

מסמך דרישות לפרויקט בקורס כריית טקסטים

תקציר הפרויקט

מטרת הפרויקט היא לישם תהליכי כריית טקסטים לפתור בעיה מוגדרת מראש ובסיסת מחקר. הפרויקט כולל שלבים של איסוף נתונים, עיבוד מקדים, מימוש אלגוריתמים מתאימים, ניתוח鄙出ים והצעת שיפורים. הפרויקט כולל חמשה שלבים עיקריים, כאשר כל שלב תורם למימוש כולל ואיכותי של המערכת.

שלבי הפרויקט

שלב א': איסוף ועיבוד מקדים של הנתונים (Data Collection & Preprocessing)

מטרות:

1. איסוף נתונים:

- השגת מידע נתונים רלוונטי מקורות אמינים (API, קבצי CSV, scraping מ אתרים).
- OIDAO רישיון נתונים ושמירה על פרטיות.

2. עיבוד מקדים (Preprocessing):

- ניקוי נתונים: הסרת תווים מיוחדים, טיפול בטקסטים כפולים, הסרת שורות ריקות.
- עיבוד טקסטים: הסרת stopwords, המרה לאותיות קטנות, סטמינג ולמנצ'יזה.
- יצירת תכונות חדשות: זיהוי ישויות (Named Entity Recognition), ניתוח רגשות (Topic Modeling), או זיהוי נושאים (Sentiment Analysis).

3. ייצוג טקסטים (Feature Engineering):

- בחירת שיטת ייצוג: Bag of Words, TF-IDF, Word2Vec, FastText, BERT.
- ניתוח השפעת תכונות (Exploratory Data Analysis) להדגשת תכונות מעניינות נתונים.

תוצריים נדרשים:

- מידע נתונים נקי ומוכן לעבוד.
- מחברת Jupyter עם ניתוחים ויזואליים (גרפים, מילימנס, word clouds).

- פסקה קצרה המתארת את ה-Dataset: מה מקורו? متى נאסף? מהן מגבלותיו? אילו קבוצות באוכלוסייה הוא מייצג?

שלב ב': מימוש אלגוריתמים מתקדמים לפתרון הבעיה (Modeling)

מטרות:

1. הגדרת בעיה למידה:

- סיווג טקסטים (Classification), ניתוח נושאים (Topic Modeling), זיהוי ישויות (NER), או ניבוי טקסט (Text Prediction).

2. בחירת אלגוריתמים:

- עבור בעיות סיווג: Logistic Regression, SVM, Naive Bayes, Decision Trees, או מודלים מתקדמים (LSTM, BERT, GPT).

- עבור בעיות ניתוח נושאים: Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA).

3. אימון מודלים:

- חלוקת הנתונים ל-Train/Validation/Test (ביחסים 70/15/15).

- שימוש בטכניות לערבול נתונים(Cross-Validation).

- כוונון היפר-פרמטרים (Hyperparameter Tuning) באמצעות Grid Search או Bayesian Optimization.

תוצרים נדרשים:

- קוד מארגן התומך בהרצה מקצה לקצה משורת הפוקודה.

- מחברת Jupyter המתארת את האלגוריתמים והבחירה שנעשו.

- שבירת המודלים המאומנים (Model Serialization) לשימוש עתידי.

שלב ג': אבלוואציה וניתוח התוצאות (Evaluation & Analysis)

מטרות:

1. מדידת ביצועים:

- שימוש במדדים מתקדמים: Accuracy, Precision, Recall, F1-Score, Confusion Matrix, AUC-ROC.
- מדדים נוספים בהתאם לאופי הבעיה (למשל, Perplexity בניתוח נושאים, BLEU בBINBO טקסט).

2. ניתוח שגיאות (Error Analysis):

- בחינת מקרי כשל (False Positives/Negatives) וזיהוי דפוסים אפשריים לשיפור.
- 3. ויזואлизציה של תוצאות:

- שימוש בגרפים להערכת ביצועי המודלים Precision-Recall Curve, Confusion Matrix Heatmap.
- ניתוח תובנות מרכזיות מתוך התוצאות.

תוצרים נדרשים:

- מחברת Jupyter עם גרפים וניתוח תוצאות.
- ניתוח שגיאות (Error Analysis): להציג דוגמאות קונקרטיות של טויות False Positives/Negatives, ולנתח מדוע המודל טעה. האם הטעות נובעת מטיקסט דו-משמעות בנתונים מסוים? מבעה בעיבוד המקדמים? זה לב ליבור של שיפור המודל.
- סיכום מסקנות והמלצות לשיפור ביצועים.

שלב ד': הצעת אלגוריתם משופר וניתוח חדש (& Re-evaluation)

מטרות:

1. שיפור המודל:

- שימוש בטכניקות מתקדמות כגון Ensemble Learning (Random Forest, Gradient Boosting), או Fine-Tuning (BERT, GPT).
- כוונון נוסף של היפר-פרמטרים על בסיס ניתוח שגיאות מהשלב הקודם.

2. אבולוציה חדשה:

- בוחינת הביצועים לאחר השיפורים בהשוואה לنتائج המקוריות.
- הערכת השפעת השיפורים באמצעות מדדים זרים לשלבים הקודמים.

תוצרים נדרשים:

- קוד מעודכן עם האלגוריתם המשופר.
- השוואה בין תוצאות המודלים לפני ואחרי השיפורים (גרפים, טבלאות).
- שלב זה צריך להתבסס ישירות על המסקנות מניתוח השגיאות. לדוגמה:
 - טענה: "המודל טועה במקרים של סרקוז. אולי הוספה תכונות רגש תעזר." -> ניסוי: הוספה Sentiment Score כתכונה ובדיקה השיפור.
 - טענה: "נראה שהמודל מתקשה עם מילים נדירות." -> ניסוי: שימוש בטכניקות Data Augmentation כמו החלפת מילים נרדפות או תרגום-חזרה (Back-translation) כדי להעשיר את הנתונים.

שלב ה': כתיבת דוח מסכם והגשת הפרויקט (Final Report & Submission)

מטרות:

1. הכנות דוח מספורט:

- מבוא: רקע, הגדרת הבעיה, מטרות הפרויקט.
- סקירה ספרות: סיכום מחקרים ופתרונות קיימים בתחום.
- מתודולוגיה: פירוט שלבי העבודה, בחירת אלגוריתמים וייצוג הנתונים.
- תוצאות וניתוח: סיכום הביצועים, ניתוח שגיאות, ותובנות.
- סיכום והמלצות: תובנות מרכזיות, מגבלות הפרויקט והמלצות להמשך.

2. הגשת הפרויקט:

- העלאת הקוד, הדוח והמחברות למאגר GitHub של המעבדה.
- ידאו שהקוד תומך בהרצה מקצת קצרה בעצרת שורת הפקודה.

דרישות נוספת:

- הדוח יוגש בפורמט PDF. מומלץ לעבוד עם Latex להיות והוא הסטנדרט המקובל במדעי המחשב בגורנלים וכנסים מתקדמים.
- שמירה על סטנדרטים לכתיבה קוד ופרויקט מתועד לקרוא.

דרישות טכניות כלליות:

- שפת תכנות: Python.
- כלי פיתוח: Jupyter Notebook, Google Colab, PyCharm/VSCode.
- ספריות מומלצות: NLTK, SpaCy, Scikit-Learn, TensorFlow, PyTorch, Hugging Face Transformers, Pandas, Matplotlib, Seaborn.
- ניהול גרסאות: שימוש ב-Git לניהול גרסאות, כאשר הפרויקט יכול להיות מאוחסן במאגר GitHub ייעודי של המעבדה.

לוח זמנים (אופציונלי)

חלוקת שלבי הפרויקט לפי לוח זמנים

שלב	כתובת	שבועות	תיאור	זמן משוער להשלמה
שלב א'	איסוף ועיבוד מקדים של הנתונים	שבוע 1-4	4 שבועות	
שלב ב'	הימוש אלגוריתמים מתקדמים	שבוע 5-8	4 שבועות	
שלב ג'	אבלוציה וניתוח תוצאות	שבוע 9-10	2 שבועות	
שלב ד'	הצעת אלגוריתם משופר וניתוח מחדש	שבוע 11-12	2 שבועות	
שלב ה'	כתבת דוח והגשת הפרויקט	שבוע 13	1 שבוע	

קריטריונים להערכת הפרויקט

חוובה לציין על שלבי הביצוע וחלוקת הפרויקט באחריות מי זה נעשה על מנת שאוכל לתת ציון אישי.

1. אינטואיטיביות ועומק העיבוד המקדים (20%).
2. בחירת אלגוריתמים והצדקות טכניות (25%).
3. מדרדי ביצועים וניתוח שגיאות (20%).
4. חידשנות ושיפור ביצועים (20%).
5. כתיבת דוח ועמידה בדרישות טכניות (15%).

אתרים רלוונטיים -

- אתר [huggingface](#) - קהילת开源 העולמית של בינה מלאכותית. אפשר למצוא שם מגוון [datasets](#) ובעיקר [מודלים](#) מכל הסוגים.
- אתר Kaggle - אתר שמציע תחרויות שונות בתחום science data. אפשר לבחור ממשם [datasets](#) או ממש להירשם [לתחרויות](#) ועל בסיסה לבצע את הפרויקט.
- [מאג'ר הNLP הלאומי](#) של ישראל שמכיל נתונים, מודלים וכל מה שצריך כדי לבצע ניתוח טקסטים בעברית וערבית.

תחרויות -

[SemEval2026](#) -

[Clef2026](#) -

[Evalita2026](#) -

הרשות מכילה הרבה מאוד מאגרים ותחרויות שלא מפורטים פה, מוזמנים לחפש.

בהצלחה בפרויקט!