מיני פרוייקט בעיבוד שפה טבעית - דו״ח מסכם לניסוי

בניית מסווג עבור מופעי המילה "אך" בתנ״ך

מגישים: יעל חדד | אורי סוכי
מנחה: ד״ר מני אדלר

משימת עיבוד השפה העברית מאתגרת גם כאשר מדובר על עברית מודרנית, עבורה מאגרי הדאטה המתוייגים הולכים ומתרבים ומודלים רבים לבעיות שונות קיימים ואף ניתנים להתאמות לבעיות ספציפיות.

עם זאת, משימה זו עבור הקורפוס התנ״כי מאתגרת אף יותר. סיבה אחת ומרכזית היא שהקורפוס הוא סגור ותיוגו משמעותו מתן פרשנות לתנ״ך, וזוהי משימה קשה בפני עצמה, ופירושים רבים ושונים קיימים ורבים מהם נתונים עד היום במחלוקת.

משימתנו בפרויקט זה הינה בניית מסווג למציאת המשמעות של המילה "אך" בתנ"ך. את המשמעויות האפשריות לקחנו מתוך המילון והן:

- אכן
- רק
- אבל
- זה עתה

תהליך העבודה החל בתיוג כלל הדוגמאות מהפסוקים השונים באופן ידני, המשיך ביצירת וקטורי מאפיינים לכל פסוק וסוכם באימון מסווג על הדוגמאות ובדיקת איכותו בעזרת הכלי Weka.

בנוסף למשימה הראשית, לאחר בחינת תוצאות הניסוי, החלטנו להוסיף שלב נוסף ומעניין, הבוחן את חשיבות גודל הדאטה להצלחת המסווג על ידי הקטנת מספר הדוגמאות בשני אופנים שונים - בחירה אקראית ובחירה המאזנת בין יחס הדוגמאות והתגים המתאימים.

ראינו באופן מובהק כי אפילו בסדרי גודל קטנים, ככל שמספר הדוגמאות גדל, כך גדל הדיוק בסיווג ביחס לתיוג הידני.

מפורטים שלבי התהליך הכוללים קישורים לקבצי דאטה וקוד באמצעותם בנינו את המסווג, ומרוכזים בעמוד github.

קריאה מהנה.

שלבי הניסוי:

- **בניית אוסף הדוגמאות (Preprocess):**
  - משכנו לקובץ csv את כל הפסוקים שהמילה "אך" מופיעה בהם, באמצעות "כלי חיפוש בתנ"ך" של Dicta.
  - במקביל, מתוך קבצי ה- TEI של ניתוח התנ"ך של Dicta משכנו את כל התגיות של הפסוקים שהמילה "אך" מופיעה בהם.
  - כיוון שהמידע בקובץ ה-csv ממוין לפי סדר הספרים בתנ"ך, בעוד ה-TEI ממוין לפי סדר לקסיקוגרפי, מיינו את שני האוספים ככה שסדר הפסוקים מקביל.
  - במידה והמילה "אך" מופיעה פעמיים באותו משפט היה צורך לשכפל את המשפט והשמת דגש על כל מופע "אך" בנפרד.
  - כיוון שהדאטה-סט קטן מאוד (161) השתמשנו בכלל הדוגמאות.

- **תיוג ידני של הדוגמאות:** [מסמך]
- המדריך שלפיו תייגנו:
  - הצבת מילים נרדפות ששומרות על הסמנטיקה.
  - השוואה לתרגום באנגלית.
  - לפי פרשני מקרא (רש"י, שטיינזלץ, מתוך "דעת מקרא" ועוד)
- הערות על התיוג:
  - בשלב התיוג הידני שמנו לב לתכונות סמנטיות כמו "מי הדובר" (למשל דמות סמכותית נוהגת לתת פקודות, ולכן השימוש היה "רק" כלומר הצבת תנאי, לעומת נתין שמדבר אל דוד המלך ו"מסביר את עצמו" ולכן הפירוש "אבל" היה המתאים - שמואל א, כ"ט ט').
  - בשלב התיוג הידני שמנו לב כי מיקום בראש המשפט (כחצי מהדוגמאות) לאו דווקא אפיין פירוש מסויים.
  - התיוג "רק" הוא הנפוץ ביותר עם למעלה מ-50% מהדוגמאות. "זה עתה" מופיע נדירות.
  - הפירוש ב-Dicta לכל 161 המופעים הוא "רק".

- **המרת הדאטה לוקטורי מאפיינים:** [קוד]
- סוגי המאפיינים שבחרנו:
  - המילים (ערך מילוני)
  - המאפיינים המורפולוגיים
  - המאפיינים התחביריים
  - ערך מילוני + מאפיינים מורפולוגיים
  - ערך מילוני + מאפיינים תחביריים
  - מאפיינים מורפולוגיים + מאפיינים תחביריים

- המרת הדאטה:
  - על מנת ליצור וקטורים בחלונות הדרושים לפי 6 סוגי המאפיינים, ראשית שלפנו את האינדקס של המילה "אך" בכל פסוק, והכנו וקטורים של התגים של המילים לפני ואחרי בהתאמה לגודל החלון. (התגים הם אלו שמכילים את המטה דאטה הדרוש ליצירת וקטורי המאפיינים.)
  - בחלק מהמקרים היה דרוש ניקוי הדאטה - על ידי ספריית regex ניקינו טקסטים מתווים לא רלוונטיים.
  - מספר הדוגמאות למילה "אך" קטן, ומספר הצירופים המורפולוגיים היה גדול ומכאן שלא מאפשר למצוא "תבנית". דיללנו את המיון המורפולוגי לחלק הראשי בלבד.
  - מרגע שהחזקנו וקטורים של "מילים" , "מאפיין מורפולוגי", "תחבירי", "למה" היה צריך ליצור "מפתח" מתאים לכל טיפוס מבין המאפיינים. לשם כך הפכנו את הרשימות שקיבלנו ל"סט" ללא כפילויות, ונתנו אינדקס ייחודי לכל טיפוס. לאחר מכן מיפינו את הוקטורים בגודל החלון מ"ערך המאפיין" למפתח המתאים לו.
  - את הוקטורים העברנו לפורמט csv תואם לדרישות Weka
  - לכל csv שרשרנו את התיוגים עצמם בעמודה האחרונה, לפי הדרוש בפורמט של Weka.

- **אימון מסווג על הדוגמאות ובדיקת איכותו:** [קוד]

לאחר הכנת הדאטה באופן שיתקבל כקלט למודלים המסווגים של Weka, הכנו סקריפט שיריץ את ה input על מודלים שונים, ושמרנו את התוצאה לקובץ csv. [**מסמך**]

- כיוון שהדאטה-סט בגודל קטן (161) השתמשנו בכל הדוגמאות, ואת הלמידה עשינו על ידי Cross-Validation.
- לאחר ניסיונות על סוגים שונים של מודלים ש-Weka מציע, אספנו את מדד ה-Accuracy עבור שלושה מודלים מסוגים שונים שנתנו את התוצאות הטובות ביותר - Naive Bayes, Multilayer perceptron, Simple logistic.
- את איכות המודל נציג לפי מטריקת Accuracy - מספר התיוגים הנכונים.
  - המטריקה F1, המבוססת על Recall ו-Precision, ברוב המקרים לא הפיקה תוצאה, זאת בשל התיוג "זה עתה" עבורו $TP = FP = 0$ ועבור תוצאה כזו מטריקת Precision אינה מוגדרת וכנגזרת F1 אינה מוגדרת גם כן.
  - לשם השלמות נצרף את ה- Confusion Matrix של תוצאות נבחרות, שבכל זאת נותנות תמונה רחבה יותר לפי כל תיוג.

תוצאות:

- הדיוק הגבוה ביותר 57.764% התקבל מהפיצ'רים syntax+lemma עם גודל חלון 4 במסווג Multilayer perceptron.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  1 15  8  0 |  a = אכן
  0 71 13  0 |  b = רק
  0 25 21  0 |  c = אבל
  0  4  3  0 |  d = זה עתה
```

- הדיוק הנמוך ביותר 43.4783% התקבל מהמאפיינים Morphologic+Lemma עם גודל חלון 4 במסווג Multilayer perceptron.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  3 11 10  0 |  a = אכן
  1 56 27  0 |  b = רק
  3 32 11  0 |  c = אבל
  0  4  3  0 |  d = זה עתה
```

בנוסף, שמנו לב כי שליש מהתוצאות מסווגות את כל התיוגים כ"**רק**", שכן זהו התיוג הנפוץ ביותר ולכן מודל כזה יחזיר דיוק גבוה מ-50% בוודאות.

מסקנות ראשוניות ובניית ניסוי המשך:
לאחר שלב הסיווג ראינו כי עבור הדאטה שלנו לא ניתן לבחור מודל ספציפי שבאמת יוכל למדל את הבעייה של סיווג המילה "אך" בתנ"ך.

בחנו מודלים מחוץ ל-Weka כדי להבין האם המשימה אפשרית בעזרת מודלים אחרים.
נציין כי מודלים אלה מאומנים על **עברית מודרנית** ואינם מותאמים לטקסט תנ"כי.

- ניסיון להריץ את המשפטים עם מודל AlefBERT כאשר המילה "אך" ממוסכת לא העלה תוצאות מתוך התיוגים האפשריים, אלא מתוך vocabulay כללי.
- ניסיון בניתוח סנטימנט של מודל heBERT החזיר תוצאה שהסנטימנט שלילי לכל סוגי התיוגים על כ-10 דוגמאות.

המסקנה העיקרית והחזקה ביותר הייתה כי הבעייה טמונה בגודל סט האימון.
כדי לאשש את הטענה בנוגע לתלות הלמידה במספר הדוגמאות - הקטנו אף יותר את הדאטה בשני אופנים:

1. אקראית - בחירה של 80 דוגמאות.
2. איזון יחס התגים - בחירה של מספר מאוזן ככל האפשר.

תוצאות בחירה אקראית:

- הדיוק <u>הגבוה</u> ביותר 53.75% התקבל מהמאפיינים Syntax+Lemma עם גודל חלון 2 במסווג Multilayer perceptron.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  0 13  1  0 |  a = אכן
  0 41  0  0 |  b = רק
  3 19  2  0 |  c = אבל
  0  1  0  0 |  d = זה עתה
```

- הדיוק <u>הנמוך</u> ביותר 31.25% התקבל מהמאפיין Morphologic עם גודל חלון 4 במסווג Naive Bayes.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  2 11  1  0 |  a = אכן
 13 23  5  0 |  b = רק
  8 15  0  1 |  c = אבל
  1  0  0  0 |  d = זה עתה
```

תוצאות איזון יחס הדוגמאות:

- הדיוק הגבוה ביותר 51.8987% התקבל מהמאפיינים Vocabulary עם גודל חלון 4 במסווג Naive Bayes.
- גם התיוג "זה עתה" מקבל התייחסות במקרה המאוזן וקיבלנו F-score = 0.511

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 13  8  2  1 |  a = אכן
  0 20  4  0 |  b = רק
  0 12  7  5 |  c = אבל
  1  3  2  1 |  d = זה עתה
```

- הדיוק הנמוך ביותר 24.0506% התקבל מהמאפיינים Morphologic+Syntax עם גודל חלון 4 במסווג Multilayer perceptron.
- גם התיוג "זה עתה" מקבל התייחסות במקרה המאוזן וקיבלנו F-score = 0.225

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  3 10  9  2 |  a = אכן
  8 11  4  1 |  b = רק
  8 10  5  1 |  c = אבל
  2  3  2  0 |  d = זה עתה
```

- כצפוי - בדאטה המאוזן מטריצת ה-Confusion מפוזרת יותר, ואינה מתרכזת רק באחד התיוגים. הצלחנו למנוע בכך את הסיווג ה"קבוע" שנוצר כאשר יש פער גדול בין מספר דוגמאות עבור תג מסויים (במקרה שלנו - "רק").
- מעניין לציין שגם בדאטה המאוזן - רוב התיוגים המוצלחים (כלומר TP) הם עבור התיוג "רק". זה מלמד אותנו שהמאפיינים שבחרנו לטובת האימון עוזרים לאפיין את התיוג הספציפי הזה, אך אולי עבור התיוגים האחרים היה שווה לנסות מאפיינים אחרים, או הרכבות שונות.
- "זה עתה" מצליח לסווג 1-2 מתוך 7 מופעים במודל Bayes בדאטה סט שבו איזנו כמות מופעים.

סיכום ומסקנות:

כפי שציפינו, הקלטים הקטנים יותר הורידו משמעותית את אחוז הדיוק:

	ממוצע דיוק
דאטה מקורי	50.69%
מחצית הדאטה - אקראי	46.42%
מחצית הדאטה - מאוזן	38.92%

שאלנו את עצמנו למה התוצאות עבור הדאטה המאוזן נמוכות בצורה כה משמעותית. ייתכן שהסיבה טמונה בבחירת המאפיינים והתאמתם לסיווג "adverbs" בצורה טובה יותר, שכן מתוך התיוגים:

- רק "adverb"
- אכן "adverb"
- אבל "conjunction"
- זה עתה "preposition"

התגים "רק" ו"אך" זכו ל-Recall גבוה משמעותית מהאחרים. כלומר ייתכן וחלקי דיבר מסויימים מקלים על הסיווג.

היפר-פרמטרים ומאפיינים:

- גודל החלון: כיוון שגם המודלים הטובים ביותר וגם הגרועים ביותר חלקו מכנה משותף של גודל החלון, נסיק כי זהו היפר-פרמטר שעדיין יש לחקור עם מאפיינים נוספים ושונים.
- במצבים בהם המילה "אך" הופיעה בתחילת המשפט שמנו תג מיוחד. ייתכן ששווה לבדוק התייחסות למשפט הקודם במצבים אלה על מנת לקבל תמונה מדוייקת יותר.
- מאפיינים ושילובים של מאפיינים: השילובים בין סוגים שונים של מאפיינים בסיסיים הניבו תוצאות טובות (באופן יחסי), ולכן יהיה מעניין לבחון קומבינציות נוספות.

הכלים בהם השתמשנו לטובת המילה הספציפית יכולים להתאים לבניית מסווגים למילים אחרות באופן כללי, פרט לשלב התיוג הידני - כל השלבים גנריים וניתנים לריצה באמצעות הקוד המצורף.
את התוצאות שבחרנו לצרף ניתן להרחיב על ידי שימוש בקוד המריץ את קבצי הcsv בWeka, וכמובן על וקטורים נוספים שניתן ליצור בהתאמה.