Процесс: “Защита проекта”

Завершено на 0%

Защита презентации начата, оставшееся время: ~10 минут

Ожидайте завершения процесса

Реализация демо-версии корпуса с возможностью поиска

Лабенская Яна

2020

Что за корпус?

- Новости газеты Московский Комсомолец, раздел “Происшествия”
- Предположительно, топикально грабежи, смерть и прочая криминальщина

Кто собирал?

- Модули `requests` и `fake_useragent`
- Трудолюбивый краулер: 120 новостей с первых 4 страниц ленты

Как предобработан корпус?

Сначала тексты и заголовки просто заливаются в датафрейм…

titles	full_texts
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...
Девочка сбежала от матери, делавшей из нее рабыню	Наверное , у каждого были прадедушки и прабабу...
Медбрат украл аппарат УЗИ у больных коронавирусом	Как стало известно « МК » , пропажу аппарата У...
Составлен рейтинг мошенничеств с коронавирусом	Лже-медики . Облапошивание начинается со звонк...
Приговор обрадовал депутата-коммуниста Шеремет...	Ранее в ходе прений прокурор потребовал для Ше...

...потом тексты дробятся на предложения по точкам...

titles	full_texts
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...
Раскрыта тайна препарата в организме утонувшей...	Девушка сразу направилась в бассейн , где око...
Раскрыта тайна препарата в организме утонувшей...	В бессознательном состоянии Софию вытащили из...
Раскрыта тайна препарата в организме утонувшей...	Смерть дочери знаменитого актера получила шир...
Раскрыта тайна препарата в организме утонувшей...	Например , у родственников погибшей возникли ...

...потом предложения дробятся по словам...

titles	full_texts	words
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	Напомним
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	,
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	23-го
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	сентября
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	вечером
...	...	...
Футболисты забили до смерти одноклубника из-за...	Ранее сообщалось , что во время футбольного м...	убила
Футболисты забили до смерти одноклубника из-за...	Ранее сообщалось , что во время футбольного м...	10
Футболисты забили до смерти одноклубника из-за...	Ранее сообщалось , что во время футбольного м...	игроков
Футболисты забили до смерти одноклубника из-за...	Ранее сообщалось , что во время футбольного м...	
Футболисты забили до смерти одноклубника из-за...		

...ПОДКЛЮЧАЕТСЯ АНАЛИЗАТОР...

titles	full_texts	words	lemma	form	POS
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	Напомним	напомнить	VERB,perf,tran plur,1per,futr,indc	VERB
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	,	,	PNCT	None
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	23-го	23-го	NOUN,inan,neut,Fixd sing,nomn	NOUN
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	сентября	сентябрь	NOUN,inan,masc sing,gent	NOUN
Раскрыта тайна препарата в организме утонувшей...	Напомним , 23-го сентября вечером дочь артиста...	вечером	вечером	ADVB	ADVB

...удаляются знаки препинания и пробелы и задаются новые индексы.

titles	full_texts	words	lemma	form	POS	ind
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	которые	который	ADJF,Apro,Subx,Anph plur,nomn	ADJF	900
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	в	в	PREP	PREP	901
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	свою	свой	ADJF,Apro,Anph femn,sing,accs	ADJF	902
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	очередь	очередь	NOUN,inan,femn sing,accs	NOUN	903
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	тоже	тоже	ADVB	ADVB	904

Как размечен корпус?

Довольно просто!
Он представляет собой файл формата csv из шести колонок: заголовок статьи, полные предложения в статье, слова, их леммы, их теговый разбор, их часть речи и индекс.

titles	full_texts	words	lemma	form	POS	ind
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	которые	который	ADJF,Apro,Subx,Anph plur,nomn	ADJF	900
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	в	в	PREP	PREP	901
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	свою	свой	ADJF,Apro,Anph femn,sing,accs	ADJF	902
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	очередь	очередь	NOUN,inan,femn sing,accs	NOUN	903
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	тоже	тоже	ADVB	ADVB	904
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	не	не	PRCL	PRCL	905
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	получат	получить	VERB,perf,tran plur,3per,futr,indc	VERB	906
Девочка сбежала от матери, делавшей из нее рабыню	Вся забота о ней будет пока полностью на плеч...	никаких	никакой	ADJF,Apro plur,gent	ADJF	907

Кто разметил корпус?

PyMorphy2!

Он умеет работать с такими тегами:

NOUN имя существительное
ADJF имя прилагательное (полное)
ADJS имя прилагательное (краткое)
COMP компаратив
VERB глагол (личная форма)
INFN глагол (инфинитив)
PRTF причастие (полное)
PRTS причастие (краткое)
GRND деепричастие
NUMR числительное
ADVB наречие
NPRO местоимение-существительное
PRED предикатив
PREP предлог
CONJ союз
PRCL частица
INTJ междометие

Пример разметки

words	lemma	form	POS	ind
которые	который	ADJF,Apro,Subx,Anph plur,nomn	ADJF	900

Как обрабатываются запросы?

Возможные запросы:

Слово без кавычек	Трактуется как запрос леммы

```
Что ищем? Не более трех слов в запросе! смерть
['смерть']
1
вы запрашиваете поиск по лемме
Это слово встретилось в предложении: Смерть дочери знаменитого актера получила широкий резонанс из-за массы несостыковок и пр
отиворечивых данных об обстоятельствах , предшествующих трагедии . | которое было в статье: Раскрыта тайна препарата в органи
зме утонувшей дочери Конкина
```

Слово в двойных кавычках	Трактуется как запрос конкретной формы

```
Что ищем? Не более трех слов в запросе! "смерти"
['"смерти"']
1
вы запрашиваете слова этой леммы в этой форме
Это слово встретилось в предложении: Признаков насильственной смерти при осмотре тела судмедэксперт не обнаружил . | которое
было в статье: Прояснились детали гибели московской студентки в яме для фонаря
```

Возможные запросы:

Слово без кавычек+тег части речи	Трактуется как запрос этой леммы в конкретной части речи

```
Что ищем? Не более трех слов в запросе! смерть+NOUN
['смерть+NOUN']
1
вы запрашиваете лемму с конкретной частью речи
Это слово встретилось в предложении: Смерть дочери знаменитого актера получила широкий резонанс из-за массы несостыковок и пр
отиворечивых данных об обстоятельствах , предшествующих трагедии . | которое было в статье: Раскрыта тайна препарата в органи
зме утонувшей дочери Конкина
```

Тег части речи	Трактуется как запрос всех слов этой части речи

```
Что ищем? Не более трех слов в запросе! ADJF
['ADJF']
1
вы запрашиваете все слова по тегу этой части речи
Это слово встретилось в предложении:Напомним , 23-го сентября вечером дочь артиста София Конкина пришла в фитнес клуб в Пресн
енском районе , что неподалёку от съёмной квартиры , в которой последние несколько месяцев она жила с дочерью и своим кавалер
ом , водителем по профессии Михаилом . | которое было в статье: Раскрыта тайна препарата в организме утонувшей дочери Конкина
```

Принцип поиска

Для первой части запроса создается новый датафрейм, содержащий те предложения из старого, в которых был найден результат запроса.

Если запрос состоит из одного слова, дальше ничего не происходит - просто распечатываются эти предложения и заголовки их статей

Если в запросе более одного слова, новый датафрейм становится областью поиска для следующего токена запроса, а индексы конкретных слов, отвечающих запросу, сохраняются.

```
main()
```

```
Что ищем? Не более трех слов в запросе! футболист
['футболист']
1
вы запрашиваете поиск по лемме
Это слово встретилось в предложении: Между футболистами снова завязалась драка , в ходе которой Овачи потерял сознание . | кото
рое было в статье: Футболисты забили до смерти одноклубника из-за ошибки на поле
```

Принцип поиска

Для второго токена запроса происходит поиск только по тем предложениям, в которых уже встретился первый.
Если найдено совпадение, проверяются индексы: если из индексов следует, что результаты первого и второго запроса стоят рядом, индекс сохраняется.
Итоговая выдача второго запроса - еще более суженный датафрейм и список его предложений, плюс список индексов, где они встретились с результатом первого.
Если токена всего два, на этом все заканчиваются и выдаются предложения со словами с этими индексами; если больше, новый датафрейм становится областью поиска для третьего токена.

```
main()
```

```
Что ищем? Не более трех слов в запросе! программа "первого"
['программа', '"первого"']
2
вы запрашиваете поиск по лемме
вы запрашиваете слова этой леммы в этой форме
Эти слова встретилось в предложении: А меж тем Яна уже прошла программу первого класса . | которое было в статье: Девочка сбежа
ла от матери, делавшей из нее рабыню
```

Принцип поиска

Третий токен ищется в датафрейме предложений, в которых уже встретились первый и второй: если найдено совпадение, его индекс сравнивается с индексом совпадения для первого и второго токена.

Выдача - распечатка предложений, где все индексы сошлись, и заголовки статей, откуда они взяты.

```
main()
```

```
Что ищем? Не более трех слов в запросе! сентябрь вечером NOUN
['сентябрь', 'вечером', 'NOUN']
3
вы запрашиваете поиск по лемме
вы запрашиваете поиск по лемме
вы запрашиваете все слова по тегу этой части речи
Эти слова встретились в предложении:Напомним , 23-го сентября вечером дочь артиста София Конкина пришла в фитнес клуб в Преснен
ском районе , что неподалёку от съёмной квартиры , в которой последние несколько месяцев она жила с дочерью и своим кавалером ,
водителем по профессии Михаилом , которое было в статье: Раскрыта тайна препарата в организме утонувшей дочери Конкина
```

Принцип поиска

Четвертый токен?

```
main()
```

```
Что ищем? Не более трех слов в запросе! в "свою" очередь ADVB
['в', '"свою"', 'очередь', 'ADVB']
К сожалению, я не обрабатываю запросы длиннее трех токенов
```

Нельзя!

Как тестировалось?

По принципу “бей палкой, пока не задергается”

Тестирование

- Ввод неизвестных слов:
  - Извещение об отсутствии их в словаре
- **Ввод знаков препинания:**
  - Проблема! На следующем слайде обсудим
- Ввод конкурирующих или неправильных запросов (“NOUN”, “”, гибель+РВЛВ, превышение количества токенов, **ввод пустого запроса**)
- Ввод капсом

Минусы:

- Не слишком-то и юзер-френдли: программа не подскажет пользователю, как именно он может формировать запросы и в каком синтаксисе она его поймет.
- Выделить бы как-то слова в предложениях…
- Головой вперед вбежали в проблему того, чтобы делить предложения только по точкам: остались вопросительные и восклицательные, и он ищет по ним как по леммам
- Частично чувствительно к регистру: маленькими буквами тег не узнается

Ссылки на материалы

Ссылка на тетрадку:

https://github.com/daneelsteel/PROJECT/blob/main/Yana%20(2).ipynb

Ссылка на базу:

https://github.com/daneelsteel/PROJECT/blob/main/clean_textbase_parsed.csv

Процесс: “Защита проекта”

Завершено на 100%

Презентация завершена!

Возможные действия:

- Задать вопрос: while question:
  Яна.ask(question)
- Ободряюще похлопать: Яна.clap()
- Отпустить с миром: Яна.sleep()