в аспекте наличия в ней текстовых заимствований предполагает формирование оценки на основе обратной зависимости, а значит, и обоснованный выбор нечетких функций специального вида.

5) Соотнесение базовых шкал оценивания при автоматизированном анализе студенческих работ на предмет наличия в них текстовых заимствований целесообразно проводить при помощи двумерных нечетких функций.

Список литературы

1. *Болкунов И. А.* Пути преодоления студенческого плагиата / И. А. Болкунов // Проблеми сучасної педагогічної освіти : [зб. ст.]. Сер.: Педагогіка і психологія / РВНЗ «Крим. гуманіт. ун-т». – Ялта, 2009. – Вип. 21. – Ч. 3. – С. 58–66.

2. *Zadeh L. A.* Fuzzy Sets as a Basis for a Theory of Possibility, Fuzzy Sets and Systems // International Journal for Fuzzy Sets and Systems. – 1978. – Vol. 1, No. 1. – PP. 3–28.

3. *Рожков Н.Н.* Система перезачета оценок успеваемости – инструмент поддержки академической мобильности // Университетское управление: практика и анализ. – 2006. – С. 104–113.

4. *Пегат А. П.* Нечеткое моделирование и управление / А. П. Пегат ; пер. с англ. – М. : БИНОМ. Лаборатория знаний, 2009. – 798 с.

5. *Круглов В. В.* Нечеткая логика и искусственные нейронные сети / В. В. Круглов, М. И. Дли, Р. Ю. Годунов. – М. : Физматлит, 2001. – 224 с.

АНАЛИЗ ТЕКСТОВ ИНТЕРНЕТА С ПОМОЩЬЮ БЕСПЛАТНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Данилевич С. Б.

Харьковский гуманитарный университет
«Народная украинская академия»
г. Харьков, ул. Лермонтовская, 27, тел. 716-44-09,
e-mail: danilevichsb@mail.ru

Целью данной работы является анализ текстов из Интернета на незнакомом читателю языке (например, чешском) с помощью бесплатного программного обеспечения.

В качестве данных для анализа были использованы первые 10 веб-страниц, полученных с помощью чешской поисковой системы *seznam.cz* по запросу «problémy v oblasti vysokoškolského vzdělávání» (*рус.*: проблемы в высшем образовании).

Для очистки от html-тегов была применена свободно распространяемая программа TextSTAT 2.9. Для составления частотного словаря воспользовались программой AntConc. К сожалению, эти программы не поддерживают кириллицу.

Частотный словарь позволяет определить, какие слова или фразы наиболее часто встречаются в большинстве работ на выбранную тему. Составление списка слов (stop list), которые не влияют на результат анализа, также является неотъемлемой частью анализа. В данном случае stop list составил 495 слов. «Вычитание» стоп-списка из полученного частотного словаря можно произвести в Excel с помощью расширенного фильтра или используя VBA. Аналогичным образом используется словарь синонимов (например, ABZ slovník českých synonym – on-line hledání – http://www.slovnik-synonym.cz/). Проблема остается с собственными именами, ошибками в самом тексте, употреблением слов на других языках и т.п.

Для удаления из списка цифр, знаков препинания, отличных от чешского способов начертания букв, удобно пользоваться в приложении MS Word подстановочными знаками команды *Найти и Заменить*. Однобуквенные и двухбуквенные выражения удаляются с помощью обычного фильтра.

Для предварительного анализа текста достаточно удобно использовать частотный словарь и конкорданс AntConc.

Последовательность действий для определения проблем, которые видят в Чешской республике в области высшего образования, следующая:

- установка программ TextSTAT, AntConc;
- отбор веб-страниц по заданной тематике;
- очистка информации этих страниц от тэгов в TextSTAT и сохранение ее в формате *.txt*;
- открытие в AntConc созданного текстового файла;
- создание частотного словаря (Word List), который может быть открыт в MS Excel;
- перевод слов полученного частотного словаря в переводчике Google;
- использование в AntConc функции *Concordance* для определения контекста использования выбранного слова;
- перевод контекста.

После проведения подобных действий было выявлено, что в данном корпусе рассматриваются в основном психологические, поведенческие проблемы, вызывающие трудности в обучении. Работодатели в государственном и частном секторах указывают на проблемы с нехваткой персонала, отвечающего потребностям современного рынка труда. Отмечается недостаточное инвестирование в высшее образование (в ЕС оно составляет лишь 1,3 % от ВВП по сравнению с 3,1 % в США и 1,5 % в Японии) и выражается надежда, что государственное финансирование будет дополнено соответствующими дополнительными ресурсами, с большим участием частного сектора.

Организация Eurydice Network (http://eacea.ec.europa.eu/-education/eurydice/), миссией которой является ознакомление лиц, ответственных за систему образования, с европейским уровнем анализа и информации с целью оказания им помощи в принятии решений, провела исследование «Модернизация высшего образования в Европе: финансирование и социальные аспекты». Решение некоторых проблем видят в росте квалификации и профессионального мастерства преподавателей, разработке и внедрении системы финансовых и нефинансовых стимулов для работодателей с целью увеличить расходы на обучение персонала и др.

Таким образом, с помощью бесплатного программного обеспечения может быть проведен предварительный анализ текстов на незнакомом языке из сети Интернет. При достаточной заинтересованности можно составить корпус большего объема, что позволит глубже анализировать информацию.

Список литературы

1. *Беленький А.* Текстомайнинг. Извлечение информации из неструктурированных текстов / А. Беленький // КомпьютерПресс. – 2008. – № 10. – С. 174–179. – Режим доступа: http://www.compress.ru/article.aspx?id=19605&iid=905.

2. *Чубукова И. А.* Data Mining : учеб. пособие. – М. : Интернет-университет информ. технологий : БИНОМ : Лаборатория знаний, 2006. – 382 с.

3. *Станкевич А. Ю.* Поиск контекстов и оценка их типичности средствами AntConc (Laurence Anthony) // Теория и практика преподавания русского языка как иностранного: достижения, проблемы и перспективы развития : матер. V Междунар. науч.-метод. конф. Минск, 16-17 июня 2011 г. – Минск : Изд. Центр БГУ, 2011. – 227 с. – С. 210-213.