

תקציר - ivrit.ai

המטרה

ivrit.ai הוא מיזם ללא מטרת רווח. שמטרתו המרכזית היא שיפור איכות השפה העברית בטכנולוגיות בינה מלאכותית, כמו מודלי שפה (כמו ChatGPT) ומודלי תמלול (Siri, Alexa, Whisper ועוד).

המאמץ המרכזי החל בהקמת תשתית של מאגרי מידע בעברית שיאפשרו לגופי מחקר ופיתוח לבנות כלי בינה מלאכותית שתומכים בעברית. כל מאגרי המידע מונגשים ללא עלות, כולל לשימוש מסחרי. כיום, לאחר כשנה וחצי של פעילות, הנגשנו את מאגרי המידע הגדולים ביותר בעברית לאימון בינה מלאכותית לתמלול דיבור, את המודלים הפתוחים האיכותיים ביותר בעברית וכן פתרונות תמלול פתוחים ונגישים לכל. כל זאת, בהתנדבות מלאה ותודות למאמץ ומרץ, זמן, כוח מחשוב ודאטה שנאספו ונתרמו בקהילת דוברי העברית.

המוטיבציה

עברית היא שפה נדירה בעולם. כ-9 מיליון דוברים ודוברות. לשם השוואה, ההערכה היא שיש כ-1.4 מיליארד דוברי אנגלית, כ-1.1 מיליארד דוברי סינית וכ-600 מיליון דוברי הינדית. כתוצאה מכך, חברות משקיעות מעט מאמצים בשפה העברית: איסוף המידע הרלוונטי יקר מאוד, ומספר המשתמשים הפוטנציאלי בשוק קטן.

פרויקט ivrit.ai נועד לצמצם את הפער הזה. אנחנו שואפים לאפשר לחברות וגופים לקבל את המידע הדרוש – מבחינת כמות, איכות, ורשיון שימוש – ובכך להקל על פיתוח טכנולוגיות מתקדמות מבוססות בינה מלאכותית לדוברי עברית.

מודלי תמלול חופשיים ומדויקים בעברית חוסכים זמן רב ברבים מענפי המשק והשירותים, במגזר הפרטי והציבורי: סיכומי פגישות, כתיבת פרוטוקולים, ראיונות, מחקרים ועוד. המודלים המשופרים יעשו טעויות מעטות, ויאפשרו לצרכנים להשתמש בתוצרים כבסיס לתהליכי המשך אוטומטיים ואנושיים נוספים של חילוץ ואחזור מידע, סיכום טקסטים, תרגום סימולטני, ניתוח נתונים ועוד ועוד. מדובר בחסכון של עשרות אלפי שעות עבודה ובתרומה משמעותית לפעילות היצרנית במשק, כמו גם ערך רב למאמצי הנגשה שונים.

ייחודיות הפרויקט

ראשית, ivrit.ai כולל את מאגרי המידע המוקלט הגדול ביותר בעברית שנגיש לציבור היום. הייחודיות השנייה היא הרישיון המשפטי, שמספק הגנה הן ליוצרי התוכן והן למשתמשים בו. יחד עם עו"ד אלי גרינבאום ממשרד עורכי הדין ארנון, תדמור לוי, פיתחנו [רשיון שימוש](#) המאפשר לבעלי זכויות בתוכן כגון פודקאסטים, תוכניות טלוויזיה, סרטים, עיתונים, ספרים ופרוטוקולים להנגיש את המידע לצורכי אימון מודלי בינה מלאכותית **בלבד**. בכך אנו מאפשרים לבעלי תוכן לתרום את התוכן שלהם לפרויקט, תוך שמירה על זכויותיהם המרכזיות, ולמפתחים שישתמשו בו – שימוש חף מדאגות הקשורות לפגיעה בזכויות יוצרים.

מה כבר השגנו?

הנגשנו מאגר של מעל 22,000 שעות תוכן מוקלט – יותר מכל מאגר ציבורי אחר בעברית. לשם השוואה, מודל התמלול Whisper הנחשב לאיכותי, אומן על כ-700 שעות בעברית.

מעל ארבעים גופים כבר אישרו לנו להנגיש את התוכן שלהם דרך המאגר, ביניהם: האקדמיה ללשון עברית, האוניברסיטה הפתוחה, הסתדרות הרופאים הישראלית, מכללת הרצוג, יד בן צבי וכן יוצרי תוכן עצמאיים כמו: גיקונומי, עידן ארץ, גיא קצוביץ', תמיר דורטל ועוד (הרשימה המלאה [פה](#)).

מתוך מאגרי המידע הרחב, למעלה מ-350 שעות אודיו, תומללו באופן ידני ושימשו לאימון מודלים של תמלול אוטומטי בעברית.

הנגשנו [לוח תוצאות](#) לדירוג מודלי התמלול הטובים ביותר בעברית, בהשוואה בין מודלים שונים - פתוחים ומסחריים.

אימנו מודלים מדויקים ומהירים לתמלול בעברית על בסיס הדאטה שנאסף. המודלים הללו הם מהטובים והמדויקים שזמינים היום בעברית, וזמינים לשימוש באופן פתוח וחופשי לגמרי.

הנגשנו [ממשק תמלול חופשי](#) שמבוסס על המודלים המדויקים ביותר שלנו וגם את בוט הוואטסאפ "אליעזר" שמסייע בתמלול הודעות מוקלטות בוואטסאפ. עד כה, בשירותים החינמיים הללו תמללנו מעל עשרת אלפים שעות (!) של אודיו עבור אלפי משתמשים שונים.

מי מאחורי הפרויקט?

- [יאיר ליפשיץ](#): מהנדס תוכנה עם ניסיון של כ-20 שנה בפיתוח וניהול ארגוני תוכנה גדולים.
- [כנרת משגב](#): מנהלת יחידת מחקרי נתונים בקרן המחקרים של בי"ח הדסה.
- [יועד סנפיר](#): מהנדס תוכנה עם ניסיון של כ-20 שנה בפיתוח.
- [יניר מרמור](#): מפתח אלגוריתמים וסטודנט במכון ויצמן.

יצירת קשר

ניתן ליצור אתנו קשר דרך info@ivrit.ai או טלפונית (050-5669187 - יניר)

כתבו עלינו

[נמאס לחכות לתמלול הודעות קוליות בוואטסאפ? הנה פתרון מצוין בחינם](#) - גיקטיים, עידן בן טובים, 20.4.25
[לאחרונה יצאו כלי בינה מלאכותית מוצלחים לתמלול, גם בעברית. הנה כמה מהם](#) - דה מרקר, רן בר זיק, 19.12.2024

[האם סוף סוף נוצר מודל AI שיודע לתמלל טקסטים בעברית](#) – מוסף הארץ, מחשבוט, אילנית שמיע 23.10.2024