Can the two languages be distinguished using a bag-of-words approach? Explain why

התשובה היא לא. הסיבה לכך היא שהשפות המוצגות בפנינו בתרגיל זה לא מגדירות כללים לגבי רצפים ומשפטים. השיטה BOW משתמשת במספר כלשהו של מילים על מנת לקבוע האם מילה היא שייכת לשפה או לא, ולכן היא לא תצליח להבין את החוקים עבור מילה אחת. לכן ניתן לסכם שבתרגיל זה גישה לא תועיל לנו על מנת להבדיל בין שתי השפות.

הסבר נוסף – כזכור מודל BOW סופר כמה מילים יש מכל סוג (במקרה שלנו בתרגיל תווים) ולפי השכיחויות מסווג. במקרה שלנו זה לא יעבוד כי למשל 2a2b2c2d2 ו-2a2c2b2d2 יש בהן את אותן השכיחויות של מילים אבל הן מסיווגים שונים. כלומר מכיוון שבמקרה של התרגיל שלנו יש חשיבות לסדר השיטה לא תעזור

Can the two languages be distinguished using a bigram or trigram based approach? Explain why

גם גישת הביגרם או הטיגרם (רצף של שתי אותיות או 3 אותיות) לא יתאים על מנת להבדיל בין שתי השפות במקרה שלנו. הסחבה לכך היא שמילה בשפה שלנו מוגדרת על ידי רצף של מספרים ואז אותיות ואז שוב רתף של מספרים ואז אותיות וכן הלאה, ועל המודל שיצליח להבדיל בין שתי השפות יהיה להצליח להבין (לאחר תהליך של למידה) שהמספרים מפרידים בין האותיות, ומה הסדר של האותיות.

מכיוון שרצף הספרות יכול להיות בכל אורך וגם רצף האותיות, יהיה קשה לגישת ביגרם או טיגרם לתת פתרון מכיוון שייתכן מאוד ורצף של שתי אותיות/3 אותיות יהיה קצר מידי על מנת שהמכונה תלמד את הקשר בין הספרות לאותיות ואת הסדר הנכון.

Can the two languages be distinguished using a convolutional neural network? Explain why

גם על ידי רשת ניורונים מסוג קונבולוציה לא נצליח להפריד בין שתי השפות, מכיוון שרשת מסוג מסווגת את השפות על ידי שימוש במשפטים או רצף של מילים. מכיוון שבתרגיל זה אין רצף של מילים, אלא כל מילה בפני עצמה גם גישה זו לא תצליח לפתור את הבעיה המוצגת בתרגיל שלפנינו.

מדוע אם ככה RNN כן יעבוד?

רשת מסוג RNN מקבלת את הקלט בשלבים, כלומר תו אחר תו, ובכל שלב הוא מקבל גם את הפלט של השכבה שלפניו, ככה שבעצם הוא מקבל מידע מהעבר ומסתמך עליו – לכן אם היה רצף של b בעבר הוא ישים לב לכך כיוון שזה יבוא לידי ביטוי במידע שיקבל מהשכבות הקודמות.