Информационное общество и проблемы прикладной информатики

Лабораторные работы №1 и №2. Инструменты для работы с текстами.

Составил М.А. Бакаев, к.т.н., доцент кафедры АСУ НГТУ.

Цели работы:

1. Получить представление об основных методах и технологиях обработки текстов на естественном языке.
2. Изучить программные инструменты (решения), существующие в данной области.
3. Научиться поиску и анализу научной информации (в т.ч. оценке качества находимых публикаций), а также использованию соответствующих электронных инструментов.
4. Продемонстрировать знание правил оформления научных работ согласно актуальным стандартам.

Введение

Современное общество характеризуется большими объемами создаваемой и обрабатываемой информации, значительная часть которой представлена на естественном языке (в том числе на веб-узлах глобальной сети Интернет, в виде гипертекста). Соответственно, начиная с 1970-х годов отмечалось бурное развитие методов и технологий, предназначенных для анализа и обработки текста на естественном языке – аннотирования, индексирования и поиска, перевода на другой язык и т.д. Сегодня вы работаете с реализациями этих технологий ежедневно, а в ходе выполнения работы вам предстоит самостоятельно ознакомиться с теорией и практикой в данной области информационного общества.

Ход работы

При выполнении ЛР1 вами должен сначала быть создан и сдан преподавателю **черновик**. Он должен включать в себя список из не менее, чем 10 **релевантных и авторитетных** публикаций, процитированных согласно ГОСТ. Из этих публикаций не менее 80% должны быть научными статьями (не учебниками или пособиями), а не менее 50% – англоязычными. Для выполнения работы рекомендуется использовать сервис Академия Гугл (scholar.google.ru).

При выполнении ЛР2 вы должны провести **самостоятельную апробацию** выбранных вами программных инструментов (онлайн или устанавливаемых на десктоп или мобильное устройство). Отчет должен сопровождаться скриншотами апробации, на которых должны быть видны системные **дата и время**[1].

[1] Для этого требования допускается исключение – для инструментов, работающих на мобильных платформах.

Задание на лабораторную работу 1

ЛР1 посвящена методам и подходам к обработке текстов на естественном языке (т.е. в основном **теоретическому и научному** аспекту данной области).

Соответственно, ваш отчет должен представлять собой **реферативный перечень** современных методов и технологий (со ссылками **по тексту** на **авторитетные** источники), организованный согласно основным задачам, выделяемым в этой области[2]:

- индексация и поиск,
- машинный перевод,
- автоматическое аннотирование,
- понимание содержания,
- распознавание речи и др.

Для каждой из задач, помимо краткого описания существующих методов и подходов, вы должны также самостоятельно обобщить тенденции в их развитии **за последние 5 лет**. При этом среди цитируемых источников для каждой задачи должно быть не менее одной релевантной публикации, у которой фамилия автора начинается на первую букву вашей фамилии или имени.

В заключительной части вашей работы вы должны самостоятельно предложить **ещё один способ классификации** рассмотренных методов и подходов, имеющий **не менее 5 и не более 10 различных признаков**.

Прочие требования к отчету по ЛР1:

- объем основного текста отчёта – не менее 20 тыс. знаков (около 8 страниц А4),
- количество цитируемых источников – не менее 15, из которых не менее половины – англоязычные,
- оформление отчета должно соответствовать требованиями соответствующих ГОСТ,
- текст отчета должен иметь приемлемый уровень оригинальности, любые прямые заимствования текстов из других источников должны **помечаться как цитаты** (в кавычках) и сопровождаться ссылкой на источник.

[2] В вашей работе должно быть не менее 7 таких задач, т.е. часть вы должны добавить **самостоятельно**.

Задание на лабораторную работу 2

ЛР2 посвящена программным инструментам, относящимся к обработке текстов на естественном языке (т.е. в основном **практическому и инженерному** аспекту данной области).

Вы должны найти и изучить, на базе информации, предоставляемой авторами, а также **самостоятельно опробовав**, не менее 3[3] существующих программных инструментов (готовых сервисов, библиотек, модулей расширения и пр.), которые:

- предназначены для решения одной из задач в сфере обработки естественного языка,
- имеют дату последнего обновления не более 10 лет назад.

Для каждого из инструментов вы должны описать:

- какая из задач обработки текста решается продуктом,
- к какому пункту классификации, самостоятельно предложенной вами в ЛР1, относится продукт,
- форматы входных и выходных данных.

[3] Из них не менее 1 продукта – иностранного авторства.

Рекомендуемая литература

1) А. В. Аграновский, Р. Э. Арутюнян. Индексация массивов документов // Журнал "Мир ПК", #06, 2003 год / Издательство "Открытые системы". Доступно по адресу http://www.osp.ru/pcworld/2003/06/049.htm
2) Семантический анализ текста онлайн (Advego). Доступно по адресу http://advego.ru/text/seo/
3) Пример решения: http://www.aot.ru/download.php#3
4) Пример решения (иностранное авторство): http://lucene.apache.org/
5) Пример модуля для продукта (CMS Drupal): https://drupal.org/project/rustemmer
6) Пример инструментария для обработки текстов: http://www.opencalais.com/
7) Материалы конференции Dialogue: http://www.dialog-21.ru/
8) Материалы конференции AINL (Artificial Intelligence and Natural Language): https://ainlconf.ru/