Отличие построения нарратива российских государственных и негосударственных СМИ

Копылов Валентин Владимирович
БПТ171

26 декабря 2018 г.

Документация проекта по курсу
«Основы программирования в Python»

Перечень используемых библиотек Python

- requests – (используется для загрузки html-страниц)
- bs4 (BeautifulSoup) – (используется для работы с html-страницами и выгрузки необходимой информации)
- re – (используется для обработки полученного текста и вытаскивания нужных паттернов слов)
- pandas – (используется для формирования списка слов, и анализа частотности и тональности)
- pymorphy2 – (используется для нормированию слов относительно начальной формы для более точного подсчета количетсва встречаемых слов)
- time (sleep) – (используется для задания интеравла выгрузки html-страниц)
- datetime – (используется для работы с временным отрезком выгружаемых новостей)
- wordcloud – (используется для построения облака встречаемых слов)
- PIL (Image) – (используется для наложения background картинки при построении облака слов)
- numpy (используется при построении визуализаций)
- matplotlib – (используется для построения визуализаций)
- tkinter – (используется для внедрения интерфейса взаимодействия внутри программы)
- squarify – (используется для построения графика типа TREEMAP)

Дополнительные ресурсы

- Linis Crowd – (данный ресурс используется для работы с тональностями слов)

Анализируемые информационные ресурсы

- Lenta.ru
- РИА Новости
- NEWSru.com

Общее описание функционала

Данная программа используется для выявления паттернов формирования нарратива в российский СМИ с помощью анализа частотности употребления слов и их тональной окраски.

На вход пользователем подается: новостной ресурс, который будет в дальнейшем использован, временной отрезок анализа, название рубрики новостного ресурса, и тип предпочитаемой визуализации

На выходе будет представлена визуализация имеющихся паттернов в определенной (выбираемой пользователем из предложенных) форме, которая будет показывать частотность употребления слов и оттенки.

Параметры на входе

- Название новостного ресурса: str
  (выбор из предложенного)
- Название рубрики: str
  (выбор из предложенного)
- Начало временного отрезка : str
  (в формате - «(Д)Д (М)М ГГГГ»: например, «7 8 2017»)
- Конец временного отрезка : str
  (в формате - «(Д)Д (М)М ГГГГ»: например, «28 8 2017»)
- Тип визуализации (название): str
  (выбор из предложенного)

Описание результатов

Формат выдачи – одна из визуализаций на основе выбранных пользователем конфигураций (новостного ресурса, временного отрезка, рубрики)

Представленные визуализации: CloudWord, TreeMap, столбчатая диаграма, показывающая наиболее встречающиеся слова и гистограмма распределения частоты слов в зависимости от тональности слова

Описание режима взаимодействия с пользователем

Пользователь взаимодействует с программой в окне (созданной с помощью tkinter), где ему будут предложено

1) выбрать новостной ресурс для взаимодействия,
2) ввести анализируемый временной отрезок,
3) выбрать рубрику (при наличии) для взаимодействия
4) выбрать способ визуализации полученной информации

Ограничения программы

У программы существует серия ограничения, которые влияют на ее работу и взаимодействие с ней:

1) Время выгрузки раздела сайта довольно большое из-за паузы, которая существует для того, чтобы новостной ресурс не заблокировал возможность взаимодействия с ним.

2) Проблемы с html-кодировкой страниц на РИА Новости, которая довольно неоднородна и нестабильна, а также не позволяет выгружать новости по имеющимся рубрикам, вследствие чего данный новостной ресурс рассматривается лишь согласно определенным датам, без деления на разделы.

3) Словарь тональностей покрывает лишь существительные и не позволяет оценить тональную окраску других частей речи, вследствие чего выборка слов значительно уменьшается.