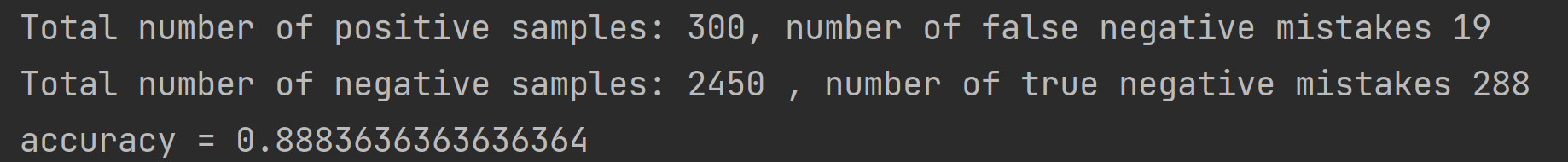
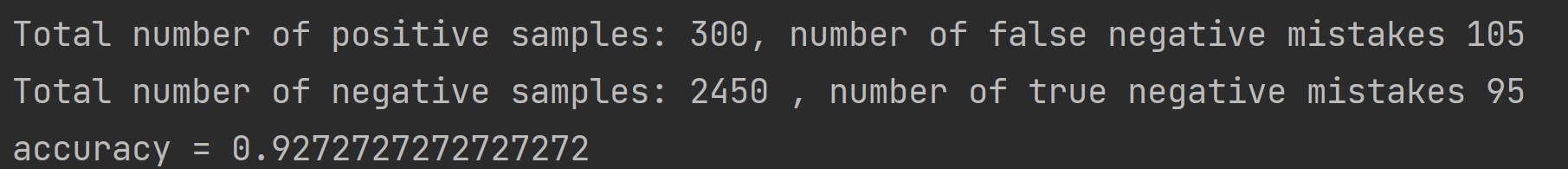
1. בביולוגיה חישובית מקובל לייצג רצפים בשיטת "one hot coding" – הפיכה של רצף באורך l המכיל אותיות מא-ב בגודל k למטריצה בגודל lxk כך שכל שורה מייצגת עמדה ברצף, וכל עמודה מייצגת אות ב-אב. בצורה זאת נקבל מטריצה בה בכל שורה מופיעה הספרה 1 בדיוק פעם אחת ושאר הכניסות הן 0. נבחר בשיטה הזאת כי היא משמרת את האינפורמציה של הרצף בצורה מושלמת, וכן משפרת זמני ריצה בשל המטריצה מאד "sparse" ושימוש בספריות שמבצעות פעולות לינארית בצורה יעילה כמו "numpy" או "pandas" .
2. במקרה שלנו, כיוון שהרצפים כולם באורך 9, ומספר החומצות האמיניות בתא הוא 20, נקבל מטריצה בגודל 20 על 9, ולכן מימד ה input של הרשת שלנו יהיה 180.   
   גרף המתאר את פונקציית הלוס עבור מודל עם שתי שכבות פנימיות בגודל 180 :   
     
   A graph with blue and orange lines

   Description automatically generated  
   כן – רשת נורונים בגודל כזה בהחלט תצור בעיה של Overfitting שינבע מכך שמספר משקולות האימון ברצף .
3. בחרנו ברשת שמקבלת קלט בגודל 180, משתמשת בפונקציה reLU ומעבירה לשכבה בגודל 21 שמשתמשת ב reLU גם היא ושולחת לשכבה בגודל 10 . השכבה הזו עושה שימוש בפונקציית tanh (מצאנו שהביצועים שלה טובים יותר מפונקציות sigmoid או softmax על בעיה זו) שמחזירה פלט בגדול 1 . הפלט הזה הוא ההסתברות ע"פ המודל שלנו שדגימה שייכת ללייבל 1, משמע תקשר לפפטיד המטרה שלנו HLA . לאחר האימון אנחנו מחפשים את ה threshold שיתן את אחוז הדיוק הטוב ביותר על גרף האימון, ובוחרים בו להיות Treshold למודל שלנו. על סט המבחן המודל שלנו השיג את התוצאות:   
   A graph with blue and orange lines

   Description automatically generated
4. בשימוש בפונקציות לינאריות בלבד, ושתי שכבות פנימיות, אחת בגודל 21 והשניה בגודל 10, השגנו את המודל הבא:  
   