**Домашнее задание № 3. Синтаксический анализ, коллокации**

На выбор варианты A, B, C, D, E (*в скобках указаны баллы оценки)*

**Срок выполнения задания** – до **10 апреля** включительно.

**Отчет высылается** **по почте** Олегу Дону на адрес [dragn613@yandex.ru](file:///C:\Users\Elena\Documents\Елена\ВШЭ\Курс_КЛ\Курс_КЛ_2023\ДЗ_КЛ_2023\dragn613@yandex.ru)

В качестве темы письма указать: «**КЛ ДЗ № 3, ФИО**»

Отчет должен быть в виде отдельного текста, код программы включается как приложение.

**A.** *(6-9 баллов)*

Провести исследование различных типов **коллокаций** на базе встроенных средств системы **RuWac** (<http://corpus.leeds.ac.uk/ruscorpora.html>), которые позволяют вычислять меры ассоциации слов на входящих в их состав корпусах текстов. Для этого:

* Рассмотреть разные типы коллокаций (фразеологические единства, полусвободные сочетания, свободные сочетания);
* Взять 5-15 коллокаций каждого типа, вычислить для них частоту встречаемости и 2-3 меры ассоциации, построить для каждой меры упорядоченный список коллокаций;
* Проанализировать полученные списки (можно ли по значениям конкретной меры отличить один тип коллокаций от другого?), а также провести сравнительный анализ рангов каждой коллокации в списках для разных мер (как отличаются ранги коллокаций в разных списках?).

Отчет: Краткая характеристика использованных средств (системы, корпуса\сов); описание исследованных коллокаций, примененных мер ассоциации и полученных списков значений мер; выводы по проведенному анализу.

**В.** *(6-9 баллов)*

Провести **сравнительный анализ** возможностей двух доступных синтаксических анализаторов (парсеров) для русского или английского языка (возможно, подключая их к своей программе). Необходимо рассмотреть предложения разных типов и сложности, со словосочетаниями различных видов. Могут быть взяты (но не только):

* *MaltParser* (<http://maltparser.org/mco/mco.html>) для английского языка
* *MaltParser* (<https://nlpub.ru/MaltParser>) для русского языка
* *Stanford Parser* (<http://nlp.stanford.edu/software/lex-parser.html>) для англ. языка
* *UDPipe* (<http://ufal.mff.cuni.cz/udpipe>) для русского языка
* *LinkGrammer Parser* (<http://slashzone.ru/parser/> ) для русского языка

Отчет: Набор проведенных тестов, краткая характеристика используемых парсеров и применяемой ими модели синтаксического анализа, выводы по тестированию, описание программы тестирования (если она использовалась).

**С.** *(7-12 баллов)*

Составить (с использованием любого модуля морфоанализа) **программу**, выполняющую **синтаксическую сегментацию** текста на русском языке на базе выбранных правил:

* сегментацию на простые предложения по знакам пунктуации **и/или**
* выделение неразрывных синтаксически связанных групп слов на основе локальных высоковероятных связей (см. правила на слайдах 33-34, 63-64 Лекции № **7**), например: *синий платок*, *очень красивый*, или даже более сложные группы: *исключительно интересный фильм, красный полосатый мяч, любит весело играть,* которыевыделяются применением нескольких правил.

Протестировать программу на нескольких текстовых фрагментах (не менее 1-2 страниц).

Отчет: описание применяемых правил сегментации, составленная и примененная программа с комментариями, результаты тестов, выводы.

**D.** *(7-12 баллов)*

Составить (с использованием любого модуля морфологического анализа) **программу**, выполняющую **извлечение** наиболее частотных неразрывных двухсловных **коллокаций** из текста на русском языке. Для этого:

* Выбрать текст среднего размера (более 5-7 страниц), или же несколько текстов одной тематики (новостных или специализированных, например, научных);
* Выбрать 1-2 синтаксических типа (грамматических образца) извлекаемых коллокаций (например, образец прилагательное-существительное: *A N* ).
* Применить для извлечения: абсолютную частоту встречаемости, меры ассоциации *MI*, *T-score* и\или др. Рассмотреть при этом статистику словоформ и/или лемм.

Проанализировать и прокомментировать топ-20 для каждого из извлеченных списков коллокаций, упорядоченных по значению меры.

***Замечание***: Согласование слов можно не учитывать (но его учет добавляет баллы).

Отчет: Описание выбранного текста/ов, синтаксических типов извлекаемых коллокаций, примененных мер ассоциации; характеристика составленной программы; списки извлеченных коллокаций с комментариями; выводы по проведенному анализу.

**E.** *(7-13 баллов)*

Составить (с использованием любого модуля морфоанализа) **программу**, выполняющую **извлечение** **терминов** из специализированного русскоязычного текста\ов.

* Выбрать текст среднего размера (более 5-7 страниц), или же несколько текстов одной предметной области (учебник, научные статьи или др.);
* Рассмотреть несколько (1-3) грамматических образцов терминов (например, образец прилагательное-существительное: *A N* ) для извлечения терминов-кандидатов;
* Применить для упорядочения извлеченных кандидатов 1-2 статистических признака: абсолютную частоту встречаемости, меру ассоциации, меру *C-value* и\или др.
* Рассмотреть топ-20 элементов каждого упорядоченного списка и вычислить для них оценку средней точности извлечения, произведя предварительно вручную оценку (разметку) термин/нетермин.
* Проанализировать и прокомментировать полученные результаты.

***Замечание***: Согласование слов в словосочетаниях можно не учитывать (но его учет добавляет баллы).

Отчет: Описание выбранного текста/ов, грамматических образцов извлекаемых терминов и примененных мер упорядочения; характеристика составленной программы; списки извлеченных словосочетаний с комментариями; оценки средней точности и выводы по проведенному анализу.