(cc) BY

Корчевская Оксана Валериевна
канд. техн. наук, доцент
Нетребко Владислав Витальевич
магистрант
ФГБОУ ВО «Сибирский государственный
технологический университет»
г. Красноярск, Красноярский край

РАЗРАБОТКА СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПО ОБРАБОТКЕ НАРРАТИВНЫХ ИСТОЧНИКОВ ДАННЫХ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Аннотация*: в данной статье рассмотрены вопросы разработки высокоэффективной системы поддержки принятия решений по обработке нарративных источников данных. Разработана высокоэффективная система поддержки принятия решений для обработки нарративных данных, позволяющая провести анализ и сделать выгрузку обработанных данных для проведения дальнейших исторических исследований, используя статистическую обработку информации.*

Ключевые слова*: нарративные данные, интеллектуальный анализ, искусственный интеллект, кластерный анализ, анализ текста.*

Большое количество эмпирических данных в социальных науках имеют нарративную форму: тексты интервью, дневники, литературные произведения, рассказы пациентов на психотерапевтических сеансах, показания свидетелей, некрологи, рекламные тексты, новости на телевиденье, тексты песен, анекдоты, любовные письма и т. д. Под нарративом понимается любой повествовательный текст, функция которого – информировать адресата о событиях.

В качестве примера были взяты данные сайта бессмертного полка (http://www.moypolk.ru/) в количестве 100 случайных нарративов и сохранены в текстовом формате. На 08.06.2015 в летопись полка записано 276837 имен, что

позволяет расширять выборку до очень больших размеров. На рисунке 1 представлены данные об одном из исторических персонажей сайта.

Проведена предварительная обработка данных:

1. Очистка данных:

– очистка текста от стоп-слов – вспомогательных слов, которые несут мало информации о содержании документа («так как», «кроме того» и т. д.);

– удаление имен, фамилий и отчеств (для этого используется база фамилий и имен с сайтов vse-familii.com и vse-imena.com, соответственно);

– исправление ошибок в орфографии;

– удаление предложений при дублировании.

Рис.1. Фрагмент с сайта бессмертного полка

2. Преобразование данных:

– стемминг – отсечение от слова окончаний и суффиксов, чтобы оставшаяся часть, называемая stem, была одинаковой для всех грамматических форм слова;

– лемматизация – для этого используется динамическая библиотека лемматизации lemmatizator.dll (lemmatizator.so). В лабораторных условиях вероятностная модель морфологии позволяет улучшить точность примерно до 98–99%. На рисунке 2 представлен пример работы лемматизатора.

```
: #lemmatize
: мы ели суп, а вдоль аллеи стояли раскидистые ели

[я] [есть] [суп] [,] [а] [вдоль] [аллея] [стоять] [раскидистый] [ель]
:
```

Рис. 2. Пример работы лемматизатора

Для анализа данных использован изменённый алгоритм *k*-средних. Модификация заключалась в том, что начальные центы кластера *k* берутся исходя из места положения «ключевых слов», а не случайным образом.

Для реализации программного средства использован язык программирования Microsoft Visual Studio C# 2012.

При запуске программы открывается главная форма с открытой вкладкой «Подготовка» (рис. 3).

Перед началом работы необходимо добавить нарративы и заполнить поле «Ключевые слова».

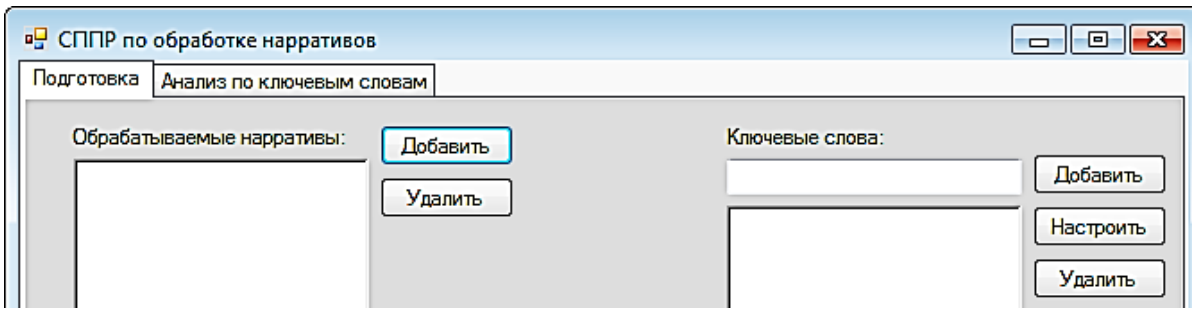

Рис. 3. Форма СППР по обработке нарративов

После нажатия кнопки «Добавить» откроется диалогое окно, в котором небходимо выбрать нужные нарративы, как показанно на рисунке 4. Программа поддерживает только текстовый(txt) формат файлов.

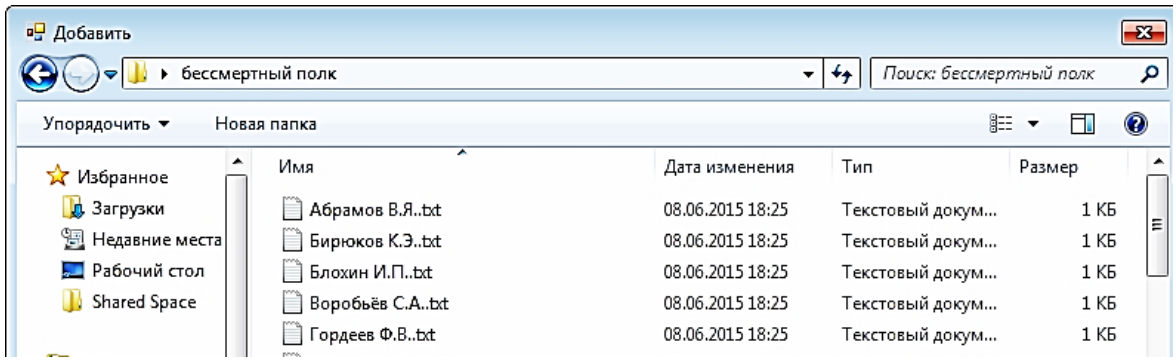

Рис. 4. Вид окна выбора нарративов

Так же в приложении предусмотрены функции сохранения и загрузки полученных результатов. Все сохраненные результаты можно загрузить в MS Office Excel и без дополнительной обработки построить диаграммы для удобства дальнейшего анализа, как показано на рисунке 5.

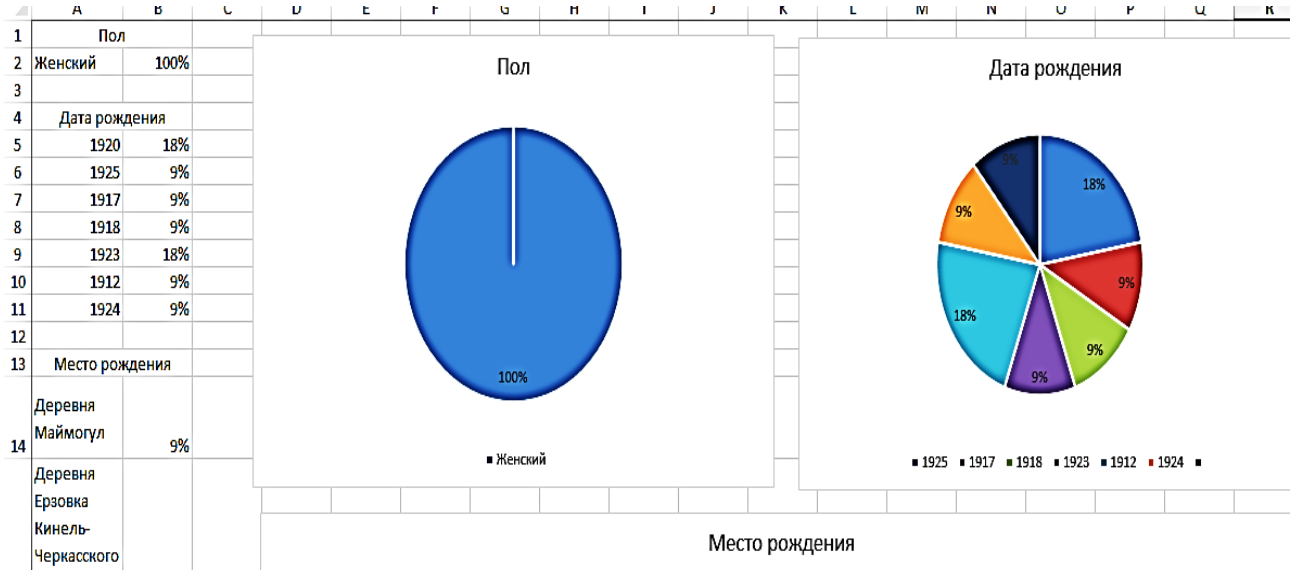

Рис. 5. Пример использования MS Office Excel для анализа данных

Разработана высокоэффективная система поддержки принятия решений для обработки нарративных данных, позволяющая провести анализ и сделать выгрузку обработанных данных для проведения дальнейших исторических исследований, используя статистическую обработку информации.

Список литературы

1. Баранов М.А. Составные ключевые термы в задаче кластеризации текстовых документов / М.А. Баранов // Информационные технологии. – 2014. – №1. – С. 3–8.

2. Бессмертный полк [Электронный ресурс]. – Режим доступа: http://www.moypolk.ru

3. Лемматизатор в грамматическом словаре [Электронный ресурс]. – Режим доступа: http://www.solarix.ru/for_developers/api/lemmatizator-api.shtml

4. Методы интеллектуального анализа данных [Электронный ресурс]. – Режим доступа: http://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques

5. Text Mining [Электронный ресурс]. – Режим доступа: http://www.statsoft.ru/home/textbook/modules/sttextmin.html#textmin_2