**Индивидуальное (итоговое) домашнее задание**

**Два** основных варианта  Выбор варианта – до **25 апреля включительно**

**I. Доклад с презентацией** *(6-10 баллов)*

Может быть сделан:

* по одной из тем, не рассмотренных в курсе КЛиАТ или же рассмотренных обзорно (в том числе: конкретные языковые или генеративные модели или др.);
* на основе 1-3 англоязычных научных статей, либо главы из англоязычной книги-монографии по КЛ, либо интернет-источников
* не менее чем из 20-25 слайдов.

Срок выполнения – **до** **16 мая включительно**

Презентация **предварительно** должна быть **выслана** (не позднее **5 мая**) **и согласована.**

**II. Создание программы-приложения по КЛ** *(7-12 баллов)*

Написать и протестировать **программу**, решающую одну из прикладных задач КЛиАТ. При ее разработке и оценке качества решения можно использовать открытые размеченные корпуса и коллекции текстов. Для ряда задач размеченные данные были созданы и применены в рамках соревнований конференции Диалог ([*http://www.dialog-21.ru*](http://www.dialog-21.ru))

Необходимо составить **отчет** (3-5 страниц), включающий: описание функций и структуры созданной программы; используемый метод и лексические ресурсы (словари /грамматики/корпуса); оценки качества/эффективности; программу с комментариями. ***Замечание:*** По запросу могут быть высланы ссылки на нужные ресурсы

Срок выполнения – **до** **25 мая включительно**

**Отчет высылается** **по почте** Олегу Дону на адрес [ordon@edu.hse.ru](file:///C:\Users\Elena\Documents\Елена\ВШЭ\Курс_КЛ\Курс_КЛ_2021\ДЗ_КЛ_21+оценки\ordon@edu.hse.ru)

В качестве темы письма указать: **КЛ Итоговое задание, ФИО**

**Варианты программы** (прикладные задачи КЛиАТ, список может быть расширен):

**1)** **Распознавание** на базе статистики букв и/или слов:

– стиля/жанра/темы текста

– схожести текстов (парафраз).

Возможно использование статистических данных, которые есть в литературе или в сети Интернет в открытом доступе.

**2)****Реферирование/аннотирование** текста на русском языке на основе эвристических правил/ статистики/обучения (возможна реализация одного из известных методов).

Подвариант**:** Построение названия (заголовка) текста на основе одного из подходов:

* статистика и эвристические правила
* машинное обучение (включая нейронные сети – см. результаты соревнования  
  [*http://www.dialog-21.ru/evaluation/2019/disambiguation/generation/*](http://www.dialog-21.ru/evaluation/2019/disambiguation/generation/) )

**3)** Автоматическое **обнаружение и исправление орфографических ошибок** в русскоязычных текстах с применением словарей/языковых моделей/редакционного расстояния (расстояния Левенштейна-Дамерау) или др. Дополнительная информация – см. результаты соревнования SpellRuEval

[*http://www.dialog-21.ru/evaluation/2016/spelling\_correction/*](http://www.dialog-21.ru/evaluation/2016/spelling_correction/)

**4) Извлечение информации из текстов**: именованных сущностей/отношений/фактов

(например, упоминаний некоторых товаров, лиц, фирм и т.п.)

* на основе правил с использованием инструментальных средств построения *IE*-приложений (см. лекцию);
* на основе обучения, с использованием размеченных данных – например, <https://www.dialog-21.ru/evaluation/2016/ner/>

**5)** **Анализ тональности** текстов или анализ мнений по отношению к какому-то объекту (лицу, продукту, товару и т.п.) путем классификации текстов на 2 класса: положительный/отрицательный или 3 класса: положительный/отрицательный/нейтральный   
 на основе одного из подходов:

* машинное обучение;
* словари и правила.

**6)** Другие прикладные задачи КЛиАТ (согласуются с преподавателем).

### Критерии оценки итогового ДЗ

### Доклад с презентацией (6-10 баллов)

- сделана презентация (+2-3)

- сложность излагаемого материала (+1-2)

- новизна/актуальность темы (+1)

- полнота описания метода/системы/модели (+1-2)

- понятность изложения (+1-2)

- сделаны выводы (+1)

### Создание программы-приложения КЛ (7-12 баллов)

- написана программа (+1-3)

- описаны функции/структура программы, инструменты (+1)

- описан используемый/ые метод/ы или система (+1-2)

- описаны тестовые данные и лексические ресурсы (+1)

- проведены эксперименты (+1-2)

- результаты проанализированы (+1-2)

- сделаны выводы (+1)