Лабораторная работа № 2. Построение и использование корпусов текстов естественного языка.

Цели работы:

1. изучить принципы построения корпусов текстов, виды разметки и способы аннотирования, инструменты работы с корпусами текстов,
2. построить корпус текстов и разработать корпусный менеджер.

Задание

1. Сформировать электронный корпус текстов по выбранной предметной области.
2. Используя результаты лабораторной работы №1 (возможность получения лингвистических сведений для произвольной лексемы естественного языка) разработать корпусный менеджер, обеспечивающий базовую функциональность работы с созданным корпусом текстов.

Объем работы и прочие требования

Корпус текстов – большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. Важнейшее свойство корпуса текстов – репрезентативность. Корпус должен максимально объективно представить разнообразие изучаемого явления и дать объективную картину использования этого явления в речевой практике носителей языка. Разметка – приписывание текстам и их компонентам специальных меток. Основным типом разметки является морфологический или частеречный. Здесь морфологические метки включают не только части речи, но и признаки грамматических категорий. Данная разметка является основой для дальнейшего анализа.

Для решения многих лингвистических задач используются так называемые текстовые корпуса — специальным образом подобранные и структурированные коллекции текстов. Наиболее информативными являются размеченные корпуса, то есть такие, в которых частям текста приписана лингвистическая информация — например, каждое слово отнесено к той или иной части речи.

Создание размеченного корпуса — очень трудоёмкий процесс, требующий времени и сил многих специалистов. По этой причине чаще всего размеченные корпуса создаются коллективами исследователей при государственных

учреждениях, и таких корпусов не очень много. Однажды созданный корпус может быть использован многими исследователями для решения различных задач. Способы применения корпуса могут быть самыми разнообразными, в том числе и такими, о которых не думали его создатели. Чтобы корпус мог приносить максимальную отдачу научному сообществу, нужно, чтобы он был доступен не только для просмотра через предусмотренный его разработчиками интерфейс, но и для скачивания целиком на компьютер пользователя.

OpenCorpora — это проект по созданию размеченного корпуса текстов силами сообщества. Корпус представляет собой хранилище, специально предназначенное для текстов с лингвистической разметкой, удобный интерфейс редактирования разметки и исправления ошибок, инструменты для контроля качества и стандарт разметки для русского языка.

pymorphy2 - морфологический анализатор, разработанный на языке программирования Python. Выполняет лемматизацию и анализ слов, способен осуществлять склонение по заданным грамматическим характеристикам слов.

Корпусный менеджер (corpus manager) - специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Методические указания:

Требуется спроектировать и программно реализовать структуры хранения данных, алгоритмы их обработки, необходимые в рамках следующих базовых требований к разрабатываемому приложению:

- входные данные – фрагмент текста (фраза или слово) на естественном языке – запрос корпусному менеджеру;
- выходные данные – частотные характеристики словоформ, лексем, грамматических категорий, леммы, морфологические характеристики словоформ и их метаданные (библиографические, типологические), конкордансные списки, согласно согласованным с преподавателем требованиям к функциональности корпусного менеджера;
- взаимодействие с пользователем посредствам графического интерфейса (интерфейс должен быть интуитивно-понятным и дружественным пользователю);
- наличие системы средств помощи пользователю;

- обеспечение возможности построения, сохранения, просмотра, редактирования, пополнения, фильтрации и поиска по заданному условию, документирования текста и/или его фрагмента в соответствии с реализуемой функциональностью корпусного менеджера;
- поддержка различных форматов представления входных данных (TXT, RTF, PDF, DOC, DOCX).

Рекомендуется использовать функциональность стандартной, а также специализированных библиотек языка программирования Python для обработки естественного языка, например, nltk.

Вариант задания выбирается студентом самостоятельно и согласовывается с преподавателем. Средства разработки выбираются студентом самостоятельно. Защита лабораторной работы предполагает демонстрацию работоспособности всех реализованных функций в соответствии с требованиями.

Требования к отчету:

В отчете представить, в том числе графически, используя такие программные средства, как Microsoft Visio или Draw.io:

- структурно-функциональную схему разработанного приложения;
- описание структур хранения данных, алгоритмов их обработки, необходимых для реализации базовых требований к разработанной программе;
- оценку быстродействия приложения;
- выводы по работе и по перспективам использования приложения.

Отчет предоставить для проверки в электронном виде.

Варианты заданий:

Номер варианта	Язык текста	Предметная область
1	**2**	**3**
1	Русский	Кинематограф
2	Русский	Кулинария
3	Русский	Растения
4	Русский	Животные
5	Русский	Медицина
6	Русский	Литература
7	Русский	Транспорт
8	Русский	Спорт

9	Русский	Музыка
10	Русский	Досуг
11	Русский	Услуги
12	Русский	Недвижимость
13	Английский	Кинематограф
14	Английский	Кулинария
15	Английский	Растения
16	Английский	Животные
17	Английский	Медицина
18	Английский	Литература
19	Английский	Транспорт
20	Английский	Спорт
21	Английский	Музыка
22	Английский	Досуг
23	Английский	Услуги
24	Английский	Недвижимость