Проектное предложение

Тип элемента практической подготовки	*Проект*
Если проект, тип проекта	*Исследовательский*
Наименование проекта	R для антиковедов. Часть 2: Токенизация и разведывательный анализ
Подразделение инициатор проекта	Школа философии и культурологии
Руководитель проекта	*Алиева Ольга Валерьевна*
Основное место работы руководителя проекта в НИУ ВШЭ	Школа философии и культурологии
Контакты руководителя (адрес эл. почты)	oalieva@hse.ru
Соруководители проекта от НИУ ВШЭ *(если имеются)*	*нет*
Контакты соруководителей от НИУ ВШЭ (адрес эл. почты)	*нет*
Основная проектная идея / описание решаемой проблемы	Проект посвящен методам токенизации в R. Токен — это отдельное наблюдение; применительно к тексту это может быть слово или сочетание слов, символ или сочетание символов, может быть даже параграф или предложение — все зависит от того, что мы намерены посчитать. Делить текст на токены мы будем с использованием различных пакетов для text-mining, научимся конвертировать данные из одного формата в другой, удалять стоп-слова, а также визуализировать результаты в ggplot2. Практическим результатом проекта станет публикация на площадке RPubs результатов анализа выбранного греческого или латинского источника (источников). Тематический план: 1. Абсолютная частотность (tf), визуализации в ggplot и wordclouds 2. Биграммы и построение сетей в ggraph; stopwords

	3. Относительная частотность (rtf) и характерные слова (tf-idf) 4. Лексические корреляции в widyr 5. Создание скользящего окна с пакетом slider 6. Pointwise mutual information (PMI) в widyr 7. Создание корпуса и работа с метаданными в пакете tm (text mining) 8. Конвертация DTM в tidy форматы и обратно 9. Зияния (стык гласных): как их посчитать? 10. Считаем длину предложений (двумя способами)
Цель и задачи проекта	Задачей проекта является количественный анализ выбранного античного источника или источников.
Проектное задание	Еженедельное выполнение заданий (просмотр обучающих видео, участие в обсуждениях на форуме группы, прохождение тестов), одна лабораторная работа и оценка своих peers
Планируемые результаты проекта, специальные или функциональные требования к результату	Публикация на площадке RPubs результатов анализа выбранного греческого или латинского источника (источников)
Дата начала проекта	*4.07.2022*
Дата окончания проекта	*11.09.2022*
Трудоемкость (часы в неделю) на одного участника	*5*
Предполагаемое количество участников (вакантных мест) в проектной команде	*5*
Названия вакансий (ролей), краткое описание задач по каждой вакансии, количество кредитов и критерии отбора для участников проекта *(если характер работ для всех участников совпадает, описывается одна вакансия)* *Кредиты на 1 участника рассчитываются по формуле:*	*Вакансия №1*: участник *Задачи:* еженедельное выполнение заданий (просмотр обучающих видео), участие в обсуждениях на форуме группы в LMS, прохождение тестов), одна лабораторная работа и оценка своих peers *Количество кредитов: 2* *Критерии отбора на вакансию:* знание древнегреческого и/или латыни, курсовая/исследовательская работа в области классической

*продолжительность в неделях * трудоемкость проекта в часах / 25*	античности; преимуществом будет опыт работы в проекте R для антиковедов. Часть 1: Извлечение данных из HTML & XML 10 недель * 5 часов/нед
Общее количество кредитов	*max 10 (5 участников по 2 кредита)*
Форма итогового контроля	*Зачет*
Формат представления результатов, который подлежит оцениванию	Публикация на RPubs
Формула оценки результатов, возможные критерии оценивания результатов с указанием всех требований и параметров	$О_т$ = оценки за тестирование (max 10, считается как среднее) $О_л$ = оценка за лабораторную работу, выставляется двумя peers и считается как среднее Итоговая = $0.5*О_т + 0.5*О_л$
Возможность пересдач при получении неудовлетворительной оценки	*Только для лабораторной работы*
Ожидаемые образовательные результаты проекта	- навыки работы с пакетами для text-mining, в том числе токенизация и подсчет частотности - визуализация данных в ggplot2
Особенности реализации проекта: территория, время, информационные ресурсы и т.п.	Дистанционно (*асинхронно*) через SmartLMS и в рабочем чате в Телеграм
Рекомендуемые образовательные программы	Филология, Философия, История, Античность, Медиевистика, Лингвистика
Требуется резюме студента	*Нет*
Требуется мотивированное письмо студента	*Да* В письме необходимо указать уровень владения древними языками (древнегреческий, латынь) и область научных интересов, включая тему курсовой работы и предполагаемые направления ее развития