תרגיל 3 – MapReduce (חלק א')

בתרגיל זה נלמד להריץ תהליך של MapReduce. נשתמש בשתי דרכי הדמיה של MapReduce ב- Hadoop: בעזרת צינור של לינוקס ובעזרת סקריפט שמשתמש בחישוב מבוזר.

נשתמש בתור קלט בתרגום לאנגלית של הספר הקלאסי "מלחמה ושלום" ע"י לב טולסטוי ש- יש בו למעלה מחצי מיליון מילים (<http://www.gutenberg.org/files/2600/2600-0.txt>). נרצה לדעת מה הן המילים הנפוצות ביותר לפי הספר הזה. תורידו את הקובץ הנ"ל (לפי ההוראות בשאלות).

**תזכורת: על מנת להריץ קבצי python ב- Ubuntu:**

* ניתן להריץ קוד מטרמינל בעזרת פקודת python, לדוגמה: python q1.py

או

* ניתן לתת הרשאות הרצה: chmod +x ./q1.py
  + ואז להריץ: ./q1.py
  + שימו לב שבקובץ הpython מופיע בראש הקובץ:

#!/usr/bin/python

* אם יש בעיה בהרצה של פקודת python, ניתן לבצע:

sudo cp /usr/bin/python3 /usr/bin/python

**שאלות:**

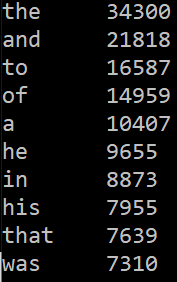
1. תיצרו תיקייה עבור התרגיל ותכנסו אליה. תורידו את הקובץ הנ"ל אל התיקייה הזאת ע"י הרצת פקודת wget עם הכתובת הנ"ל בתור ארגומנט. ניתן להחליף את שם הקובץ. בהמשך אני מניח ששם הקובץ הוא war\_and\_peace.txt.

**שלב א: בעזרת צינור של לינוקס**

1. עליכם להשתמש בקוד עבור בעיית Word Count שראיתם במצגת (ניתן להוריד אותו מ- מודל) ולהריץ הדמיה של MapReduce בעזרת צינור ופקודת sort (כמו שמפורט במצגת) עבור הקלט הנ"ל. בנוסף, יש למיין את שורות הפלט (כלומר, הזוגות של מפתח-ערך) לפי הערכים ולקחת את 10 השורות הראשונות בעזרת המשכת הצינור ע"י הפקודות: sort –grk 2 | head. יש לעיין בדפי man כדי להבין את המשמעות של הפקודות האלו.

**מקרים לבדיקה (Test Cases):**

הפלט צריך להיות (יתכנו סטיות קטנות בערכים מסיבות טכניות):



**הערה**: ייתכן שתקבלו שגיאה BrokenPipeError. ניתן להתעלם ממנה. אם תרצו לדעת את הסיבה ואיך לפתור אותה, ניתן לקרוא את [התשובה המקושרת](https://stackoverflow.com/a/26738736/2725810).

1. נמדוד את הזמן שה- Reducer הנ"ל לוקח. בשביל זה נוסיף פקודת time לפני הפעלת Reducer בצינור. גם כן, נזרוק את הפלט (time מדפיס לפלט מופרד של שגיאות), כך שסוף פקודת ההרצה תיראה כך:

| time ./reducer.py > /dev/null

**מקרים לבדיקה (Test Cases):**

הפלט צריך להתחיל כך (עם זמן אחר כמובן):



אנחנו מעוניינים בזמן הראשון מצד שמאל (0.26 שניות בדוגמה) שהוא זמן מעבד.

1. עליכם לשפר את הביצועים של MapReduce עבור בעיית Word Count. בשביל זה נשים לב ש Reducer מקבל קלט ממוין לפי מפתח, כך שאפשר לחסוך מספר גישות למילון. תשפרו את ה Reducer.

**מקרים לבדיקה (Test Cases):**

* תריצו MapReducer עם ה Reducer המשופר על קלט קטן (לדוגמה, בעזרת פקודת echo) ותוודאו שהפלט נכון.
* תריצו MapReducer עם ה Reducer המשופר כמו שעשיתם בשאלה 1 ותוודאו שהפלט זהה לפלט שקיבלתם אז.
* תחזרו על מדידת הזמן כמו בשאלה 3 עם ה Reducer המשופר. הביצועים צריכים להשתפר ב- כ- 20% לכל הפחות.

**שלב ב: בעזרת סקריפט שמשתמש בחישוב מבוזר**

1. שימו לב שהדמיה שביצענו היא מאד לא ריאליסטית. ראשית היא לא משתמשת בחישוב מבוזר. אך חשוב יותר ש- יכולים להיות באגים בקוד והיא לא תתפוס אותם, כי היא מניחה ש- ל- Mapper ול- Reducer יש גישה לכל הנתונים. בשאלה זאת נתקין סקריפט להדמיה ריאליסטית יותר של MapReduce.
   1. תתקינו תוכנות git ו- parallel (מומלץ לחפש אותם ב- Google ולקרוא עליהן, ובפרט על git, כי היא מייצגת משפחה שלמה של תוכנות חשובה ביותר בעולם הנדסת תוכנה):

sudo apt update

sudo apt install git parallel

* 1. תורידו את הסקריפט בעזרת הפקודה:

git clone --depth 1 https://github.com/d2207197/local-mapreduce.git

**מקרים לבדיקה (Test Cases):**

האם הצלחתם לבצע את כל הפקודות ו- ווידאתם שהן אכן הצליחו (כלומר, שלא קיבלתם שגיאות)?

1. תריצו את תהליך ה MapReduce עם שני Reducers משופרים בעזרת הסקריפט. בהנחה שהסקריפט נמצא בתת תיקייה של התיקייה הנוכחית בשם local\_mapreduce, הפקודה בשביל ההרצה נראית כך:

cat war\_and\_peace.txt | local-mapreduce/lmr -k 300k 10 ./mapper.py ./reducer.py output 2>err

חשוב שתבינו את הארגומנטים שמסרנו ל- סקריפט:

* -k 300k – הדמיה של HDFS עם 300 אלף תווים בכל בלוק. זה שווה ערך ל- -k 300000.
* 10 – כמות Reducers.
* output – תיקייה עבור פלטים של Reducers.

תשתמשו בפקודות לינוקס כדי לחבר פלטים של כל ה- Reducers ולקבל פלט כמו בשאלה 2.

שאלות חשיבה:

* כמה Mappers ירוצו עבור הקובץ war\_and\_peace.txt כאשר משתמשים ב –k 300k?
* תעיינו בפלט של ה- Reducers. תוודאו שהוא הגיוני לפי מה למדתם על MapReduce.
* שימו לב שנוצרה תיקייה עם שם המתחיל ב- mapper\_tmp. בתיקייה זאת ניתן לראות מה הם הנתונים שכל Reducer קיבל מכל Mapper. תוודאו ש- מה שאתם רואים כאן הגיוני לפי מה שלמדתם על MapReduce.

**מקרים לבדיקה (Test Cases):**

האם הצלחתם לקבל פלט כמו בשאלה 2 ולבצע תרגילי חשיבה?

# אופן הגשה:

1. יש להגיש למודל את ה Reducer המשופר.
2. יש למלא את הבוחן.