**Рабочая программа курса «Работа с платформой PolyAnalyst. Компоненты текстового анализа».**

**Наименование программы:** Работа с платформой PolyAnalyst. Компоненты текстового анализа.

**Количество часов:** 25 астрономических часов

**Формат обучения:** очная форма обучения

**Материально-техническое оснащение:** для выполнения практических заданий необходимы компьютеры (ноутбуки) с установленной программой PolyAnalyst, проектор для преподавателя.

|  |  |  |  |
| --- | --- | --- | --- |
| **Тема** | **Осваиваемые компетенции** | **Практические задания** | **Узлы** |
| **Введение** | | | |
|  | Примеры задач, которые можно решить с помощью PA |  |  |
| **Занятие 1 «Загрузка и предварительная обработка данных»** | | | |
| Знакомство с платформой PolyAnalyst | * Запуск программы * Создание проекта * Импорт и экспорт проекта * Знакомство с рабочими панелями аналитика * Знакомство с базовыми действиями в платформе PolyAnalyst (добавление, копирование и удаление узлов, создание «скрипта») * Знакомство с базовыми узлами импорта данных * Знакомство с базовыми операциями с таблицей (переименование, удаление столбцов) * Изменения интерфейса программы | * Запуск PolyAnalyst * Создание нового проекта * Создание простого скрипта, состоящего из нескольких «узлов» * Экспорт/импорт проекта * Одновременная работа с несколькими проектами * Копирование узлов в другой проект |  |
| Загрузка и предварительная обработка данных | * Использование узлов предобработки информации для объединения, фильтрации данных, удаления дубликатов, создания выборки и тд | * Импорт данных из разных источников * Объединение данных из разных источников * Определение языка документа * Удаление дубликатов документов * Фильтрация документов * Создание выборки * Переименование/удаление колонок | * Узлы импорта (Excel, CSV, Internet Source, Files, …) * Sample * Distinct Texts * Distinct * Language Detection * Filter Rows * Merge * Join * Modify Columns * Filter Columns |
| Создание новой колонки | * Создание новой колонки * Функции расчета значений новой колонки | * Добавление новой колонки к существующим данным * Использование логических, арифметических и условных функций для заполнения колонки | * Derive |
| Предварительная обработка текстов | * Предварительная обработка и очистка данных * Знакомство с регулярными выражениями | * Удаление из текста элементов html-разметки, гиперссылок * Извлечение метаинформации документа с помощью регулярных выражений | * Extract Terms * Replace Terms |
| Знакомство с узлами текстового анализа | * Проверка орфографии * Использование узла индекс * Знакомство со словарями морфологии и спелчека * Извлечение ключевых слов | * Проверка орфографии документа * Изменение настроек узла индекс * Редактирование словарей морфологии и спелчека * Настройка параметров извлечения ключевых слов | * Spell Check * Index * KE |
| **Занятие 2 «Узлы текстового анализа»** | | | |
| Классификация и кластеризация документов | * Классификация и кластеризация текстов * Настройки предикторов (лексемы, ключевые слова, сущности) * Определение тематики текста | * Обучение классификатора * Применение обученной модели к новому набору данных | * Text Classifier * Text Clustering * Text Thematic * Score |
| Таксономия | * Различные режимы работы таксономии * Создание пользовательской таксономии | * Создание таксономии для классификации документов | * Taxonomy * Score |
| Словари. Менеджер словарей. | * Создание словаря * Редактирование словаря * Экспорт и импорт словаря | * Создание словаря, содержащего список слов. |  |
| Язык поисковых запросов PDL | Создание поисковых запросов PDL, в т.ч. с использованием:   * операторов PDL: or, and, not * функций работы с морфологией (lemma, form) * функций поиска последовательностей (sentence, optional, phrase, near, repeat) * именованных параметров функций * функций работы с пунктуацией * функций обращения к словарю * функций поиска стандартных сущностей | * Написание запросов на языке PDL для заполнения таксономии и поиска информации в текстах. * Использование запросов в узлах Taxonomy, Search Query, Filter Rows | * Search Query * Taxonomy * Filter Rows |
| Извлечение стандартной информации из текстов | * Извлечение стандартных сущностей (имена людей, названия компаний, организаций, географические объекты и тд) * Извлечение фактов * Анализ тональности * Анонимизация и восстановление сущностей | * Извлечение стандартных именованных сущностей и фактов из текста * Отображение связей между сущностями | * Entity Extraction * Sentiment Analysis * Fact Extraction * Entity Anonymization * Generic Dataset * Link Terms |
| **Занятие 3 «Извлечение произвольной информации из неструктурированного текста»** | | | |
| Способы извлечения произвольной информации в PA | * Понятие пользовательской сущности * Добавление пользовательской сущности в узел Entity Extraction | * Добавление пользовательской сущности в узел Entity Extraction | * Entity Extraction |
| Язык извлечения информации XPDL | * Знакомство с языком XPDL * XPDL и PDL * Структура правил языка XPDL * Работа с редактором правил | * Создание правила на языке XPDL * Отображение результатов | * Entity Extraction |
| Возможности языка XPDL | * Вложенные правила * Фильтрующие правила * Форматирующие функции * Использование ранее извлеченных сущностей | * Создание набора правил на языке XPDL с использованием вложенных и фильтрующих правил * нормализация и форматирование извлеченной информации | * Entity Extraction |
| **Занятие 4 «Создание решения по извлечению информации из неструктурированного текста»** | | | |
| Разработка пользовательской сущности | * Самостоятельная разработка сущности на языке XPDL * Валидация результатов правил | * Самостоятельное написание набора правил на языке XPDL | * Entity Extraction |
| Анализ извлеченной информации | Создание отчета на основе извлеченной информации:   * Фильтрация строк * Создание и заполнение новой колонки * Агрегирование результатов * Экспорт результатов в таблицу * Замена терминов и категорий * Рассылка таблицы при помощи почтового сервера | Работа с результатами узла Извлечение сущностей:   * создание таблицы * фильтрация строк * выгрузка результатов в документ | * Generic Dataset * Derive * Aggregate * Join * Filter Rows * Send Mail |
| **Занятие 5 «Визуализация результатов»** | | | |
| Знакомство с узлами визуализации | * Знакомство с различными способами визуализации информации (графики, графы, диаграммы, облако тэгов) * Визуализация результатов на карте | * Добавление визуализации в ранее созданный проект | * Graph * Charts * GIS * Trend Graph * Tag cloud |
| Веб-отчёты | * Создание веб-отчёта * Создание связей между элементами веб-отчёта | * Добавление веб-отчета в ранее созданный проект | * Веб-версии узлов визуализации |
| Многомерный анализ | * Знакомство с узлами многомерного анализа в PA | * Подготовка данных для многомерного анализа * Построение OLAP таблицы и многомерной матрицы | * OLAP * Dimension Matrix * Link Analysis |