Программная система терминологического анализа научной периодической литературы в области химии

Докладчик: **Альперин Борис Львович**

Научный руководитель: к.х.н., с.н.с. **Кузьмин Андрей Олегович**, Институт катализа им. Г.К. Борескова СО РАН

к.ф-м.н., н.с. **Саломатина Наталья Васильевна,** Институт математики им. С.Л. Соболева

НГУ

ИК СО РАН

Новосибирск, 2014

Описание работы

Цель – разработка методов и средств для анализа терминологической базы предметной области.

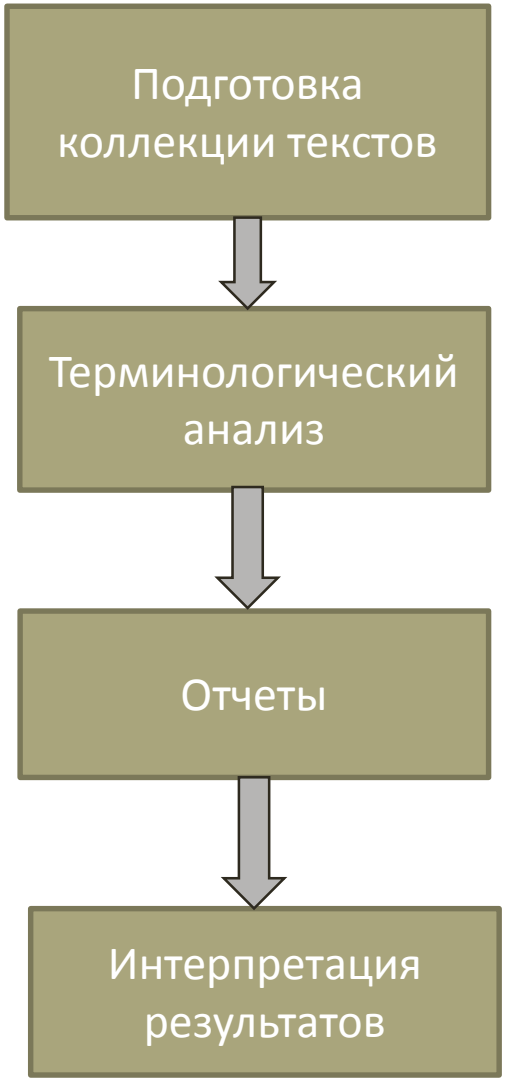
Сферы применения:

- Построение и пополнение быстроменяющегося тезауруса ПО с установлением связей между терминами;
- Поиск "горячих" направлений или тенденций развития ПО;
- Отслеживание смысловых связей между текстами путем сопоставления их терминологических спектров; таксономия текстов.

Предметная область – химический катализ.

Корпус текстов – тезисы конфереции «Европейский конгресс по катализу» за 2005, 2007, 2009, 2011, 2013 г. (ок. 6000 документов).

Схема работы

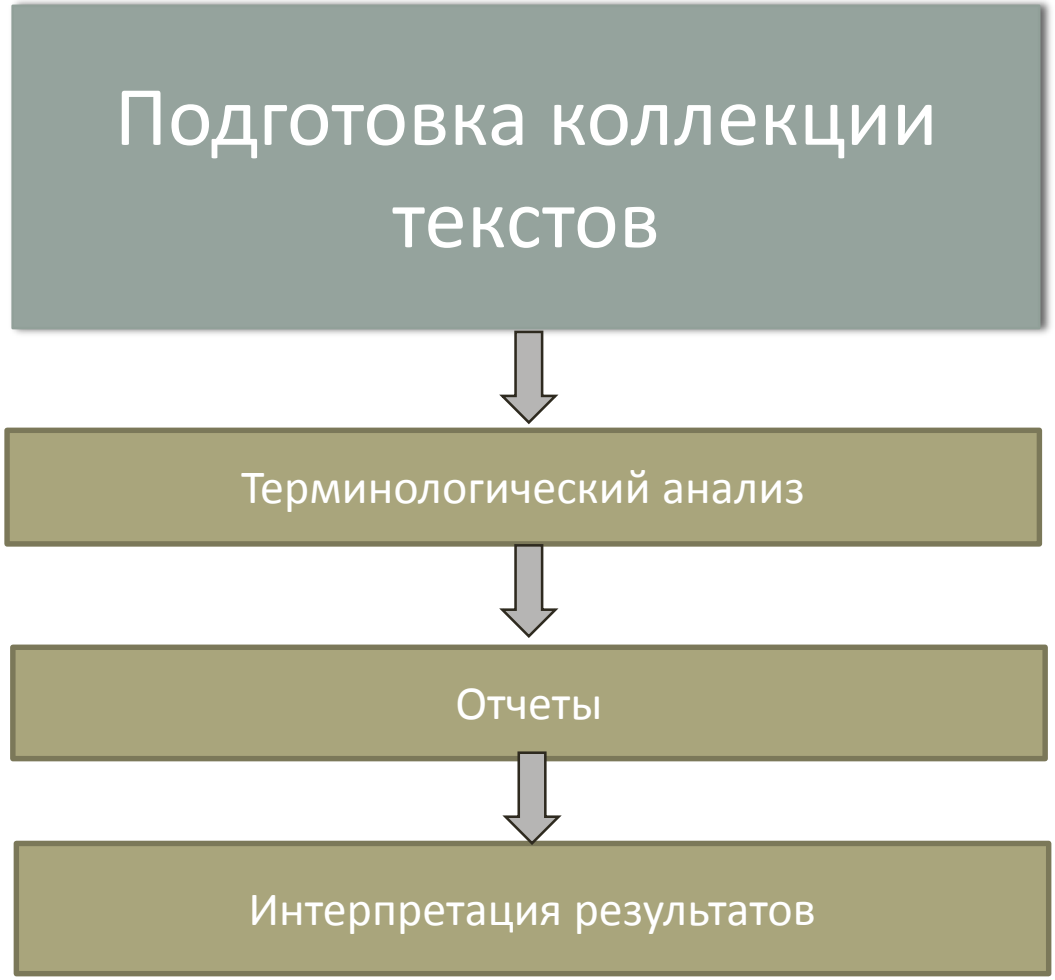


Перевод pdf-документов подборки в текстовое представление для дальнейшего анализа. Предварительная коррекция.

Обработка и фильтрация терминологического спектра

Отчеты о трендах, аномалиях, статистике распределения терминов

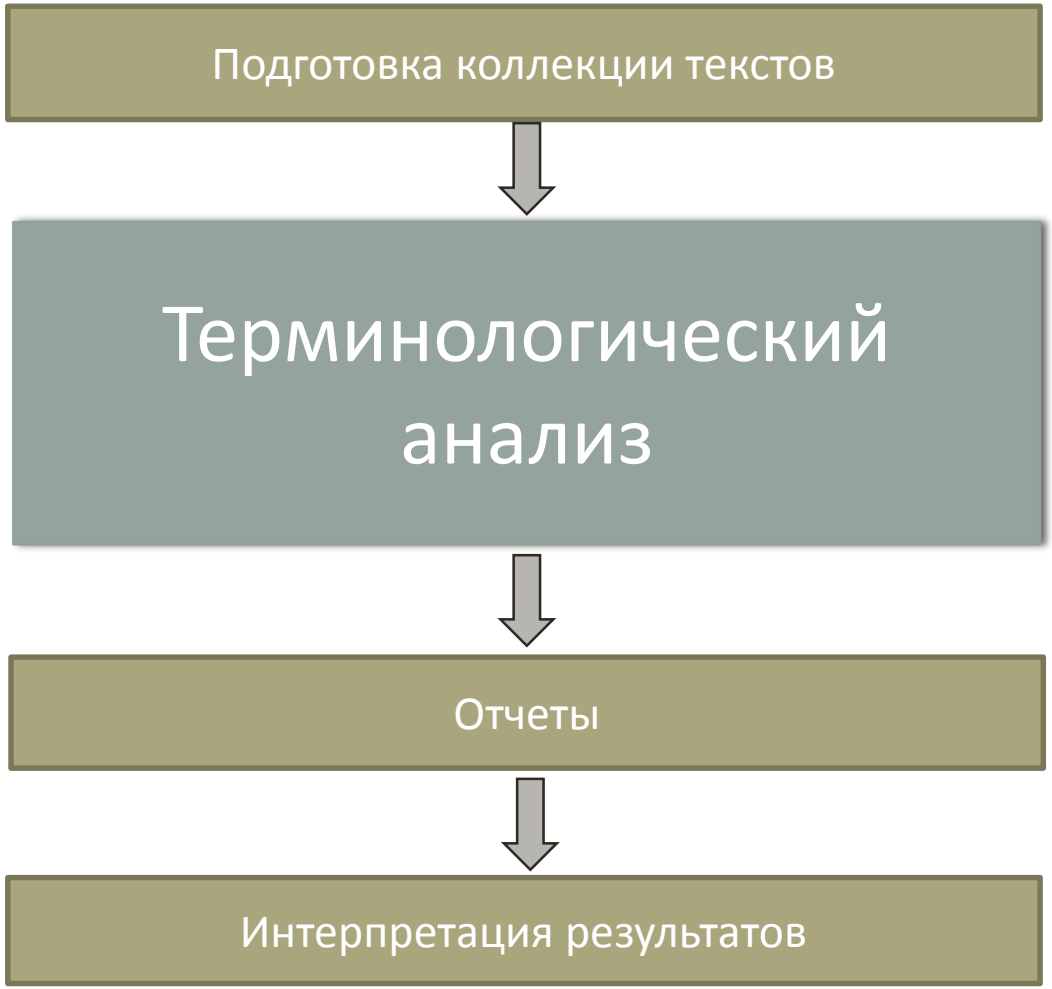
Поиск нужных терминов по различным критериям


Подготовка коллекции текстов
Терминологический анализ
Отчеты
Интерпретация результатов

Подготовка коллекции текстов

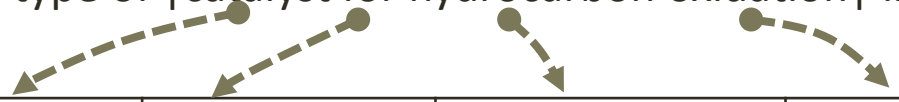
- Структура pdf-файла – набор блоков с указанием их размера и положения. Каждая буква – отдельный блок.
- Необходимо выделить логическую модель документа – название, авторы, организации, основная часть, ссылки и др.

- Текст разбивается на блоки с одинаковыми стилистическими параметрами
- Классификация блоков (название, авторы, организации, ...)
- Построение итогового документа и его сохранение в БД.


Подготовка коллекции текстов
Терминологический анализ
Отчеты
Интерпретация результатов

Модель терминологического анализа – общая схема

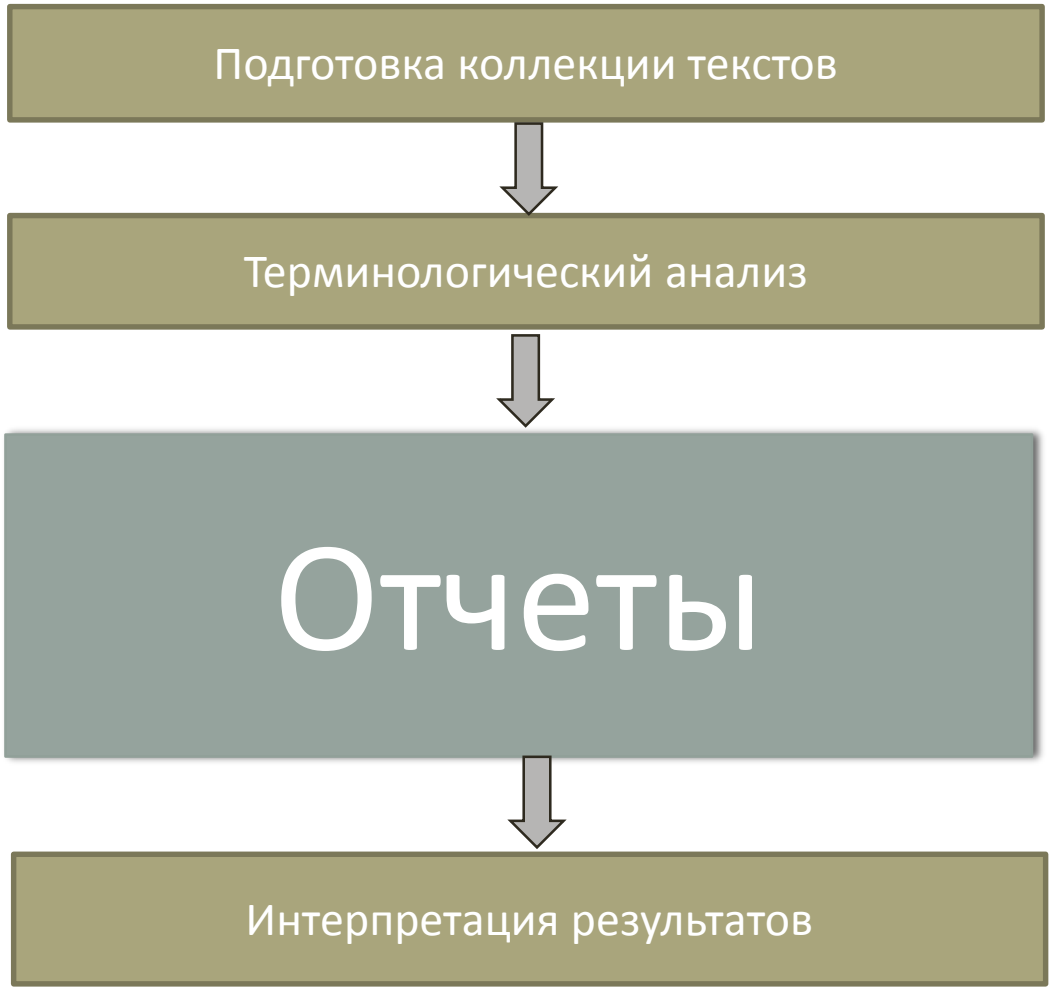
… development of new type of |catalyst for hydrocarbon oxidation| …



Слово	catalyst	for	hydrocarbon	oxidation
Часть речи	сущ-е	предлог	сущ-е	сущ-е

Этапы обработки:

1. разметка текста;
2. выделение терминоподобных L-грамм;
3. постобработка.


Подготовка коллекции текстов
Терминологический анализ
Отчеты
Интерпретация результатов

Отчеты

- Просмотр текстов
- Просмотр терминологической разметки
- Характеристики L-грамм:
  - Общие характеристики
  - Частотное распределение по годам
  - Тренды – L-граммы с монотонно возрастающими частотами
  - Общенаучные термины

Пользовательский интерфейс

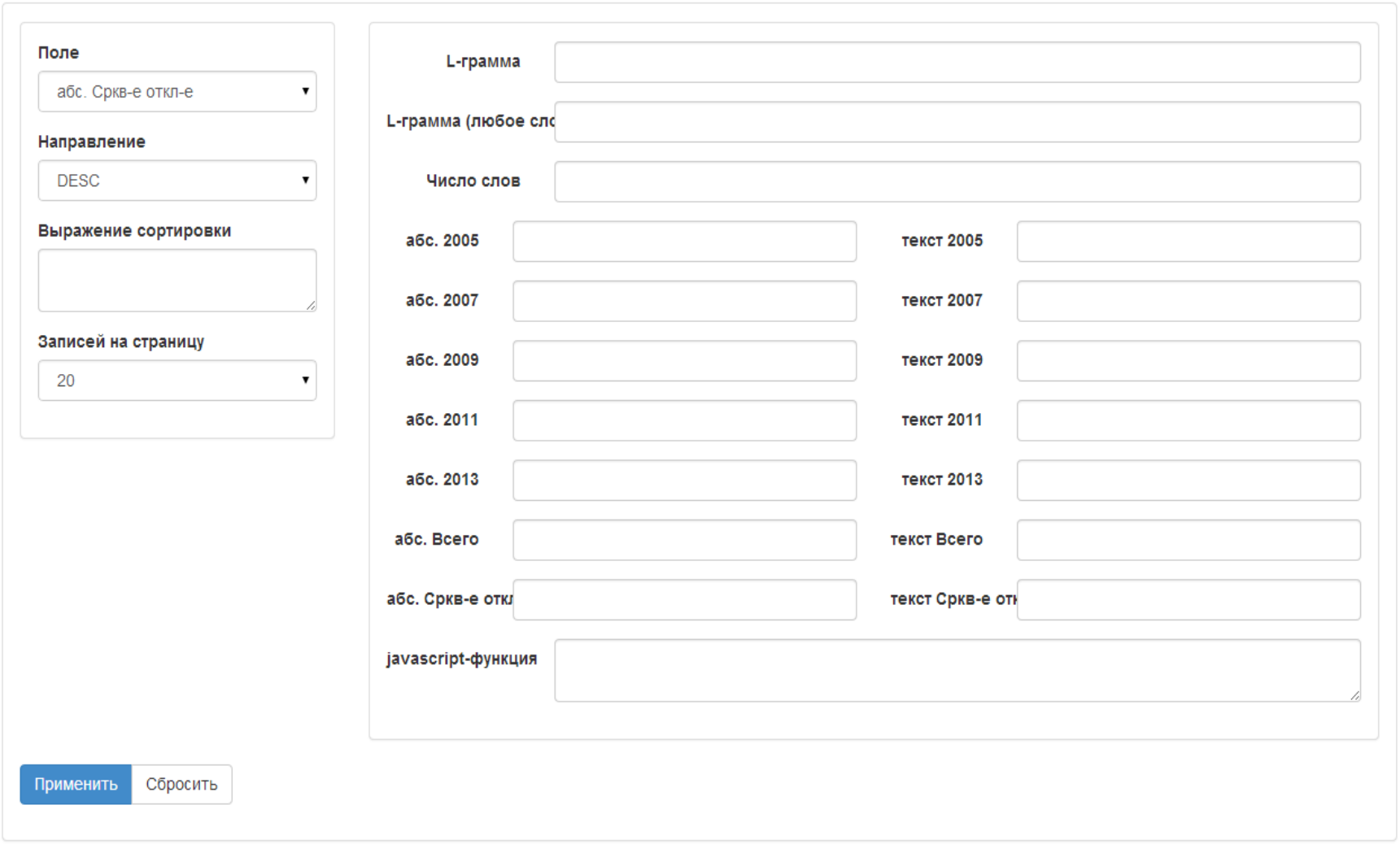
Linshun Xu, et al. – Conversion Between Different Types of Carbon Species and Their Influences on the Surface Reactivity of Co(0001) (EuropaCat X - Glasgow, Scotland 28 August - 2 Sept 2011)

Introduction. The accumulation of carbonaceous deposits on transitional metal catalysts is inevitable in the catalytic conversion of hydrocarbon and is recognized to considerably affect the catalytic activity and selectivity, but the exact nature and role of active carbonaceous species remain ambiguous [1]. The growth of graphene on transitional metal surfaces also involves the conversion between different carbon species [2]. The underlying microscopic mechanisms of carbon-induced changes in reactivity of metal catalysts have recently been addressed using the surface science approach for the selective hydrogenation of alkynes [3] and the isomerization and hydrogenation of alkenes [4]. In this presentation, we will present our recent results on the effects on different carbon species on the surface reactivity of Co(0001). We observed the temperature-controlled transformation between carbide-like carbon cluster and graphite species on Co(0001) formed by the ethylene decomposition. We also prepared an ordered Co(0001)- (2□2)-carbide surface by the thermal decomposition of C1 fragments.

Блоки терминоподобных L-грамм

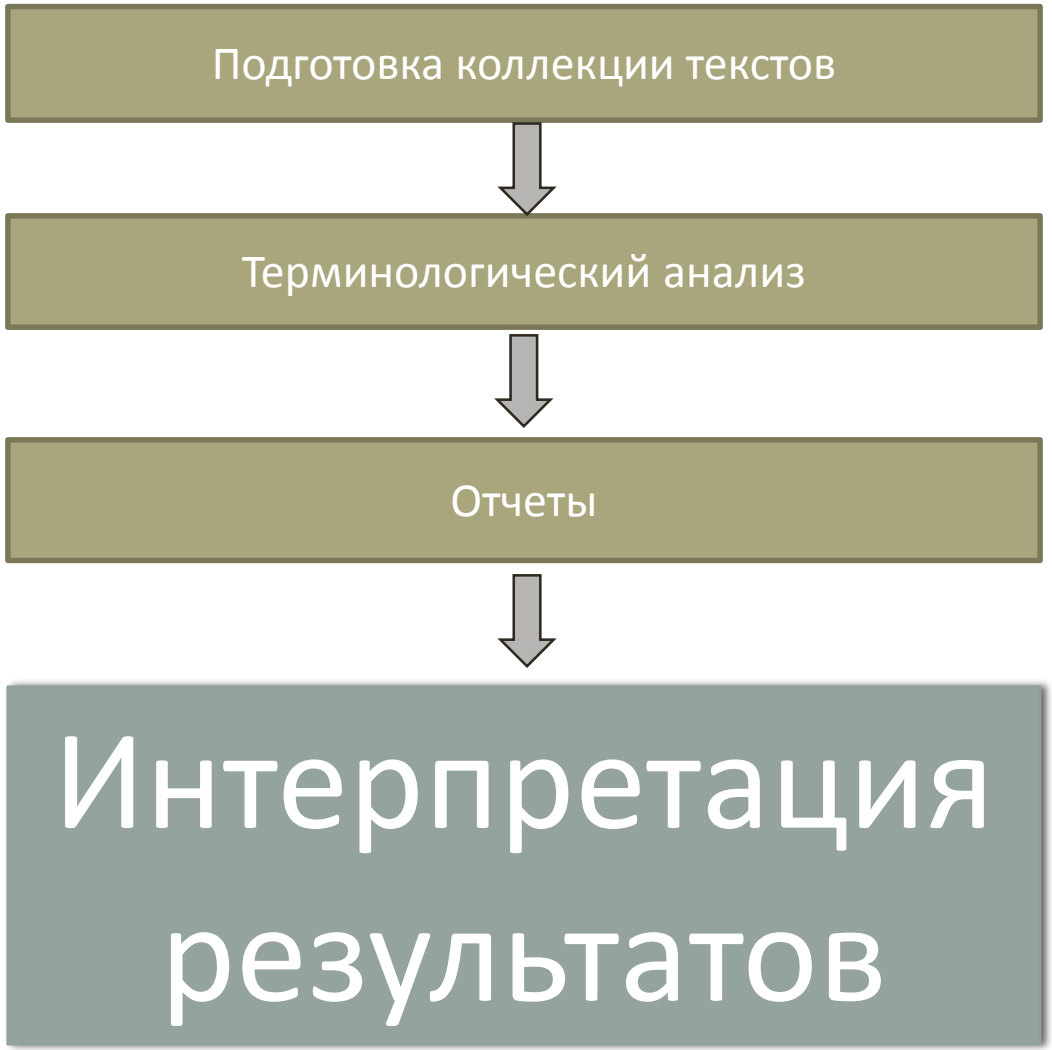
Термины из словаря UIPAC GoldBook

Пользовательский интерфейс



Пользовательский интерфейс – фрагмент положительного тренда

L-грамма	2005	2007	2009	2011	2013	Всего	откл-е
FIXED BED REACTOR	31	58	66	73	96	324	19.282
REFORMING OF METHANE	28	47	59	64	90	288	18.624
OXIDATION OF BENZYL	0	4	5	5	29	43	9.462
COUPLING OF METHANE	1	5	5	9	29	49	9.0627
CONVENTIONAL IMPREGNATION METHOD	1	3	5	9	11	29	3.386
AB INITIO SIMULATION	0	7	7	7	13	34	3.759
HYDROLYSIS OF CELLULOSE	0	0	3	9	21	33	7.224


Подготовка коллекции текстов
Терминологический анализ
Отчеты
Интерпретация результатов

Интерпретация результатов

Coupling of methane – конденсация метана

- Процесс открыт в 1980-х
- Из-за низкого выхода не удалось коммерциализировать

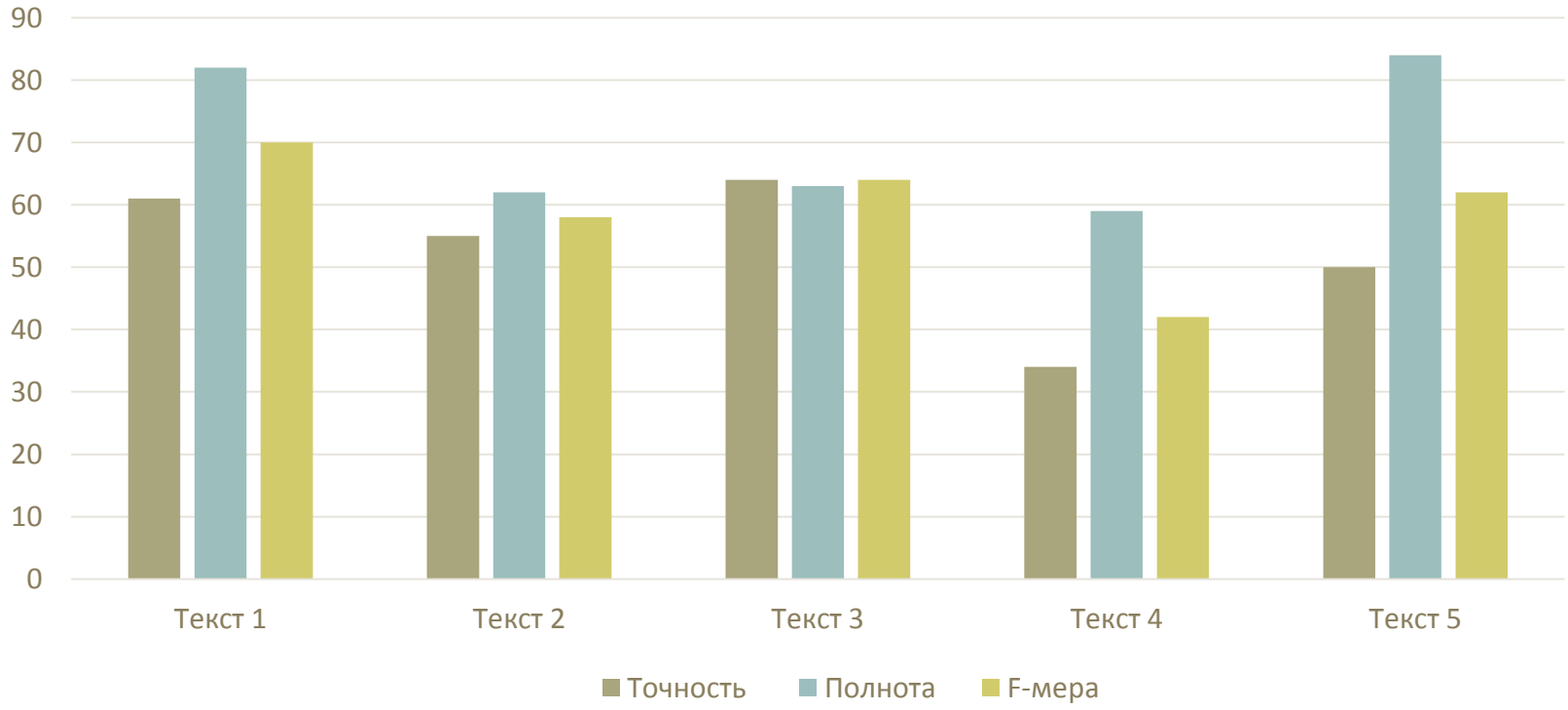
Упоминания:

Применение новых катализаторов	16
Пример применения нового катализатора	2
Улучшение процесса	2
Альтернативный метод	1
Косвенное упоминание	6
Исследование механизмов реакции	1

Новые проекты:

OCMOL, Oxidative Coupling of Methane followed by Oligomerization to Liquids (впервые встречается в 2011 г.)

Оценка эффективности



Средняя точность – 53%
Средняя полнота – 70%
Средняя F-мера – 59%

Полученные результаты

- Разработан подход к извлечению структурной информации из pdf-документов публикаций
- Разработана модель терминологического анализа с учетом специфики предметной области
- Проведены эксперименты по уточнению элементов модели
- Выполнена программная реализация
- Разработан пользовательский интерфейс для просмотра результатов анализа

Спасибо за внимание