pa - build dictionary

June 2, 2019

In [1]:
```
# Start logging process at root level
import logging
log_filename = "logs/pa-build-dictionary.log"
logging.basicConfig(filename=log_filename, format='%(asctime)s : %(levelname)s : %(mes
logging.root.setLevel(level=logging.INFO)
```

In [ ]:
```
# Build dictionary
from gensim import corpora
from gensim.corpora import WikiCorpus
import gensim.downloader as api

datafile_src = "datasets/enwiki-latest-pages-articles.xml.bz2"
#datafile_name = "datasets/chunks/enwiki-chunks-999/enwiki-chunk-999-1.xml.bz2"
#unfiltered_dictionary_name = "dictionaries/enwiki-chunk-999-1_lem.txt.bz2"
#filtered_dictionary_name = "dictionaries/enwiki-filtered-20-10-100000.txt.bz2"
unlemmatized_dictionary_name = "dictionaries/enwiki-20190409-dict-unlemmatized.txt.bz2
lemmatized_dictionary_name = "dictionaries/enwiki-20190409-dict-lemmatized.txt.bz2"

must_lemmatize = True

if must_lemmatize:
    dictionary_name = lemmatized_dictionary_name
else:
    dictionary_name = unlemmatized_dictionary_name

#datafile_name = api.load("text8")
#datafile_name = [wd for wd in datafile_name]
#unfiltered_dictionary_name = "dictionaries/text8.txt.bz2"
#filtered_dictionary_name = "dictionaries/text8-filtered-20-10-100000.txt.bz2"

#dct = corpora.Dictionary(datafile_name)
wiki = WikiCorpus(datafile_src, lemmatize=must_lemmatize) #False to no use lemmatizati
wiki.dictionary.save_as_text(dictionary_name)
#dct.save_as_text(unfiltered_dictionary_name)

# Remove words occuring less than 20 times, and words occuring in more than 10% of the
#wiki.dictionary.filter_extremes(no_below=20, no_above=0.1, keep_n=100000)
```

```
        #wiki.dictionary.save_as_text(filtered_dictionary_name)

        del wiki
        #del dct
```

In [ ]:

In [ ]:

In [ ]:
```
# Load dictionary from file
datafile_name_2 = "datasets/chunks/enwiki-chunks-9999/enwiki-chunk-9999-1.xml.bz2"
unfiltered_dictionary_name_2 = "dictionaries/enwiki-chunk-9999-1_2.txt.bz2"

from gensim.corpora import Dictionary
dictionary = Dictionary.load_from_text(unfiltered_dictionary_name)

wiki_2 = WikiCorpus(datafile_name_2)
dictionary.add_documents(wiki_2)

dictionary.save_as_text(unfiltered_dictionary_name_2)

del wiki_2
```

In [ ]: