Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра информатики

Комплекс программ для автоматизации анализа популярности технологических областей в корпусе текстов русскоязычных электронных медиа

Антон Михайлович Алексеев, группа 532

Руководитель — Т.В. Тулупьева, доц., доц., к. пс. н.
Рецензент — А.Е. Пащенко, н. с., к. т. н.

4 июня 2014 г.

Структура доклада

1. Введение
2. Объект автоматизации
3. Цель и задачи
4. Актуальность
5. Данные и выбранный инструментарий
6. Методы и комплекс программ
7. Формальные признаки
8. Результаты

Введение

- Маркетологические Интернет-исследования на основе текстов на ествественном языке
- Объём данных слишком велик для «ручного» анализа
- Богатый инструментарий для английского языка

Объект автоматизации

Подсчёт случаев совместных упоминаний технологических областей и наименований организаций в разное время в электронных медиа

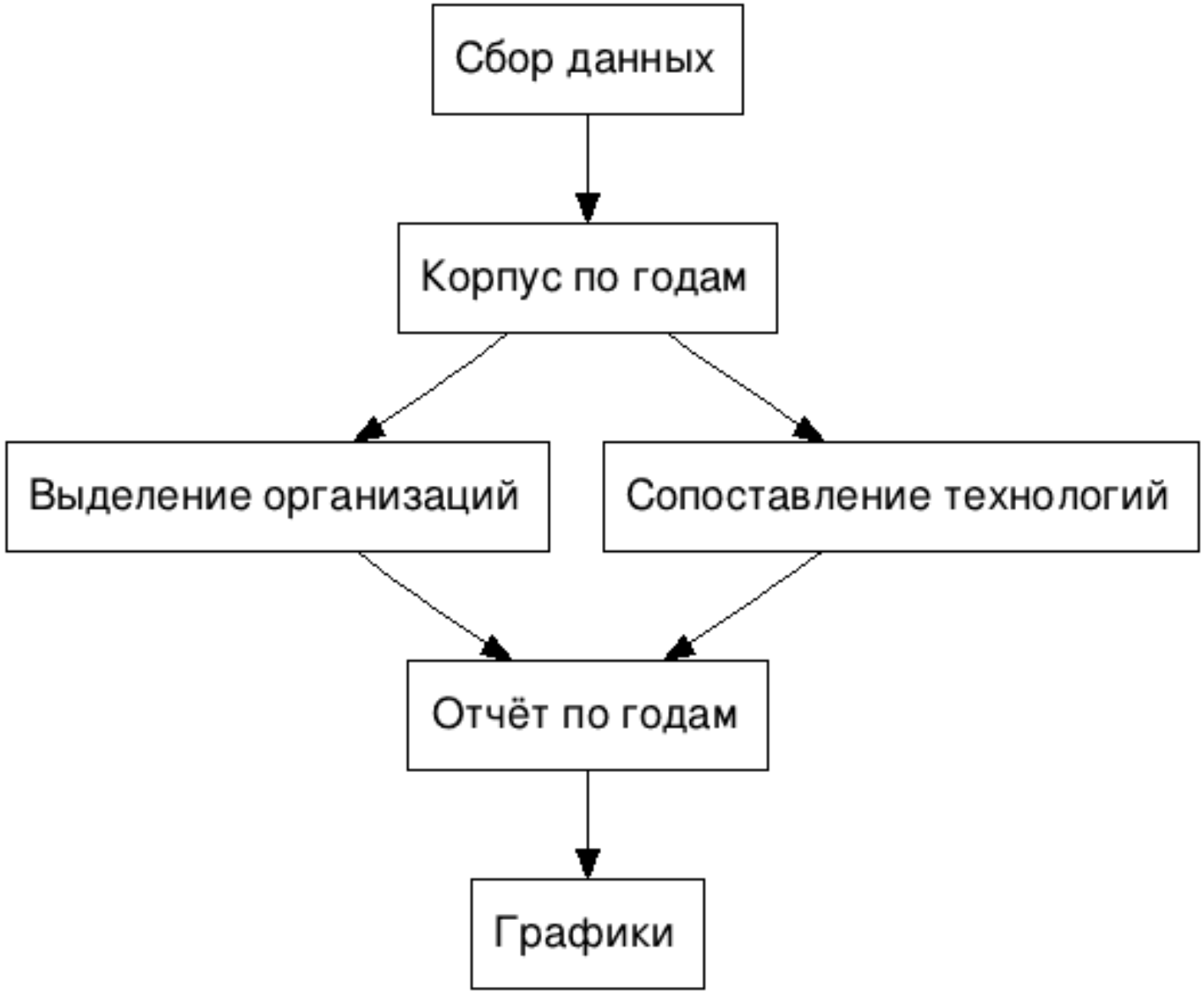
Цель

Разработка комплекса программ, способного установить взаимосвязь упоминаний технологических областей и наименований организаций в блогах или новостях на русском языке на основе корпуса текстов электронных медиа

Задачи

1. Сбор тестовых данных
2. Модуль, выделяющий из текста наименования организаций
3. Модуль, выделяющий из текста технологические области
4. Модуль, осуществляющий построение табличных отчётов по результатам двух последних задач
5. Модуль, осуществляющий визуализацию отчётов в виде графиков

Схема обработки данных

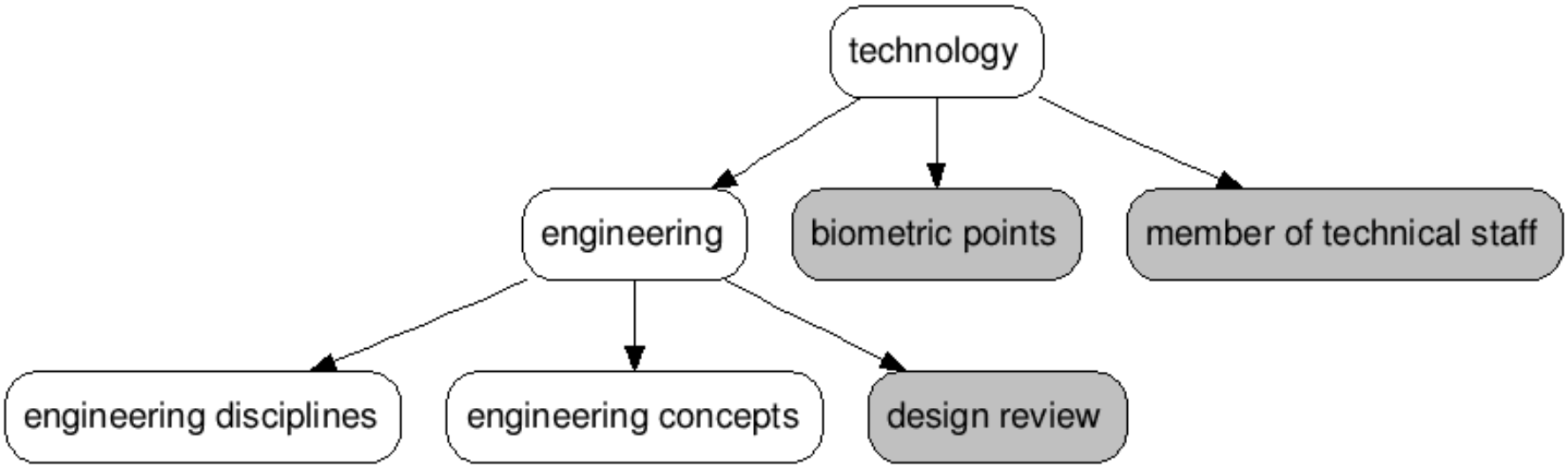


Источники данных

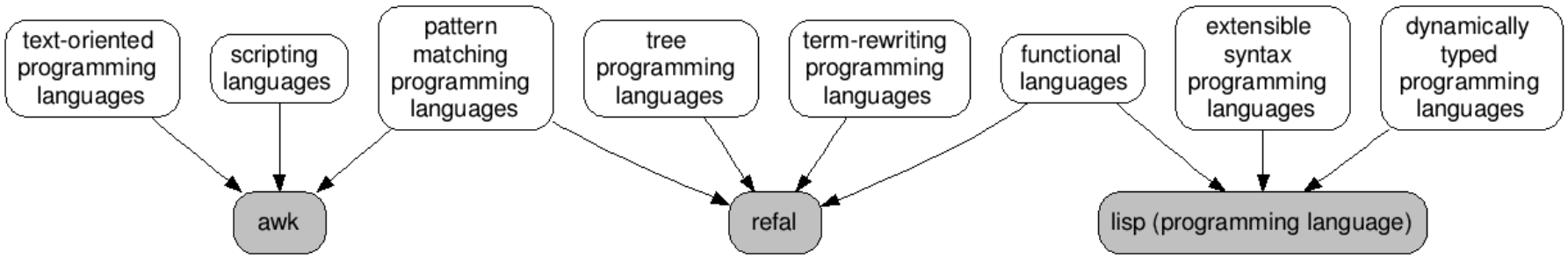
- CrunchBase
- Habrahabr.ru
- Lenta.ru: «Наука и техника»
- Русскоязычная и англоязычная версии Википедии

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

Стемминг — один из видов нормализации токенов; как правило, под этим термином понимают эвристический процесс «отрезания» окончаний слов

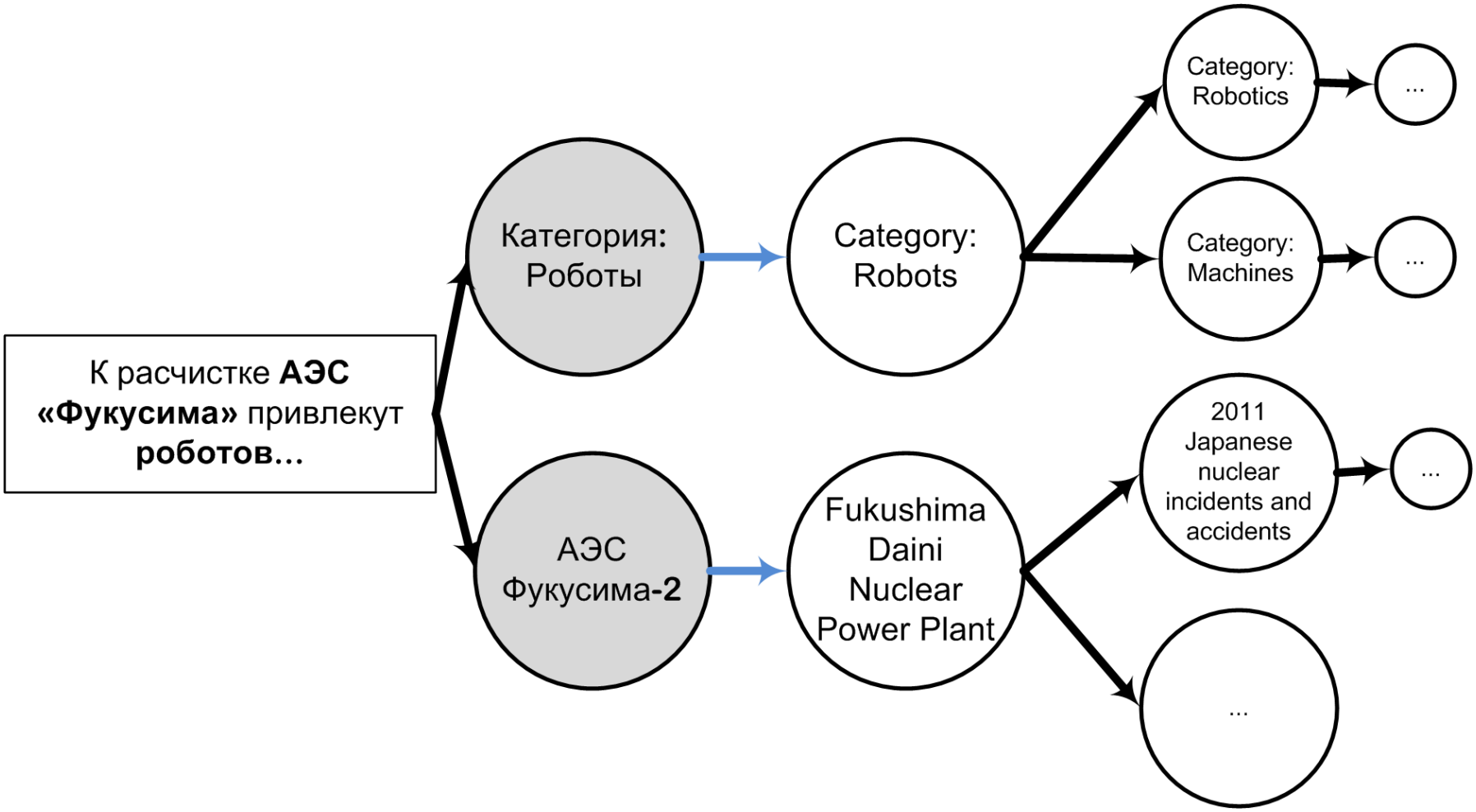
Токенизация — задача разбиения текста на семантически ценные последовательности символов, возможно, с удалением некоторых символов (например, пунктуации)

Извлечение технологических областей [1]

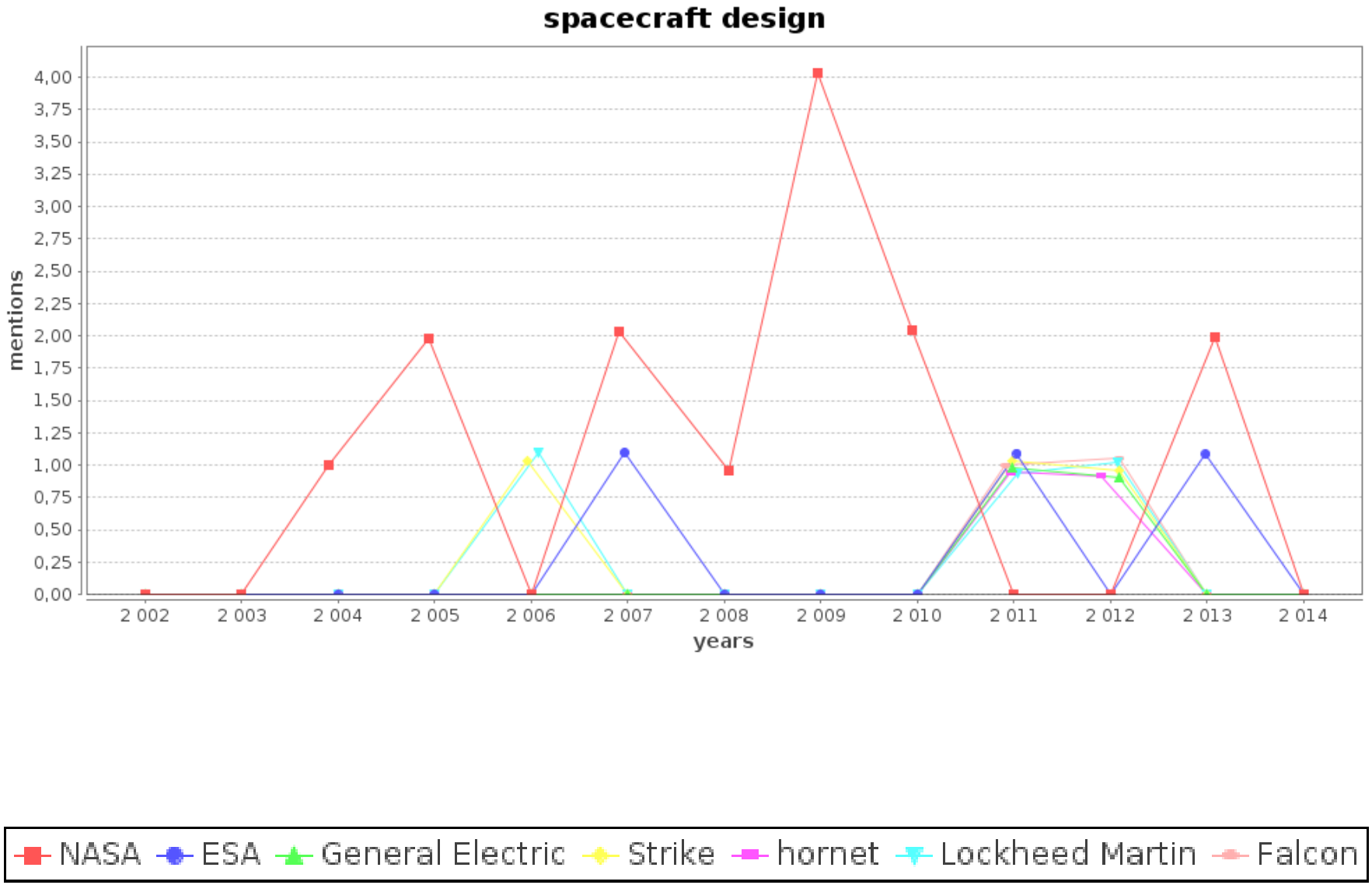
Первый этап — предобработка

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

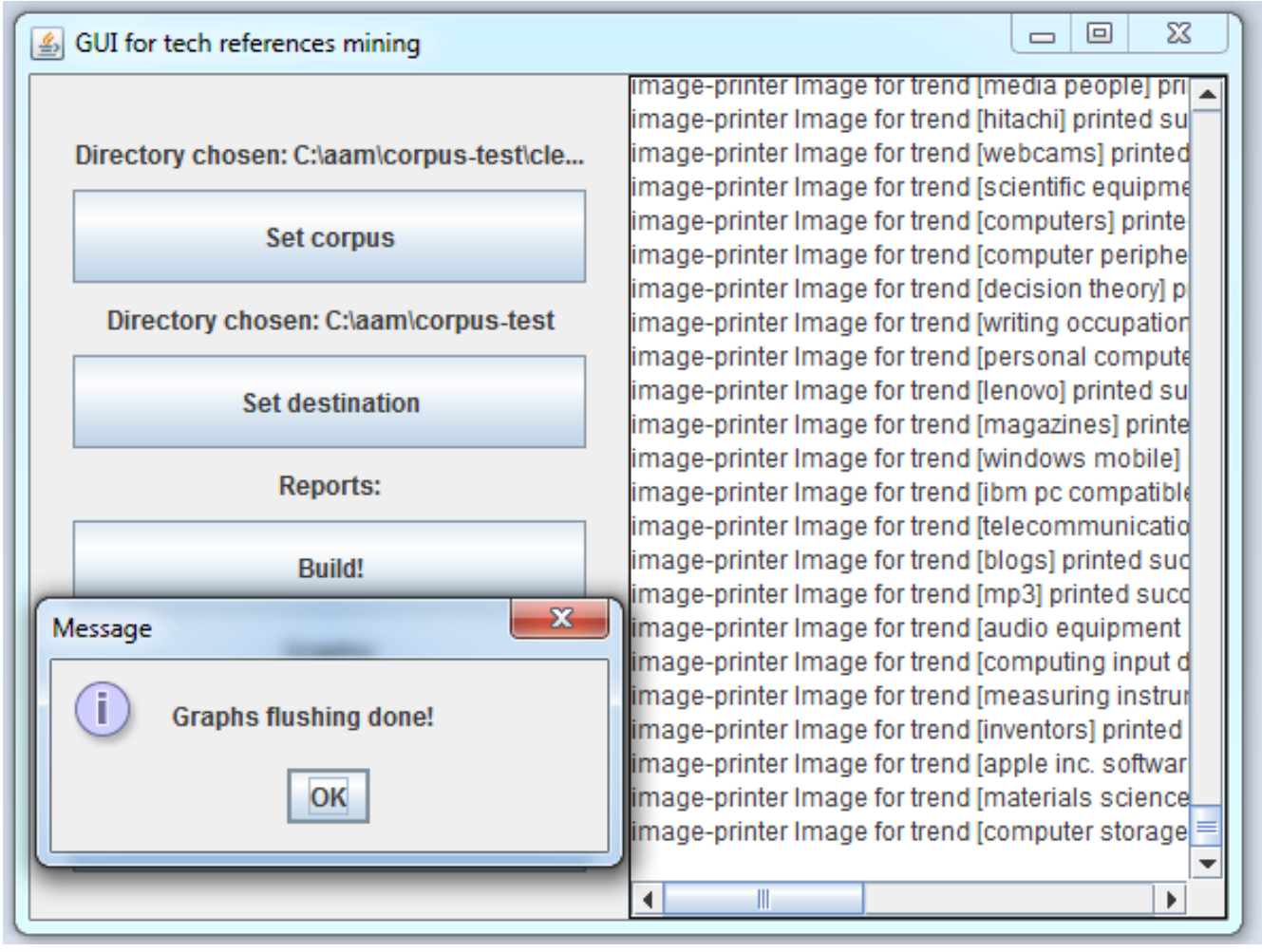
Извлечение технологических областей [2]



Пример результата работы



Пример графического интерфейса



Использованный инструментарий

- Scala
- Java
- Python
- WikiXMLJ
- slf4j + logback
- Apache Commons
- Apache Lucene
- Apache Maven

Результаты

- Автоматизированы
  - извлечение технологических областей и наименований организаций
  - построение отчётов и их визуализация
- Более 5000 строк программного кода на Scala, Java и Python и XML-разметки
- Подход представлен на «СПИСОК-2014»
- Доклад принят на «НСМВ-2014»

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра информатики

Комплекс программ для автоматизации анализа популярности технологических областей в корпусе текстов русскоязычных электронных медиа

Антон Михайлович Алексеев, группа 532

Руководитель — Т.В. Тулупьева, доц., доц., к. пс. н.
Рецензент — А.Е. Пащенко, н. с., к. т. н.

4 июня 2014 г.