НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Разработка приложения для изучения русского языка как иностранного на базе корпуса политической журналистики

Выполнила:
Студентка группы 12ФПЛ
Ванина Наталья Валерьевна
Научный руководитель:
доктор полит. наук, проф.
Гронская Наталья Эдуардовна

- *Цели*: создание, обработка и разметка корпуса политических статей, моделирование и ввод в эксплуатацию упражнения на базе полученного корпуса.
- *Актуальность:* в настоящее время существует достаточно мало электронных приложений в свободном доступе (особенно тематических) для изучения русского языка как иностранного. Данное упражнение может быть полезно как студентам, так и журналистам, изучающим русскоязычные СМИ.
- *Объект* исследования: корпусная лингвистика, методология создания корпуса текстов; технология создания приложений в сфере обучения.
- *Предмет* исследования: исследовательская база по созданию корпуса текстов и учебных приложений. Работы в области автоматизированного извлечения информации (Information Retrieval (IR)), естественной обработки языка (Natural Language Processing(NLP)), статьи на тему «политика» с сайта Lenta.ru (Россия, мир).

Задачи:

1. Проанализировать литературу, посвященную компьютерной лингвистики, в частности методику создания корпуса текстов и его применение в образовательной среде;
2. Написать код на языке программирования Python, позволяющий автоматически пополнять корпус новыми и уже размеченными с помощью программы Mystem статьями на тему «политика» с сайта Lenta.ru;
3. Спроектировать и написать код, используя работы в области NLP, позволяющий создать упражнение на базе полученного ранее корпуса статей;
4. Изучить технологию создания электронных приложений, выбрать наиболее подходящий формат для данного исследования;
5. Ввести упражнение в эксплуатацию;
6. Сделать выводы о проделанной работе.

- *Методы* исследования: поиск и теоретический анализ литературы в области обработки и использования корпуса текстов в образовательных целях; анализ и подбор наиболее подходящих инструментов для создания электронного обучающего приложения.
- *Структура* исследования: введение, три главы, заключение, список литературы и источников, приложение.

1 Глава. Корпусная лингвистика. Разработка корпуса политических статей

- Корпусная лингвистика и её применение в области преподавания иностранного языка
- Методология создания корпуса текстов
  - Автоматизированное извлечение текстов для корпуса политических статей с сайта Lenta.ru
  - Обработка и разметка полученной коллекции текстов. Грамматический парсер MYSTEM (библиотека «pymystem3» для языка программирования Python)

Пример разметки статьи из корпуса:


```
<?xml version="1.0"?>
- <articles>
  - <atricle>
      <url>/news/2016/05/04/caralone/</url>
      <title>В Госдуме предложили ужесточить наказание за оставление детей в машине</title>
      <date> 16:33, 4 мая 2016</date>
      <author>lenta.ru</author>
      <body>В Госдуме поддержали идею об ужесточении наказания для родителей, оставляющих своих дет
        комитета Госдумы по вопросам семьи, женщин и детей Ольга Красильникова, передает агентство г
        бороться путем введения штрафов. При этом Красильникова не считает возможным помещать винс
        несовершеннолетним. «Я не хочу, чтобы родителей сажали на 15 суток, это принесет вред ребенку
        ощутимая сумма», — отметила парламентарий.Красильникова также сообщила, что в настоящее вр
        Кодекс об административных правонарушений (КоАП РФ) с тем, чтобы оставление ребенка в автом
        жизни». Депутат подчеркнула, что поддерживает эту инициативу.Согласно действующему законод
        родителями или иными законными представителями несовершеннолетних обязанностей по содерж
        ответственность в виде предупреждения или штрафа от 100 до 500 рублей.3 мая сотрудники моско
        автомобиля, припаркованного в запрещенном месте в районе ВДНХ. Родителей ребенка удалось на
        президенте России Павел Астахов заявлял о необходимости наказывать родителей, оставляющих д
      <lemmas>в госдума поддерживать идея об ужесточение наказание для родитель, оставлять свой ребен
        госдума по вопрос семья, женщина и ребенок ольга красильникова, передавать агентство городско
        введение штраф. при это красильникова не считать возможный помещать виновный под администр
        хотеть, чтобы родитель сажать на 15 сутки, это приносить вред ребенок. вначале нужно ограничив
        отмечать парламентарий.красильников также сообщать, что в настоящий время мвд россия разраб
        административный правонарушение (коап рф) с то, чтобы оставление ребенок в автомобиль «стан
        депутат подчеркивать, что поддерживать этот инициатива.согласно действующий законодательств
        или иной законный представитель несовершеннолетний обязанность по содержание и воспитание
        предупреждение или штраф от 100 до 500 рубль.3 май сотрудник московский служба эвакуация об
        запрещать место в район вднх. родитель ребенок удаваться находить лишь спустя час.лето 2015 го
        необходимость наказывать родитель, оставлять ребенок в возраст до шесть год один. </lemmas>
      <analyze>[{"text": "В", "analysis": [{"gr": "PR=", "lex": "в"}]}, {"text": " "}, {"text": "Госдуме", "analys
        "поддержали", "analysis": [{"gr": "V,пе=прош,мн,изъяв,сов", "lex": "поддерживать"}]}, {"text": "
        " "}, {"text": "об", "analysis": [{"gr": "PR=", "lex": "об"}]}, {"text": " "}, {"text": "ужесточении", "a
        "наказания", "analysis": [{"gr": "S,сред,неод=(вин,мн|род,ед|им,мн)", "lex": "наказание"}]}, {"tex
        {"text": "родителей", "analysis": [{"gr": "S,муж,од=(вин,мн|род,мн)", "lex": "родитель"}]}, {"text"
        (непрош,пр,мн,прич,полн,несов,действ|непрош,род,мн,прич,полн,несов,действ|непрош,вин,мн,при
        "analysis": [{"gr": "APRO=(пр,мн|род,мн|вин,мн,од)", "lex": "свой"}]}, {"text": " "}, {"text": "детей
        {"text": "без", "analysis": [{"gr": "PR=", "lex": "без"}]}, {"text": " "}, {"text": "присмотра", "analysi
        "analysis": [{"gr": "PR=", "lex": "в"}]}, {"text": " "}, {"text": "машине", "analysis": [{"gr": "S,жен,н
```

Глава 2. Проектирование и разработка и упражнения для изучения политических терминов

- Проектирование упражнения. Создание списка основных политических терминов. Поиск и разметка терминов в полученном корпусе
- Разработка упражнения. Метод «расстояние Левенштейна или Дамерау – Левенштейна» как основа для упражнения

Распределение статей по категориям:

Вертикаль власти

Власть, Государство, Презеденг, Правительство, Премье-министр, Государственная Дума, Депутат, Армия, Милиция, Губернатор, Мэр, Чиновник

Articles »

Процедура выборов

Выборы, Кандидат, Предвыборная программа, Агитация, Партия, Оппозиция, Избиратель

Articles »

Полномочия власти

Закон, Конституция, Контроль, Бюджет, Оборона, Свобода слова, Цензура

Articles! »

Образ власти

Патриотизм, Безпасность, Социальная защита, Привелегии, Коррупция, Бюрократия, Олигархия

Articles! »

Глава 3. Создание и ввод в эксплуатацию сайта, как платформы для веб-приложения

1)Проектирование и разработка сайта

- Средства разработки

Язык гипертекстовой разметки HTML

Язык разметки XML

Каскадные таблицы стилей CSS

- Веб-дизайн, создание интерфейса

Шаблон Bootstrap3

- Адаптация страниц под все виды мониторов (включая мобильную версию) и браузеры

2)Создание веб-приложения с помощью микрофреймворка Flask

3)Структура сайта

4)Создание веб-сервера

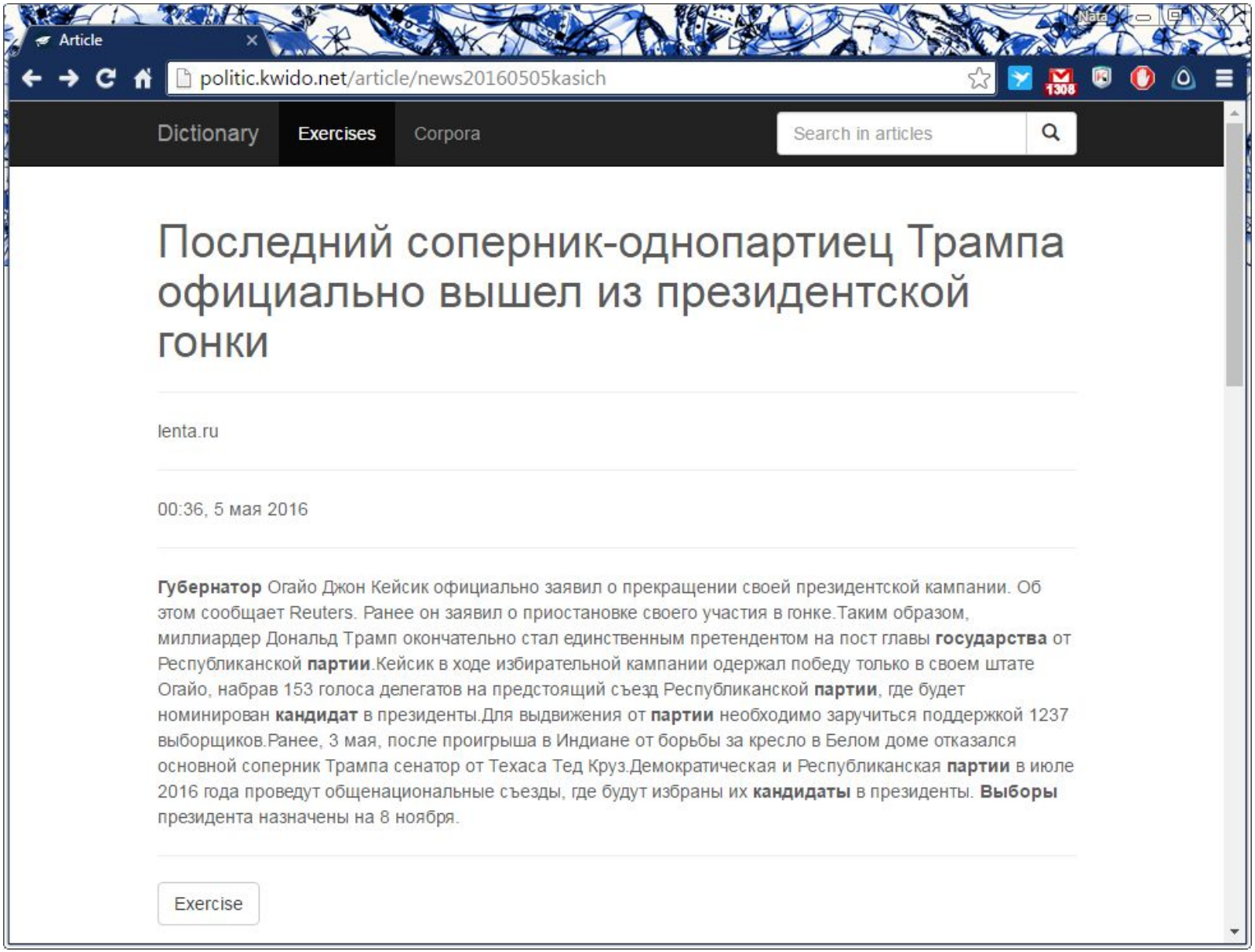
Адаптация сервера под файлы с расширением «.py» (программы на языке Python)

Количество статей, список статей категории «Процедура выборов»

Articles count:565
Вертикаль власти: 422
Процедура выборов: 200
Полномочия власти: 246
Образ власти: 33

- США и Россия договорились распространить режим прекращения огня на Алеппо
- Самарский губернатор отказался от участия в праймериз «Единой России»
- Бывшая глава Петрозаводска собралась вернуть кресло мэра через Верховный суд
- Миронов остался лидером «Справедливой России»
- Мадуро запустил военные учения по отражению иностранной агрессии
- В Госдуме предложили укоротить весеннюю сессию на три недели
- Сеул ответил на предложение КНДР о возобновлении военных переговоров
- Трамп согласился в виде исключения пустить лондонского мэра-мусульманина в США
- Исследование указало на поддержку Путина сторонниками оппозиции
- Распил Ленина

Пример упражнения:



[0] Огайо Джон Кейсик официально заявил о прекращении своей президентской кампании. Об этом сообщает Reuters. Ранее он заявил о приостановке своего участия в гонке.Таким образом, миллиардер Дональд Трамп окончательно стал единственным претендентом на пост главы **[6]** от Республиканской **[1]** .Кейсик в ходе избирательной кампании одержал победу только в своем штате Огайо, набрав 153 голоса делегатов на предстоящий съезд Республиканской **[2]** , где будет номинирован кандидат в президенты.Для выдвижения от **[3]** необходимо заручиться поддержкой 1237 выборщиков.Ранее, 3 мая, после проигрыша в Индиане от борьбы за кресло в Белом доме отказался основной соперник Трампа сенатор от Техаса Тед Круз.Демократическая и Республиканская **[4]** в июле 2016 года проведут общенациональные съезды, где будут избраны их **[7]** в президенты. **[5]** президента назначены на 8 ноября. Verify

[0] Губернатор Огайо Джон Кейсик официально заявил о прекращении своей президентской кампании. Об этом сообщает Reuters. Ранее он заявил о приостановке своего участия в гонке.Таким образом, миллиардер Дональд Трамп окончательно стал единственным претендентом на пост главы **[6]** от Республиканской **[1]** партия .Кейсик в ходе избирательной кампании одержал победу только в своем штате Огайо, набрав 153 голоса делегатов на предстоящий съезд

[0] Count of errors: **0**
[1] Count of errors: **1**
[2] Count of errors: **6**
[3] Count of errors: **6**

Заключение:

- Целью написания данной выпускной квалификационной работы являлось создание и введение в эксплуатацию упражнения, основанного на размеченном корпусе политических статей.
- В ходе исследования, возникли небольшие проблемы с разметкой и нумеровкой слов в тексте, но в большинстве случаев программа срабатывает правильно.
- Дальнейшие перспективы исследования состоят в развитии сайта (создание регистрации для сохранения достижений, обратной связи с преподавателем и онлайн-переводчик терминов), создание новых упражнений на базе полученного корпуса. Также, данное приложение позволяет увеличить количество и расширить тематику изучаемых терминов.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!