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. Данные и выбранный инструментарий
5. Комплекс программ
6. Формальные признаки
7. Результаты

Введение

- Маркетологические Интернет-исследования на основе текстов на ЕЯ
- 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 во влиятельных СМИ

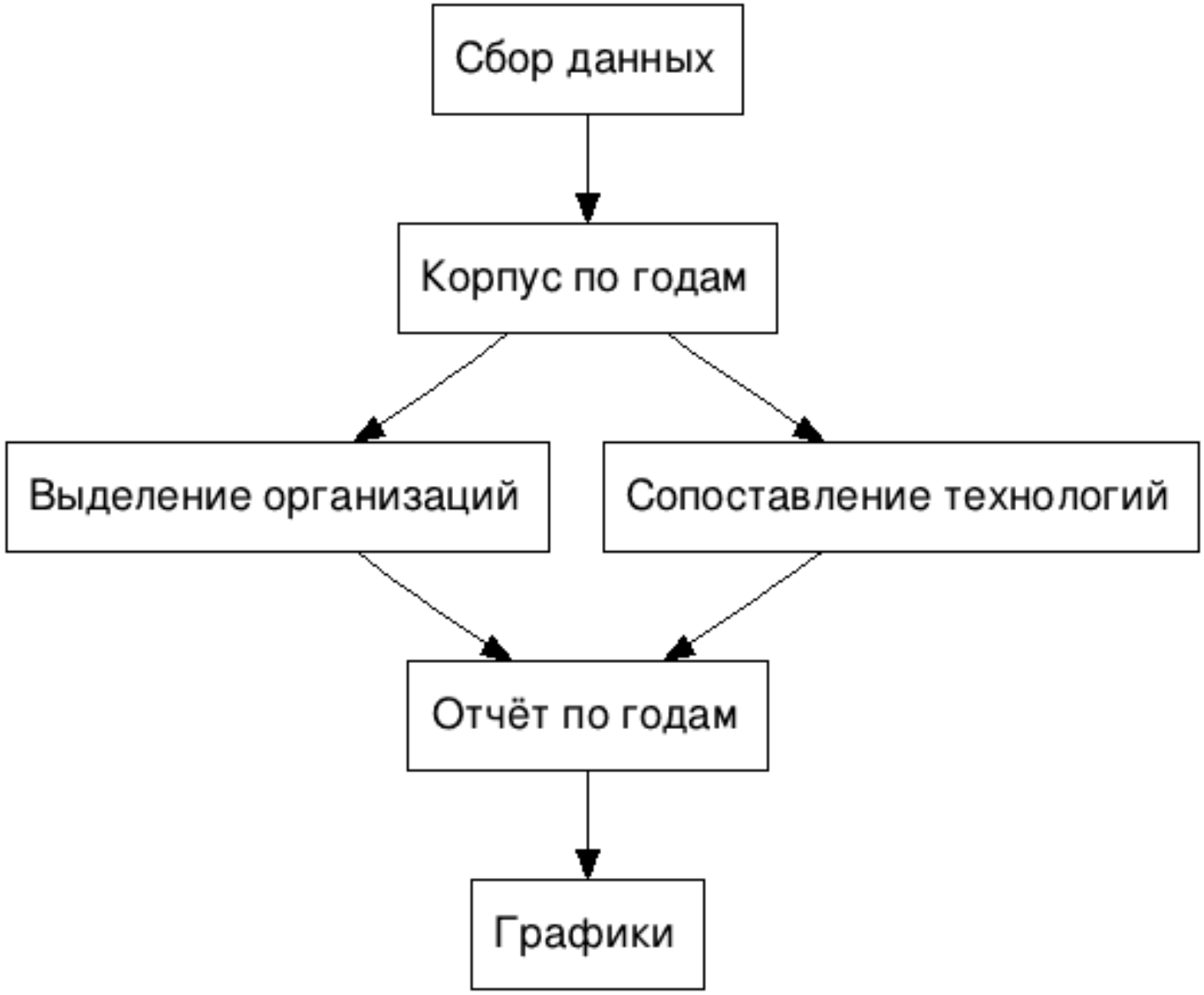
Цель

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

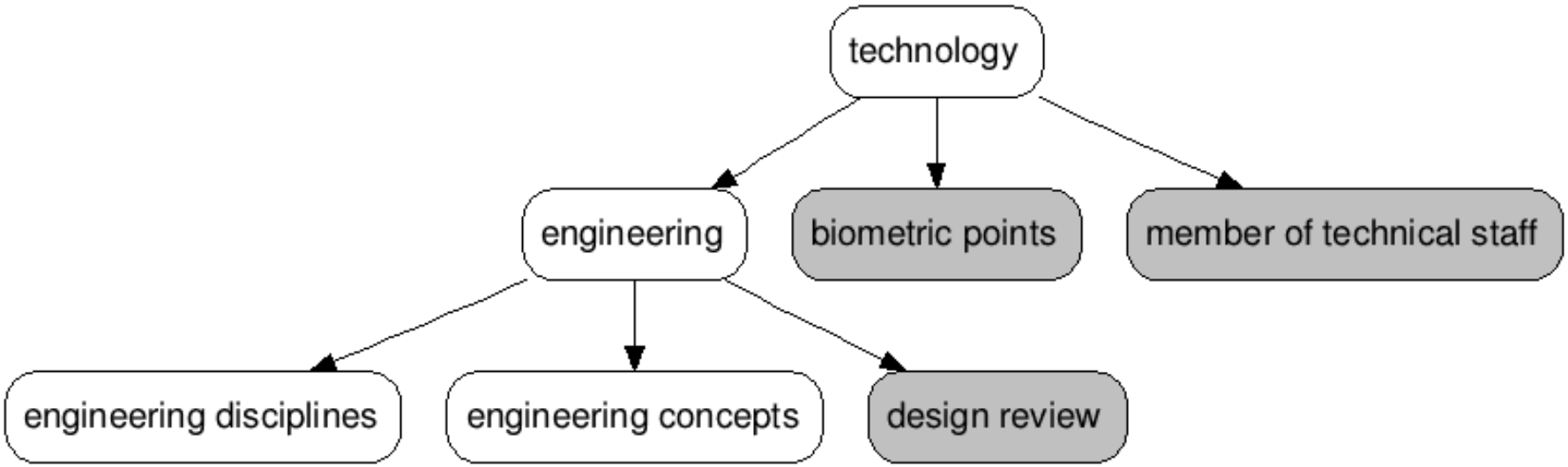


Источники данных

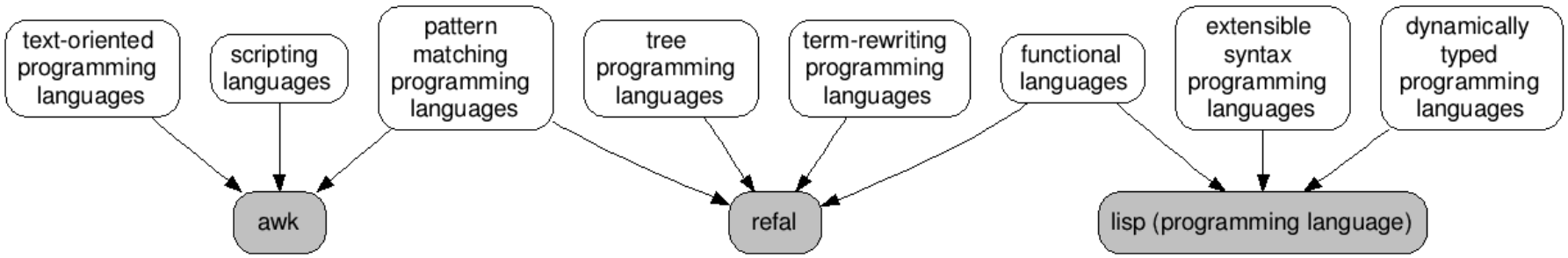
- CrunchBase
- Habrahabr.ru
- Lenta.ru: «Наука и техника»
- Русскоязычная и англоязычная версии Википедии

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 нормализованных наименований организаций

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

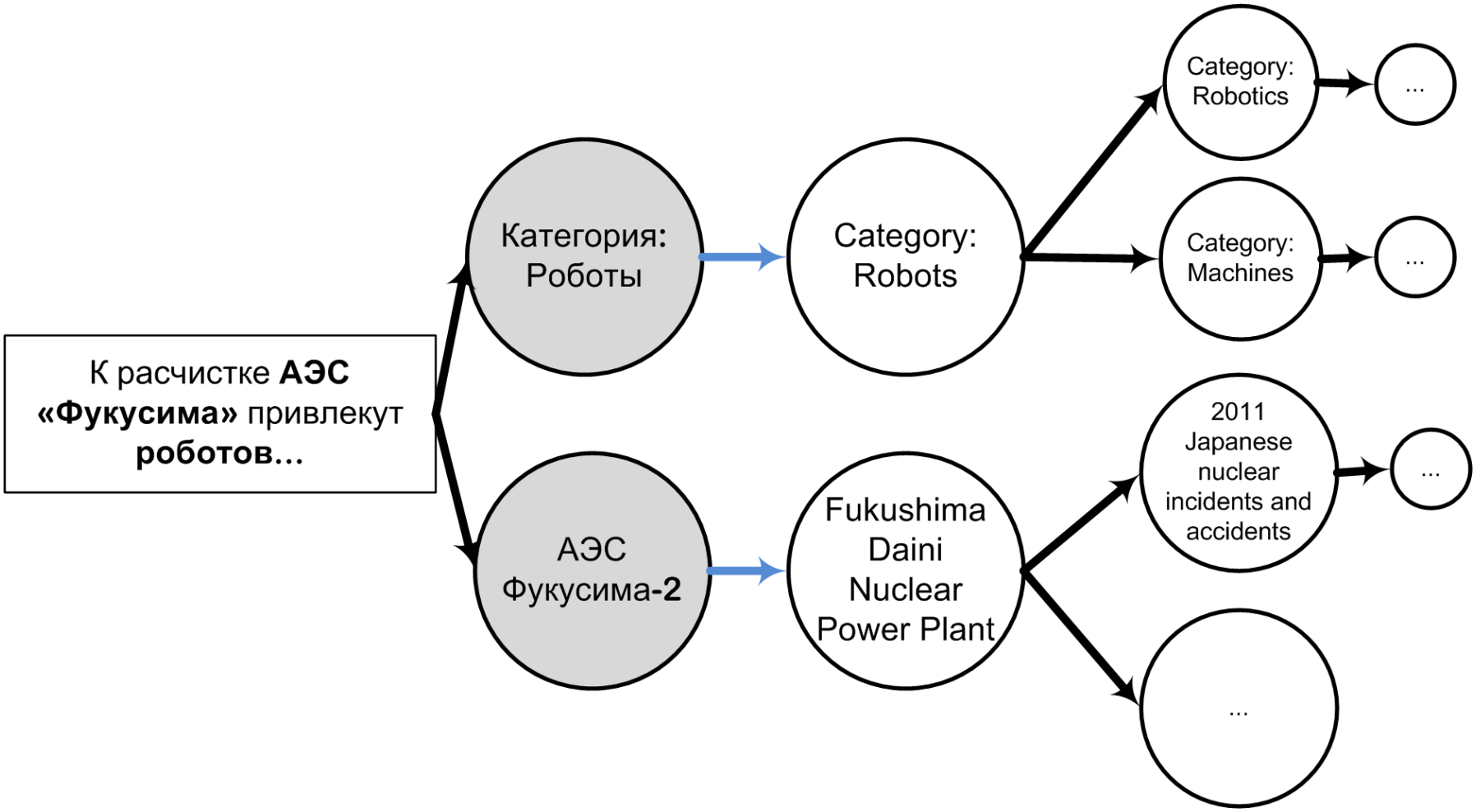
Извлечение технологических областей [1]

1. Предобработка
   - 1.1 Токенизация
   - 1.2 Фильтрация по списку стоп-слов
   - 1.3 Фильтрация по списку нормализованных слов, встречавшихся в текстах ссылок и заголовков Википедии

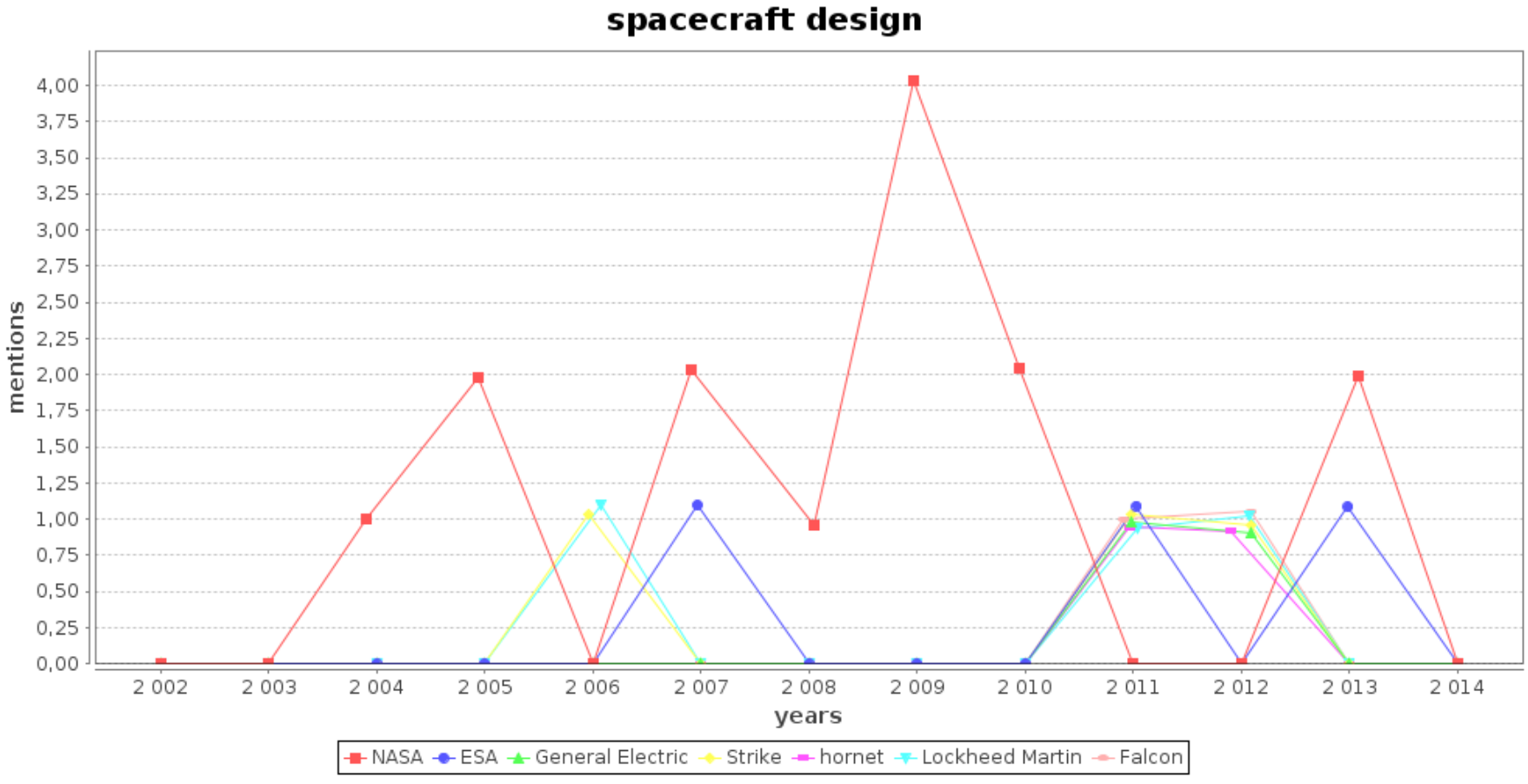
Извлечение технологических областей [2]

2. Поиск технологических областей
   - 2.1 Поиск в тексте заголовков русскоязычной Википедии с помощью инвертированного индекса Lucene
   - 2.2 Переход к англоязычным версиям найденных статей
   - 2.3 «Восхождение» по BFS-дереву категории **Technology** в графе категорий англоязычной версии Википедии
   - 2.4 Запоминание всех посещённых вершин графа как технологических областей

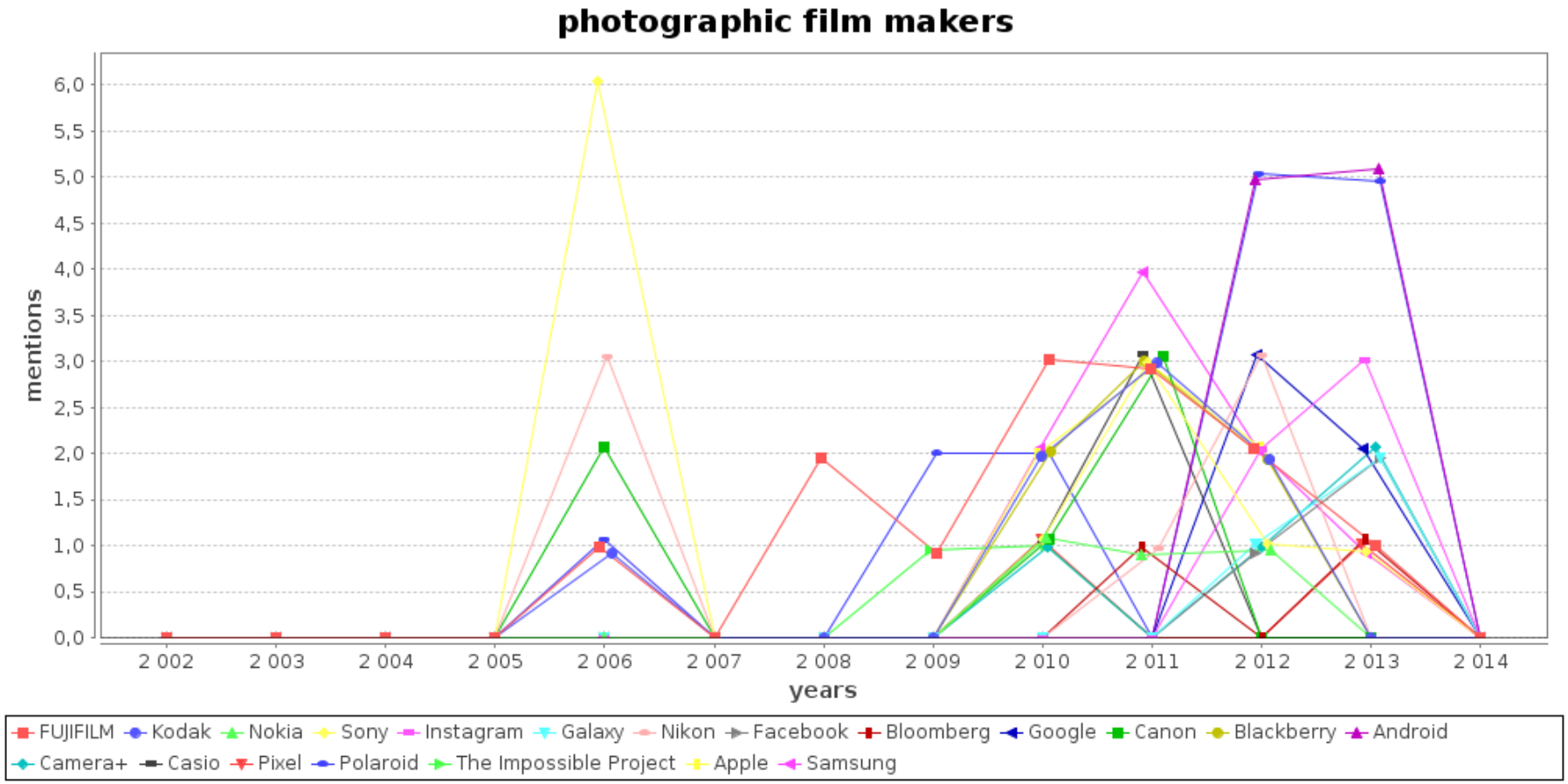
Извлечение технологических областей [3]



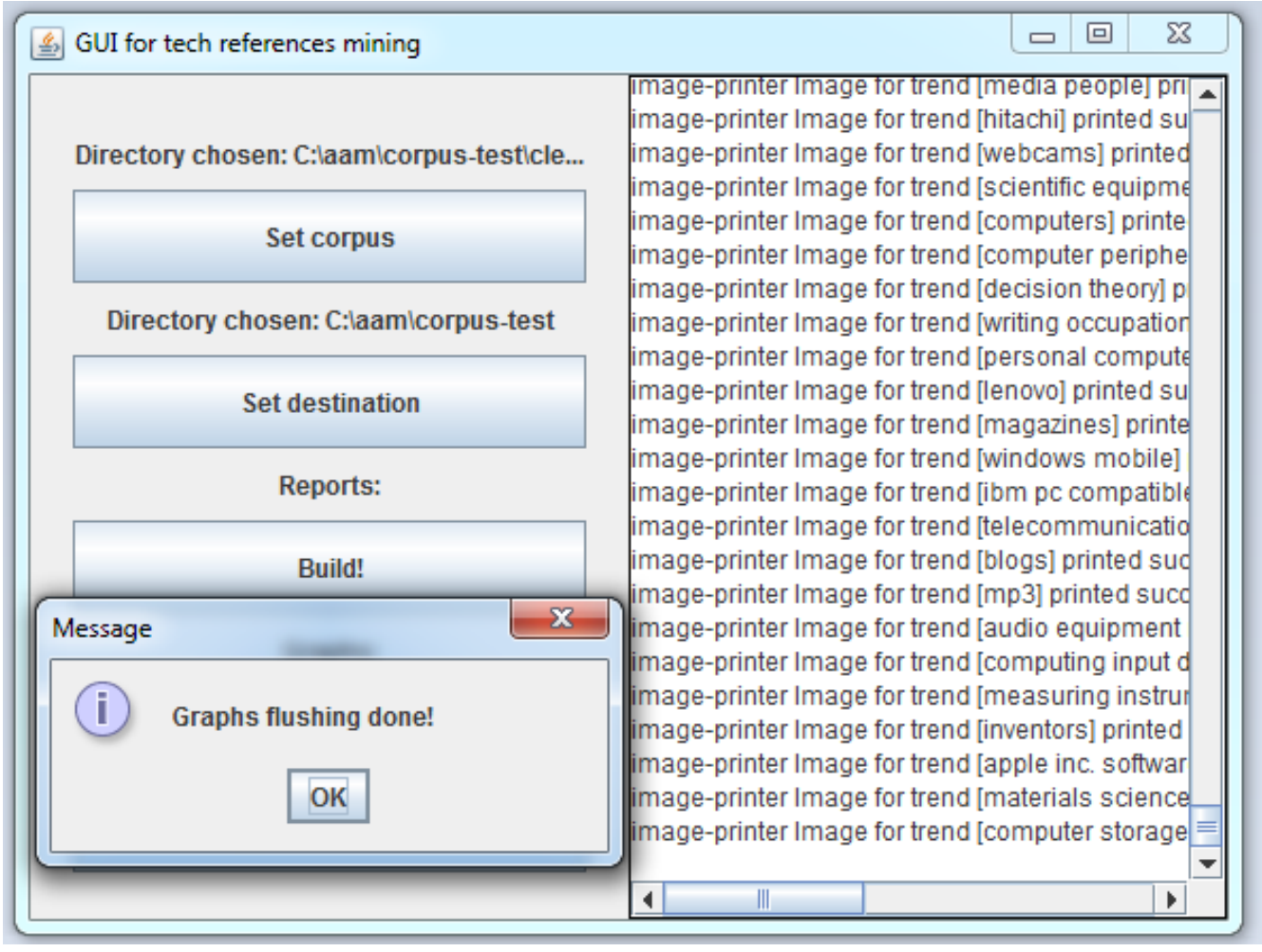
Примеры графиков [1]



Примеры графиков [2]



Пример графического интерфейса



Использованный инструментарий

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

Формальные признаки

- Доклад на всероссийской конференции «СПИСОК-2014»
- Более 5000 строк программного кода на Scala, Java и Python и XML-разметки

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.