

# INSTITUTO TECNOLÓGICO DE COSTA RICA

## TAREA #2

*Saúl Zamora*

profesor  
Kevin Moraga

March 25, 2017

## 1 Introducción

Utilizando los datos de Google Books n-gram viewers; los cuales son tuplas de tamaño fijo, que en este caso son palabras extraídas de los libros existentes en Google Books. La  $N$  especifica el número de elementos en la tupla, as un 5-gram tiene 5 palabras. Los datos están en texto plano en el siguiente formato:

- *ngram TAB year TAB match\_count TAB page\_count TAB volume\_count  
NEWLINE*

El objetivo principal es hacer uso de un cluster de procesamiento utilizando *Kubernetes* y *Docker* sobre el cual se lanzarán las instancias de *Hadoop* que utilizan técnicas de *MapReduce* para resolver consultas.

## 2 Ambiente de desarrollo

- Sistema operativo utilizado: Linux Ubuntu 16.04 LTS
- Python versión 2.7.12
- Docker versión 1.5.2
- Hadoop versión 2.7

## 3 Estructuras de datos usadas y funciones

### 3.1 Mapper

Se realizan operaciones sobre strings para obtener ngrams línea por línea. Luego se realiza una impresión del ngram en la salida estándar con el formato:

- *ngram ngram\_count*

Con *ngram\_count*  $\bar{I}$ . Dichos datos son usados como la entrada del *reducer*.

### 3.2 Reducer

Se leen las impresiones y nuevamente se utilizan operaciones sobre strings para realizar el conteo de los ngram repetidos y luego presentar un conteo final.

## 4 Instrucciones de ejecución'

Asumiendo que el ambiente de desarrollo está listo, hay que seguir los siguientes pasos:

- Descargar el repo desde <https://github.com/aleks279/200835773-tarea2>
- Navegar al folder *200835773-tarea2*

- El comando siguiente corre el proceso de MapReduce sin Hadoop:

```
– cat googlebooks-eng-all-4gram-20120701-zz | python map_reduce/mapper.py | sort –  
k1,1 | python map_reduce/reducer.py
```

- El comando siguiente corre el proceso de MapReduce con Hadoop:

```
– hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-  
streaming-2.7.3.jar –mapper map_reduce/mapper.py –reducer map_reduce/reducer.py –  
input data/googlebooks-eng-all-4gram-20120701-zz –  
output out
```

## 5 Bitácora de trabajo

- 15-03-2017:
  - 1 hora - instalación de Docker.
  - 4 horas - configuración local de aplicación en Ruby on Rails, Docker y Heroku.
- 17-03-2017:
  - 4 horas - configuración local de aplicación en Ruby on Rails y Docker.
- 18-03-2017:
  - 2 horas - configuración local de Hadoop.
  - 1 hora - documentación.
  - 4 horas - investigación en MapReduce, Hadoop y Ruby con Rubydoop
- 20-03-2017:
  - 2 horas - refactor del app. Instalación de JRuby.
- 22-03-2017:
  - 2 horas - refactor. Sin éxito. App compila y crea el .jar, pero genera errores al correr. Prueba con un contador de palabras simple.
- 24-03-2017:
  - 3 horas - start over con Python. Configuración de Docker y Python en Distelli (para deploy).
  - 2 horas - aclimatarme a Python. Cambiar contador de palabras de ejemplo a contador de ngrams.
- 25-03-2017:
  - 2 horas - pruebas con Hadoop y Python.

Total de horas trabajadas: 27 horas.

## 6 Comentarios finales

- Debido al horario laboral, la falta de tiempo fue una limitante y no fue posible realizar el cluster de Kubernetes.

## 7 Conclusiones

- La configuración inicial de Docker es *overly complicated*.

## References

- [1] Bourgau, P. (2017). How to boot a new Rails project with Docker and Heroku - Philippe Bourgau's blog. [online] Philippe.bourgau.net. Available at: <http://philippe.bourgau.net/how-to-boot-a-new-rails-project-with-docker-and-heroku/>
- [2] Digitalocean.com. (2017). How to Install Hadoop in Stand-Alone Mode on Ubuntu 16.04 — DigitalOcean. [online] Available at: <https://www.digitalocean.com/community/tutorials/how-to-install-hadoop-in-stand-alone-m>
- [3] GitHub. (2017). iconara/rubydoop. [online] Available at: <https://github.com/iconara/rubydoop>
- [4] Noll, M. (2017). Writing An Hadoop MapReduce Program In Python - Michael G. Noll. [online] Michael-noll.com. Available at: <http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python>
- [5] Distelli.com. (2017). How to Build and Deploy a Python Application on Docker — Distelli. [online] Available at: <https://www.distelli.com/docs/tutorials/build-and-deploy-python-with-docker>