בס"ד

איכות תוצאות החיפוש בגוגל עבור תוכן מדעי

דו"ח פרויקט

הקדמה:

בעידן שבו התרגלנו לזמינות של מידע חופשי במהירות ובקלות, עולה השאלה מה האיכות של המידע אליו אנחנו נחשפים והאם הוא מושפע מפרמטרים שלא חשבנו עליהם.
המחקר עומד לבחון מחדש את ההערכות האקדמיות לכך שמיקום גאוגרפי ושפה מהווים חסמים לתוכן מדעי איכותי ובכך לתרום למיקוד המאמצים הנעשים בתחום.

מאגר הנתונים עליו המחקר יתבסס הוא באדיבות הפקולטה לחינוך למדע וטכנולוגיה בטכניון, כחלק מקבוצת מחקר העוסקת בתקשורת המדע. הקבוצה החלה בתהליך ניתוח הנתונים ואחד הקשיים המרכזיים הוא השוואה של איכות התוכן – מכאן שהאתגר הצפוי לנו במענה על שאלת המחקר טמון בהתמודדות עם משתנים רבים המרכיבים את האיכות והשוואתם על פני מספר חתכים שונים המורכבים משפות, מדינות וקטגוריות של מונחים מדעיים.

שאלת המחקר המרכזית שאותה נבחן היא השפעת המיקום הגיאוגרפי על מדדים שונים של תוצאות החיפוש החוזרות בגוגל עבור מונחים מדעיים שונים.
בנוסף נבחן את האפשרות להעמיד מודל רגרסיה המחשב את מדד האיכות ע"י מדדי נגישות.
המוטיבציה בבניית המודל היא לעשות שימוש בסריקת מדדי נגישות של אתרים (כמו למשל צירוף גרפים ומולטימדיה) לטובת אמידה של רמת איכות התוכן של האתרים.

שאלות המחקר שלנו ייחודיות ביחס למחקר המקביל בטכניון בכך שהן בוחנות תלויות חדשות, כמו למשל איכות אל מול נגישות, ובנוסף בוחנות את השפעת המיקום הגיאוגרפי שעד כה לא נעשה חקירה לגביו.
יתרה מזאת, מצאנו מאמר [1] הקובע כי העולם המדעי העכשווי מושפע מהטייה שמקורה בידיעת השפה האנגלית. המאמר קובע כי הדבר גורם לפגיעה ביכולת של אזרחים וחוקרים להשתלב בעולם המדעי ללא ידיעת השפה האנגלית. בפרויקט שלנו נתעמק במגוון רחב יותר של שפות ובפערים שלהם בהיבט המדעי.

הגישה המתוכננת היא לערוך מבחני השערות, להפעיל רגרסיה מרובת-משתנים וכלים נוספים שיתמכו במענה לשאלות.

סקירת נתונים:

ישנו קובץ אקסל יחיד המכיל 2 לשוניות:
- **Coded_srs** - מכילה את טבלת הנתונים הנדגמים
- **The codebook** - מכילה מקרא על נתונים קטגוריאליים

הטבלה מכילה נתונים שנאספו ממדינות שונות ברחבי העולם , מכל מדינה נאספו כמה דגימות של חיפוש מילה מדעית בגוגל בשפת האם של המדינה ועל הדוגמים היה נדרש למלא את הנתונים (לפי סולם ערכים שנקבע מראש) עבור מדדים של איכות, נגישות, קונספרטיביות וסוציו-מדעיים .

מבנה הטבלה הינו כ 3000 שורות ו 61 עמודות.
בנוסף הטבלה מחולקת לקטגוריות (על פי צבעים) שמחלקים את המידע הנבדק לכל רשומה:
- **צבע כתום –** מידע כללי על הקישור והמידע המצוי בו.
- **צבע צהוב –** קריטריונים שבודקים את רמת **האיכות** המדעית שנדגמה.
- **צבע ירוק –** קריטריונים שבודקים את רמת **הנגישות** המדעית שנדגמה.
- **צבע ורוד –** קריטריונים שבודקים מאפיינים של **סוגיות קונספרטיביות**.
- **צבע כחול –** קריטריונים שבודקים **סוגיות סוציו- מדעיים.**

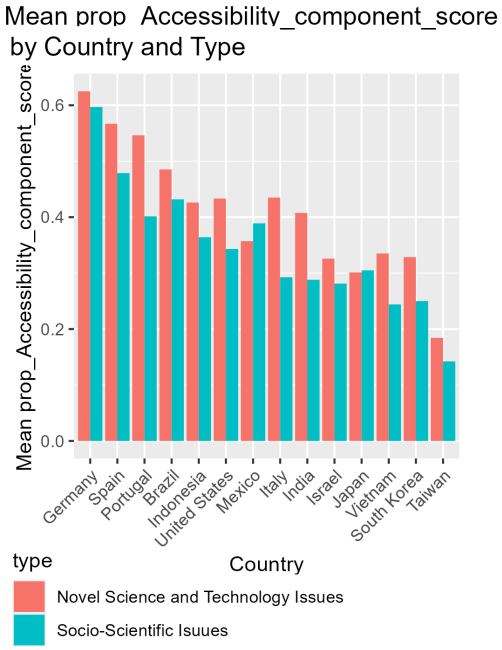
1- Márquez, M. C., and Porras, A. M. (2020). Science Communication in Multiple Languages Is Critical to Its Effectiveness. Front. Commun. 5:31. doi: 10.3389/fcomm.2020.00031.

שיטות ותוצאות:

1- פערים בתוכן המדעי בגוגל בין המדינות השונות

בדקנו האם יש הבדלים בין המדינות השונות עבור המדדים המשוקללים בכל קטגוריה.
בשלב הראשון עשינו מבחן F שנועד לבדוק אם יש לפחות מדינה אחת שהתוחלת שלה שונה משאר התוחלות. בשלב השני השתמשנו במבחן Least Significant Difference כדי לבדוק מי הן המדינות שיוצרות יחד קבוצה הומוגנית ושונות באופן מובהק משאר המדינות (בשני המבחנים השתמשנו ב alpha=0.05).

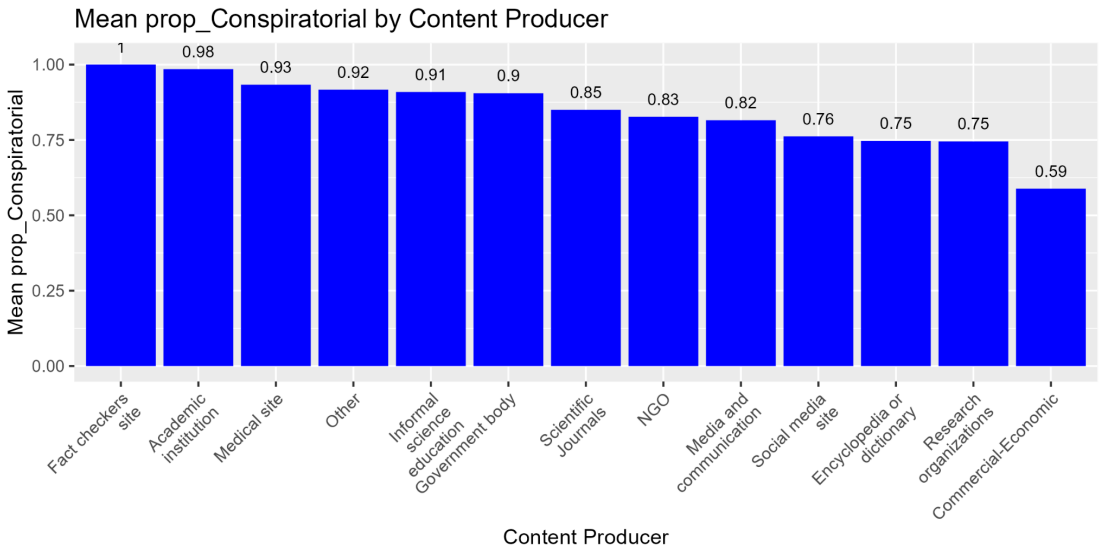
- בבדיקת **מרכיב האיכות**** מצאנו כי המדינות מתחלקות ל-2 קבוצות עיקריות, כאשר המדינות בעלות האיכות הגבוהה ביותר הן- מקסיקו, איטליה, ספרד, ארה"ב, גרמניה וטיוואן. [1]
  - הפער המשמעותי ביותר הוא במונחים מסוג Socio-Scientific Issues, עבורם הקבוצה הראשונה איכותית באופן מובהק **ב- 12% יותר** משאר המדינות [2]
  - בקבוצת המדינות הראשונה יש באופן מובהק **9.5% יותר ציטוטים והפניות לספרות** [3]



- בבדיקת **מרכיב הנגישות*** מצאנו כי המדינות בעלות הביצועים הטובים ביותר הן גרמניה וספרד, בזמן שטיוואן נמצאת בתחתית הרשימה.[4]
  - כפי שניתן לראות בגרף משמאל, עיקר הפער במונחים המדעיים בטיוואן נובע מקטגוריות של **מונחים Socio-Scientific Issues ו- Novel Science and Technology Issues** [5]

- בבדיקת **המדד הסוציו-מדעי** מצאנו כי אינדונזיה היא המדינה עם הביצועים הטובים ביותר בזמן שקבוצת המדינות שחולקות את הביצועים הנמוכים ביותר היא- דרום קוריאה, טיוואן, גרמניה ויפן.[6]
  - מצאנו שהמדד 'האם התוכן המדעי מכיל מידע על סיכונים או יתרונות אל מול חסרונות', גדול באופן מובהק באינדונזיה **ב-39% יותר מאשר ממוצע שאר המדינות**.[7] לאור פער כה משמעותי, היינו ממליצים לוודא שזה לא נובע בגלל **טעויות בדאטא או באיכות המתייגים**.

- בבדיקת **מדד הקונספירציה** מצאנו כי קבוצת המדינות שהמידע שלהן הוא באופן יחסי הכי קונספירטיבי היא- ארה"ב, ספרד, ישראל ויפן.[8] (מדד בין 0-1, כך ש-1 אינו קונספרטיבי)
  - עבור קבוצת המדינות הקונספרטיביות ביותר, **מצאנו שמפיק התוכן הקונספרטיבי ביותר הוא מסוג מסחרי-כלכלי** (למשל חברות פארמה) [9]



*המספרים מתייחסים למספר של הקוד הרלוונטי בנספחים
** מרכיב האיכות משוקלל ע"י שלושה פיצ'רים - 'שימוש במידע מספרי', 'אזכור רשימת מקורות' ו-'שימוש בציטוטים או הפניות ספרותיות'
*** מרכיב הנגישות משוקלל ע"י ארבעה פיצ'רים - 'מקום להשאיר תגובות', 'גרפים', 'מולטימדיה' ו-'צירוף לינקים'

2- האם ניתן לאמוד את איכות האתר בעזרת נגישות האתר?

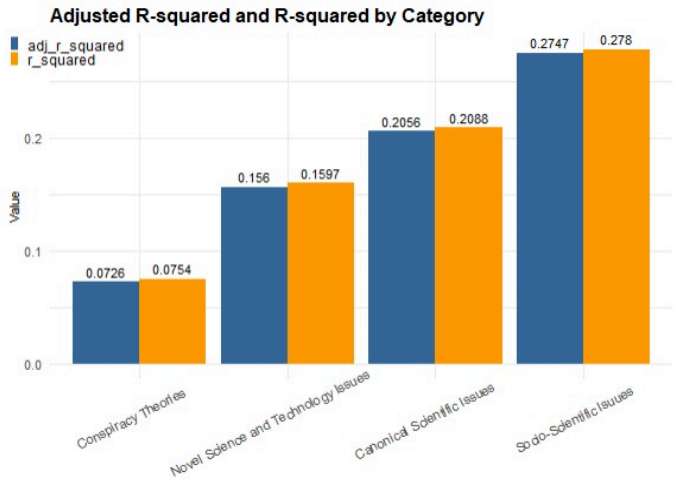
בנינו מודל רגרסיה מרובת משתנים המחשבת את איכות האתר בעזרת מדדי נגישות האתר.
בדקנו את ההנחה שההפרעה המקרית מתפלגת נורמלית במודל. לפיכך חישבנו את residuals[10] ולפי הגרף (מצורף בנספחים) נוכחנו להסיק שהשגיאות מתפלגות קרוב להתפלגות נורמלית ולכן המשכנו בחישוב משוואת הרגרסיה הלינארית והסקת המסקנות לפיה.

משוואת הרגרסיה אליה הגענו היא:

prop_Quality_mean[11] = **0.629** + **0.373** * prop_Accessibility_component_score - **0.535***prop_Accessibility_mean + **0.334***prop_Jargon_score

לאחר המודל הראשוני פעלנו ב-2 היבטים שונים כדי לשפר את המודל ולקבל תמונת יותר מלאה, אמינה ואיכותית:[12]

- **משקול נתונים-** כדי לשפר את אמינות המודל, ביצענו משקול על פי היחס של כמות הנתונים מכל מדינה. כלומר למדינה בעלת כמות רשומות גדולה יותר יהיה משקל גדול יותר וכך תגדל השפעתה על תוצאות המודל.
- **פילוח נתונים-** בנינו מודלים נפרדים עבור כל קטגוריה של מונחים במטרה להבחין בשוני של הקורלציה בין מדדי הנגישות לאיכות ולהסיק מכך מסקנות. לדוגמא- עבור מונח מקטגוריה Socio-Scientific Issues ניתן יהיה לאמוד את רמת איכות האתר ע"י מדדי הנגישות שלו, בצורה יותר מדוייקת משאר הקטגוריות.



בגרף משמאל ניתן לראות את המדדים R^2 ו-adj R^2 של המודלים לאחר פילוח ומשקול הנתונים כפי שהוזכר לעיל.

מגבלות ועבודה עתידית:

הנתונים מושפעים באופן ישיר **מהטייה אנושית** משום שמדובר בתיוגים של מתנדבים, ולכן כדי לקבל מסקנות עלינו לוודא את מהימנות התיוגים ע"י בדיקה חיצונית או **בדיקות סטטיסטיות**. הנתונים אמנם עברו סינון ראשוני לאחר מבחנים סטטיסטים של החוקרים מהטכניון אך עדיין תיתכן שונות שמקורה בהבדל בין מתייגים. מעבר לכך, המונחים המדעיים הם בשפת המקור של כל מדינה ולכן **לא יכולנו להפעיל אגרגציה** ולנתח את הנתונים פר מונח.

בטווח הארוך היינו משקיעים משאבים באיסוף נתונים **ממדינות נוספות** מאחר ומאגר הנתונים שלנו מכיל **8.3% בלבד ממספר המדינות בעולם**, וכן יתכנו פערים מדעיים במדינות שטרם נדגמו. הדבר יתרום לייצוג מקיף יותר של מדינות ויאפשר לנו לזהות **טרנדים ברמת יבשת**. יתרה מזאת, כחלק מעבודת ה - Feature engineering ניתן לבצע scraping על עמודת "Link" במטרה לאתר פיצ'רים נוספים כגון : מס' המילים, וסמנטיקה באמצעות כלי NLP, וזאת כדי ליצור מהימנות גבוהה יותר של מדדי איכות ונגישות.
מעבר לכך, היינו אוספים תיוגים נוספים **משפות נוספות מאותו מקום גאוגרפי**. כמו למשל, איסוף נתוני חיפוש בשפה האנגלית, המהווה שפה עיקרית לחיפוש מדעי ברוב מדינות העולם, לצורך התמקדות בהבדל הגיאוגרפי. צעד נוסף להרחבת מאגר השפות יכול להיעשות במדינות אשר להן יותר משפה רשמית אחת או שפות מקומיות נוספות.

קישור לקובץ הקוד + README:
https://github.com/Liorwaknin22/FinalProject_AdvanedProgramming.git