In the following assignments you will write Hadoop Streaming applications. You will be working with Wikipedia articles dump and auxiliary datasets:

Wikipedia dataset location in HDFS: */data/wiki/en_articles_part*

"Stop words" dataset is located in *'/datasets/stop_words_en.txt'* file in **local** filesystem.

Format: *article_id <tab> article_text*

Below, you can find starter template (code sample) to preprocess Wikipedia articles for MapReduce. You can use it for programming (LTI) assignments in the following week.

```
#!/usr/bin/env python

import sys
import re

reload(sys)
sys.setdefaultencoding('utf-8')

for line in sys.stdin:
    try:
        article_id, text = unicode(line.strip()).split('\t', 1)
    except ValueError as e:
        continue
    text = re.sub("^\W+|\W+$", "", text, flags=re.UNICODE)
    words = re.split("\W*\s+\W*", text, flags=re.UNICODE)

    # your code goes here

```

Remarks:

- English Wikipedia contains a lot of characters from other languages. So, you should parse Unicode correctly.
- you need to remove punctuation marks and transform words to lowercase, to get correct quantities.

If you want to deploy the environment on your own machine, please use bigdatateam/yarn-notebook Docker container.

✓ Completado