### Московский авиационный институт

**Факультет прикладной математики и физики**

**Лабораторная работа №1**

**по курсу:**

**«Информационный поиск»**

**по теме:**

**«Формирование текстового корпуса»**

**2 семестр**

Студент: Ахмед С. Х.

Преподаватель: Калинин А. Л.

Группа: 8О-106М

**Москва, 2019 г**

Постановка задачи

Необходимо подготовить корпус документов, который будет использован при выполнении остальных лабораторных работ:

• Скачать его к себе на компьютер. В отчёте нужно указать источник данных.

• Ознакомиться с ним, изучить его характеристики. Из чего состоит текст? Есть ли дополнительная мета-информация? Если разметка текста, какая она?

• Разбить на документы.

• Выделить текст.

• Найти существующие поисковики, которые уже можно использовать для поиска по выбранному набору документов (встроенный поиск Википедии, поиск Google с использованием ограничений на URL или на сайт). Если такого поиска найти невозможно, то использовать корпус для выполнения лабораторных работ нельзя!

• Привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче.

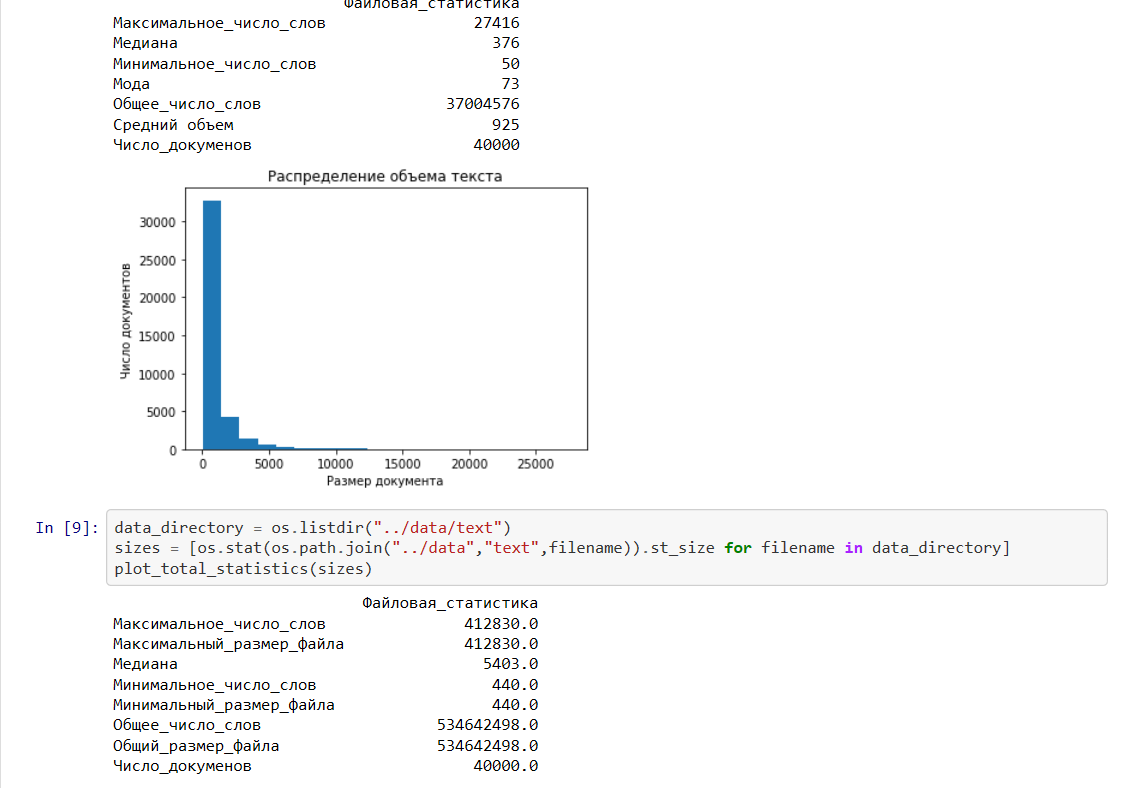
Оборудование:

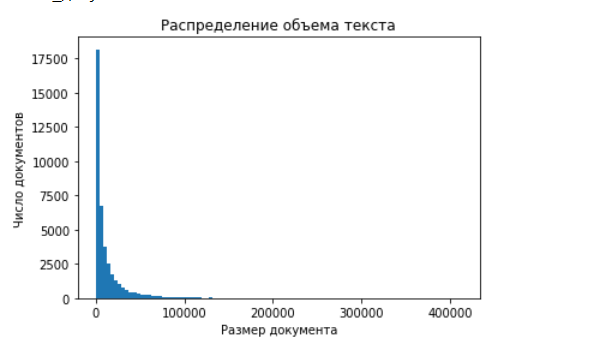
Компьютер HP Omen 15 под управлением операционной системы Windows 10, Intel Core i5-7300HQ 2.50 GHz, 12 Gb RAM

Программное обеспечение

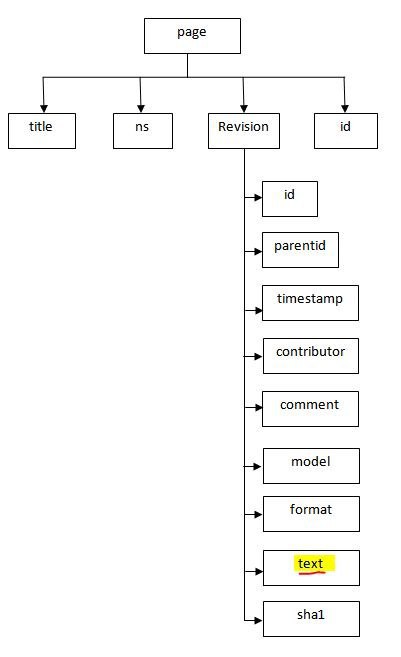
|  |  |
| --- | --- |
| Язык программирования | Python 3.6 |
| Среда программирования | Anaconda, Jupyter Notebook |
| Использованные сторонние модули | Re,nltk |

Статистические данные по полученному результату





Структура одной статьи из Wiki Dump



Тематика и часть текста

В качестве тематики была выбрана выборка статей русскоязычной википедии периода 20.11.2018. Пример текста:

'литва официальное название лито вская респу блика государство расположенное северной европе одна из стран балтии столица страны вильнюс площадь км² протяжённость севера на юг км запада на восток км население составляет человек по этим показателям является крупнейшим прибалтийским государством имеет выход балтийскому морю расположена на его восточном побережье береговая линия составляет всего км наименьший показатель среди прибалтийских государств на севере граничит латвией на юго востоке белоруссией на юго западе польшей калининградской областью россии член оон года ес нато года оэср мая года входит шенгенскую зону еврозону географические данные карта литвы поверхность равнинная со следами древнего оледенения поля луга занимают территории леса кустарники болота внутренние воды высшая точка над уровнем моря холм аукштояс или аукштасис калнас юго восточной части страны км от вильнюса крупнейшие реки неман вилия более тыс озёр территории крупнейшее из них друкшяй на границе латвии литвы белоруссии площадь км² самое глубокое таурагнас самое длинное асвея длина км местечка дубингяй климат переходный от морского средняя температура зимой минус летом выпадает мм осадков год полезные ископаемые торф минеральные материалы строительные материалы история древнейшая история территория современной литвы была заселена людьми конца ix тысячелетия до жители занимались охотой рыболовством использовали лук стрелы кремнёвыми наконечниками скребки для обработки кожи удочки сети конце неолита iii ii тыс до на территорию современной литвы проникли индоевропейские племена они занимались земледелием скотоводством при этом охота рыболовство оставались основными занятиями местных жителей вплоть до широкого распространения железных орудий труда индоевропейцы заселившие земли между устьями вислы западной двины выделились отдельную группу названную учёными балтами традиционно считается что этническая основа литвы сформирована носителями культуры курганов веке на территории современных восточной литвы северо западной белоруссии около vii века литовский язык отделился от латышского зарождение государства становление на территории современной литвы относят xiii веку при этом само название литва впервые упомянуто кведлинбургских анналах под годом сообщении об убийстве язычниками миссионера бруно на границе руси литвы по наиболее версии топоним возник от названия небольшой реки летавка притока няриса согласно более современной гипотезе название страны могло произойти от этнонима леты или лейти которым жители окрестных земель называли дружинников литовских князей начале xiii века земли балтов язычников запада началось вторжение немецких рыцарей крестоносцев они покорили пруссию ливонию это же время юга началась экспансия галицко волынского княжества середине xiii века многие литовские земли были объединены под властью князя миндовга принявшего году католическое крещение коронованного году через несколько лет миндовг отрёкся от христианства до начала xiv века литовские земли оставались языческими несмотря на то что уже году миндовг был свергнут его правление положило начало более чем пятисотлетнему существованию великого княжества литовского великое княжество литовское карта границ великого княжества литовского речи посполитой литовской республики xiv начале xv веках территория великого княжества литовского стремительно росла основном за счёт присоединения земель западной руси включение состав государства славянских земель многократно превышающих по площади количеству населения собственно литовские земли привело перениманию литовскими князьями получившими во владение русские земли православной культуры западнорусского языка со временем западнорусский язык стал официальным языком канцелярии великих князей собственно литовский язык до xvi века оставался бесписьменным хотя продолжал использоваться на этнически литовских землях году великий князь литовский ягайло заключил кревскую унию королевством польским по условиям унии ягайло обязался присоединить великое княжество литовское королевству польскому крестить литовские земли по католическому обряду сам становился королём польши сохранял титул великого князя литовского однако вскоре он вынужден был уступить власть великом княжестве литовском своему двоюродному брату витовту витовт хотя признал себя вассалом ягайло проводил самостоятельную внешнюю политику таким образом полное объединение государств не состоялось годы правления витовта великое княжество литовское достигло наивысшего расцвета его территория составила примерно тысяч км² этот же период литовские князья из династии гедиминовичей вели тяжёлую борьбу тевтонским орденом который был разбит году грюнвальдской битве объединёнными войсками великого княжества литовского королевства польского году состав великого княжества литовского окончательно вошла жемайтия великий князь казимир одновременно бывший королём польским расширил влияние династии ягеллонов подчинил пруссию посадил своего сына на чешский венгерский троны годах существовала политическая система государств ягеллонов охватывавшая польшу вассалами пруссией молдавским княжеством великое княжество литовское чехию венгрию правовой основой государства являлся стату изданный трёх редакциях отражающих социально экономические политические изменения статут регламентировал вопросы гражданского уголовного процессуального права на территории великого княжества третья редакция статута действовала до года речи посполитой году люблине была заключена новая уния польшей результате которой образована речь посполитая согласно акту люблинской унии литвой польшей правил совместно избираемый король государственные дела решались общем сейме однако правовые системы армия чиновники оставались раздельными xvi xviii веках литве по польскому образцу сложилась политическая система известная как шляхетская демократия она наличием широких прав шляхты дворянства управлении государством одновременно этим происходила полонизация шляхты выраженная перенимании правящим сословием великого княжества литовского польского языка культуры идентичности на сословия полонизация столь значительного влияния не оказала составе российской империи xviii веке результате опустошительных войн всеобъемлющего кризиса речь посполитая пришла упадку попала под влияние российской империи годах состоялись разделы речи посполитой между россией пруссией австрией почти вся территория бывшего великого княжества литовского была присоединена российской империи попытках восстановить польско литовское дворянство приняло сторону наполеона году также неоднократно поднимало восстания которые однако окончились поражением стремлении ликвидировать польское влияние литве российские власти предприняли широкую кампанию деполонизации русификации году была запрещена литовская печать латиницей литовское население особенно католическое духовенство сопротивлялись русификации кирилличные издания игнорировали книги напечатанные латиницей книгоноши нелегально ввозили из соседней пруссии году запрет на литовскую латиницу был отменён во время первой мировой войны файл rzeczpospolita lithuania claims png thumb right px литва польша после первой мировой войны похода желиговского на вильно начавшаяся первая мировая война быстро по территории литвы концу года все этнически литовские земли германией литовцы потеряли все политические права вначале даже запрещены литовские периодические издания однако литовская интеллигенция попыталась воспользоваться геополитической ситуацией начала искать возможности для восстановления независимости литвы сентября года вильнюсе была проведена литовская конференция во время которой была избрана литовская тариба совет литвы ходе конференции было принято решение необходимости создания независимого литовского государства этнографических границах со столицей вильнюсе председателем совета был избран антанас сметона декабря года было провозглашено восстановление литовского государства марта года император вильгельм ii признал независимость литвы на основании акта признании литовской тариба была преобразована государственный совет литвы июля года государственный совет принял решение установить литве конституционную монархию предложить вюртембергскому принцу вильгельму фон ураху корону впрочем ноября года это решение было отозвано были приняты основные положения временной конституции литвы ноября года президиум совета утвердил первое временное правительство литвы из шести министров под руководством аугустинаса вольдемараса тем самым дав начало созданию аппарата литвы литовская республика после ухода основных немецких частей начала боёв большевистской красной армией декабря года была образована литовская советская республика февраля года вильнюсе состоялось объединённое заседание циков литвы белоруссии где было провозглашено образование литовско белорусской сср литбел феврале марте года войска литовской тарибы поддержанные немецкими гарнизонами начали военные действия против литбела апреле года ним присоединилась польская армия результате территория литбела была занята польскими частями для борьбы польшей советской россии требовался нейтралитет литвы для чего июля года москве был заключён советско литовский договор литбел прекратил своё существование советская россия признала независимость литвы передачу ей спорного виленского края после поражения красной армии под варшавой советского отступления польские части под командованием генерала люциана желиговского инсценировали мятеж якобы самовольно заняли территорию виленского края октября года было объявлено создании на территории края государства срединная литва однако уже году оно вошло состав польской республики качестве воеводства литовские власти продолжали считать вильнюс столицей литвы хотя фактически руководство осуществлялось из каунаса году литве введена должность президента первым президентом государства был избран антанас сметона мая года состоялось первое заседание демократически избранного учредительного собрания году страна была принята лигу наций году была принята постоянная конституция приведены реформы области земельных ресурсов финансов образования введена литовская валюта лит открыт литовский университет клайпедский край мемельланд населённый основном прусскими литовцами немцами по решению лиги наций находился под временным управлением французской администрации году результате восстания местных литовцев при негласном участии литовской полиции клайпедский край был присоединён литве на правах автономии французская администрация не предприняла никаких шагов для борьбы восстанием февраля года страны антанты признали присоединение клайпедского края литве декабре года литве произошёл военный переворот вернувший власти лидера националистов антанаса сметону началась так называемая авторитарная фаза управления государством году была принята конституция расширяющая президентские полномочия оппозиционные партии были запрещены цензура ужесточена права национальных меньшинств урезаны марта года польша предъявила литве ультиматум требованием признать виленский край неотъемлемой частью польского государства год спустя марта года литва получила ультиматум германии требованием вернуть ей клайпедский край оба ультиматума литва была вынуждена принять вторая мировая война присоединение ссср согласно секретному протоколу заключённому августе года пакту молотова риббентропа литва была включена сферу интересов германии сентября германия начала вторжение польшу сентября ссср осуществил вторжение по итогам которого присоединил восточные земли польши том числе вильно сентября ссср инициировал переговоры об отказе германии от претензий на литву обмен на территории варшавского люблинского воеводств польши октября года москве был подписан договор передаче литовской республике города вильно виленской области взаимопомощи между советским союзом литвой.

Стоить заметить, при парсинге часть разметки осталась, так же в качестве артефакта парсинга можно заметить сплит некоторых слов.

В качестве основного поисковика для данного поисковика принимается внутренний поисковик Wikipedia, но так же подходит Google.

В случае с поиском Wiki, их алгоритм стремится выдать конкретную статью свзяанную с введеным названием, что в некоторых случаях не то, что мы ожидаем увидеть(особенно если помним, что это слово входит в часть статьи).

В случае с поиском Google , он выдает отранжированный список статей, но статья из википедии, которой принадлежит термин может не попасть в топ лист раздаточной ленты, что является своего рода проблемой