בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

Assignment 3 - Understanding the Challenge

1. Bag of words -
בשיטה זו לומדים את השפה על ידי k המילים שסביב המילה המבוקשת. כלומר מתחשבים בקונטקסט הלוגי של השפה. בשפה שלנו אין משמעות למילים ואין תחביר וקשרים ביניהן. כל מילה היא בפני עצמה ושייכת לשפה רק על פי המבנה שלה בלי קשר לשאר המילים בשפה. לכן שיטה זו לא תניב למידה נכונה של השפה, ולא תתן תוצאות טובות.
2. Bigram or trigram -
בשיטה זו מפרקים כל משפט בשפה לזוגות או שלישיות של אותיות, ובהתאם לשכיחות האותיות בשפה מסווגים את הקלט האם הוא בשפה או לא. בשפה שלנו מה שמבדיל בין מילים שבשפה למילים שלא בשפה הוא סדר רצפי האותיות שבכל מקרה מופרדות על ידי רצף מספרים. כך שכשמסתכלים ברזולוציות של צמדי אותיות בתוך הקלט, לא ניתן להסיק על הקלט האם הוא בשפה או לא, שהרי שיטה זו לא מייחסת לחשיבות הסדר של הצמדים או השלישיות בתוך הקלט.
3. CNN -
במודל למידה זה עובדים עם פילטר בגדלים משתנים. שיטה זו נועדה להתמקד בפיצ'רים מסוימים ואזורים שונים בקלט שנותנים כובד משמעותי יותר לסיווג הנכון. גם שיטה זו מאבדת חשיבות לסדר ולרצף שמתקבל בקלט. בשפה שלנו כל הסיווג נופל על שאלת יחס הסדר בין האותיות. לכן יש חשיבות להסתכל על כל הרצף ולא להתמקד רק בחלקים מסוימים. לכן מודל זה לא מתאים למידת השפה שלנו.
היינו יכולים לנסות לקבוע פילטר בדיוק באזור "האמצעי" של הקלט בו ניתן לקבוע האם הגיעו רצפי b לפני רצפי c או להפך. אך בפועל לא ניתן להגדיר גודל ספציפי לפילטר שכזה כי הרצפים הם באורך אינסופי ולא מוגדר מראש ולכן גם שיטה זו לא תעזור ללמידה נכונה.