מגישים: יהונתן שאקי 204920367, אור שחר, 209493709.

- לא ניתן להבדיל בין השפות בעזרת bag of words. זאת מאחר ו-bag of words לא נותן חשיבות לסדר המילים (במקרה שלנו, התווים), ואילו שתי השפות יכולות להכיל בדיוק אותם תווים (מספרים בין 1 ל-9 ו-a,b,c,d).
- גם בעזרת ביגרם וטריגרם (ולמעשה, בעזרת כל n-gram) לא ניתן להפריד בין השפות, זאת בעיקר בגלל הריפוד שהמספרים יוצרים. לכל n, ניתן להסתכל על המילה $1a(1)^{n+1}b(1)^{n+1}c(1)^{n+1}d1$ ששייכת לשפה הראשונה ו-$1a(1)^{n+1}c(1)^{n+1}b(1)^{n+1}d1$ ששייכת לשפה השניה. ניתן לראות שכל תתי הרצפים באורך n קיימים בשני המילים (בגלל שכל תת רצף יכיל רק אות אחת ו-n-1 אחדות, או לחלופין רק n אחדות) ולכן השפות לא יכולות להיות מופרדות בעזרת n-gram ובפרט ביגרם או טריגרם.
- השפות לא יכולות להיות מופרדות בעזרת רשת קונבולוציה מהסיבה שהן לא יכולות להיות מופרדות בעזרת n-gram; הפילטרים יהיו בגודל קבוע כלשהו (נאמר n) ומאחר וממש אותם דברים יופיעו לכל פילטר, יתקבלו בהכרח אותם ציונים. אומנם ניתן להגיד שהסדר של הציונים יהיה שונה, אבל אם רצפי ה-1 המפרידים יהיו ארוכים מספיק גם הציונים יהיו מופרדים בעצמם (בהרבה רצפים של 1) ולכן הבעיה לא באמת תפתר מאותם סיבות בדיוק. בנוסף, רשתות קונבולוציה פחות מתאימות לטיפול באורכים משתנים של קלט, בעיקר כאלו לא חסומים – כנראה שיהיה צורך לעשות maxpooling כלשהו על מנת שהרשת תוציא תשובה אחת ויחידה בסופו של דבר. בכל מקרה, אני סבור לכל ארכיטקטורת רשת כזו יהיה ניתן לבחור מספר אחדות מפריד שאותו הרשת לא תדע לפתור, גם אם יהיו כמה שכבות קונבולוציה.