Проект создания газетного Интернет-корпуса

Михаил Дубов, Никита Левицкий, Егор Моренко,
Екатерина Черняк, Артем Шаль, Андрей Шестаков

Научно-учебная группа «Методы анализа и визуализации текстов»
Национальный исследовательский университет «Высшая школа экономики»

http://ami.hse.ru/vitext

Почему газетные публикации?

- ► Тексты собираются не со всего веба;
- ► Фиксированное число источников:
  - ⇒ Качественное решение задачи удаления html-разметки и рекламы;
  - ⇒ Известная структура публикаций в каждом источнике.
- ► Высокое качество собранных данных:
  - ■ Язык – достаточно формален, опечаток – мало;
  - ■ Содержательно тексты – на высоком уровне;
  - ■ Достаточно узкий набор тем;
  - ■ Многие тексты категоризированы.

Источники текстов

Российская газета
Известия
Независимая
Коммерсантъ
МК Московский комсомолец

Реализация

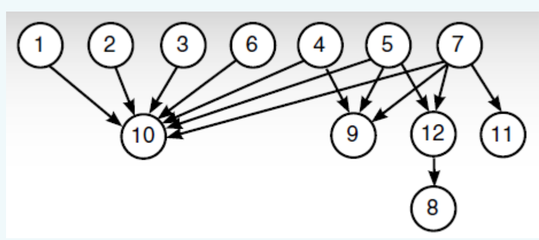
Основные принципы развития ПО для сбора корпуса

- ► Полная автоматизация процесса пополнения корпуса новыми текстами и их разметки;
- ► Простота расширяемости ПО парсерами для новых источников текстов;
- ► Возможность добавления в систему новых модулей по обработке текстов.

Анализ собранных текстов

- ► Построение таксономий и онтологий на основе данного корпуса текстов:

- ► Анализ связей между ключевыми словосочетаниями на данном корпусе текстов[1]:



1. Ввод автоматизированного производства.
2. Выпуск пресс-релизов.
...
4. Изменение уровня концентрации собственности.
5. Повышение квалификации персонала.
...
9. Изменение организационно-правовой формы.
10. Повышение эффективности управления затратами.
...

[1] *Миркин Б. Г., Черняк Е. Л., Чугунова О. Н. Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы. Бизнес-информатика, №3(21), 2012. с. 31-41.*

В данной работе использованы результаты, полученные в ходе выполнения проекта «Методы визуализации текстовой информации с помощью построения суффиксных деревьев, мультифасетных классификаций и иерархических онтологий: алгоритмическое и программное обеспечение», выполненного в рамках Программы «Научный фонд НИУ ВШЭ» в 2013 году под руководством д.т.н., профессора Б.Г. Миркина, грант № 13-05-0047.