In this assignment you will use Spark to compute various statistics for word pairs. At the same time, you will learn some simple techniques of natural language processing.

Dataset location: */data/wiki/en_articles*

Format: *article_id <tab> article_text*

While parsing the articles, do not forget about Unicode (even though this is an English Wikipedia dump, there are many characters from other languages), remove punctuation marks and transform words to lowercase to get the correct quantities. Here is a starting snippet:

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

from pyspark import SparkConf, SparkContext
sc = SparkContext(conf=SparkConf().setAppName("MyApp").setMaster("local[2]"))

import re

def parse_article(line):
    try:
        article_id, text = unicode(line.rstrip()).split('\t', 1)
    except ValueError as e:
        return []
    text = re.sub("^\W+|\W+$", "", text, flags=re.UNICODE)
    words = re.split("\W*\s+\W*", text, flags=re.UNICODE)
    return words

wiki = sc.textFile("/data/wiki/en_articles_part1/articles-part", 16).map(parse_article)
result = wiki.take(1)[0]

for word in result[:50]:
    print word
```

You can use this code as a starter template. Now proceed to LTI assignments.

If you want to deploy the environment on your own machine, please use bigdatateam/spark-course1 Docker container.

Marcar como completo