A red and blue logo

Description automatically generated with low confidenceGraphical user interface, text

Description automatically generatedמשרד הביטחון – ויקיפדיה

 ‏13/09/2021

**מדינת ישראל - משרד הביטחון**

**מפא"ת/מו”פ/תקשוב/ענף בינה מלאכותית**

**הרשות הלאומית לחדשנות טכנולוגית**

**זירת תשתיות טכנולוגיות**

**משרד החדשנות, המדע והטכנולוגיה**

****

**תכנית לאומית למחקר, פיתוח בניית תשתיות לאומיות ל NLP  בעברית**

**בקשה לקבלת מידע (RFI)**

**הזמנה להבעת עמדה וקבלת מידע בנוגע לפרויקט לאומי בתחום עיבוד שפה טבעית (NLP) ממוקד במחקר ופיתוח בבינה מלאכותית**

**משרד הביטחון (להלן: "מפא"ת"), הרשות הלאומית לחדשנות טכנולוגית (להלן: "רשות החדשנות") ומשרד החדשנות, המדע והטכנולוגיה ( להלן "מחמ"ט"), מנהלי המיזם, מזמינים את הציבור למסור מידע ולהביע עמדה בנוגע לפרויקט לאומי בתחום עיבוד שפה טבעית (NLP) בבינה מלאכותית במדינת ישראל.**

1. **רקע**:
   1. פורום תל"ם (ועדה וולונטרית לתשתיות לאומיות למדע) מינה ועדת בדיקה לבחינת הצורך בהתערבות ממשלתית לשם האצת התפתחות תחום הבינה המלאכותית ומדע הנתונים (להלן: "**הוועדה**"). הוועדה, בראשותה של ד"ר ארנה ברי ונציגים מרשות החדשנות, האקדמיה, מפא"ת והתעשייה המליצו על תכנית ייזום לאומית שתתווה פתרון מערכתי הממוקד, בשלב ראשון, בארבעה תחומים: תשתיות, הון אנושי, רגולציה ונגישות לנתונים.
   2. מיזם "HebNLP" הינו מיזם לאומי ליצירת תשתית, מחקר ופיתוח יכולות מתקדמות לקידום תחום ה NLP  בעברית. המיזם מבוצע כחלק מיישום המלצות הבדיקה של פורום תל"מ (תשתיות מחקר לאומיות) בנושא.
   3. הוועדה קבעה כי נדרשת התערבות ממשלתית להאצת ההתפתחות בתחום הבינה המלאכותית ומדע הנתונים. כן נקבע כי נדרשת השקעה משמעותית בתשתיות וכלים להנגשת NLP  בעברית, נקבעה תכנית דו שנתית והוקצו משאבים בהתאם.
   4. עקרונות מובילים למימוש התכנית:
      1. רישיון שימוש מלא בתוצרים ובקוד לגופי הממשלה
      2. הנגשת תוצרים לציבור בקוד פתוח (למשל - github  ברשיון MIT )
      3. ניהול מרכזי של התכנית והפרויקטים בה ע"י מפא"ת, הרשות לחדשנות ומשרד החדשנות, המדע והטכנולוגיה (כולם או חלקם) ויכולת שילוב נציגים מטעמם בכל שלבי הפרויקט בגופים החוקרים ו/או המפתחים.
      4. בכל פרויקט יוכלו מפא"ת, הרשות לחדשנות ומשרד המדע להשתתף בתהליך כ design partner ולבצע הטמעה ויישום של הפיתוחים והמוצרים בסביבת העבודה שלהם, על מנת להביא לבשלות טובה יותר עבור כלל המשתמשים ולצרכים ספציפיים. החברה תידרש להיערך לכך בהתאם.
   5. מנהלי המיזם פונים בזאת למוסדות, גופים וחברות רלוונטיים העוסקים בתחום טכנולוגי זה בבקשה לקבלת מידע (RFI)  הנדרש למימוש הפרויקט (להלן "המשיבים").
   6. בנוסף מבקשים מנהלי המיזם לקבל מידע לגבי ניסיון ויכולות המשיבים לפיתוח תשתיות וכלי AI  בדגש על תחום ה NLP  ויכולות מימוש בקוד פתוח.
   7. מודגש ומובהר, כי השותפים טרם החליטו אם וכיצד לפעול בנושא, ופנייה זו אינה בבחינת הזמנה להגיש הצעות ואינה חלק מהליכי מכרז או מסלול הטבה או הליך אחר לקבלת הצעות, אלא פנייה לקבלת מידע ועמדה בלבד ובעקבותיה ישקלו השותפים את המשך פעולתם. מבלי לגרוע מהאמור במסמך זה, השותפים שומרים לעצמם את הזכות לפנות בשאלות הבהרה או בפנייה לקבלת מידע או נתון נוסף מכל גורם שהוא, לרבות מהנמענים, כולם או חלקם, אולם אין בפניה כאמור כדי לייצר מחויבות כלשהי כלפי מי שהשותפים או מי מטעמם יפנו אליו. יובהר ויודגש, כי בקשה זו אינה מהווה הצעה או הזמנה להגשת בקשה או ליצירת התקשרות עפ"י תקנות חובת המכרזים, התשנ"ג-1993, או על-פי כל דין או נוהל אחר.

1. **יעדים עיקריים של תהליך ה-RFI** 
   1. מיפוי המוסדות, הגופים והחברות בעלי ידע וניסיון רלוונטי בתחום למחקר ופיתוח תשתיות וכלי AI  בדגש על NLP בעברית ו/או ערבית, ובטכנולוגית קוד פתוח ומעוניינים להשתתף בתהליך כפי שמתואר בהמשך המסמך.
   2. הצגת משמעויות ראשונות ניהוליות וכלכליות.
   3. הצגת יכולות הרלוונטיות בתחום זה, כפי שיתואר בנספח א'.
2. **המידע המבוקש**:
   1. להלן פירוט המידע אשר מנהלי המיזם מעוניינים לקבל במסגרת המענה. האמור להלן אינו בגדר רשימה סגורה, וניתן בהחלט להרחיב גם על נושאים שאינם כלולים ברשימה זו, ושלדעת מגישי המענה יש חשיבות להתייחסות אליהם.
   2. המענה ל RFI  יוגש ב- 3 חלקים:
      1. מענה טכני.
      2. מענה ניהולי.
      3. מענה כלכלי.
   3. המידע יהיה שלם ויענה על כל הבקשות שבמסמך הסימוכין (נספחים) וכמפורט להלן. כמו כן מתבקש המשיב להגיש תמצית מנהלים.
   4. המענה  הטכני
      1. יש לפרט את הפתרונות הטכניים המוצעים על ידי המשיב ביחס לדרישות הקיימות בפרק מפרט הדרישות.
      2. לכל דרישה בה ניתן, יפרט המשיב את התכולות הנדרשות ביותר למיטב הבנתו. לדוגמה, יגדיר אילו קורפוסים יוכל לאסוף, מה מציע לתייג ומה גודל המידע המוערך. דוגמה נוספת - בסעיף ביגור פארסר בביצועים גבוהים שבפרק 3 (קוד פתוח לחבילות תשתית), המשיב יפרט מהם בראייתו הפארסרים המובילים אותם נכון לבגר.
      3. בנוסף, אם יש דרישות חסרות לדעת המשיב, נא ציינו את הדרישה ופרטו ביחס אחיה.
   5. המענה הניהולי
      1. יצוין האם המשיב מעוניין להשתתף בתהליך הפיתוח המשוער כפי שמתואר במסמך זה (בחלקו או כולו).
      2. ניתן וכדאי להוסיף הערות/הארות/השגות על אופן ביצוע המיזם (ככל שיוחלט על ביצועו) בשלבים הבאים המתוארים במסמך זה וכל נושא אחר שהמשיב רוצה להתייחס אליו.
      3. המשיב נדרש לפרט ניסיון בכל תחום רלוונטי בשלוש שנים האחרונות.
      4. ככל שיוחלט על ביצוע המיזם, יתכן הגורם המבצע יידרש להתחייב לתחזק פרוייקט אותו יפתח כפרויקט "קוד פתוח" עם קהילה פעילה תחת GitHub למשך שלוש שנים לפחות מיום אספקת הפרויקט, על המשיב לקחת בחשבון סעיף זה בתמחור הפרויקט הכללי.
      5. ככל שיוחלט על ביצוע המיזם, המשיב יידרש לשתף את הקוד המפותח בפרויקט זה (על כלל רכיביו, כולל רכיבי צד שלישי או אריזתם, Data Sets וכו'). במידה ונדרש כחלק מהפרויקט רכיב שאינו קוד פתוח (בין אם של צד ג' ובין אם מוצר של המשיב) יש לציין זאת בהבלטה ולכלול את התשלום על רישיון שנתי בתמחור הפרויקט בסעיף נפרד.
   6. המענה הכלכלי
      1. המשיב מתבקש לכלול במענה הערכת מחיר למחקר ופיתוח פרוייקטים במספר נושאים במסגרת המיזם כמפורט בנספח הדרישות הטכניות. יש לפרט הערכת כ"א צפויה ועלות לכל סעיף בטבלת התמחור בסעיף 5. אין לתת מחיר כולל לפרק בלבד, אך ניתן לתמחר חלק  מהפרויקטים (סעיפים) בפרק מסוים - פרויקטים אשר רק אותם מעוניין הגורם המבצע לפתח. על המשיב להביא בחשבון כי יתכן והגורם המבצע יידרש לבצע רק חלק מהפריטים שהציע, וזאת תוך תיאום עם ישויות אחרות.
      2. יש להגיש הערכה מפורטת לאומדן העלויות לביצוע כל פרויקט. ככלל יש להעריך את כל הנושאים עבורם יידרש תשלום גם אם הם לא מופיעים ברשימת התכולה כולל כל העמלות, היטלים, הנחה, מסים, קבלני משנה (כולל מע"מ – אם יש). כמו כן יש לפרט ולתמחר אופציות וכן חלופות נוספות שהוצעו ע"י המשיב.
      3. המשיב ישמור את הספרים, הדו"חות, המסמכים וכל מידע כתוב אחר ששימש להכנת המענה ויציגם לעיון לנציגי מנהלי המיזם על פי בקשתם, וכן ימסור להם כל מידע הבהרה או פרטים בכתב ובע"פ שידרשו על ידם לשם בחינת המידע.
      4. בלי לגרוע מכלליות האמור לעיל, המשיב ימסור למנהלי המיזם כל מידע בע"פ ובכתב כפי שיידרש על ידו לשם בדיקה של הערכות העלות בהתאם לקבוע בהמ"ב 40.060-40.066.
      5. כל פנייה לקבלני משנה לשם קבלת מידע למרכיבים מסוימים של המערכות, אין בה כדי ליצור יחסים חוזיים בין מנהלי המיזם לקבלני המשנה ו/או לגרוע מאחריותה והתחייבותו של המשיב בכל הנוגע לקבלני משנה.
      6. העלויות יוגשו בהתאם למבנה פירוט העבודה (WBS) המפורט בנספח הטכני.
      7. קונספט אפשרי (שנבחן ע"י השותפים ואינו מחייב) באשר לאופן ההתקשרות:
         1. ההתקשרות לביצוע המיזם תבוצע בסבבי מימוש של "סלי פרויקטים".
         2. בכל סבב יוגדר "סל" של פרויקטים אותו מעוניינים מנהלי המיזם לבצע במסגרת אותו סבב. תקציב הסל יוגדר ויפורסם מראש.
         3. יפורסם מסמך מפורט לדרישות טכנולוגיות (SOW) והגדרת היעד הכמותי אליו נדרש להגיע וכיצד ייבדק הפרויקט (DOD).
         4. מנהלי המיזם יתקשרו לביצוע אחד או יותר מהפרויקטים בסל כלשהו עם כל גורם מבצע שיהיה מסוגל טכנולוגית וזמין לבצעם.
         5. התמורה לכל משיב תהיה הסכום שהוקצה לנושא, לחלק למספר הגופים המבצעים המשתתפים. יתכן גם שכמות הגופים המבצעים המבצעים פרויקט מסוים תוגבל משיקולי תקציב, הגעה למסה קריטית או שיקולים אחרים של השותפים.
      8. המיזם מוגדר כמאמץ משותף של המדינה, האקדמיה והתעשייה. בהתאם לכך, הגופים יידרשו להשתתף במימון פעילות המחקר והפיתוח, בהיקף ששיעורו יקבע בהמשך.
3. **הגשת תגובות/עמדות**:
   1. המעוניינים לתרום מידע ולהציג את עמדותיהם בנושא פנייה זו, מתבקשים להעביר את התייחסותם בכתב לכתובת המייל: nlp@nationalplan.ai , לא יאוחר מיום **15 לנובמבר 2021**.
   2. ניתן לפנות לכתובת המייל הנ"ל בבקשה להבהרות ושאלות, עד ה 31.10.21. בתחילת נובמבר (מועד מדויק יפורסם בנפרד) יבוצע כנס ספקים פתוח בו יינתן מענה לכלל השאלות שיתקבלו עד למועד הכנס.
4. **מידע כללי ושמירת זכויות**
   1. מבלי לגרוע מהאמור בבקשה זו, מנהלי המיזם או מי מטעמם שומרים לעצמם את הזכות לפנות בשאלות הבהרה או בפנייה לקבלת מידע או כל נתון אחר למשיבים לפנייה זו, כולם או חלקם, החל ממועד פרסום פנייה זו. בנוסף, מנהלי המיזם או מי מטעמם שומרים לעצמם את הזכות לקיים מפגש עם לצורך הצגת פנייה זו, אשר יתקיים בחודש אוקטובר 2021 ואשר המועד המדויק שלו יפורסם בנפרד.
   2. מנהלי המיזם או מי מטעמם רשאים להשתמש בתוכן ניירות העמדה ונספחיהם (לרבות המידע והנתונים שיצוינו בהם) או כל מסמך אחר שיוגש לצורך גיבוש המלצות לתכנית עתידית, או לשם כל שימוש אחר, ושומרים לעצמם את הזכות להשיב או לא להשיב לתגובות שיתקבלו לפנייה זו, ככל שיתקבלו.
   3. מובהר כי מנהלי המיזם רשאים להשתמש בכל מידע שיימסר על ידי המשיבים לפניה זו (ובכלל זה לצורך פרסום פנייה לקבלת הצעות בדרך של מכרז ו/או משא ומתן, מפרטים ו/או אפיונים אשר יתבססו על הידע והניסיון שיצטברו מהמענים שיוגשו) ולאור כך המשיבים יהיו מנועים מלטעון כנגד מנהלי המיזם או מי מטעמם טענה ו/או דרישה ו/או תביעה כלשהי בדבר זכויות כלשהן הקשורות ו/או הנובעות מתשובתם לפנייה זו ו/או מהחומר שיצורף לפנייה זו (לרבות זכויות יוצרים של מוסרי המידע).
   4. אין בפנייה זו כדי ליצור מחויבות כלפי מי מהמשיבים לפנייה זו ומנהלי המיזם או מי מטעמם יהיו רשאים לשקול צעדים בהתאם לשיקול דעתם הבלעדי. מנהלי המיזם או מי מטעמם רשאים להיפגש בכל שלב שהוא עם כל גוף או משיב שימצאו לנכון, בהתאם להוראות כל דין, ומבלי שיחולו עליהם מגבלות מכל סוג שהוא.
   5. מודגש בזאת מפורשות כי טרם התקבלה החלטה כלשהי בדבר ביצוע התקשרות בכל היבט הקשור  לפנייה זו, ואין לראות בפרסום פנייה למידע זו, או בכל מעשה בגינה משום התחייבות ו/או מחויבות כלשהי של ממשלת ישראל, מנהלי המיזם או מי מטעמם, כלפי אדם או גוף כלשהו בנושא (לרבות התחייבות לכל פעילות עתידית באשר היא).
   6. כל ההוצאות הכרוכות להכנת והגשת המידע בתגובה לפנייה זו הן באחריות ובמימון הבלעדיים של הנמענים או המשיבים לפנייה זו ומנהלי המיזם לא יישאו בכל תשלום או הוצאה בגין הפנייה או כל היבט ועניין הנוגעים ו/או קשורים אליה. בכל מקרה, אין בפנייה זו משום משא ומתן ו/או יצירת התחייבות ו/או מחויבות מכל סוג של מנהלי המיזם.
   7. מענה לפנייה זו אינו תחליף להגשת בקשה מסודרת לקבלת סיוע ממי ממסלולי ההטבה השונים של רשות החדשנות, או הקולות הקוראים של מחמ"ט הקיימים במועד פרסום פנייה זו או שיהיו בתוקף בעתיד, על-פי ההוראות, הכללים והנהלים של מסלולי הטבה אלו.
   8. הגשת מענה לבקשה למידע זו אינה תנאי להשתתפות בהליכים עתידיים אם וככל שיהיו כאלה.
   9. למען הסר ספק, מובהר כי אין בפנייה זו, כדי לחייב את מנהלי המיזם בכל אופן שהוא, לרבות בעניין האפשרות להמשיך בתהליך זה, ואין בה כדי ליצור התחייבות של מנהלי המיזם כלפי הנמענים או המשיבים לפנייה זו, והיא מהווה פנייה מקדמית לצורך קבלת מידע בלבד ובעקבותיה ישקלו מנהלי המיזם את המשך פעולותיהם.
   10. על המשיבים לפנייה זו חל איסור לעשות שימוש בידע המפורט במסמך זה (לרבות בנספח), ללא אישור כלל הגורמים הרלוונטיים אצל מנהלי המיזם מראש ובכתב.
   11. הנספח המצורף לפנייה זו כפוף לאמור בה, ובכל מקרה לא יפורש בניגוד לאמור בה. במקרה של סתירות בדרישות בפנייה זו (לרבות בנספח), על המשיב להביאן לידיעת מנהלי המיזם לפני הגשת המענה ומנהלי המיזם יקבעו את החלופה הישימה.
5. הוראות הנוגעות ו/או הקשורות למשרד הביטחון (להלן בסעיף זה: "**המשרד**") בלבד:
   1. המשרד שומר לעצמו את הזכות לפנות לחלק מהגופים שישיבו לפנייה זו ו/או לחברות אחרות להמשך פעילות בנושא (ככל שתתקיים על-ידו), כל זאת על-פי החלטתו בלבד.
   2. המשרד שומר לעצמו את הזכות לראיין את האנשים המוצעים על-ידי המשיבים לפנייה, כולם או חלקם, כחלק מהמשך הפעילות בנושא (ככל שתבוצע על-ידו).
   3. המשרד שומר לעצמו את הזכות לבקש ממי מהמשיבים לפנייה זו שמות ופרטי קשר של 2 ממליצים לכל הפחות בארגונים ממשלתיים או בגודל של מעל ל-200 עובדים עבורם ביצע המשיב פעילות מתחומי העניין במהלך 4 השנים האחרונות.
   4. היה ויוחלט בהמשך על-ידי המשרד על פניה למשיבים בבקשה לקבלת הצעת מחיר (RFP) לפיתוח והצטיידות, רשאי יהיה המשרד להפיץ בקשה לקבלת הצעת מחיר למשיבים, על פי שיקול דעתו. הגשת מענה לבקשה זו לא תחייב את המשיב להגיש מענה לבקשה להצעת מחיר.
   5. למשרד שמורה הזכות להשתמש בכל ידע נשוא בקשה זו במלואו או בחלקו, לצרכיו הוא, ללא כל הגבלה, לרבות הוצאת בקשה להצעות מתחרות (RFP) לתעשיות, ו/או בחירת ספק לביצוע העבודה, ו/או העברת הידע לצד שלישי ובתנאי שהמשתמש הסופי הינו המשרד.
   6. על המשיבים לשמור את הספרים, הדו"חות, המסמכים וכל מידע כתוב אחר ששימש להכנת המענה ולהציגם לעיון נציגי המשרד על פי בקשתם, וכן ימסרו להם כל מידע הבהרה או פרטים בכתב ובעל-פה שיידרשו על ידם לשם בחינת המידע.
   7. כל פנייה למי מקבלני המשנה של המשיבים לפנייה לשם קבלת מידע בנוגע למרכיבים מסוימים של המערכות, אין בה כדי ליצור יחסים חוזיים בין המשרד לקבלני המשנה ו/או לגרוע מאחריותו והתחייבותו של המשיב בכל הנוגע לקבלני משנה.
   8. מבלי לגרוע מכלליות האמור לעיל, המשיב ימסור למשרד כל מידע בעל-פה ובכתב כפי שיידרש על-ידו לשם בדיקה של הערכות העלות בהתאם לקבוע בהוראות משרד הביטחון 40.060-40.066.

**נספח א' – מפרט דרישות טכני משוער למיזם הנבחן**

**מובהר כי אין באמור בנספח זה כדי ליצור מחויבות ו/או התחייבות ו/או הצעה ו/או הזמנה ו/או התקשרות מטעם מי מנהלי המיזם ביחס לנושא הנבחן האמור.**

**בנוסף, מובהר כי האמור בנספח זה מהווה תיאור ראשוני של הדרישות הטכניות למיזם הנבחן ע"י השותפים. משכך, ככל שיוחלט לבצע את המיזם, ייתכנו שינויים בדרישות הטכניות, ובכלל זה שינוי בדרישות המפורטות להלן ו/או הוספת דרישות או תנאים נוספים.**

1. **רקע**
   1. HebNLP הינו מיזם לאומי המשלב יצירת תשתית, מחקר ופיתוח ויכולות טכנולוגיות מתקדמות בתחום ה NLP בעברית ויתכן גם בערבית (בין אם במקביל או באופן מדורג). המיזם יאפשר יכולות פיתוח קלות ומהירות משמעותית מהקיים כיום, וכל הידע והתוצרים יהיו נגישים בקלות וללא עלות למערכת הביטחון, משרדי הממשלה, לאקדמיה ולתעשייה הישראלית בכלל.
   2. כלל הידע והיכולות במיזם יפותחו בקוד וארכיטקטורה פתוחים, כך שתאפשר הנגשה קלה ומהירה של יכולות אלו.
   3. משך ביצועו של המיזם נדרש להיות עד שנתיים לכל היותר לכלל התכולות המאופיינות להלן.
2. **מרכיביו המרכזיים של המיזם** 
   1. ‏13/09/2021פרקים ראשיים ומספר סעיפים (פרוייקטים) בכל פרק. כאמור בגוף המסמך, ניתן להציע נושאים ופעילויות נוספות ככל וכאלה עשויות לקדם את תחום ה-HebNLP.
   2. פרק 1: פיתוח גנרי של אלגוריתמים מבוססי NLP ספציפי לשפות שמיות במגוון מתארים, מורפולוגיה, סגמנטציה, ניתוח תחבירי, תלויות, ישויות, קשרים בין ישויות, צ'אטבוט בעברית ברמת הבנת הדיאלוג וניתוח הקשרים.
      1. מיפוי מחקרים נוכחיים באקדמיה הישראלית ובתעשייה בנושא NLP מתקדם לשפות עשירות במורפולוגיה (MRLs) ובנושאים משיקים.
      2. מו"פ יעודי בנושא מורפולוגיה וסגמנטציה לשפות שמיות (MRLs).
      3. מו"פ יעודי בנושא ניתוח תחבירי ותלויות.
      4. מו"פ יעודי בנושא חילוץ וזיהוי ישויות כולל חילוץ וניתוח קשרים בין ישויות.
      5. מו"פ יעודי בנושא תימצות (Summarization) בשפות שמיות.
      6. מו"פ יעודי של Hebrew Conversational AI (יאפשר פיתוח צ'אטבוט בעברית).
      7. מו"פ יעודי בנושא איחזור ודירוג מסמכים בשפות שמיות.
      8. מיצור (Productization) של כיוונים שהתפתחו במו"פ והבשילו - באמצעות פיתוח תוכנה בסטנדרט של Production בקוד פתוח.
      9. התמודדות עם טקסטים מסומלצים.
   3. פרק 2: אימון מודל שפה ענק בעברית, מבוסס בלשנות שפות שמיות, כולל מחקר יישומי לשפות שמיות.
      1. סקירת חלופות, בחירת ארכיטקטורה, חומרה, ודאטהסטים.
      2. היערכות טכנית חומרתית.
      3. היערכות טכנית - תוכנה ואלגוריתמיקה.
      4. היערכות טכנית - ניקוי דאטהסט.
      5. אימון.
      6. ניתוח תוצאות ושימושים במשימות Downstream שונות.
      7. בניית ופיתוח API מול ממשקים שונים נדרשים.
      8. Domain adaptation רפואי, משפטי, ןכו' (ניתן גם שילוב של כלי שיזהה את הדומיין)
   4. פרק 3: קוד פתוח לחבילות תשתיות למימוש גנרי של NLP כקוד פתוח - שימוש לאימון מודלי שפה - UD Universal Dependencies, הנגשה כקוד פתוח של Datalakes, העברה לחוקרים ושיתוף קהילת מפתחים.
      1. סקירת חבילות קיימות, יתרונות, חסרונות, פערים.
      2. עדכון וטיוב UD Universal Dependencies בעברית - כולל Treebank מטוייב.
      3. פיתוח/ביגור פארסר בביצועים גבוהים כולל תמיכה טובה בעברית בקוד פתוח (דוגמת YAP).
      4. פיתוח/ביגור תוסף בקוד פתוח לחיפוש טקסטואלי איכותי בעברית (מעל תשתית ביג דאטה שתיבחר כגון Elastic).
      5. פיתוח/ביגור מנוע תרגום בקוד פתוח לתרגום בין שפות היעד - עברית, ערבית.
      6. פיתוח / ביגור מסווג מסמכים (Supervised and Unsupervised) בעברית ובערבית.
   5. פרק 4: יכולות מעבר משמע לטקסט בעברית - כולל פיתוח מדדים לשילוב בין ניתוח טקסט לניתוח קול ותובנות למודלים משולבים - מולטי מודליטי, טקסט, קול ותמונה.
      1. סקירת חלופות, בחירת ארכיטקטורה, חומרה, ודאטהסטים.
      2. מו"פ ייעודי - עיבוד קול בעברית (זיהוי דובר, זיהוי מילות מפתח, Voice To Text).
      3. מו"פ ייעודי - OCR בעברית במגוון מתארים.
      4. מחקר משולב - מולטי מודליטי קול, טקסט, תמונה.
      5. מיצור (Productization) של המו"פ - באמצעות פיתוח תוכנה בסטנדרט של Production בקוד פתוח.
   6. פרק 5: מודלים היסקיים ופיתוח מודלים גנריים של בחינת מודלים היסקיים מבוססי NLP.
      1. מו"פ גרף הידע הלאומי.
      2. מו"פ חיבור ישויות – אונטולוגיה.
      3. מו"פ אבני בניין למנוע היסקים לישויות טקסטואליות.
   7. פרק 6: יצירה והנגשה של קורפוסים ודאטהסטים - משפחות העניין, כמות, איכות, רמות ניתוח - מהרמה הבסיסית של מבנה המשפט Surface Level ועד רמת הסמנטיקה העמוקה
      1. סקירת קורפוסים קיימים באקדמיה ובתעשייה.
      2. הקמת / בחירת תשתית לתיוג, לשיתוף ולבחינה.
      3. ריכוז / רכש / הבנייה של מאגר טקסטואלי לא מתויג בכמות מספקת, בדומיין רחב מספיק בעברית וערבית.
      4. איזה דאטהסטס היו ממליצים?
      5. הכנת חמישה דאטהסטים מתויגים
      6. יצירת בנצ'מארק (כולל חלוקת Train/Test וכולל סביבת Test ל"הגשות") לבחינת ביצועי אלגוריתמים על משימות שיוגדרו מעל הדאטהסטים שהוכנו.
   8. פרק 7: הסברתיות והגנה על מודלי בינה מלאכותית - בהקשרי מידול שפות ו- nlp.
      1. מו"פ בנושא Explainability.
      2. מו"פ בנושא הגנה מפני מתקפות כגון Adversarial.
   9. פרק 8 - NLP עיבוד נתונים ברמת החיישנים - ממוקד בהקשרי NLP ושפות שמיות, כולל הבנייה והרזיית רשתות נוירונים, Federated Learning, התייחסות ל- Edge Privacy.
      1. מו"פ בנושא עיבוד בקצה.
      2. מו"פ בנושא Federated Learning.
      3. מו"פ בנושא הרזיית רשתות בטכניקות כגון Pruning.
3. **הפרוט נדרש ברמת הפרק והסעיף**
   1. המשיב יפרט עבור כל פרק ולכל סעיף בו הוא מעוניין להתמודד, ככל שיוחלט על ביצוע המיזם.
   2. תיאור כללי של הפרק, תפיסת פעולה, שימושים, טכנולוגיות וכו', וכן פרוט ספציפי לכל סעיף.
   3. הנושאים בהם להבנתו של המשיב הפרק /התחום חסר ונדרש להשקיע בו .
   4. תכולות בהם נכון להתמקד בראייתו והסיבות לכך. למשל- במידה ותפותח תכולה X, מי הגורם (סוג הגוף, חברות ו/או מערכות) שיעשו שימוש ביכולות שיפותחו?
   5. הישג כמותי. לדוגמה, בפרק המתאר ביגור אלגוריתמים יצוין כי ניתן לבגר  X אלגוריתמים ואת פרוט האלגוריתמים אותם נכון לבגר.
   6. כלים/אלגורתמים קיימים מובילים במידה וקיימים.
   7. ידע נוסף וכל נתון אחר שרלוונטי לדעת המשיב.
4. **דרישות מהספק - קוד פתוח**
   1. במידה ולמשיב יש ניסיון בפיתוח ותחזוקה של פרויקטים ב"קוד פתוח" – יש להגיש תיאור של עד 3 פרויקטים ראשיים, חלקם של האנשים הרלוונטים אצל המשיב לפיתוח הפרויקט וקישור לקוד הפתוח של המערכת.
   2. המשיב יפרט ויפנה לפרסומי קוד פתוח שפרסם בעבר – למשל בפלטפורמת GITHUB (רישיון MIT או APACHE ) או פלטפורמות מקבילות.
5. טבלת תמחור
   1. המשיב יעריך השקעת כ"א צפויה (בשנות אדם) ועלות (בש"ח). יש לפרט בעלי מקצוע/תפקידים שונים (תכניתן, data scientist וכו'). יצוין כי ניתן לערב קבלני משנה. יש לפרט קבלנים עיקריים ( מעל 10% מהתקציב).
   2. אין לתמחר עלות לפרק באופן כללי אלא עלות וכ"א לכל סעיף (שורה בעלת סיעוף משנה, לדוגמה – 1.2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WBS #** | תיאור המשימה | הערכת עלות כספית | פרוט השקעת כ"א צפוי  (מקצוע ושעות סה"כ) | השקעה כספית של החברה בפרויקט |
|  |  |  |
|  |  |  |
|  |  |  |
| **פרק 1** | פיתוח גנרי של אלגוריתמים מבוססי **NLP** ספציפי לשפות שמיות במגוון מתארים, מורפולוגיה, סגמנטציה, ניתוח תחבירי, תלויות, ישויות, קשרים בין ישויות, צ'אטבוט בעברית ברמת הבנת הדיאלוג וניתוח הקשרים |  |  |  |
| 1.1 | מיפוי מחקרים נוכחיים באקדמיה הישראלית ובתעשייה בנושא NLP מתקדם לשפות עשירות במורפולוגיה (MRLs) ובנושאים משיקים |  |  |  |
| 1.2 | מו"פ יעודי בנושא מורפולוגיה וסגמנטציה לשפות שמיות (MRLs) |  |  |  |
| 1.3 | מו"פ יעודי בנושא ניתוח תחבירי ותלויות |  |  |  |
| 1.4 | מו"פ יעודי בנושא חילוץ וזיהוי ישויות כולל חילוץ וניתוח קשרים בין ישויות |  |  |  |
| 1.5 | מו"פ יעודי בנושא תימצות (Summarization) בשפות שמיות |  |  |  |
| 1.6 | מו"פ יעודי של Hebrew Conversational AI (יאפשר פיתוח צ'אטבוט בעברית) |  |  |  |
| 1.7 | מו"פ יעודי בנושא איחזור ודירוג מסמכים בשפות שמיות |  |  |  |
| 1.8 | מיצור (Productization) של כיוונים שהתפתחו במו"פ והבשילו - באמצעות פיתוח תוכנה בסטנדרט של Production בקוד פתוח |  |  |  |
| 1.9 | התמודדות עם טקסטים מסומלצים |  |  |  |
| פרק 2 | אימון מודל שפה ענק בעברית, מבוסס בלשנות שפות שמיות, כולל מחקר יישומי לשפות שמיות |  |  |  |
| 2.1 | סקירת חלופות, בחירת ארכיטקטורה, חומרה, ודאטהסטים |  |  |  |
| 2.2 | היערכות טכנית חומרתית |  |  |  |
| 2.3 | היערכות טכנית - תוכנה ואלגוריתמיקה |  |  |  |
| 2.4 | היערכות טכנית - ניקוי דאטהסט |  |  |  |
| 2.5 | אימון |  |  |  |
| 2.6 | ניתוח תוצאות ושימושים במשימות Downstream שונות |  |  |  |
| 2.7 | בניית ופיתוח API מול ממשקים שונים נדרשים |  |  |  |
| 2.8 | Domain adaptation רואי, משפטי, ןכו' (ניתן גם שילוב של כלי שיזהה את הדומיין) |  |  |  |
| פרק 3 | קוד פתוח לחבילות תשתיות למימוש גנרי של **NLP** כקוד פתוח - שימוש לאימון מודלי שפה - **UD Universal Dependencies**, הנגשה כקוד פתוח של **Datalakes**, העברה לחוקרים ושיתוף קהילת מפתחים |  |  |  |
| 3.1 | סקירת חבילות קיימות, יתרונות, חסרונות, פערים |  |  |  |
| 3.2 | עדכון וטיוב UD Universal Dependencies בעברית - כולל Treebank מטוייב |  |  |  |
| 3.3 | פיתוח/ביגור פארסר בביצועים גבוהים כולל תמיכה טובה בעברית בקוד פתוח (דוגמאת YAP) |  |  |  |
| 3.4 | פיתוח/ביגור תוסף בקוד פתוח לחיפוש טקסטואלי איכותי בעברית (מעל תשתית ביג דאטה שתיבחר כגון Elastic) |  |  |  |
| 3.5 | פיתוח/ביגור מנוע תרגום בקוד פתוח לתרגום בין שפות היעד - עברית, ערבית |  |  |  |
| 3.6 | פיתוח / ביגור מסווג מסמכים (Supervised and Unsupervised) בעברית ובערבית |  |  |  |
| פרק 4 | יכולות מעבר משמע לטקסט בעברית - כולל פיתוח מדדים לשילוב בין ניתוח טקסט לניתוח קול ותובנות למודלים משולבים - מולטי מודליטי, טקסט, קול ותמונה |  |  |  |
| 4.1 | סקירת חלופות, בחירת ארכיטקטורה, חומרה, ודאטהסטים |  |  |  |
| 4.2 | מו"פ יעודי - עיבוד קול בעברית (זיהוי דובר, זיהוי מילות מפתח, Voice To Text) |  |  |  |
| 4.3 | מו"פ יעודי - OCR בעברית במגוון מתארים |  |  |  |
| 4.4 | מחקר משולב - מולטי מודליטי קול, טקסט, תמונה |  |  |  |
| 4.5 | מיצור (Productization) של המו"פ - באמצעות פיתוח תוכנה בסטנדרט של Production בקוד פתוח |  |  |  |
| פרק 5 | מודלים היסקיים ופיתוח מודלים גנריים של בחינת מודלים היסקיים |  |  |  |
| 5.1 | מו"פ גרף הידע הלאומי |  |  |  |
| 5.2 | מו"פ חיבור ישויות - אונטולוגיה |  |  |  |
| 5.3 | מו"פ אבני בניין למנוע היסקים לישויות טקסטואליות |  |  |  |
| פרק 6 | יצירה והנגשה של קורפוסים ודאטהסטים - משפות העניין, כמות, איכות, רמות ניתוח - מהרמה הבסיסית של מבנה המשפט **Surface Level** ועד רמת הסמנטיקה העמוקה |  |  |  |
| 6.1 | סקירת קורפוסים קיימים באקדמיה ובתעשייה |  |  |  |
| 6.2 | הקמת / בחירת תשתית לתיוג, לשיתוף ולבחינה |  |  |  |
| 6.3 | ריכוז / רכש / הבנייה של מאגר טקסטואלי לא מתויג בכמות מספקת, בדומיין רחב מספיק בעברית, ערבית ורוסית |  |  |  |
| 6.4 | הכנת חמישה דאטהסטים מתויגים |  |  |  |
| 6.9 | יצירת בנצ'מארק (כולל חלוקת Train/Test וכולל סביבת Test ל"הגשות") לבחינת ביצועי אלגוריתמים על משימות שיוגדרו מעל הדאטהסטים שהוכנו |  |  |  |
| **פרק 7** | **הסברתיות והגנה על מודלי בינה מלאכותית - בהקשרי מידול שפות ו- NLP** |  |  |  |
| 7.1 | מו"פ בנושא Explainability |  |  |  |
| 7.2 | מו"פ בנושא הגנה מפני מתקפות כגון Adversarial |  |  |  |
| **פרק 8** | **עיבוד נתונים ברמת החיישנים - ממוקד בהקשרי NLP ושפות שמיות, כולל הבנייה והרזיית רשתות נוירונים, Federated Learning, התייחסות ל- Edge Privacy** |  |  |  |
| 8.1 | מו"פ בנושא עיבוד בקצה |  |  |  |
| 8.2 | מו"פ בנושא Federated Learning |  |  |  |
| 8.3 | מו"פ בנושא הרזיית רשתות בטכניקות כגון Pruning |  |  |  |