TEL'2012
«Корпусы национальных языков: модели и технологии»

ТЕХНОЛОГИЯ ПОЛНОТЕКСТОВОГО ПОИСКА В МУЛЬТИЯЗЫЧНЫХ СЕТЕВЫХ РЕСУРСАХ

Д.В. Ландэ[1,2], д.т.н., В.В. Жигало[2]

[1]Институт проблем регистрации информации НАН Украины
[2]Информационный центр «ЭЛВИСТИ»

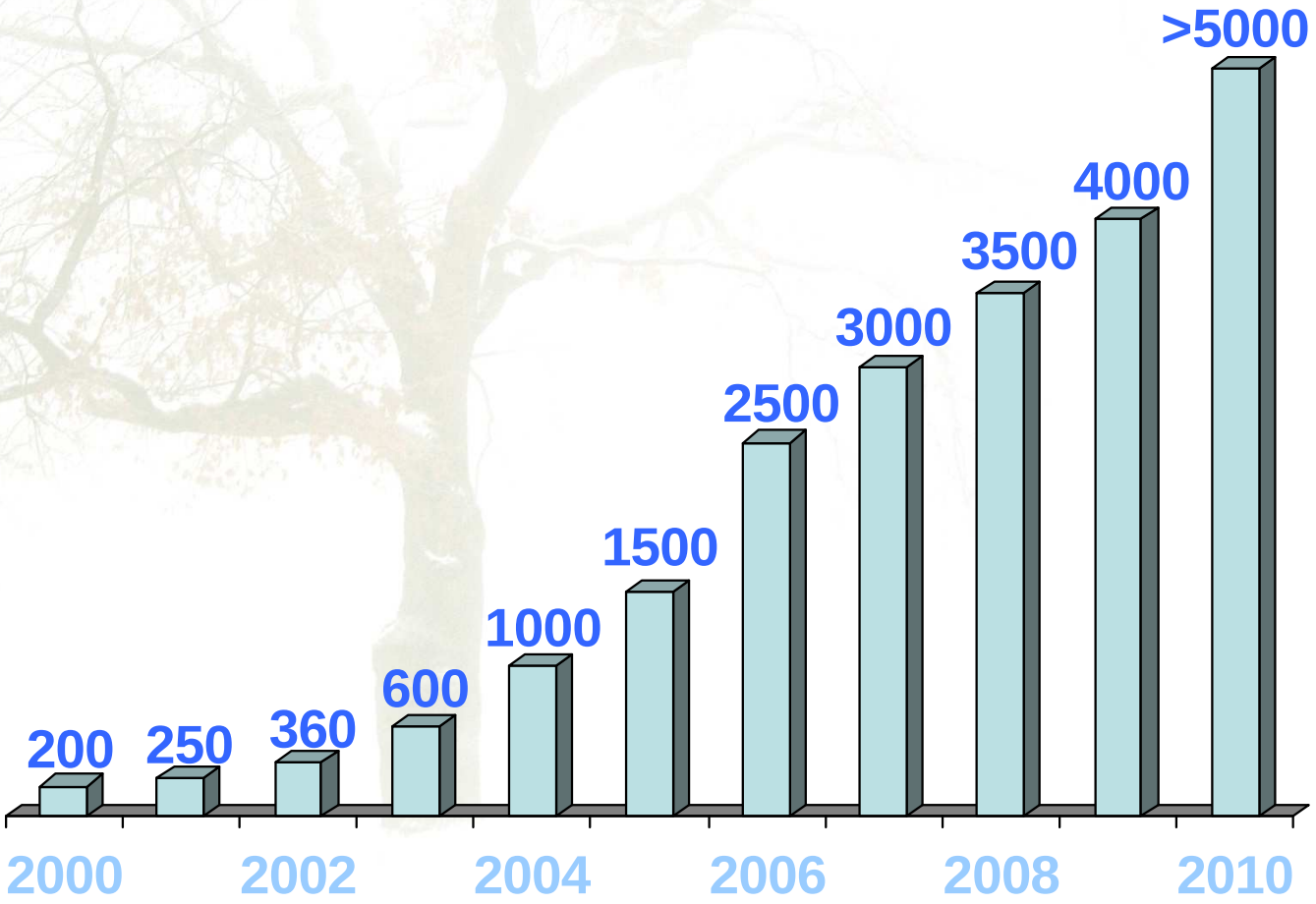
Казань-2012

Несколько слов о технологии контент-мониторинга

InfoStream

Основные характеристики

InfoStream

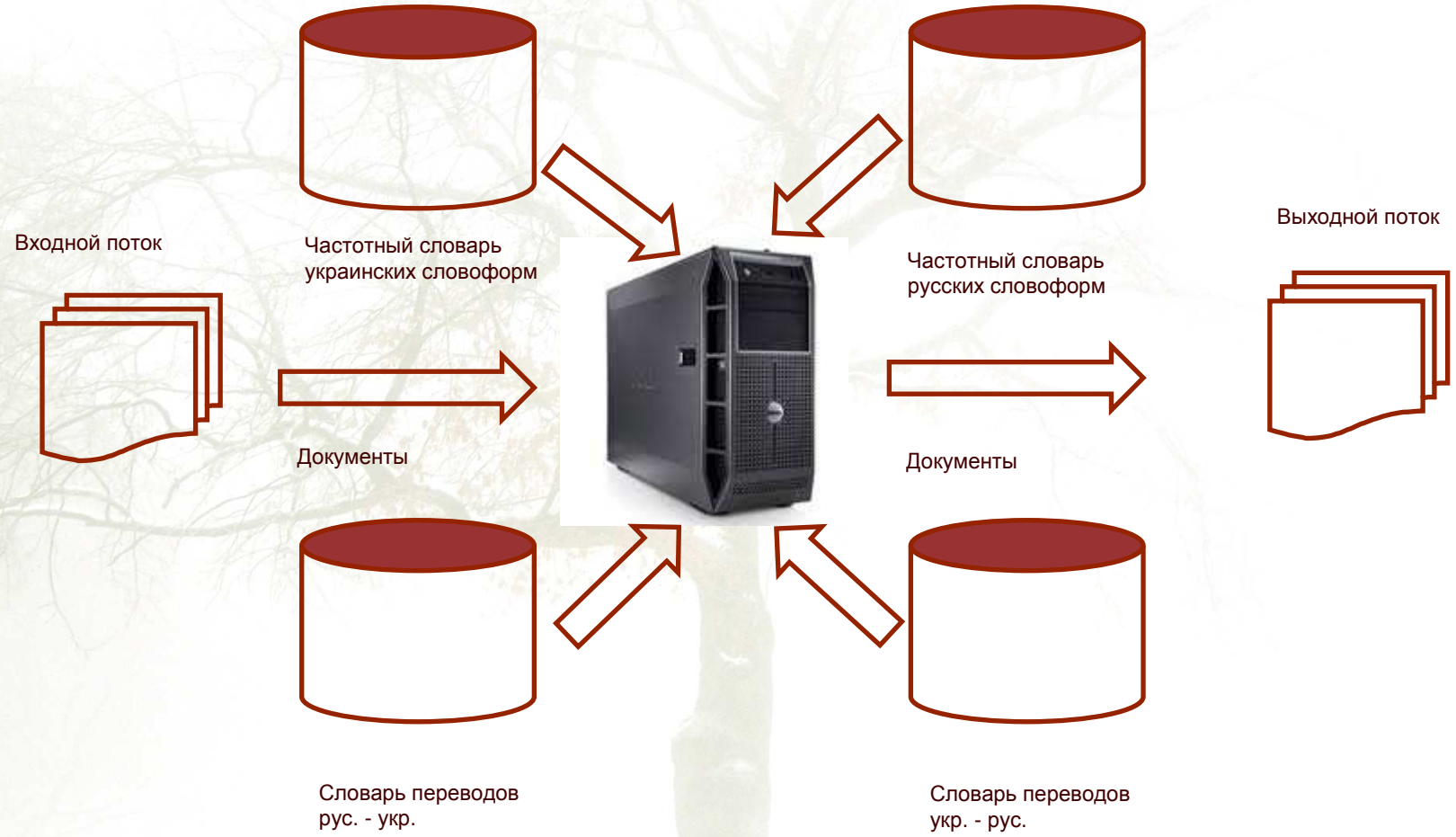


СТАТИСТИЧЕСКИ-ЛЕКСИКОГРАФИЧЕСКИЙ АЛГОРИТМ ВЫЯВЛЕННЯ РАЗНОЯЗЫЧНЫХ ДУБЛИКАТОВ

Процедура выявления дубликатов:

- подключение морфологических словарей;
- создание частотных словарей - обучение системы;
- подключение словарей переводов;
- выявление опорных слов в документах;
- сравнение опорных слов.

ИЗВЛЕЧЕНИЕ И ПЕРЕВОД ОПОРНЫХ СЛОВ



МОРФОЛОГИЧЕСКИЕ СЛОВАРИ

Для русского и украинского языков были использованы свободно доступные электронные словари: ispell с набором более 1 млн. словоформ и «Словники України», с набором более 4 млн. словоформ, а также словарь Зализняка, который насчитывает порядка 100 тыс. слов.

Эксперты дополнили морфологические словари неологизмами, названиями известных фирм, брендов и известными фамилиями, которых не было в исходных словарях.

ЧАСТОТНЫЕ СЛОВАРИ

Для обучения частотных морфологических словарей взяты электронные публикации новостей, полученные из Интернет с помощью системы контент-мониторинга InfoStream.

«Обучение» словарей проводится в несколько этапов. Первый этап - разделение документов на словоформы и сохранение полученных словоформ и номеров соответствующих документов.

На втором этапе подсчитывается количество вхождений каждой словоформы, и количество документов в которых она встретилась. Определяется вероятная нормальная форма каждого слова.

Для выявления омонимии сохраняются все нормальные формы соответствующие словоформе, т. е. если одной словоформе соответствует сразу несколько нормальных форм, сохраняются подсчитанные частоты со всеми найденными нормальными формами. На третьем этапе происходит заключительный подсчет количества нормальных форм и сохранение результатов в частотный словарь.

«ОБУЧЕНИЕ» ЧАСТОТНОГО СЛОВАРЯ

Словоформа	Количество	Индекс нормальных форм
села	20	садиться → +20 село →+20
село	50	садиться → +50 село → +50
сели	10	садиться → +10
селом	30	село → +30
		село = 100 садиться = 80

ОПРЕДЕЛЕНИЕ ОПОРНЫХ СЛОВ

При реализации алгоритма происходит считывание текстового документа из входного потока, после чего выполняется выделение словоформ и поиск нормальной формы для каждой из них. В случае контекстной неоднозначности, выбирается наиболее частотная (с наибольшим индексом) по словарю нормальная форма словоформы.

После вычисления соответствующих весовых коэффициентов с помощью формулы Okapi BM25 происходит ранжирование нормализованных слов и выбирается двенадцать наиболее «весомых».

Использовался лишь относительно небольшой, но, по-видимому, самый существенный для данной задачи срез - множество имен существительных, дополненное некоторыми фамилиями, аббревиатурами, названиями компаний.

Полученные двенадцать опорных слов переводятся на другой язык с помощью словарей переводов. Все опорные слова и слова-переводы приписываются к документу.

Okapi BM25

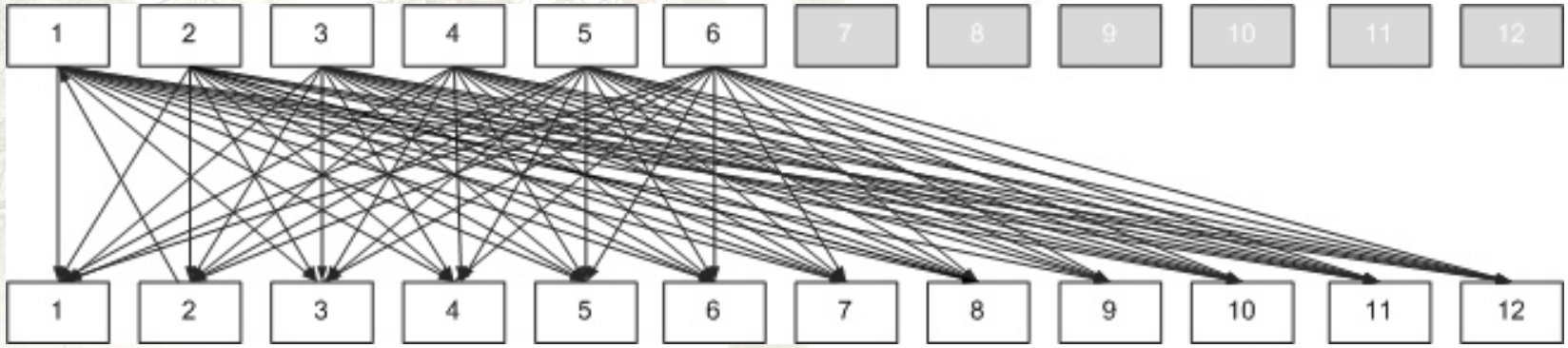
В предложенной процедуре индексирования для выделения наиболее значимых термов использовался статистический метод, базирующийся на применении общеизвестного подхода TF IDF, а точнее его модификации Okapi BM25, в которой каждому терму из документа приписывается вес по формуле:

$$W(t,D) = \frac{f(t,D)(k+1)}{f(t,D)+k(1-b+b\,|D|/L)} \cdot \log \frac{N-n(t)+0.5}{n(t)+0.5},$$

где f(t,D) - частота встречаемости терма t в документе D, |D| - длина документа D, L - **средняя длина документа в коллекции текстов**, общее количество которых - N, n(t) - количество документов в коллекции, содержащих данный терм, k, b - параметры, выбираемые экспертами.

ВЫЯВЛЕНИЕ ДУБЛИКАТОВ

В системе InfoStream используется механизм поиска дубликатов, который позволяет с помощью опорных слов находить подобные документы, представленные на одном языке. В этом механизме 6 опорных слов исследуемого документа, сравниваются с 12-ю опорными словами каждого из документов корпуса.



Процедура сравнения была дополнена рядом эвристических критериев, например:

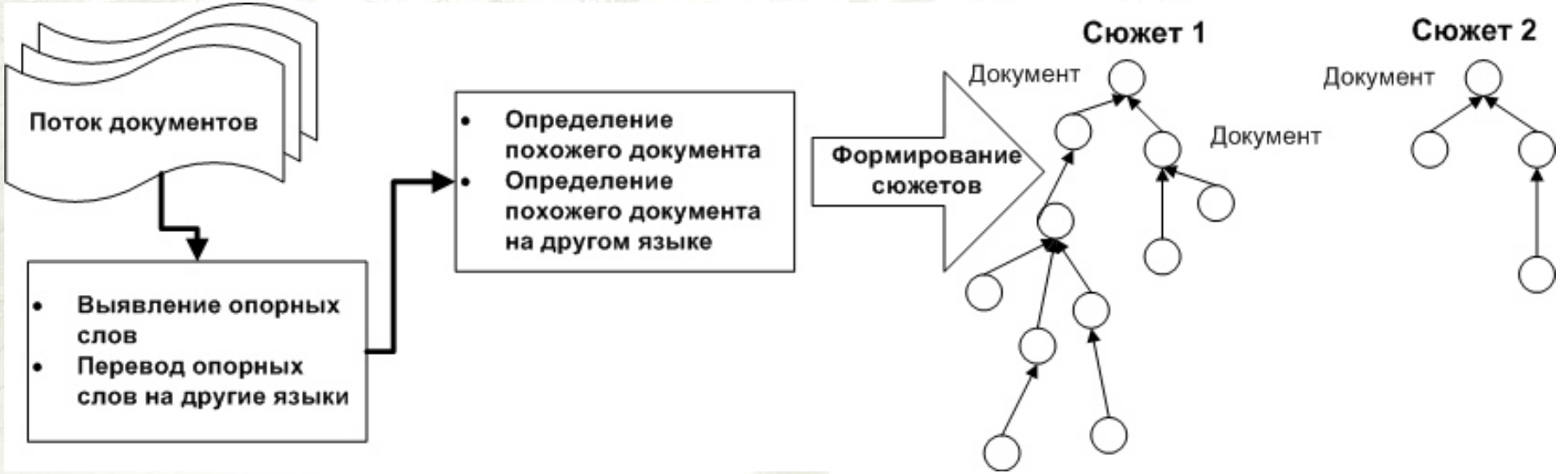
- общее количество слов в переведенном варианте не должно отличаться от оригинала более чем на 10%;
- количество чисел в документах не должно отличатся больше чем на два.

ХАРАКТЕРИСТИКИ КОРПУСА

Общее количество слов в корпусе составляет более 192,7 млн., из которых 96 млн. из украинских документов, 96.7 млн. – из русских документов. Средняя длина документа в корпусе составляет 195 слов для украинского и 196 слов для русского.
Количество источников документов на украинском языке содержащихся в корпусе – 997. Количество источников документов на русском языке – 1768.

№ п.п.	Украино-язычные источники	Количество публикаций	Русскоязычные источники	Количество публикаций
1.	ForUm	33547	ForUm	30903
2.	УНІАН	31573	УНИАН	26509
3.	РБК-Екраїна	21517	УКРИНФОРМ	25838
4.	УТРО-Україна	20019	РБК-Украина	21849
5.	УКРІНФОРМ	19031	Корреспондент.net	21646
6.	Оглядач	18460	УТРО-Украина	19769
7.	ProUa	14090	ICTV	19719
8.	Кореспондент.net	13505	ProUa	15189
9.	Укроп	12346	Обозреватель	14844
10.	ГлавРед	8905	ГлавРед	10475
11.	Новинар	8159	NewsRu.ua	6284
12.	NewsRu.ua	7377	Форпост	5621
13.	УКРИНФОРМ	6518	Подробности	4204
14.	Форпост	6017	Київ-Прес-Інформ	3385
15.	Вголос	5535	Zaxid.net	3081

ФОРМИРОВАНИЕ ОСНОВНЫХ СЮЖЕТОВ



ПОИСКОВЫЙ ИНТЕРФЕЙС – ОБЗОР ОСНОВНЫХ СЮЖЕТОВ

Активная база данных: Главная | Система интеграции интернет-ресурсов | Вход | Выход | InfoStream Online

Помощь | Кабинет | Источники | Статистика | Новости проекта

снег

Период: 7 дней | Убрать дубли | Морфология

Найти | Динамика | Дайджест | Очистить | События | Сюжеты | Язык запросов Примеры

Обзор основных сюжетов

снег;

документов - 3000, сюжетов - 1201

В виде графа (Java) | Распечатать

1. **На выходные Украину заметет и заморозит**
   {1-} В Киеве завтра -8/-3 °С ночью, -3/+2 днем, пройдет снег. В воскресенье также осадки. Ночью -8/-3, днем -3/+2 °С. В понедельник ожидается мокрый снег, ночью -5/0, днем от -3 до +2 °С.
   Сюжет полностью (436)

   2012.01.18 14:12 Завтра в Киеве снег, -4...-6 Левый берег
   436
   2012.01.20 17:49 Погода в Киеве на выходных: снег и гололед Моє Місто

2. **Попов поручил срочно очистить Киев от снега**
   Глава Киевской городской государственной администрации Александр Попов провел экстренное совещание по вопросам преодоления последствий непогоды. Об этом говорится в сообщении КГГА. А.Попов поручил всем коммунальным службам столицы вывести на улицы города всю снегоуборочную технику. Глава КГГА выразил неудовлетворение сегодняшней работой коммунальщиков в условиях снегопада и заверил, что лично будет контролировать ситуацию с уборкой снега.
   Сюжет полностью (97)

   2012.01.18 18:26 Снегоуборочная техника стала причиной пробок в столице Автоновости AVTOINDEX
   97
   2012.01.20 17:42 ГАИ призывает водителей ездить с включенными фарами ближнего света Портал ексклюзивних новин

3. **В Борисполе сняты ограничения на прием-отправку воздушных судов**
   В аэропорту Борисполь работают обе ВПП В аэропорту Борисполь сняты все ограничения на прием-отправку воздушных судов. Как сообщили Корреспондент.net, с 14:30 по киевскому времени были сняты ограничения на отправление, а с 16:08 отменены ограничения на прием всех видов воздушных судов. "На данный момент действуют обе взлетно-посадочные полосы.
   Сюжет полностью (83)

   2012.01.20 14:34 Снегопад блокировал работу аэропорта "Борисполь" Комментарии:proUA
   83
   2012.01.20 18:00 В аэропорту "Борисполь" сняты ограничения на прием-отправку воздушных судов ┣ Транспортный бизнес

4. **В Киеве не справляются со снегом**
   За прошедшие сутки с улиц Киева вывезено 617 тонн снега, что почти вдвое больше, чем в течение предыдущих суток. Об этом сообщили в Главном управлении контроля за благоустройством КГГА. В частности, по информации, поступившей от районных в городе Киеве госадминистраций, коммунальной корпорации Киевавтодор и коммунального объединения Киевзеленстрой, к уборке столицы от снега было задействовано
   Сюжет полностью (44)

   2012.01.18 15:16 Снег в Киеве убирают семь тысяч коммунальщиков и более полутысячи единиц техники Корреспондент.net
   44
   2012.01.20 16:56 617 тонн снега вывезли с улиц столицы за прошедшие сутки Наш Киев

Фрагмент параллельного корпуса

500-win.txt - Блокнот

Файл Правка Формат Вид Справка

```
<item>
        <rus>"Charge-Coupled Device" позволяет получать изображение,
преобразуя фотоны света в электроны (электрический ток).</rus>
        <ukr>"Charge-Coupled Device" дозволяє одержувати зображення,
перетворюючи фотони світла в електрони (електричний струм).</ukr>
</item>
<item>
        <rus>"Chevrolet" после столкновения раскрутило и он,
неуправляемый, задевает еще один "Nissan".</rus>
        <ukr>"Chevrolet" після зіткнення розкрутило і він, некерований,
зачіпає ще один "Nissan".</ukr>
</item>
<item>
        <rus>"Children-UA" это организация волонтеров из Европы и Канады,
которых объединяет неравнодушие к современной проблемы украинских
детей-сирот в Донецкой области.</rus>
        <ukr>"Children-UA" це організація волонтерів з Європи та Канади,
яких об'єднує небайдужість до сучасної проблеми українських дітей-сиріт в
Донецькій області.</ukr>
</item>
<item>
        <rus>"Chinese Democracy" станет первой полноценной пластинкой Guns
N'Roses с начала 1990-х годов.</rus>
        <ukr>"Chinese Democracy" стане першою повноцінною платівкою Guns
N'Roses з початку 1990-х років.</ukr>
</item>
```

Онлайн-интерфейс – сайт http://ling.infostream.ua

InfoStream - Украинско-русский параллельный текстовый корпус - Mozilla Firefox

http://ling.infostream.ua/

InfoStream - Украинско...

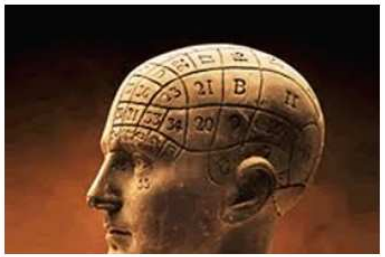
Поиск в параллельном корпусе:

Русский
Украинский

Морфология

Найти

Украинско-русский параллельный текстовый корпус



В Информационном центре ElVisti создан выровненный на уровне предложений украинско-русский параллельный текстовый корпус из веб-публикаций. Объем корпуса - более **2,6 млн.** пар уникальных предложений.

Метод построения корпуса базируется на использовании "опорных слов" в тестовых документах, а также средствах их автоматического перевода. Опорные слова в рамках данного подхода выделяются с использованием русского и украинского морфологических словарей, а также словарей переводов имен существительных для русского и украинского языков. Кроме того, для вычисления весов терминов в документах используются некоторые дополнительные эмпирико-статистические правила. Для выравнивания параллельного корпуса на уровне предложений использовались преимущественно статистические методы.

Алгоритмы были реализованы в виде программного комплекса, который интегрирован с системой контент-мониторинга InfoStream, благодаря чему корпус постоянно пополняется.

Предполагается дальнейшее использование данного лингвистического ресурса для создания системы автоматического перевода новостных сообщений.

Язык запросов

Запросы состоят из поисковых слов и операторов. В качестве поисковых слов могут использоваться слова естественного языка или их правые усечения. По умолчанию, при отключенной морфологии, каждое слово воспринимается как усечение (слова менее 3 символов ищутся как точное совпадение). Для поиска по полному слову, а не усечению, необходимо дописать к нему специальный символ "**]**". Система не чувствительна к регистрам букв.

В системе используется следующий набор операторов:

~, **-** - оператор контекстного следования;
@ - оператор контекстной близости;
!, **^** - логическое **И-НЕТ**;
&, **+** - логическое **И**;
|, , - логическое **ИЛИ**;
Оператор контекстного следования (**~**) отбирает пары поисковых терминов, которые в тексте документа расположены друг за другом, причем учитывается порядок следования терминов.

Оператор контекстной близости (**@**) отбирает пары поисковых терминов, которые находятся рядом друг с другом, причем порядок следования не важен.

Различные уровни определяются с помощью круглых скобок.

Режим поиска – сайт http://ling.infostream.ua

Поиск в параллельном корпусе:

вертолет

Морфология

Русский
Украинский
Найти

вертолет

Найдено документов - **3472**, страница 1 из 348
Статистика слов:
ВЕРТОЛЕТ - 5651,

1. **Международная аэрокосмическая выставка "Фарнборо-2010" пройдет в Лондоне**
   Около 1350 участников из 52 стран примут участие в международной аэрокосмической выставке "Фарнборо-2010", которая пройдет в пригороде Лондона 19-25 июля, сообщает BBC Russia со ссылкой на пресс-службу Федеральной службы по военно-техническому сотрудничеству (ФСВТС) России.

   Міжнародна аерокосмічна виставка "Фарнборо-2010" пройде в Лондоні
   Близько 1350 учасників з 52 країн візьмуть участь у міжнародній аерокосмічній виставці "Фарнборо-2010", яка пройде в передмісті Лондона 19- 25 липня, повідомляє BBC Russia з посиланням на прес-службу Федеральної служби з військово-технічного співробітництва (ФСВТС) Росії.

2. **"Си Бриз-2010": украинские десантники одели немцев в тельняшки**
   В рамках учений "Си бриз 2010" в Украине немецкие десантники, за плечами которых не один десяток прыжков с парашютом, впервые осуществили прыжки с украинского вертолета Ми-8 под украинскими куполами.

   "Сі Бриз-2010": українські десантники одягли німців у тільники
   У рамках вчень "Сі бриз 2010" в Україні німецькі десантники, за плечами яких не один десяток стрибків з парашутом, вперше здійснили стрибки з українського вертольота Мі-8 під українськими куполами.

3. **Пэрис Хилтон учится водить вертолет**
   Пэрис Хилтон платит за один урок, на котором ее учат водить вертолет, 7 тысяч долларов.

   Періс Хілтон навчається водити гелікоптер
   Періс Хілтон платить за один урок, на якому її навчають водити гелікоптер, 7 тисяч доларів.

4. **Под Берлином взрываются снаряды**
   Борьба с огнем поке не очень успешна: пожарные не могут зайти в лес, чтобы приступить к тушению пожара, так как снаряды продолжают взрываться.

   Під Берліном вибухають снаряди
   Боротьба з вогнем поки що не дуже успішна: пожежники не можуть зайти в ліс, щоб приступити до гасіння пожежі, так як снаряди продовжують вибухати.

Описание ресурса – сайт http://ling.infostream.ua

Для скачивания доступен заархивированный фрагмент параллельного корпуса размером в **100 тысяч** пар уникальных предложений (в ZIP-архиве ~ 9 МБ).

Формат представления данных приближен к XML:

```
<item>
    <rus>предложение</rus>
    <ukr>речення</ukr>
</item>
... 99 998 раз ;)
<item>
    <rus>предложение</rus>
    <ukr>речення</ukr>
</item>
```

Информация представлена в кодировке CP1251 (Windows).

Использование этого фрагмента корпуса в научных и учебных целях - свободное.

Подробности - в статье Д.Ландэ и В.Жигало

Препринт: arXiv:0807.0311, PDF

TEL'2012
«Корпусы национальных языков: модели и технологии»

Спасибо за внимание!

Казань-2012