

***Концептуальный анализ
(на основе методов
семантического анализа текстов)***

Задачи исследований по теме

- ▣ 1. Разработать новые методы, алгоритмы и технологии решения задачи создания декларативных средств для автоматической кластеризации текстовых документов СМИ.
- ▣ 2. Исследовать и разработать методы и алгоритмы выделения из текстов сущностей (значимых понятий) для задачи кластеризации.
- ▣ 3. Исследовать и разработать алгоритмы формирования частотных словарей слов и словосочетаний и представить их в табличном виде.
- ▣ 4. Исследовать и разработать технологии и процедуры назначения элементам формализованного представления документа весовых коэффициентов их смысловой значимости.
- ▣ 5. Выполнить анализ полученных результатов при различных исходных данных.
- ▣ 6. Разработать общую технологическую схему процесса создания декларативных средств для автоматической кластеризации текстовых документов СМИ.

Теоретическая концепция фразеологического концептуального анализа текстов

Основной идеей этой концепции является обоснование использования в качестве основных единиц смысла устойчивых фразеологических и терминологических словосочетаний, обозначающих понятия и отношения между понятиями, представленные в предметной области.

Иерархия единицы смысла:

- *Наименование понятия (сущность) – выражено словом или словосочетанием*
- *Предложение – его смысловой структурой является предикатно-актантная структура*
- *Сверхфразовое единство – фрагмент текста, объединенный общей темой*

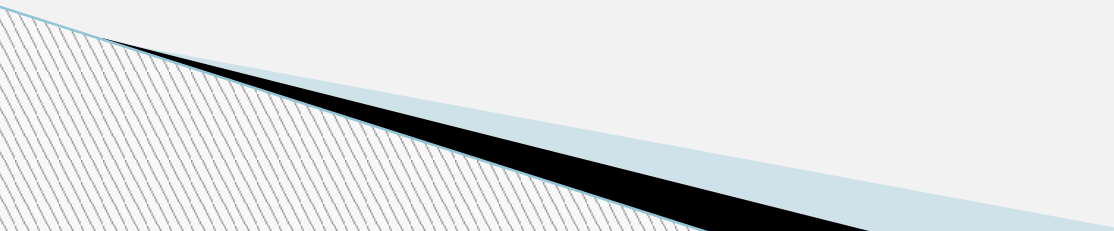
Смысловое представление содержания текста - концептуальный образ документа (КОД) - совокупность взаимосвязанных наименований понятий текста, расположенных в нем строго определенном порядке)

Семантическая карта документа – концептуальный граф, в котором вершины – нормализованные наименования понятий, дуги – унифицированные смысловые отношения между понятиями

Основные технологии автоматической обработки неструктурированной текстовой информации

- 1. Технологии создание декларативных средств по тематическому корпусу текстов.*
- 3. Технологии кластеризации, классификации и рубрицирования текстов на основе анализа из содержания.*
- 4. Автоматическое реферирование текстов.*
- 5. Автоматическое установление смысловой близости документов.*
- 6. Технологии извлечения фактографической информации.*
- 7. Технологии семантического поиска информации.*
- 8. Построение и динамический анализ семантической структуры текстов.*
- 9. Выделение ключевых тем и информационных объектов.*
- 10. Определение общей и объектной тональности сообщений.*
- 11. Исследование частотных характеристик текстов.*

Технология автоматической обработки текстов документов (на этапе их ввода в хранилище)

- ▣ *Формально-логический контроль текста*
 - ▣ *Морфологический анализ*
 - ▣ *Семантико-синтаксический анализ*
 - ▣ *Концептуальный анализ*
 - ▣ *Дистрибутивно-статистический анализ*
 - ▣ *Автоматическое смысловое структурирование документов на предложения и сверхфразовые единства (последовательность контекстно-связанных предложений)*
 - ▣ *Формирование различных форм формализованного представления текста*
 - ▣ *Автоматическая классификация текста*
 - ▣ *Формирование смысловой структуры текста в виде графа понятий и их отношений*
- 

Концепция установления смысловой близости фрагментов документов

- *В качестве базовой теоретической концепции использовалась концепция фразеологического концептуального анализа констатирующая, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав.*
- *Наиболее устойчивой единицей смысла является понятие, определяемое как социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания.*
- *Понятия занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней.*
- *Смысл понятия проявляется в полной мере только через всю систему его отношений со всеми другими понятиями языка.*
- *Второй по значимости единицей смысла является предложение. Из предложений формируются различного рода сверхфразовые единства, которые представляются в виде последовательностей связного текста.*

Концепция установления смысловой близости фрагментов документов (продолжение)

- **В связном тексте** предложения выступают в тесной смысловой связи. В основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда порождает текст.
- **Локальная связность** обеспечивает раскрытие смысла понятия на основе его контекста. Под смысловой связанностью текста или его фрагмента будем понимать совокупность наименований понятий, расположенных в тексте в определённом порядке и отражающих основное смысловое содержание текста или его фрагмента.
- **Локальная смысловая схожесть наименований понятий** текста определяется как сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах.
- **Глобальная смысловая схожесть текстов** или их фрагментов определяется как сходство состава идентичных наименований понятий и порядка их следования в текстах или их фрагментах. Каждое понятие этого фрагмента также должно удовлетворять условию локальной смысловой схожести.

Принципы установления смысловой близости фрагментов документов

- *Преобразование текстового представления в его **формализованное смысловое представление** дает возможность сопоставления текстов по их смысловому содержанию.*
- *Такое сопоставление смыслового содержания текстов, обеспечивающее **выявление идентичных по смыслу фрагментов текстов** должно удовлетворять следующим условиям:*
 1. *В двух текстах должна быть **пересекающаяся совокупность наименований понятий**. Число понятий этой совокупности должно быть равно или превышать число наименований понятий, входящих в состав единичного высказывания.*
 2. *В двух таких текстах должны быть фрагменты, в которых **концентрация пересекающихся наименований понятий** превышает пороговое значение. Эти фрагменты должны иметь соизмеримые размеры.*
 3. *Эти фрагменты текстов должны быть **сходными по составу наименований понятий и порядку их следования**.*

Методы выделения наименований понятий

- ▣ *Выделение наименований понятий выполняется на этапе концептуального анализа текстов*
- ▣ *Концептуальный анализ текстов - это лингвистическая процедура, обеспечивающая выявления их понятийного (концептуального) состава, формализации наименований понятий и установления смысловых связей между ними*
- ▣ *Методы концептуального анализа текстов*
 - *Концептуальный анализ с контролем по эталонному концептуальному словарю (ЭКС) объемом 1.5 млн. наименований понятий*
 - *Концептуальный анализ текстов на основе “логической шкалы” словаря ЭКС*
 - *Концептуальный анализ текстов на основе синтаксических структур словаря ЭКС*
 - *Концептуальный анализ на основе обобщенных синтагм*
 - *Гибридный метод выявления наименований понятий из текстов СМИ*

Фрагмент частотного словаря синтаксических структур словосочетаний в словаре ЭКС

Частота	Процент.	Синтагма	Словосочетание-представитель синтагмы
00519917	34.4	AN	Информационные технологии
00280589	18.6	NN	Поиск информации
00140454	9.3	NAN	Система информационного поиска
00110726	7.3	AAN	Автоматизированная информационная система
00103044	6.8	ANN	Автоматизированная обработка текстов
00086080	5.7	NNN	Вестник Академии предпринимательства
00017339	1.1	ANNN	Американская ассоциация инженеров транспорта

Алгоритм №1 концептуального анализа текстов с контролем по словарю ЭКС

- ▣ **1.Идея алгоритма:** *если некоторому отрезку текста соответствует в эталонном словаре хотя бы одно наименование понятия, имеющее такую же длину и такую же синтаксическую структуру, то этот отрезок текста с большой вероятностью также является наименованием понятия.*
- ▣ **2.Условие:** *текст должен быть разделен на синтаксические предложения и каждое предложение должно быть разделено на всевозможные фрагменты последовательностей контактно расположенных слов*

Алгоритм №1 концептуального анализа текстов с контролем по словарю ЭКС

- ▣ Шаг 1. Членение входного текста на предложения;
- ▣ Шаг 2. Морфологический анализ текста;
- ▣ Шаг 3. Пословная нормализация текста;
- ▣ Шаг 4. Членение предложений текста на отдельные слова и отрезки текста длиной от 1-х до 5-ти слов;
- ▣ Шаг 5. Формирование поисковых образов слов и словосочетаний;
- ▣ Шаг 6. Поиск в словаре ЭКС нормализованных текстовых фрагментов.
- ▣ Шаг 7. Исключение из результатов поиска слов и словосочетаний, которые на одних и тех же отрезках текста входят в состав других, более длинных словосочетаний.
- ▣ Шаг 8. Преобразование полученных результатов в структуру метаданных.

Алгоритм №2 концептуального анализа текстов на основе “логической шкалы” словаря ЭКС

- ▣ **1.Идея алгоритма:** *если известна информация о длине словосочетания и о всех словах, входящих в состав этих словосочетаний, а также о месте каждого слова в словосочетании (первое, последнее или промежуточное место), то при наложении представления этой информации на аналогичной представлении структуры реальных текстов, то можно с высокой степенью вероятности выделить в текстах словосочетания, аналогичные по своей структуре и лексическому составу словосочетаниями, содержащемуся в эталонном словаре, по которому было выполнено это информационное представление структуры и лексического состава словосочетания.*
- ▣ **2.Условие:** *текст должен быть разделен на синтаксические предложения и каждое предложение должно быть разделено на всевозможные фрагменты последовательностей контактно расположенных слов*

Алгоритм №2 концептуального анализа текстов на основе “логической шкалы” словаря ЭКС

- ▣ Шаг 1. Членение входного текста на предложения;
- ▣ Шаг 2. Морфологический анализ текста;
- ▣ Шаг 3. Пословная нормализация текста;
- ▣ Шаг 4. Членение предложений текста на отдельные слова и отрезки текста длиной от 2-х до 5-ти слов;
- ▣ Шаг 5. Формирование поисковых образов слов и словосочетаний, выделенных в п.4;
- ▣ Шаг 6. Формирование по словосочетаниям соответствующих им обобщенных синтагм;
- ▣ Шаг 7. Поиск обобщенных синтагм, построенных по фрагментам текста, в эталонном словаре обобщенных синтагм;
- ▣ Шаг 8. Исключение из массива п. 4 отрезков текста, обобщенные синтагмы которых не находятся в эталонном словаре обобщенных синтагм;
- ▣ Шаг 9. Исключение из массива, сформированного в п.8, таких слов и словосочетаний, которые на одних и тех же отрезках текста входят в состав других, более длинных словосочетаний (факультативный пункт).

Алгоритм №3 концептуального анализа текстов на основе синтаксических структур словаря ЭКС

- ▣ **1.Идея алгоритма:** если некоторому отрезку текста соответствует в эталонном словаре хотя бы одно наименование понятия, имеющее такую же длину и такую же синтаксическую структуру, то этот отрезок текста с большой вероятностью также является наименованием понятия. **2.Условие:** текст должен быть разделен на синтаксические предложения и каждое предложение должно быть разделено на всевозможные фрагменты последовательностей контактно расположенных слов

Алгоритм №3 концептуального анализа текстов на основе синтаксических структур словаря ЭКС

- ▢ Шаг 1. Членение входного текста на предложения;
- ▢ Шаг 2. Морфологический анализ текста;
- ▢ Шаг 3. Пословная нормализация текста;
- ▢ Шаг 4. Членение предложений текста на всевозможные фрагменты текста длиной от 2-х до 5-ти слов;
- ▢ Шаг 5. Формирование поисковых образов фрагментов, выделенных в п.4;
- ▢ Шаг 6. Формирование по фрагментам соответствующих им синтаксических структур;
- ▢ Шаг 7. Поиск синтаксических структур, построенных по фрагментам текста, в эталонном словаре обобщенных синтагм;
- ▢ Шаг 8. Исключение из массива п. 4 отрезков текста, синтаксических структур которых не находятся в эталонном словаре обобщенных синтагм;
- ▢ Шаг 9. Исключение из массива, сформированного в п.8, таких слов и словосочетаний, которые на одних и тех же отрезках текста входят в состав других, более длинных словосочетаний (факультативный пункт).
- ▢ Шаг 10. Преобразование полученных результатов в структуру

Модернизированный алгоритм №3 концептуального анализа текстов на основе синтаксических структур словаря ЭКС

- ▣ **1.Идея алгоритма:** *если сформированной последовательности обобщенных символов грамматических классов слов некоторого отрезка текста соответствует какой-либо элемент словаря обобщенных синтагм (словаря ОС), и этот отрезок текста не совпадает ни с одним из элементов словаря малоинформативных словосочетаний (словаря МС) и при этом все его слова совпадают со словами словаря значимых слов (словарь ЗС), то этот отрезок текста с большой вероятностью является наименованием понятия.*
- ▣ **2.Условие:** *текст должен быть разделен на синтаксические предложения и каждое предложение должно быть разделено на всевозможные фрагменты последовательностей контактно расположенных слов*

Модернизированный алгоритм №3 концептуального анализа текстов на основе синтаксических структур словаря ЭКС

- Шаг 7. Поиск синтаксических структур, построенных по фрагментам текста, в эталонном словаре обобщенных синтагм;*
- Шаг 8. Исключение из массива п. 4 отрезков текста, синтаксических структур которых не находятся в эталонном словаре обобщенных синтагм;*
- Шаг 9. Исключение из массива п. 4 отрезков текста, синтаксических структур которых совпадает с одним из элементов словаря малоинформативных словосочетаний (словаря МС);*
- Шаг 10. Исключение из массива п. 4 отрезков текста, опорные и зависимые слова которых не совпадают со словами словаря значимых слов ;*
- Шаг 11. Исключение из массива, сформированного в п.8, таких слов и словосочетаний, которые на одних и тех же отрезках текста входят в состав других, более длинных словосочетаний (факультативный пункт).*
- Шаг 12. Преобразование полученных результатов в структуру метаданных.*

Алгоритм №4 концептуального анализа текстов на основе обобщенных синтагм словаря ЭКС

- ▣ **1.Идея алгоритма:** *если фрагменту сформированной последовательности обобщенных синтагм предложения соответствует какой-либо элемент словаря обобщенных синтагм, представляющий собой последовательность синтагм, отражающих конкретные формы слов и наборы их грамматических признаков, то можно с большой вероятностью утверждать, что в составе этого предложения имеется отрезок текста, представляющий собой наименование понятия в контекстном окружении.* **2.Условие:** *текст должен быть разделен на синтаксические предложения и для каждого предложения должно быть построено его представление в виде последовательности обобщенных синтагм.*

Алгоритм №4 концептуального анализа текстов на основе обобщенных синтагм

- ▢ Шаг 1. Членение входного текста на предложения;
- ▢ Шаг 2. Морфологический анализ текста;
- ▢ Шаг 3. Построение синтаксического представления предложения в виде последовательности обобщенных синтагм с казанием позиций в предложении;
- ▢ Шаг 4. Членение последовательности обобщенных синтагм на всевозможные фрагменты текста длиной от 3-х до 16-ти элементов;
- ▢ Шаг 5. Поиск синтагм фрагментов текста в эталонном словаре обобщенных синтагм;
- ▢ Шаг 6. Исключение из массива, сформированного в п.8, таких синтагм, которые на одних и тех же отрезках входят в состав других, более длинных словосочетаний .
- ▢ Шаг 7. Получение текстовых словосочетаний, на основе информации о позициях найденных синтагм в предложении.
- ▢ Шаг 8. Преобразование полученных результатов в структуру метаданных.

Результаты работы алгоритм №4 (выделения словосочетаний основе обобщенных синтагм)

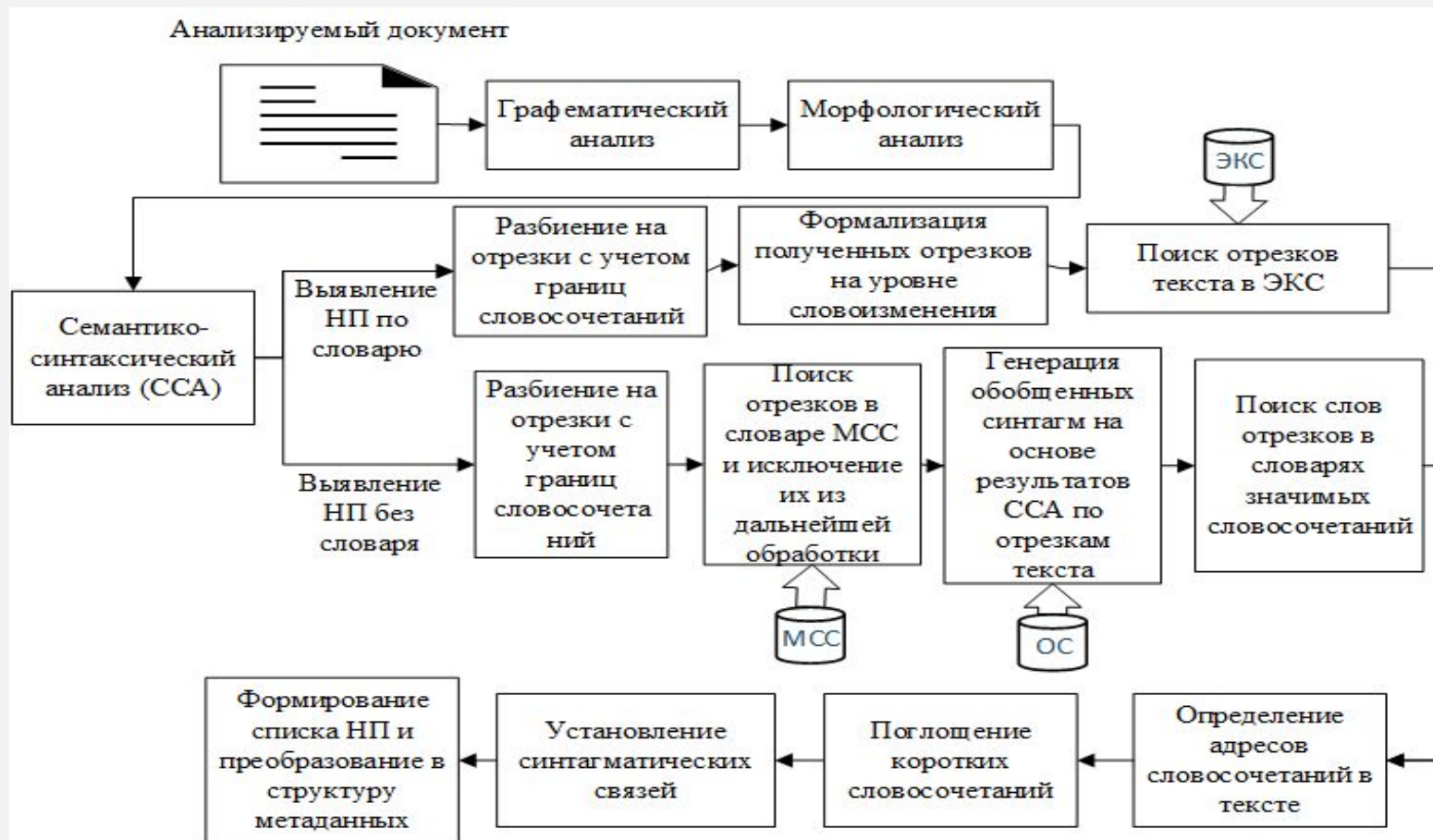
1. **Обобщенная синтагма именного словосочетания = ФSAA**
Частота встречаемости в корпусе текстов = 429
Текстовые словосочетания, соответствующие обобщенной синтагме;

Анонимный блогер, 30-градусный мороз, 569-страничный документ, Апелляционный суд, Атомный ледокол, Безмешковый пылесос, Винный мир, Властный вакуум, Военный бюджет, Военный вариант, Гендерный дисбаланс, Глобальный фонд, Горный массив, Западный вектор, Западный проект, Зеленый свет...

- 2 **Обобщенная синтагма именного словосочетания = ФгЙГ**
Частота встречаемости в корпусе текстов = 348
Текстовые словосочетания, соответствующие обобщенной синтагме;

Изначальное пожертвование, Классовое сознание, Кумулятивное воздействие, Машинное обучение, Новое мышление, Персонафицированное обучение, Полное молчание, Программное обеспечение, Финальное движение, Эффективное управление, Ядерное оружие, авторитарное правление, агрессивное выступление, адаптивное поведение, акционерное соглашение, алкогольное отравление, атомное оружие, безобидное голосование, бесперебойное обеспечение, беспорядочное вращение.

Гибридный алгоритм №5 выявления наименований понятий в текстах документов



Исходные статистические данные по массиву сообщений СМИ

- ▣ Кол. Документов в массиве = 3 004 документов
- ▣ Всего слов в массиве документов= 523 810 слов
- ▣ Разных слов (на уровне словоизменения) = 88 925
- ▣ Среднее число слов в документе = 174.4 слов/док
- ▣ Среднее число разных слов в документе = 29.5 слов/док

- ▣ Всего словосочетаний в массиве (по словарю ЭКС)= 1 106 355 словосоч.
- ▣ Разных словосочетаний (на уровне словоизменения слов) = 67 571 словосоч.
- ▣ Кол. разных главных слов (на уровне словоизменения слов) = 5 577слов
- ▣ Среднее число словосочетаний в документе = 368.3 словосоч./док
- ▣ Среднее число разных словосочетаний в документе = 22.5 словосоч./док

Сравнительные характеристики объемов частотных словарей, полученных по корпусу текстов сообщений СМИ различными методами концептуального анализа (КА)

Накоп- ленные частоты	Объемы частотных словарей, полученные различными методами КА			
	КА №1	КА №2	КА №3	КА №4
00010	10 541	10 994	13 977	5 709
00009	11 568	11 986	15 574	6 677
00008	12 786	13 215	17 588	7 897
00007	14 356	14 727	20 188	9 490
00006	16 380	16 921	23 026	17 827
00005	19 143	19 823	29 198	21 316
00004	23 205	24 909	38 047	26 459
00003	29 340	32 752	54 850	33 585
00002	39 864	58 838	102 599	68 140
00001	67 571	279 716	664 053	414 611

Назначение элементам формализованного представления документа весовых коэффициентов их смысловой значимости

- Для реализации статистической меры TF-IDF (TF — term frequency, IDF — inverse document frequency) и ряда подобных мер и поисковых функций необходимо для каждого понятия множества определить следующие параметры:
 - 1. *Общее число документов массива*
 - 2. *Общее число понятий в массиве документов*
 - 3. *Частоту встречаемости понятия в массиве документов*
 - 4. *Частоту встречаемости понятия в конкретном документе*
 - 5. *Число документов, в которых встречается понятие*
 - 6. *Длина (в понятиях) документа*
 - 7. *Средняя длина документа в массиве*

Фрагмент частотного словаря наименований понятий, полученный по корпусу текстов сообщений СМИ текстов

- ▢ 00000084 глава государство * Глава государства
- ▢ 00000084 чемпионат мир * Чемпионат Мира
- ▢ 00000065 русский язык * Русский язык
- ▢ 00000061 казахский язык * Казахский язык
- ▢ 00000059 круглый стол * КРУГЛЫМ СТОЛОМ
- ▢ 00000056 государственный орган власть * Государственные органы власти
- ▢ 00000052 сборный казахстан * Сборная Казахстана
- ▢ 00000052 социальный сеть * Социальные сети
- ▢ 00000051 президент страна * Президент страны
- ▢ 00000049 сельский хозяйство * Сельское хозяйство
- ▢ 00000048 южный корея * Южная Кореи
- ▢ 00000047 олимпийский чемпион * Олимпийский чемпион
- ▢ 00000046 казахстанский клуб * Казахстанские клубы
- ▢ 00000045 футбольный клуб * Футбольному клубу
- ▢ 00000044 министерство культура * Министерства культуры
- ▢ 00000043 местный власть * Местная власть

Пример формализованного смыслового представления содержания документа (представление документ - понятия)

- **Doc-934.t** = 0003 Олимпийские игры в Лондоне / 0002 двукратный олимпийский чемпион / 0003 Ильин / 0002 перепроверка допинг-пробы / 0003 Олимпиада-2008 / 0003 успешное выступление / 0003 юниор / 0003 спортивный клуб / 0003 Международная Федерация тяжелой атлетики / 0003 Майя / 0002 пресс-конференция / 0002 положительный результат / 0002 олимпийский чемпион / 0002 мировой рекорд / 0002 исполнительный директор / 0002 Пекин / 0001 чемпион мира / 0001 допинг / 0001 весовая категория / 0001 Федерация тяжелой атлетики / 0001 Международная Федерация / 0001 МОК / 0001 Лондон / 0001 Азиатские Игры

Пример представление «понятия - документ»

Кол.док.=00036 Понятие - «неправительственная организация»

*1342.t*1347.t*1353.t*1358.t *1367.t *1369.t *1374.t *1398.t *1442.t *1456.t *1499.t
 *1543.t *1547.t *1554.t *1562.t *1559.t *1561.t *1567.t *1573.t *1578.t *1634.t
 *1685.t *1734.t *1753.t *1834.t *1873.t *2313.t *2442.t *2461.t *2477.t *2667.t
 *2689.t *2846.t *2874.t *2972.t *. 2986.t

Кол.док.=00036 Понятие - «налоговые поступления»

*1347.t*1349.t*1354.t*1358.t *1368.t *1369.t *1371.t *1398.t *1441.t *1453.t *1495.t
 *1543.t *1546.t *1556.t *1563.t *1559.t *1566.t *1567.t *1574.t *1578.t *1634.t
 *1648.t *1684.t *1733.t *1831.t *1872.t *2311.t *2442.t *2461.t *2477.t *2667.t
 *2687.t *2846.t *2873.t *2972.t *. 2984.t

Кол.док.=00036 Понятие - «министр национальной экономики»

*1143.t*1147.t*1253.t *1358.t *1368.t *1369.t *1371.t *1398.t *1441.t *1453.t
 *1495.t *1543.t *1546.t *1556.t *1563.t *1559.t *1566.t *1567.t *1574.t *1578.t
 *1634.t *1648.t *1684.t *1733.t *1831.t *1872.t *2311.t *2442.t *2442.t *2461.t
 *2477.t *2667.t *2689.t *2846.t *2874.t *2972.t

Фрагмент частотного словаря главных слов словосочетаний, полученный по сокращенному частотному словарю сообщений СМИ текстов

- ▣ Объем сокращенного частотного словаря (полученного по словарю ЭКС)=23 205 Объем частного словаря разных главных слов словосочетаний = 2 514**

- ▣ Фрагмент частотного словаря главных слово словосочетаний**
- ▣ 00000107 система * Система
- ▣ 00000096 уровень * Уровень
- ▣ 00000086 страна* Стран
- ▣ 00000082 работа * работа
- ▣ 00000081 развитие * развитие
- ▣ 00000069 политика * политик
- ▣ 00000063 государство * Государство
- ▣ 00000062 проблема * проблем
- ▣ 00000058 организация * Организация
- ▣ 00000058 отношение* отношение
- ▣ 00000057 орган * Органы
- ▣ 00000057 рынок * Рынки

Фрагмент словаря парадигматических отношений наименований понятий типа «род - вид»

Родовое понятие = «спорт*4208.00»

Видовые понятия =

*спорт велосипедный - велосипедный спорт *4208.01*

*спорт высших достижений - спорт высших достижений *4208.02*

*спорт горнолыжный - горнолыжный спорт *4208.03*

*спорт детский - детский спорт *4208.04*

*спорт детско-юношеский - детско-юношеский спорт *4208.05*

*спорт казахский - казахский спорт *4208.06*

*спорт конный - конный спорт *4208.07*

*спорт конкобежный - конкобежный спорт *4208.08*

*спорт лыжный - лыжный спорт *4208.09*

*спорт массовый - массовый спорт *4208.10*

*спорт отечественный - отечественный спорт *4208.11*

*спорт профессиональный - профессиональный спорт *4208.12*

Фрагмент частотного словаря наименований понятий с информацией типа «род - вид»

- ▢ 00002 спортивная машина *1994.00 *1994.09
- ▢ 00002 спортивная команда *1624.00 *1624.07
- ▢ 00002 спортивная карьера *1519.00 *1519.08
- ▢ 00002 спортивная инфраструктура *1411.00 *1411.04
- ▢ 00002 спортивная журналистика *1100.00 *1100.03
- ▢ 00002 спортивная жизнь *1087.00 *1087.03
- ▢ 00002 спортивная деятельность *0911.00 *0911.02
- ▢ 00002 спортивная гимнастика *0719.00 *0719.02
- ▢ 00002 спорт *4208.00
- ▢ 00002 спорт высших достижений * 4208.00 * 4208.03
- ▢ 00002 спорный тезис *4427.00 *4427.07
- ▢ 00002 спорное утверждение *4732.00 *4732.05
- ▢ 00002 спорное решение *3791.00 *3791.05
- ▢ 00002 спорное предложение *3249.00 *3249.08
- ▢ 00002 спорное положение *3249.00 *3249.06

Фрагмент частотного словаря глагольных наименований понятий с информацией типа «род - вид»

- ▣ 00001 являлся *9878.00
- ▣ 00001 являлся аграрным донором *9878.00 *9878.01
- ▣ 00001 являлся акционером этой компании *9878.00 *9878.02
- ▣ 00001 являлась бы именно ликвидация элитарности столицы *9878.00 *9878.03
- ▣ 00001 являлся бы центром притяжения туристов *9878.00 *9878.04
- ▣ 00001 являлся в рукопашных поединках на коне *9878.00 *9878.05
- ▣ 00001 являлся важным фактором в контексте образовательной политики в целом *9878.00 *9878.06
- ▣ 00001 являлся веком казахского рыцарства *9878.00 *9878.07
- ▣ 00001 являлся Верховным комиссаром по делам беженцев *9878.00 *9878.08
- ▣ 00001 являлась витриной интеллектуального потенциала страны *9878.00 *9878.09
- ▣ 00001 являлась государственной идеологией казахского ханства *9878.00 *9878.10
- ▣ 00001 являлась государственной религией караханидов *9878.00 *9878.11

Результаты выполненных исследований по теме (период 01.10 2018 – 15.01.2019)

- ▣ 1. Разработаны новые методы, алгоритмы и технологии решения задачи создания декларативных средств для автоматической кластеризации текстовых документов СМИ.
- ▣ 2. Исследованы и разработаны методы и алгоритмы выделения из текстов сущностей (значимых понятий) для задачи кластеризации.
- ▣ 3. Разработаны алгоритмы формирования частотных словарей слов и словосочетаний и представления их в табличном виде.
- ▣ 4. Разработан алгоритм формирования смыслового представления документов.
- ▣ 5. Разработаны технологии и процедуры назначения элементам формализованного представления документа весовых коэффициентов их смысловой значимости.
- ▣ 6. Выполнен предварительный анализ полученных результатов при различных исходных данных.

Задачи исследований по теме на период 15.01.2019 - 12.30. 2019

- ▣ *1. Исследовать и разработать методы и алгоритмы установления смысловой близости между наименованиями понятий, представленных различным лексическим составом.*
- ▣ *2. Исследовать и разработать методы и алгоритмы выделения фактов в документах сообщений СМИ.*
- ▣ *3. Исследовать и разработать методы и алгоритмы формализации и отождествления фактов, представленных различным лексическим составом.*
- ▣ *4. Исследовать и разработать методы и алгоритмы сравнения смыслового содержания документов сообщений СМИ.*
- ▣ *5. Исследовать и разработать методы и алгоритмы установления тональности и достоверности фактов, методами концептуального анализа текстов.*
- ▣ *6. Исследовать и разработать методы, алгоритмы и технологии автоматизированного построения онтологического представления понятийного состава предметной области.*