עבודה מסכמת ניתוח שפות טבעיות תשפ"ה

תאריך ההגשה: 30/6/2025

הערה מהסילבוס:

<u>הערות מחייבות:</u>

- במקרה של ציון נכשל בבחינה/פרוייקט סופי - לא ייחשב ציון התרגילים.
- בעבודה המסכמת שהיא פרוייקט או פרוייקט בחלקים, הסטודנט יגיש את העבודה, שהיא בעצם כתיבת קוד. תהיה בחינת הגנה על הפרויקט המהווה את הציון של ה-70%.

רקע:

לאחר שסיימתם בהצטיינות את לימודיכם, התקבלתם לעבודה חדשה בצוות הדאטא של רשת מסעדות גדולה הפועלת בכל רחבי הארץ. הבוס החדש שלכם פונה אליכם עם משימה דחופה – לקוחות רבים משאירים ביקורות על המסעדות, והוא רוצה לפרסם בדף הנחיתה של כל מסעדה את הביקורות החיוביות בלבד – באופן אוטומטי. כדי לעשות זאת, תצטרכו לבנות מודל שמזהה ביקורות חיוביות מתוך קובץ דאטה קיים של ביקורות מתויגות. הבוס מצפה לראות תוצאות כבר בסוף השבוע – אז קדימה לעבודה!

חומרים:

יש להשתמש בקובץ הבא:

https://www.kaggle.com/datasets/ziadmostafa1/restaurant-reviews

חלק ראשון: הכנה והכרות עם הדאטה:

1. הקובץ התקבל בפורמט tab separated value - tsv
   כתוב סקריפט פייטון אשר ממיר אותו לפורמט csv
2. כתוב קוד שמוצא כמה ביקורות חיוביות יש וכמה שליליות
3. כתוב קוד אשר מוצא את האורך הממוצע לביקורת חיובית ולביקורת שלילית
4. האם מתוך המידע שקיבלתם בתשובות שבסעיפים 2 ו3 ניתן להסיק שקבוץ המידע טוב או לא טוב לשימוש? (תשובה מילולית בבלוק טקסט במחברת)
5. כתוב קוד אשר מוצא מתוך הביקורות השליליות מה הם 5 מילות המפתח שחוזרות הכי הרבה פעמים ומה ניתן ללמוד עליהן מהדברים שחשובים למבקרים

6. כתוב קוד אשר מחזיר את הביקורת הארוכה ביותר והאם היא חיובית ושלילית
7. כתוב קוד שמבצע NER (Recognition Entity Named) לביקורת הארוכה מהסעיף הקודם. לאחר מכן, הצג את הישויות המוכרות ואת סוגי הישויות

חלק שני : ייצוג וניתוח ביקורות על בסיס ישויות מזוהות(NER)

שלב א: זיהוי ישויות בכל ביקורת

הריצו Named Entity Recognition על כל הביקורות בקובץ.

שמרו את הישויות (entities) שנמצאו בכל ביקורת – לדוגמה: שמות של מסעדות, מנות, ערים, תאריכים וכו.'

שלב ב: יצירת ייצוג המבוסס על ישויות

עבור כל ביקורת, צרו טקסט חדש שמכיל רק את הישויות שזוהו.
לדוגמה, אם בביקורת הופיעו "the manager at Tel Aviv branch was rude" הטקסט החדש יהיה.["Tel Aviv", "manager"] :

בנו מטריצת תכונות (features) מהטקסטים החדשים באמצעות TF-IDF או.Bag of Words

שלב ג: אימון מודלים על בסיס ייצוג הישויות

חלקו את הדאטה ל־.Train/Test

אמן את כל האלגוריתמים הבאים על ייצוג ה־:NER

Logistic Regression

Random Forest Classifier

Support Vector Machine (SVC)

Multinomial Naive Bayes

K-Nearest Neighbors (KNN)

שלב ד: הערכת ביצועים

מדדו עבור כל מודל את:

Accuracy

Precision

Recall

F1 Score

הציגו מטריצת בלבול (Confusion Matrix) לכל מודל.

שלב ה: ניתוח ראשוני

האם ייצוג מבוסס-ישויות מספק ביצועים טובים?

אילו סוגי ישויות תורמות לדעתכם להבנה של ביקורת חיובית?

חלק שלישי : בניית מודלים על בסיסBag of Words

שלב א: הסרת סימני פיסוק

מה לעשות:

- עברו על כל הביקורות והסירו מהן סימני פיסוק (כגון נקודות, פסיקים, סימני קריאה ושאלה, גרשיים וכו').
- הסבירו את החשיבות של הסרת סימני פיסוק

שלב ב: החזרת מילים לצורתן הבסיסית

מה לעשות:

- בצעו המרה של כל מילה לצורתה הבסיסית בעזרת אחת מהשיטות lemmatization : או stemming
- שמרו את הטקסט החדש בקובץ CSV חדש.

הסבירו בכתיבה:

- איזו שיטה בחרתם ולמה?
- למה חשוב להחזיר מילים לצורתן הבסיסית?

שלב ג: יצירת ייצוג בשיטתBag of Words

מה לעשות:

- השתמשו בשיטת **Bag of Words** כדי להמיר את כל הטקסטים למערך מספרי.
- שמרו את הפלט כקובץ CSV נוסף.

הסבירו בכתיבה:

מה היא שיטת bag of words ומה החשיבות שלה בעבודה של ניתוח הטקסט

שלב ד: אימון מודלים

משימה:

- חלקו את הדאטא ל־Train/Test.
- אמן את כל המודלים הבאים:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Support Vector Machine (SVC)**
4. **Multinomial Naive Bayes**
5. **K-Nearest Neighbors (KNN)**

שלב ה: הערכת ביצועי המודלים

מה לעשות:

- לכל מודל הציגו את המדדים הבאים:
  - **Accuracy**

- **Precision**
- **Recall**
- **F1 Score**

- הציגו **מטריצת בלבול (Confusion Matrix)** לכל מודל.

שלב ו: בחירת מודל והסקת מסקנות

מה לעשות:

- נתחו את ביצועי המודלים.
- בחרו את המודל שהציג את התוצאות הטובות ביותר
- הסבירו מדוע בחרתם במודל זה למשימה בפועל.

הוראות הגשה:

1. יש לעשות את כלל העבודה במחברת COLAB/JUPYTER
2. העבודה צריכה להיות מסודרת:

-לכל שלב צריך להיות כותרת בבלוק טקסט

-שלבים בהם יש קוד, הקוד נמצא בבלוק קוד

-שלבים בהם צריך לכתוב מלל הכתיבה תהיה בבלוק טקסט

-המחברת תתחיל עם שמות המגישים

-תשובה מילולית יכולה להיות בעברית

-תיעוד והסברים בתוך הקוד חייבים להיות באנגלית

3. יש להגיש:

-קובץ ipynb עם כלל התשובות המילוליות והקוד

-קובץ csv עם הדאטה המקורי

-קובץ txt עם שמות כלל הספריות שצריך להוריד כדי שהקוד יעבוד