pa - wikidump splitter

June 2, 2019

```
In [1]: # Start logging process at root level
        import logging
        logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.I
        logging.root.setLevel(level=logging.INFO)
```

```
In [2]: total_lines = 1092633438
        chunk_pages = 99999 # 999999: ~2.3GB, 99999: ~300MB, 9999: ~89MB, 999: ~8MB
        chunks_folder = "datasets/chunks/"
        folder_name = chunks_folder+"enwiki-chunks-"+str(chunk_pages)+"/"
        chunk_basename = "enwiki-chunk-"+str(chunk_pages)+"-"
```

```
In [3]: import os
        # Check and create chunk diretory
        if not os.path.exists(chunks_folder):
            print("Chunks folder was not present.")
            os.mkdir(chunks_folder)
        if not os.path.exists(folder_name):
            print("Data chunk folder was not present.")
            os.mkdir(folder_name)
```

```
Data folder was not present.
```

```
In [4]: # Based on:
        # https://stackoverflow.com/questions/6184912/how-to-split-large-wikipedia-dump-xml-bz

        import os
        import bz2
        from timeit import default_timer as timer

        #print("expecting: " +  + " parts")

        def split_xml(filename):
            ''' The function gets the filename of wiktionary.xml.bz2 file as  input and create
            smallers chunks of it in a the diretory chunks
            '''
            # Counters
            pagecount = 0
            filecount = 1
            total_pages = 0
            # open chunkfile in write mode
            chunkname = lambda filecount: os.path.join(folder_name,chunk_basename+str(filecoun
            chunkfile = bz2.BZ2File(chunkname(filecount), 'w')
            # Read line by line
            bzfile = bz2.BZ2File(filename)
            #print(sum(1 for _ in bzfile))
            print("Chunking...")
            start = timer()
            for line in bzfile:
                chunkfile.write(line)
                # the </page> determines new wiki page
                if b'</page>' in line:
                    pagecount += 1
                    total_pages += 1
                if pagecount > chunk_pages:
                    chunkfile.write(b'</mediawiki>') # add end tag
                    end = timer()
                    print(datetime.datetime.now(),":",filecount,"->", round(end - start,2), "se
                    chunkfile.close()
                    pagecount = 0
                    filecount += 1
                    chunkfile = bz2.BZ2File(chunkname(filecount), 'w')
                    start = timer()
                    # add start tag
                    chunkfile.write(b'<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.
            try:
                chunkfile.close()
            except:
                print('Files already close')

            print("Done.")

In [ ]: split_xml('datasets/enwiki-latest-pages-articles.xml.bz2')

In [ ]:
```