



Thesaurus on Islam: Development and Current State

Natalia Loukachevitch

Lomonosov Moscow State University

louk_nat@mail.ru

Text Analytics

- A lot of text data
 - News flows
 - Social network messages
- Information-retrieval and information-analytical systems
 - Analysis of information in a specific domain requires special resources
- This presentation
 - Thesaurus on Islam for automatic document processing

Outline

- Lexical and terminological resources in natural language processing and information retrieval
- Family of RuThes-like thesauri
- Thesaurus on Islam
 - Scope
 - Terms and relations
 - Applications

Traditions of Knowledge Representation in Broad Domains

❖ Information-Retrieval thesauri

- ❖ Controlled vocabulary with formalized relations for improving of information retrieval

❖ WordNet-like thesauri

- ❖ Hierarchical net of lexical concepts - synsets
- ❖ Describes the lexical system of the general lexicon
- ❖ There were several projects of development domain-specific WordNet-like thesauri

❖ Formal ontologies

- ❖ Formal description of the domain in form on concepts and relations between them
 - ❖ Concepts, instances
 - ❖ Attributes, relations
 - ❖ Axioms

Accounting Traditions of Knowledge Representation

❖ Information-Retrieval thesauri

- ❖ + representation of terms, concepts, multiword expressions, small set of relations
- ❖ - oriented to manual work, absence of ambiguous words, weak formalization

❖ WordNet-like Thesauri

- ❖ + detailed description of synonyms, senses, multilevel hierarchy
- ❖ - problems with description of multiword expressions, deficiency of relations ,
- ❖ - model is not for terminology representation

❖ Formal Ontologies

- ❖ + concepts, formal principles of relation description
- ❖ - formalized description is difficult to be matched with language units
- ❖ - it is difficult to create large resources for broad domains

RuThes Thesaurus

- Linguistic Ontology - most concepts are based on senses of real language expressions
 - Developed more than 20 years
 - Corporate-owned, now partially published
 - <http://www.labinform.ru/pub/ruthes/index.htm>
 - Used in various IR and NLP projects (i.e. in news processing system of Central Bank of Russia)
- Unified representation – net of concepts
 - For different parts of speech
 - For lexical units and domain terms
 - Words and multiword expressions
- Current size
 - 55 thousand concepts, 4.1 relations per concept
 - 167 thousand Russian words and multiword expressions.

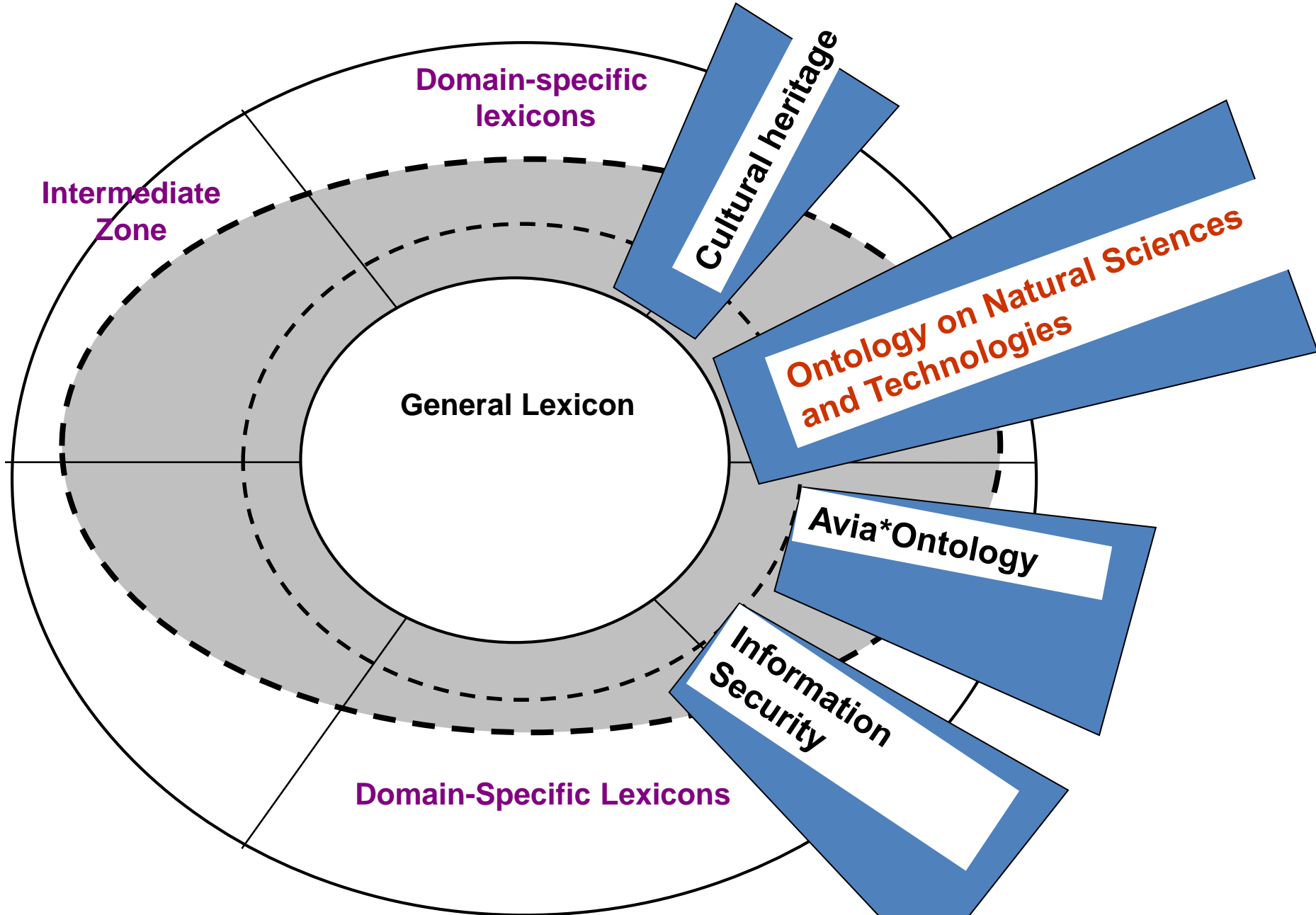
RuThes units

- Main principles
 - Distinguishable concepts – distinctions with neighbour concepts on denotational level
 - Concept should have an unambiguous and concise name
 - Text entries should be equivalent in respect to concept relations
- A concept unites the following language expressions (ontological synonyms):
 - words that belong to different parts of speech (*stabilization, stabilize, stabilized*)
 - linguistic expressions relating to different linguistic styles, genres
 - single words, idioms, free multiword expressions, which senses correspond to the concept

Conceptual relations

- **Small set of relations**
 - Class – subclass
 - Transitivity, inheritance
 - Part-whole
 - Transitivity of part-whole relations
 - Described parts are strictly related to their wholes
 - External ontological dependence (Gangemi et al., 2001; Guarino, 2009)
 - **Forest** depends on **Tree**
 - **Forest Fire** depends in **Forest**
- **Main principle for establishing relations – reliable relations**
 - Concepts of lower levels of the hierarchy should be rigidly related to upper concepts

RuThes and Domain-Specific Resources



Thesaurus on Islam: Main Stages of Development

- Resource for automatic document processing of Internet-pages, news articles, or social network messages
- The main stages of the Islam Thesaurus development:
 - to find an important Islam or Muslim-related concept analyzing the results of automatic term extraction, Islam dictionaries or Islam-related texts;
 - to introduce the concept into the thesaurus providing it with a unique **understandable name**,
 - to provide the concept with text entries that can express it in texts (synonyms);
 - to describe relations of the introduced concept to existing thesaurus concepts.

Scope of Islam Thesaurus

- basic concepts of Islam; various branches of Islam, Muslim law
- religious literature, suras of Qur'an
- forbidden actions in Islam
- Islam buildings, cult objects, worship actions, religious attributes
- Muslim calendar and holidays
- Islamic finance
- famous spiritual and military leaders of Islam
- Islam rituals, types of Muslim clothing
- Muslim education
- Islamic organizations and communities
- general concepts existing in different religions
- conflict with other religions, concepts and specific entities involved in current conflicts, various attacks, attitudes towards Islam
- the political structure of Islamic countries;
- the geography of regions with a predominantly Islamic population etc.

Collecting various text variants for a concept

Отношения на концептах

МУСУЛЬМАНСКИЙ

Название концепта
МУСУЛЬМАНСКАЯ УММА (СООБЩЕСТВО)
МУСУЛЬМАНСКАЯ ШКОЛА
МУСУЛЬМАНСКИЕ ЧЕТКИ
МУСУЛЬМАНСКИЙ ГОЛОВНОЙ ПЛАТОК
МУСУЛЬМАНСКИЙ ДРЕСС-КОД
МУСУЛЬМАНСКИЙ КАЛЕНДАРЬ
МУСУЛЬМАНСКИЙ КОЛЛЕДЖ

Добавить
Изменить
Удалить

Отношение	Асп	Название концепта
ВЫШЕ		ЛУННЫЙ КАЛЕНДАРЬ
ЦЕЛОЕ		ИСЛАМ
ЧАСТЬ		МЕСЯЦ МУСУЛЬМАНСКОГО КА

1 +
8 -

MUSLIM CALENDAR

Фильтр

Текстовый вход
АРАБСКИЙ ЛУННЫЙ КАЛЕНДАРЬ
ГОД ХИДЖРЫ
ИСЛАМСКИЙ КАЛЕНДАРЬ
ИСЛАМСКИЙ ЛУННЫЙ КАЛЕНДАРЬ
КАЛЕНДАРЬ ХИДЖРЫ
ЛУННАЯ ХИДЖРА

Добавить
Изменить
Удалить
Изменить синоним

Текстовый вход
ИСЛАМСКИЙ МЕСЯЦ
МЕСЯЦ ИСЛАМСКОГО КАЛЕНДАРЯ
МЕСЯЦ МУСУЛЬМАНСКОГО КАЛЕНДАРЯ
МЕСЯЦ ПО ХИДЖРЕ
МЕСЯЦ ХИДЖРЫ
МУСУЛЬМАНСКИЙ МЕСЯЦ

Добавить
Изменить
Удалить

Перейти к синонимам Фрагменты текстов

Закреть

Description of ambiguity

- *Hegira (Hijrah)*
 - 1) migration of Muhammad in the year 622,
 - 2) Islamic calendar, which was set from the date of the migration,
 - 3) mass migration of Muslims to Islamic countries from non-Islamic countries
- Introduced concepts
 - *MUHAMMAD'S HEGIRA: hegira, Muhammad's Hegira, Hijrah,*
 - *ISLAMIC CALENDAR: Muslim calendar, Islamic calendar Hijri calendar, Muslim moon calendar, Islamic moon calendar, year of hijri, hijri, Arab moon calendar,*
 - *MIGRATION TO MUSLIM LANDS: Hijrah, Hijrah mass Muslim migration.*

Two senses of Sunna

Отношения на концептах

СУННА

Название концепта
СУННА ГАЙРИ-МУАККАДА
СУННА ЗАВАИД
СУННА МУАККАДА
▶ СУННА ПРОРОКА
СУННИЗМ
СУННИТ
СУННИТСКИЙ БОГОСЛОВ

Добавить
Изменить
Удалить

Отношение	Асс	Название концепта
ВЫШЕ		БИОГРАФИЯ
ВЫШЕ		ИСТОЧНИК ИСЛАМСКОГО ПРАВА
ВЫШЕ		ЛЕГЕНДА, ПРЕДАНИЕ
ЧАСТЬ		ХАДИС (ИЗРЕЧЕНИЕ)
▶ АССОЦ	2	ЖЕЛАТЕЛЬНОЕ ДЕЙСТВИЕ В ИСЛАМЕ
АССОЦ	1	ПРОРОК МУХАММЕД

1 +
..

Добавить
Изменить
Перейти
Удалить

1 +
..

Фильтр

Текстовый вход
▶ МУСУЛЬМАНСКОЕ СВЯЩЕННОЕ ПРЕДАНИЕ
ПРЕЧИСТАЯ СУННА
СУННА
СУННА МУХАММАДА
СУННА МУХАММЕДА
СУННА ПРОРОКА

--->
<---

Добавить
Изменить
Удалить
Изменить синоним

Текстовый вход
▶ ЖЕЛАТЕЛЬНОЕ ДЕЙСТВИЕ В ИСЛАМЕ
СУННА

Добавить
Изменить
Удалить

Перейти к синонимам Фрагменты текстов

Закреть

Use of Dependence Relations

Отношения на концептах

ПРОРОК М

Название концепта
ПРОРОК ИСЛАМА
ПРОРОК ИСМАИЛ
ПРОРОК МУСА
ПРОРОК МУХАММЕД
ПРОРОК НУХ
ПРОРОНИТЬ (ПРОИЗНЕСТИ)
ПРОРОЧИТЬ

Добавить

Изменить

Удалить

Отношение	Асп	Название концепта
АССОЦ	2	ВОЛОС ПРОРОКА МУХАММЕДА
АССОЦ	2	ДОЧЬ ПРОРОКА МУХАММЕДА
АССОЦ	2	ЖЕНА ПРОРОКА МУХАММЕДА
АССОЦ	2	КОРАН
АССОЦ	2	МАВЛИД (МУСУЛЬМАНСКИЙ П
АССОЦ	2	МЕДИНСКИЙ ПЕРИОД
АССОЦ	2	МОГИЛА ПРОРОКА МУХАММЕДА

Добавить

Изменить

Перейти

Удалить

1 + -

КОРАН

Фильтр

Текстовый вход
ГОСПОДИН ПРОРОКОВ
МИР ЕМУ И БЛАГОСЛОВЕНИЕ
МИР ЕМУ И БЛАГОСЛОВЕНИЕ АЛЛАХА
МУХАММАД, ПРОРОК ИСЛАМА
МУХАММЕД, ПРОРОК ИСЛАМА
ОСНОВАТЕЛЬ ИСЛАМА

Добавить

Изменить

Удалить

Изменить синоним

Текстовый вход
КОРАН
КОРАНИЧЕСКИЙ
МУДРЫЙ КОРАН
СВЯЩЕННАЯ КНИГА ИСЛАМА
СВЯЩЕННАЯ КНИГА МУСУЛЬМАН
СВЯЩЕННЫЙ КОРАН

Добавить

Изменить

Удалить

Перейти к синонимам

Фрагменты текстов

Закреть

Description of Prohibited in Islam

Отношения на концептах

ЗАПРЕТНОЕ

Название концепта
ЗАПРЕТ ПОЛИТИЧЕСКОЙ ПАРТИИ
ЗАПРЕТ РОСТОВЩИЧЕСТВА
▶ ЗАПРЕТ УПОТРЕБЛЕНИЯ АЛКОГОЛЯ
ЗАПРЕТ УПОТРЕБЛЕНИЯ МЕРТВЕЧИНЫ
ЗАПРЕТ УПОТРЕБЛЕНИЯ СВИНИНЫ
ЗАПРЕТИТЬ
ЗАПРЕТНЫЙ МЕСЯЦ

Добавить

Изменить

Удалить

1 +

-

Отношение	Асг	Название концепта
▶ ВЫШЕ		ПИЩЕВЫЕ ЗАПРЕТЫ В ИСЛАМЕ
АССОЦ	1	АЛКОГОЛЬНЫЙ НАПИТОК
АССОЦ	2	ХАДД ЗА УПОТРЕБЛЕНИЕ АЛКОГ

Добавить

Изменить

Перейти

Удалить

1 +

-

Фильтр

Текстовый вход
▶ АЛКОГОЛЬ ПОД ЗАПРЕТОМ
АЛКОГОЛЬНЫЕ НАПИТКИ ПОД ЗАПРЕТ
ЗАПРЕТ АЛКОГОЛЬНЫХ НАПИТКОВ
ЗАПРЕТ АЛКОГОЛЯ
ЗАПРЕТ НА АЛКОГОЛЬ
ЗАПРЕТ НА АЛКОГОЛЬНЫЕ НАПИТКИ

--->

<---

Добавить

Изменить

Удалить

Изменить синоним

Текстовый вход
▶ ЗАПРЕЩЕНО ЕСТЬ МУСУЛЬМАНАМ
ИСЛАМСКИЕ ПИЩЕВЫЕ ЗАПРЕТЫ
НЕЛЬЗЯ ЕСТЬ МУСУЛЬМАНАМ
ПИЩЕВЫЕ ЗАПРЕТЫ В ИСЛАМЕ

Добавить

Изменить

Удалить

Перейти к синонимам

Фрагменты текстов

Закреть

Description of Relations with Other Peoples

Отношения на концептах

РЕЛИГИОЗНАЯ

Название концепта
РЕЛИГИОЗНАЯ АТТРИБУТИКА
РЕЛИГИОЗНАЯ ВОЙНА
РЕЛИГИОЗНАЯ ДЕЯТЕЛЬНОСТЬ
РЕЛИГИОЗНАЯ ДИСКРИМИНАЦИЯ
РЕЛИГИОЗНАЯ ЕРЕСЬ
РЕЛИГИОЗНАЯ ЗАПОВЕДЬ
РЕЛИГИОЗНАЯ ИДЕОЛОГИЯ

RELIGIOUS DISCRIMINATION

Фильтр

Текстовый вход
ГОНЕНИЯ НА ВЕРУЮЩИХ
ДИСКРИМИНАЦИЯ ВЕРУЮЩИХ
ДИСКРИМИНАЦИЯ ГРАЖДАН ПО РЕЛИГ
ДИСКРИМИНАЦИЯ ПО КОНФЕССИОНАЛЬ
ДИСКРИМИНАЦИЯ ПО РЕЛИГИОЗНОЙ П
ДИСКРИМИНАЦИЯ ПО РЕЛИГИОЗНОМУ

Перейти к синонимам Фрагменты текстов

Отношение	Асп	Название концепта
ВЫШЕ		ДИСКРИМИНАЦИЯ
ВЫШЕ		РЕЛИГИОЗНАЯ НЕТЕРПИМОСТЬ
НИЖЕ		ДИСКРИМИНАЦИЯ МУСУЛЬМАН
НИЖЕ		ДИСКРИМИНАЦИЯ ХРИСТИАН
АССОЦ	1	РЕЛИГИЯ
АССОЦ	1	РЕЛИГИОЗНОЕ МЕНЬШИНСТВО
АССОЦ	2	РЕЛИГИОЗНАЯ ЭМИГРАЦИЯ

Текстовый вход

Текстовый вход
ГОНЕНИЯ МУСУЛЬМАН
ГОНЕНИЯ НА ИСЛАМ
ГОНЕНИЯ НА МУСУЛЬМАН
ДИСКРИМИНАЦИЯ В ОТНОШЕНИИ МУСУ
ДИСКРИМИНАЦИЯ МУСУЛЬМАН
ДИСКРИМИНАЦИЯ МУСУЛЬМАНОК
ДИСКРИМИНАЦИЯ МУСУЛЬМАНСКОГО Н

Закреть

Text Processing: Thesaurus Matching

<input checked="" type="checkbox"/> ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫЕ С					Справка	Новая обработка
<input checked="" type="checkbox"/> Опознанные	<input checked="" type="checkbox"/> Снять все	<input checked="" type="checkbox"/> Т	<input checked="" type="checkbox"/> Т_А	<input checked="" type="checkbox"/> Т_М		
	<input type="checkbox"/> ФИО	<input type="checkbox"/> Организации				
<input type="checkbox"/> Сантимент	<input type="checkbox"/> NECRF	<input checked="" type="checkbox"/> Подсвечивать фон				
■ СПИСОК НАЙДЕННЫХ РУБРИК						
■ СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА						
■ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ						
■ АННОТАЦИЯ						
▼ <input checked="" type="checkbox"/> ОБРАБОТАННЫЙ ТЕКСТ						

Шиитские террористы **наступают** на **Киркук**
16 октября в 17:00 **Икрамутдин Хан** 2

Шиитские террористические формирования "**Хашд аш-Шааби**" в составе т.н. **Иракской армии** уже примерно сутки **ведут наступление** на **Киркук**, к которому они стягивали силы в предшествовавшие несколько дней.

Вооруженные силы (иракского) Курдистана - Пешмерга - оказывают им ожесточенное **сопротивление**.

Определенные **успехи**, которых **шиитам** удалось достигнуть на западе **Киркука**, связаны с тем, что свои позиции там оставили части **Пешмерги**, подконтрольные **оппозиционной партии ПСК**, судя по всему, вошедшей в сговор с **Багдадом**. В связи с этим в настоящий момент **правительство Курдистана осуществляет мобилизацию** и стягивают к **городу** силы, чтобы удержать его.

Руководство Курдистана заявило, что не стремится к **ведению боевых действий**, но и **уступать город шиитским бандам** не намерено.

Лидеры ряда **стран мира** призвали **Эрбиль (столица Курдистана)** и **Багдад прекратить боевые действия и начать** переговоры.

Automatic Text Categorization

ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫЕ С

[Справка](#)

[Новая обработка](#)

<input checked="" type="checkbox"/> Опознанные	<input checked="" type="checkbox"/> Снять все	<input checked="" type="checkbox"/> Т	<input checked="" type="checkbox"/> Т_А	<input checked="" type="checkbox"/> Т_М			
	<input type="checkbox"/> ФИО	<input type="checkbox"/> Организации					
<input type="checkbox"/> Сантимент	<input type="checkbox"/> NECRF	<input checked="" type="checkbox"/> Подсвечивать фон					

СПИСОК НАЙДЕННЫХ РУБРИК

Этно-религиозный

V002001220 Шииты	64
V002001210 Сунниты	64
V002001120 Католицизм	64
V002001110 Православие	64
V002001000 И	64

СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА

ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

АННОТАЦИЯ

ОБРАБОТАННЫЙ ТЕКСТ

Как **распределяют высшие государственные посты** в **Ливане**

16.10.2017, 17:13

В 1943 году **президент Ливана Бишар аль-Хури** и **премьер-министр Рияд Сольхом** заключили устное **соглашение**, согласно которому **высшие государственные должности** в **стране должны распределяться** по **религиозному принципу**. Оно получило название **Национальный пакт** и **считается** неписаной частью **ливанской конституции**.

Согласно **документу**, **премьер-министром Ливана назначают суннита**, **председателем парламента — шиита**, а их **заместителями — православных**. **Президентом Ливана традиционно становится католик-маронит**. Это стало **возможным**, так как во **время заключения Национального пакта христиане** были **ливанским большинством**.

Когда **мусульмане** получили **демографическое преимущество**, **полномочия христианского президента** были сокращены **Таифским соглашением** (заключено в 1989 году по итогам **гражданской войны**). Но **принцип пропорционального представительства до**

ЗАКЛЮЧЕНИЕ ДОГОВОРА
Термин:
A210130 ЗАКЛЮЧЕНИЕ ДОГОВОРА

Content Analysis of Islam Site

- Golosislama.ru
 - Islam-related news and Islam basics
- More than 26 thousand pages were extracted
- Used thesaurus: RuThes + specialized Islam thesaurus, more than 5000 Islam related terms
- Topic models
 - 100 and 200 topics
 - Manual labeling of topics: understability of topics

Unigram topic	Phrase-enriched topic	Thesaurus-enriched topic
Syria topic (Run 1) Relation coherence 0.11	Syria topic (Run 5) Relation coherence 0.13	Syria topic (Run 12) Relation coherence 0.36
<p>сирия (Syria)</p> <p>сирийский (Syrian)</p> <p>асад (Assad)</p> <p><u>оон</u> (UN)</p> <p>оппозиция (opposition)</p> <p>башар (Bashar)</p> <p><u>страна</u> (country)</p> <p>дамаск (Damask)</p> <p>президент (President)</p>	<p>сирия (Syria)</p> <p>башар асад (Bashar al-Assad)</p> <p>сирийская оппозиция (Syrian opposition)</p> <p>сирийский (Syrian)</p> <p>режим асада (al-Assad regime)</p> <p>асад (Assad)</p> <p>сирийский режим (Syrian regime)</p> <p>режим башара асада (Bashar al-Assad regime)</p> <p>сирийская власть (Syrian authorities)</p>	<p>сирия (Syria)</p> <p>сирийский (Syrian)</p> <p>асад (Assad)</p> <p>дамаск (Damask)</p> <p>башар асад (Bashar al-Assad)</p> <p>сирийская оппозиция (Syrian opposition)</p> <p>оппозиция (opposition)</p> <p>режим асада (al-Assad regime)</p> <p>режим (regime)</p>

Conclusions

- The Thesaurus on Islam for automatic document processing is presented
 - comprises the concepts of Islam and related concepts of social life.
 - contains more than 5 thousand terms.
- The thesaurus is created by the model of the thesaurus RuThes thesaurus
 - It is also assumed that the Thesaurus on Islam will be compatible with the published version of the RuThes thesaurus,
 - Various applications processing a wide range of texts, including news reports, specialized sites, posts in social networks.