ЛР1: Токенизация

Задание

Реализовать процесс разбиения текстов документов на токены, который потом будет использоваться при индексации. Для этого потребуется выработать правила, по которым текст делится на токены. Необходимо описать их в отчёте, указать достоинства и недостатки выбранного метода. Привести примеры токенов, которые были выделены неудачно, объяснить, как можно было бы поправить правила, чтобы исправить найденные проблемы.

Метод решения

1. Изучение способов токенизации текста с использованием разных библиотек python.
2. Написание и отладка кода, выполняющего разделение текста на отдельные токены.
3. Сбор и анализ статистических данных.

В качестве инструмента по токенизации текста была выбрана библиотека nltk, которая обладает словарем стоп-слов русского языка. Дополнительно текст был очищен он мусорных знаков: знаков пунктуации, знаков операций и других ascii символов, которые не представляют буквы русского алфавита.

Результаты выполнения

Файлы	Размер до токенизации	Размер после токенизации	Средняя длина токена	Время
все статьи в формате json	11Gb	9.9Gb	6.33	1:29:40

Исходные данные

{"id": "7", "url": "https://ru.wikipedia.org/wiki?curid=7", "title": "Литва", "text": "Литва\n\nЛитва́ ( ), официальное название — Литóвская Респýблика () — государство, расположенное в северной части Европы. Площадь — км². Протяжённость с севера на юг — 280 км, а с запада на восток — 370 км. Население составляет человек (январь, 2021). Занимает 137-е место в мире по численности населения и 121-е по территории. Имеет выход к Балтийскому морю, расположена на его восточном побережье. Береговая линия составляет всего 99 км (наименьший показатель среди государств Балтии). На севере граничит с Латвией, на юго-востоке — с Белоруссией, на юго-западе — с Польшей и Калининградской областью России. По площади и населению является самым крупным прибалтийским государством.\n\nСто-ли-ца — Вильнюс. Официальный язык — ли-тов-ский. Де-неж-ная еди-ни-ца — евро.\n\nВосстановление независимости страны провозглашено 11 марта 1990 года. 6 сентября 1991 года Государственный совет СССР признал независимость Литвы. \n\nЛитва — член ООН (1991), ОБСЕ (1991), Со-ве-та Ев-ро-пы (1993), ВТО (2001), Европейского союза (2004), НАТО (2004) и ОЭСР (2018). Входит в Шенгенскую зону и Еврозону.\n\nЭтимология слова «Литва» точно не известна, при этом существует множество версий, ни одна из которых не получила всеобщего признания. Корень «лит» и его варианты «лет»/«лют» допускают различные толкования как в балтских и славянских, так и в других индоевропейских языках. Так, например, существуют созвучные топонимы на территории Словакии «\"Lytva\"» и Румынии «\"Litua\"», известные с XI–XII веков. По мнению Е. Поспелова, топоним образован от древнего названия реки Летава (Lietavà от «лить», русское «Летаука»). Феодальное княжество, по землям которого протекала эта река, со временем заняло ведущее положение и название было распространено на всё государство. В «Повести временных лет» (XII век) упоминается этноним «литва», полностью совпадающий с названием местности «Литва» и по смыслу (территория, где живёт литва), и по форме.\n\nПоверхность — равнинная, со следами древнего оледенения. Поля и луга занимают 57 % территории, леса и кустарники — 30 %, болота — 6 %, внутренние воды — 1 %.\n\nВысшая точка — 293,84 м над уровнем моря — холм Аукштояс (или Аукштасис калнас) в юго-восточной части страны, в 23,5 км от Вильнюса.\n\nКрупнейшие реки — Неман и Вилия.\nБолее 3 тыс. озёр (1,5 % территории): крупнейшее из них — Друкшяй на границе Латвии, Литвы и Белоруссии (площадь 44,8 км²), самое глубокое — Таурагнас, 61 м), самое длинное — Асвея длиной в 30 км у местечка Дубингяй.\n\nКлимат переходный от морского к континентальному. Средняя температура зимой −5 °C, летом +17 °C. Выпадает 748 мм осадков в год.\n\nПолезные ископаемые: торф, минеральные материалы, строительные материалы.\n\nТерритория современной Литвы была заселена людьми с конца X–IX тысячелетия до н. э. Жители занимались охотой и рыболовством, использовали лук и стрелы с кремнёвыми наконечниками, скребки для обработки кожи, удочки и сети. В конце неолита (III–II тысячелетия до н. э.) на территорию современной Литвы проникли индоевропейские племена. Они занимались земледелием и скотоводством, при этом охота и рыболовство оставались основными занятиями местных жителей вплоть до широкого распространения железных орудий труда. Индоевропейцы, заселившие земли между устьями Вислы и Западной Двины, выделились в отдельную группу, названную учёными балтами.\n\nТрадиционно считается, что этническая основа Литвы сформирована носителями археологической культуры восточнолитовских курганов, сложившейся в V веке н. э. на территории современных Восточной Литвы и Северо-Западной Белоруссии. Около VII века литовский язык отделился от латышского.\n\nСтановление государственности относят к XIII веку, при этом само название «Литва» впервые упомянуто в Кведлинбургских анналах под 1009 годом в сообщении об убийстве язычниками миссионера Бруно на границе Руси и Литвы (— косв.п.). По наиболее распространённой, но аргументированно опровергнутой, версии, топоним возник от названия небольшой реки Летаука, притока Няриса. Согласно более современной гипотезе, название страны могло произойти от этнонима «леты» или «лейти», которым жители окрестных земель называли дружинников литовских князей.\n\nВ начале XIII века в земли балтов-язычников с запада началось вторжение немецких рыцарей-крестоносцев. Они покорили Пруссию и Ливонию. В это же время с юга началась экспансия Галицко-Волынского княжества. К середине XIII века многие литовские земли были объединены под властью князя Миндовга, принявшего в 1253 году католическое крещение и коронованного в 1253 году. Через несколько лет Миндовг отрёкся от христианства и до начала XIV века литовские земли оставались языческими. Несмотря на то, что уже в 1263 году Миндовг был свергнут, его правление положило начало более чем пятисотлетнему существованию Великого княжества Литовского.\n\nВ XIV — начале XV веках территория Великого княжества Литовского стремительно росла, в основном за счёт присоединения земель Западной Руси. Включение в состав государства густо населённые православными русинами обширных территорий, многократно превышающих по площади и количеству населения собственно литовские земли, привело к перениманию литовскими князьями, ставшими наместниками на русских землях, православной культуры, а формирующаяся русинская письменность стала официальным языком документов ВКЛ. Собственно литовский язык до XVI века оставался бесписьменным, хотя и продолжал использоваться на этнически литовских землях.\n\nВ 1385 году великий князь литовский Ягайло заключил Кревскую унию с Королевством Польским. По условиям унии, Ягайло обязался присоединить Великое княжество Литовское к Королевству Польскому и крестить литовские земли по католическому обряду, а сам становился королём Польши и сохранял титул великого князя литовского. Однако вскоре он вынужден был уступить власть в Великом княжестве Литовском своему двоюродному брату Витовту. Последний, хотя и признал себя вассалом Ягайло, проводил самостоятельную внешнюю политику и таким образом полное объединение государств не состоялось. В годы правления Витовта (1392–1430) Великое княжество Литовское достигло наивысшего расцвета, а его территория составила примерно 930 тыс. км².\n\nВ этот же период литовские князья из династии Гедиминовичей вели тяжёлую борьбу с Тевтонским орденом, который был разбит в 1410 году в Грюнвальдской битве объединёнными войсками Великого княжества Литовского и Королевства Польского. В 1422 году в составе великого княжества окончательно вошла Жемайтия, долгое время служившая основным предметом споров с крестоносцами.\n\nВеликий князь Казимир, одновременно бывший и королём польским, расширил влияние династии Ягеллонов — подчинил Пруссию, посадил своего сына на чешский и венгерский троны. В 1492–1526 годах существовала политическая система государств Ягеллонов, охватывавшая Польшу (с вассалами Пруссией и Молдавским княжеством), Великое княжество Литовское, Чехию и Венгрию.\n\nПравовой основой государства являлся статут, изданный в трёх редакциях (1529, 1566, 1588), отражающих социально-экономические и политические изменения. Статут регламентировал вопросы гражданского, уголовного и процессуального права. На территории Великого княжества третья редакция статута действовала до 1840 года.\n\nВ 1569 году в Люблине была заключена новая уния с Польшей, в результате которой образована Речь Посполитая. Согласно акту Люблинской унии Литвой и Польшей правил совместно избираемый король, а государственные дела решались в общем Сейме. Однако правовые системы, армия и чиновники оставались раздельными.\n\nВ XVI–XVIII веках в Литве по польскому"}

Результат

{"id": "7", "url": "https://ru.wikipedia.org/wiki?curid=7", "title": "Литва", "text": "литва литва́ официальное название лито́вская респу́блика государство распоположенное северной части европы площадь км² протяжённость севера юг 280 км запада восток 370 км население составляет человек январь 2021 занимает 137 е место мире численности населения 121 е территории имеет выход балтийскому морю расположена восточном побережье береговая линия составляет 99 км наименьший показатель среди государств балтии севере граничит латвией юго востоке белоруссией юго западе польшей калининградской областью россии площади населению является самым крупным прибалтийским государством сто-ли-ца вильнюс официальный язык ли-тов-ский де-неж-ная еди-ни-ца евро восстановление независимости страны провозглашено 11 марта 1990 года 6 сентября 1991 года государственный совет ссср признал независимость литвы литва член оон 1991 обсе 1991 со-ве-та ев-ро-пы 1993 вто 2001 европейского союза 2004 нато 2004 оэср 2018 входит шенгенскую зону еврозону этимология слова литва точно известна существует множество версий одна которых получила всеобщего признания корень лит варианты лет лют допускают различные толкования балтских славянских других индоевропейских языках например существуют созвучные топонимы территории словакии lytva румынии litua известные xi xii веков мнению е поспелова топоним образован древнего названия реки летава lietavà лить русское летаука феодальное княжество землям которого протекала эта река временем заняло ведущее положение название распространено всё государство повести временных лет xii век упоминается этноним литва полностью совпадающий названием местности литва смыслу территория живёт литва форме поверхность равнинная следами древнего оледенения поля луга занимают 57 территории леса кустарники 30 болота 6 внутренние воды 1 высшая точка 293 84 м уровнем моря холм аукштояс аукштасис калнас юго восточной части страны 23 5 км вильнюса крупнейшие реки неман вилия 3 тыс озёр 1 5 территории крупнейшее друкшяй границе латвии литвы белоруссии площадь 44 8 км² самое глубокое таурагнас 61 м самое длинное асвея длиной 30 км месте чка дубингяй климат переходный морского континентальному средняя температура зимой −5 °c летом 17 °c выпадает 748 мм осадков год полезные ископаемые торф минеральные материалы строительные материалы территория современной литвы заселена людьми конца x ix тысячелетия н э жители занимались охотой рыболовством использовали лук стрелы кремнёвыми наконечникам и скребки обработки кожи удочки сети конце неолита iii ii тысячелетия н э территорию современной литвы проникли индоевропейские племена занимались земледелием скотоводством охота рыболовство оставались основными занятиями местных жителей вплоть широкого распространения железных орудий труда индоевропейцы заселившие земли устьями вислы западной двины выделились отдельную группу названную учёными балтами традиционно считается этническая основа литвы сформирована носителями археологической культуры восточнолитовских курганов сложившейся v веке н э территории современных восточной литвы северо западной белоруссии около vii века литовский язык отделился латышского становление государственности относят xiii веку само название литва впервые упомянуто кведлинбургских анналах 1009 годом сообщении убийстве язычниками миссионера бруно границе руси литвы → косв п наиболее распространённой аргументированно опровергнутой версии топоним возник названия небольшой реки летаука притока няриса согласно современной гипотезе название страны могло произойти этнонима леты лейти которым жители окрестных земель называли дружинников литовских князей начале xiii века земли балтов язычников запада началось вторжение немецких рыцарей крестоносцев покорили пруссию ливонию это время юга началась экспансия галицко волынского княжества середине xiii века многие литовские земли объединены властью князя миндовга принявшего 1251 году католическое крещение коронованного 1253 году несколько лет миндовг отрёкся христианства начала xiv века литовские земли оставались языческими несмотря 1263 году миндовг свергнут правление положило начало пятисотлетнему существованию великого княжества литовского xiv начале xv веках территория великого княжества литовского стремительно росла основном счёт присоединения земель западной руси включение состав государства густо населённые православными русинами обширных территорий многократно превышающих площади количеству населения собственно литовские земли привело перениманию литовскими князьями ставшими наместниками русских землях православной культуры формирующаяся русинская письменность стала официальным языком документов вкл собственно литовский язык xvi века оставался бесписьменным хотя продолжал использоваться этнически литовских землях 1385 году великий князь литовский ягайло заключил кревскую унию королевством польским условиям унии ягайло обязался присоединить великое княжество литовское королевству польскому крестить литовские земли католическом у обряду становился королём польши сохранял титул великого князя литовского однако вскоре вынужден уступить власть великом княжестве литовском своему двоюродному брату витовту последний хотя признал вассалом ягайло проводил самостоятельную внешнюю политику таким образом полное объединение государств состоялось годы правления витовта 1392 1430 великое княжество литовское достигло наивысшего расцвета территория составила примерно 930 тыс км² период литовские князья династии гедиминовичей вели тяжёлую борьбу тевтонским орденом который разбит 1410 году грюнвальдской битве объединёнными войсками великого княжества литовского королевства польского 1422 году состав великого княжества окончательно вошла жемайтия долгое время служившая основным предметом споров крестоносцами великий князь казимир одновременно бывший королём польским расширил влияние династии ягеллонов подчинил пруссию посадил своего сына чешский венгерский троны 1492 1526 годах существовала политическая система государств ягеллонов охватывавшая польшу вассалами пруссией молдавским княжеством великое княжество литовское чехию венгрию правовой основой государства являлся статут изданный трёх редакциях 1529 1566 1588 отражающих социально экономические политические изменения статут регламентировал вопросы гражданского уголовного процессуального права территории великого княжества третья редакция статута действовала 1840 года 1569 году люблине 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 польскому образцу сложилась политическая система известная шляхетская демократия характеризовалась наличием широких прав шляхты дворянства управлении государством одновременно этим происходила полонизация шляхты выраженная перениманиями правящим сословием великого княжества литовского польского языка культуры идентичности непривилегированные сословия полонизация столь значительного влияния оказала xviii веке результате опустошительных войн всеобъемлющего государственного кризиса речь посполитая пришла упадку попала влияние российской империи 1772 1793 1795 годах состоялись разделы речи посполитой россией пруссией австрией вся территория бывшего великого княжества литовского присоединена российской империи попытках восстановить государственность польско литовское дворянство приняло сторону напполеона 1812 году также неоднократно поднимало восстания 1830 1831 1863 1864 которые однако окончились поражением стремлении ликвидировать польское влияние литве российские власти предприняли широкую кампанию деполонизации русификации 1864 году частично запрещена литовская печать латиницей литовское население особенно католическое духовенство оппозиционировались русификации кириллические издания игнорировали книги напечатанные латиницей книгоноши нелегально ввозили соседней пруссии 1904 году запрет литовскую латиницу отменён начавшаяся первая мировая война быстро распространилась территории литвы концу 1915 года этнически литовские земли контролировались германией литовцы потеряли политические права вначале запрещены литовские периодические издания однако литовская интеллигенция попыталась воспользоваться геополитической ситуацией начала искать возможности восстановления независимости литвы 18 22 сентября 1917 года вильнюсе проведена литовская конференция время которой избрана литовская тарыба совет литвы ходе конференции принято решение необходимости создания независимого литовского государства этнографических границах столицей вильнюсе председателем совета избран сметона 11 декабря 1917 года ещё 16 февраля 1918 года провозглашено восстановление литовского государства отличие принятой 11 декабря диктовку германских властей декларации документ 16 февраля говорит полной независимости литвы россии германии однако документ 16 февраля даёт самостоятельность бумаге заключения брест литовского мирного договора германия игнорирует декларацию 16 февраля ссылаясь резолюцию 11 декабря взвешивает возможность создать литовское королевство германским монархом 23 марта 1918 года император вильгельм ii признал независимость литвы основании акта признании литов

Как видно из результата токенизации, не все токены были выделены удачно, к примеру ‘°c’, ‘iii’, ‘→’, знаки ударения все еще присутствуют. Добиться лучших результатов можно более тщательной обработкой текста.

Из достоинств выбранного метода токенизации можно выделить:

- использование готового словаря стоп-слов русского языка

Из недостатков:

- невысокая скорость обработки

Исходный код

```python
import os
import json
from pathlib import Path
from trash import trash
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
nltk.download("stopwords")

russian_stopwords = stopwords.words("russian")

def replacer(text, dic):
    for i, j in dic.items():
        text = text.replace(i, j)
    return text


def preprocess_text(text):
    text = text.lower()
    text = replacer(text, trash)

    tokens = word_tokenize(text, language="russian")
    tokens = [token for token in tokens if token not in russian_stopwords]

    text = " ".join(tokens)
    text = " ".join(text.split())
    return text


if __name__ == '__main__':

    raw_dir_path = "/Users/denis/MAI/IR/lab3/wikipedia_articles_one_by_one"
    raw_dir_list = [(raw_dir_path + "/" +  i) for i in sorted(
        os.listdir(raw_dir_path),
        key=lambda x: int(x.replace(".json", "").replace("wiki_article_", ""))
        )]
    new_path = "/Users/denis/MAI/IR/lab3/wikipedia_articles_tokenized"

    for filepath in raw_dir_list:
        with open(filepath) as file:
            data = json.load(file)
            data["text"] = preprocess_text(data["text"])
            new_name = f"wiki_article_{data['id']}.json"
            with open(new_path + "/" + new_name, "w", encoding='utf8') as new_file:
                json.dump(data, new_file, ensure_ascii=False)

```

Выводы

В процессе выполнения данной лабораторной работы была выполнена токенизация текстов статей википедии с использованием библиотеки nltk на Python. Было интересно поискать готовые средства для предобработки текста и создания токенов. Наличие встроенного и постоянно обновляемого словаря стоп-слов в библиотеке nltk было приятным открытием.