресе выделена улица
- *StreetFound* — название улицы встречается в адресе из отзыва

Отзыв считаем сматчившимся с кандидатом, если

$$(Numbers \rightarrow Intersection)\&(StreetIsDefined \rightarrow StreetFound)$$

Стало ли лучше?

	old matching	new matching
$Recall_{total}$	0.51	0.515
$Recall_{important}$	0.759	0.801
$Precision_{important}$	0.793	0.946

Эпик вин?

Да, но

- По-прежнему плохи школы, больницы и всё, что похоже на

$$Hospital\ №\ 6, Moscow$$

- Нужна гибкость: налёрнить функцию $f(review, organization)$, принимающую решение?

Что понял и узнал

- Улучшить продакшеновое решение — не всегда простая задача
- Scala рулит
- Lucene рулит

Questions time?

Questions time?

- Но это ещё не всё

Подсчёт совместных упоминаний организаций и технологий

Антон Алексеев, CSC
anton.m.alexeyev@yandex.ru

При участии HP Labs: А. Уланов, С. Серебряков

5 июня 2014 г.

Содержание прошлой серии

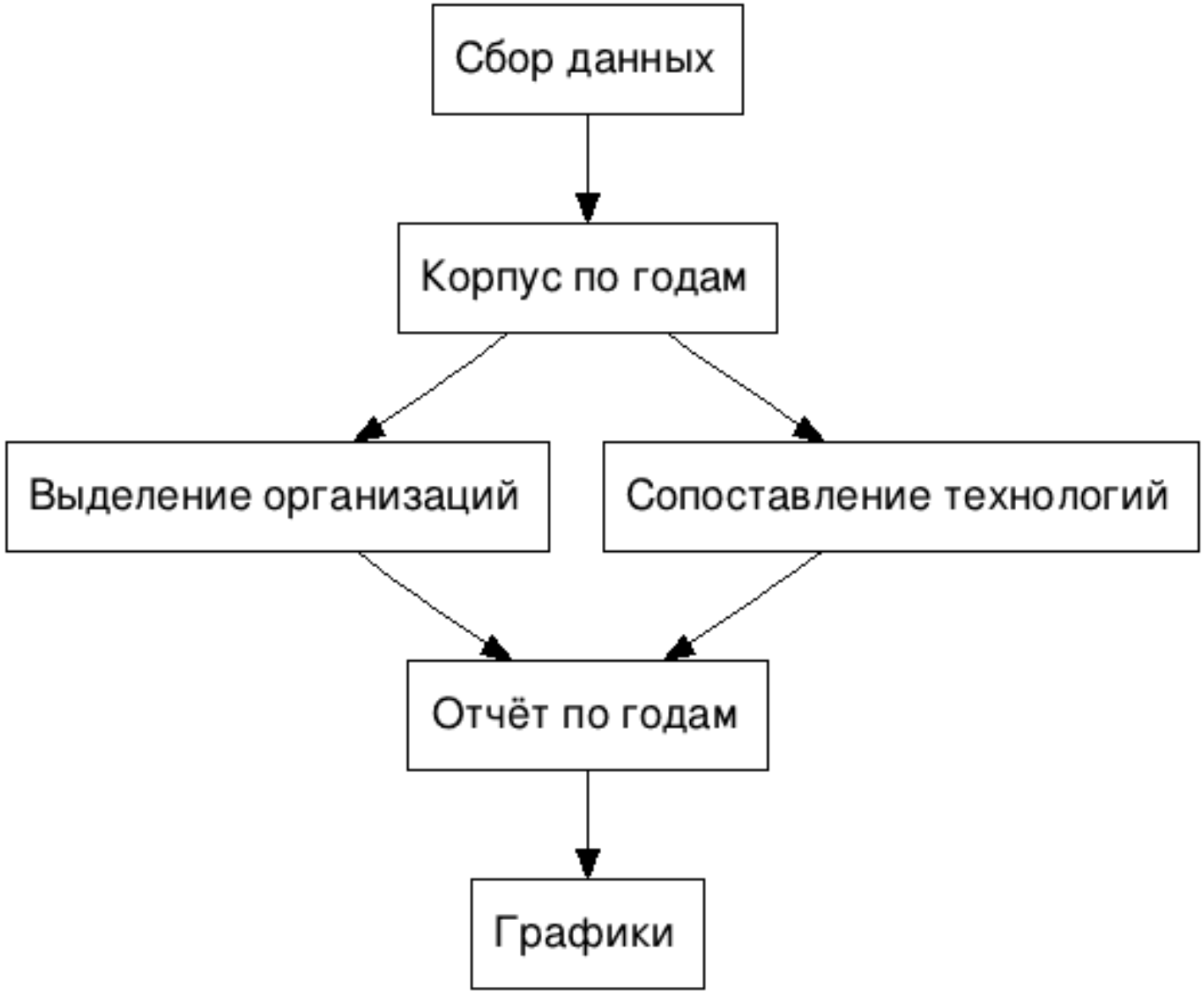
Тексты, Википедия, тренды, вот это всё

- «А для русского языка будете делать?»

План

- Общее понятие
- Источники данных
- Извлечение организаций
- Извлечение трендов

Схема обработки данных

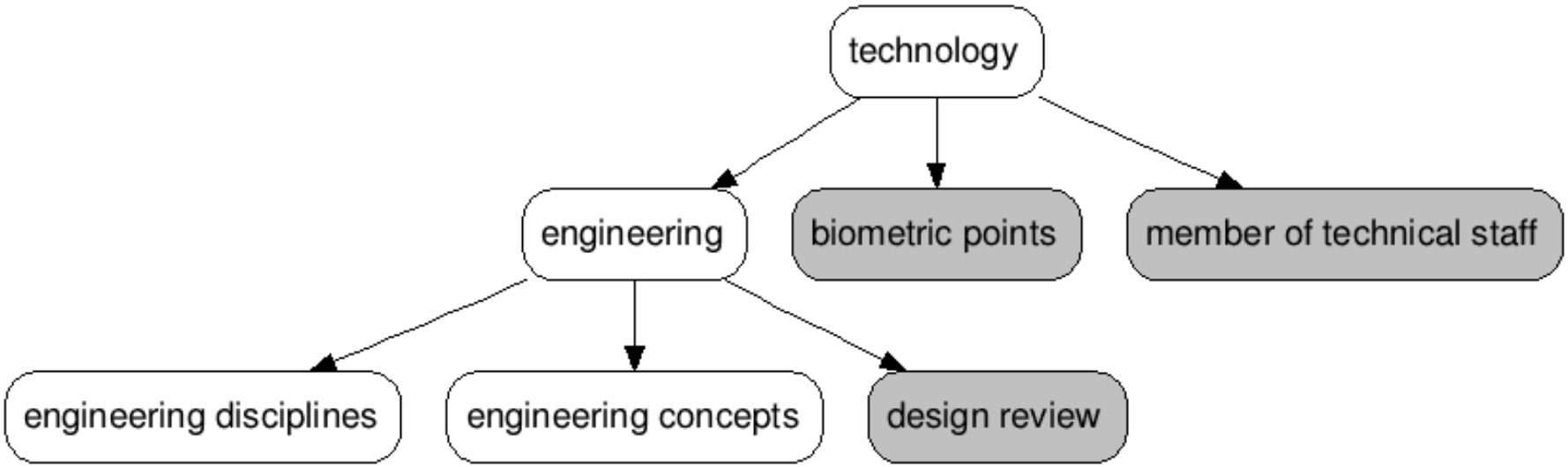


Источники данных

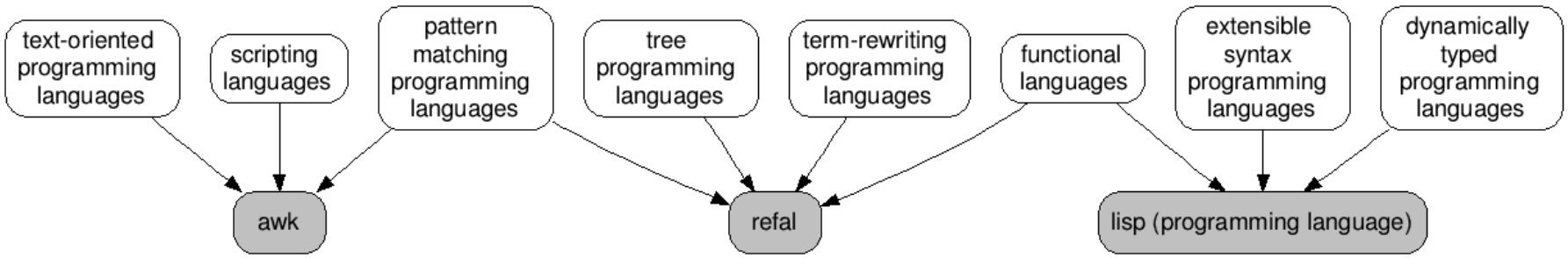
- CrunchBase
- Habrahabr.ru
- Lenta.ru: «Наука и техника»
- Русскоязычная и англоязычная версии Википедии

Структура Википедии [1]

Помимо статей, текстовых ссылок и заголовков, пользователями Википедии поддерживается «дерево категорий»



Структура Википедии [2]



Извлечение наименований организаций

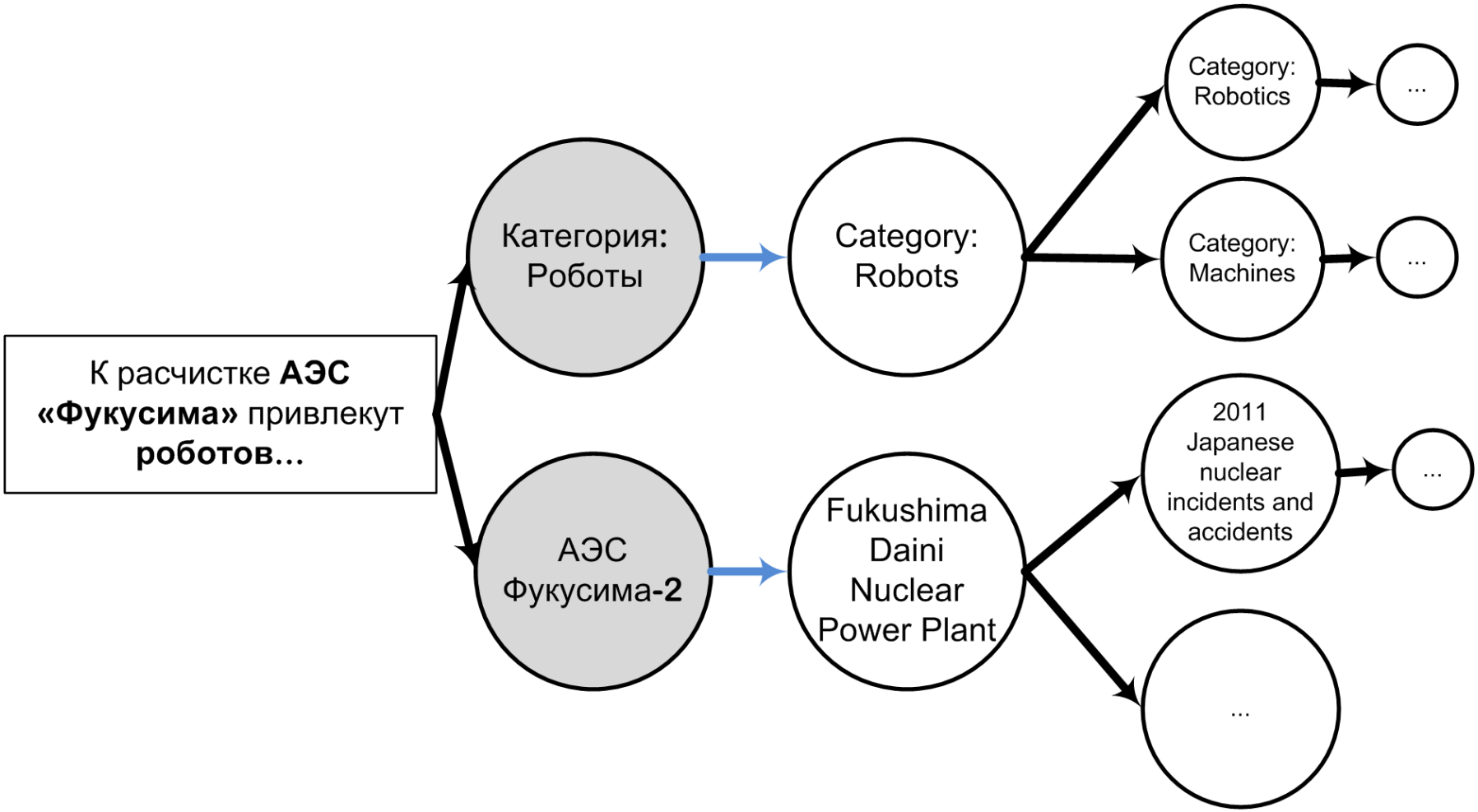
1. Токенизация
2. Стемминг
3. Поиск подстрок из сформированного списка предобработанных наименований организаций

Извлечение технологических областей [1]

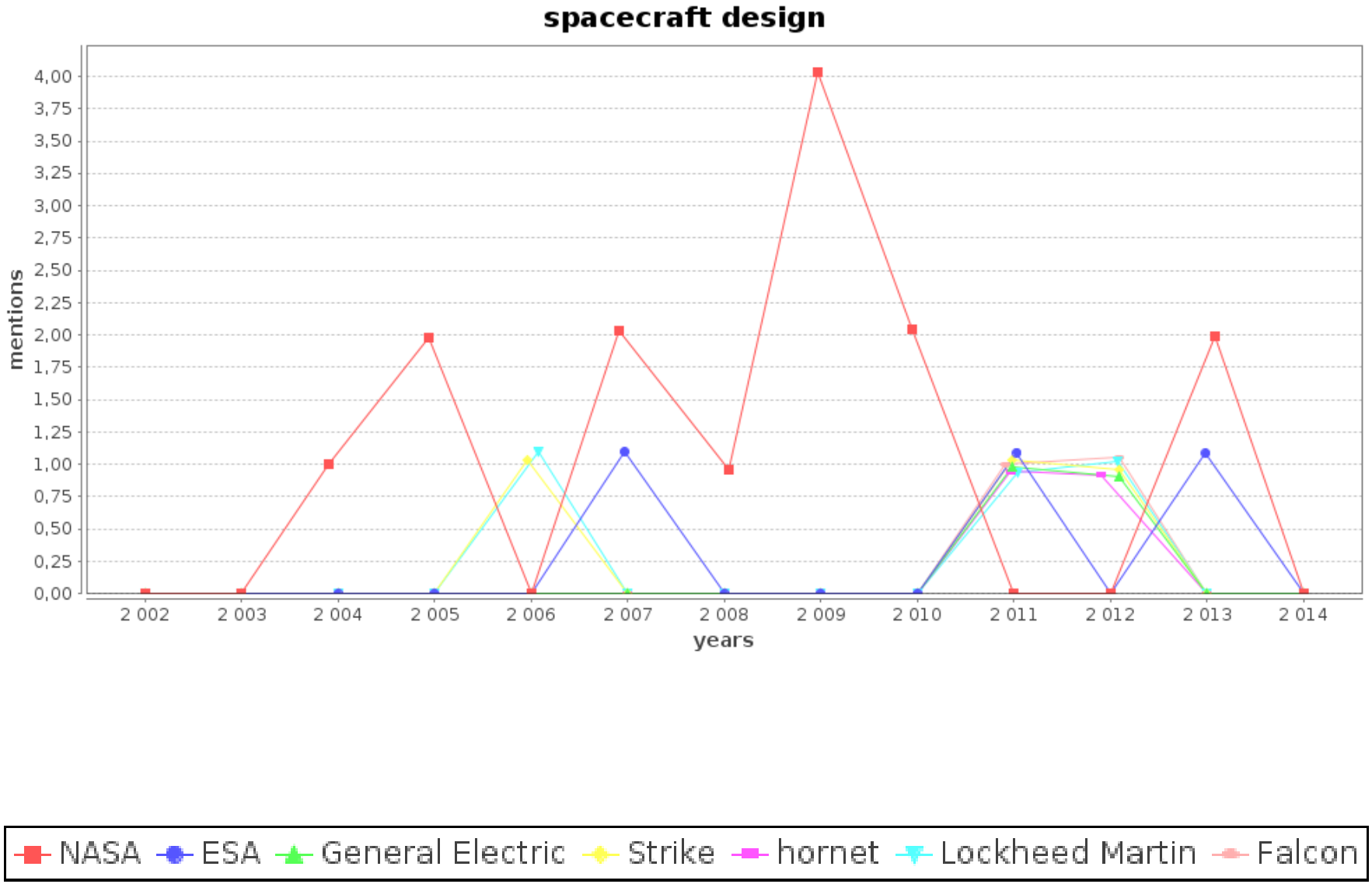
Первый этап — предобработка

1. Токенизация
2. Фильтрация по списку стоп-слов
3. Фильтрация по списку слов, встречавшихся в текстах ссылок и заголовков Википедии

Извлечение технологических областей [2]



Пример результата работы



Использованный инструментарий

(логотипы технологий нужны для того, чтобы размещать их на слайдах)

ЧСВ

- Подход представлен на «СПИСОК-2014»
- Доклад принят на «НСМВ-2014»

Благодарности

- CSC, Яндекс, руководители практик

Благодарности

- CSC, Яндекс, руководители практик
- А ещё спасибо Papeeria за шаблон презентации

Матчинг + извлечение трендов

Антон Алексеев, CSC
anton.m.alexeyev@yandex.ru

Причастны: Яндекс, HP Labs

5 июня 2014 г.