**מדעי הרוח הדיגיטליים – מיני פרויקט**

ליאור מואב – 205618663.

התבקשתי לנתח את אנציקלופדית חכמי תימן שנכתב ונערך ע"י ד"ר משה גברא, אשר חוקר ומחבר ספרים בנושא יהדות תימן באוניברסיטת בר-אילן.

את קבצי האנציקלופדיה (המחולקת לשני חלקים) ניתחתי בעזרת הבנת מבנה הקבצים, הכניסות של כל אחד מהערכים והשוואות בין האינדקסים השונים לערכים.

ראשית חילקתי את חלקי האנציקלופדיה, כל אחד מהם, לקבצים שונים, כאשר קובץ אחד מכיל את האינדקסים של אותו חלק והאחר מכיל את הערכים שלו.

לאחר חילוץ האינדקסים, מצאתי את המיקום בקובץ הערכים אם קיים ומשם שמרתי עבור כל אדם (ערך) את שמו, תקופת חייו, מקצועו ומקום מולדתו בקובץ txt פשוט, עם \t כ-delimiter בין התכונות של כל ערך. את הקובץ העברתי ל-OpenRefine שנלמד בכיתה אשר מוציא טבלה מסודרת ממנה ניתן לנתח מידע בצורה פשוטה וברורה.

לבסוף, דיווחתי על כל האינדקסים שלא קיים להם ערך (או קיים אך ישנן אי התאמות בשמות), כמו גם את מספרם.

התוכנית רצה בצורה די פשוטה, אך עם הבנה לעומק של מבנה קובץ האנציקלופדיה כ-docx. הצורה בה כתובה התוכנית מאפשרת לכרות עוד נתונים ומידע מעניין לגבי כלל האנשים שקיים להם ערך באנציקלופדיה.

התוכנית מקבלת כקלט את קבצי ה-docx המפורקים לאינדקסים וערכים עבור שני חלקי האנציקלופדיה ואת רשימת הערים המוזכרים בה. התוכנית מפרסרת את הקבצים בעזרת ספריית docx2txt ב-python. התוכנית עוברת באופן איטרטיבי על תוכן הקבצים ומבצעת עליו מניפולציות תחת תנאים מסוימים, ונעזרת בנוסף בספריות re, codecs.

הקבצים שהתוכנית מוציאה כפלט:

* Missing values – part1: מכיל את כלל האינדקסים שלא קיים עבורם ערך בחלק הראשון של האנציקלופדיה. כמו כן, את מספר הערכים שניתן להצליב עבורם אינדקס ומספר האינדקסים הכולל.
* Missing values – part2: כנ"ל, עבור החלק השני.
* Persons info – part1/2.txt: מכילים טבלאות עבור כל אחד מחלקי האנציקלופדיה, המכילות את הערכים עם השם, עיר מגורים, תפקיד ותקופה. אם אחד מהנ"ל חסר, המחרוזת "---" תחליפו.
* אבחנות שלי לגבי האנציקלופדיה:
* קיימים הרבה אינדקסים **שאכן** קיים להם ערך, אך הכותרת שלהם פגומה, מכילה תווים שונים או כתובה באופן שונה מהאינדקס. במקרה כזה, האינדקס יופיע בפרויקט כאינדקס ללא ערך. כך למשל, עבור המון ערכים במקום המיועד ל"תקופה" מופיעה המחרוזת "מאח כ'" במקום "מאה כ'". בנוסף, ישנם הרבה תווים שהופכים קטעים מסוימים לא קריאים (בלעז, ג'יבריש) או שהשפה התימנית התערבבה עם חלק מהטקסט. ניסיתי לתקן חלק מהבעיות אך לא ניתן לתקן מקרים "פרטיים" יותר.
* כפועל היוצא מכך, מתוך סה"כ האינדקסים בשני החלקים (2923) ישנם 1365 חכמים עם ערך מדויק. כלומר 47% הצלחה.