**פרויקט מסכם – Sequence Generation with Keras**

**רקע כללי:**

בפרויקט זה תתנסו במשימה מתקדמת יותר, על כל שלביה, תוך שימוש ברבים מן הכלים והידע שרכשתם במהלך הקורס, ותראו כמה רחוק אפשר להגיע איתם.

בפרויקט תנתחו נתוני רצפים טקסטואליים, תלמדו מודל שפתי עבורם (כזה שיודע לחזות בכל שלב מה המילה הבאה), ותשתמשו במודל שאימנתם כדי לייצר רצף טקסטואלי חדש המתנהג בצורה דומה. למשל, אם עולם הנתונים הנבחר הוא הודעות ווטסאפ, אז הרעיון הוא לקחת הודעות ווטסאפ משולחים שונים, לאמן מודל נפרד לכל שולח, ואז לייצר הודעות חדשות מכל שולח, שאמורות להישמע דומות לאופן שבו השולח כותב.

**שלבי הפרויקט:**

1. כל זוג סטודנטים יבחר במשימה מתוך אוסף האפשרויות המפורט בהמשך. לא תתאפשר בחירה של אותה משימה ע"י יותר מזוג אחד, ולכן ננהל קובץ משותף שאליו תזינו את בחירותיכם. עליכם לאסוף את הנתונים המתוארים במשימה שבחרתם.
   * הקישור לקובץ בחירת המשימות (כנסו מתוך משתמש הBGUMAIL שלכם): <https://docs.google.com/spreadsheets/d/1PHW1kbHeAAY79h4B2bKf-zsoWQnBnkOgMvtET2R3xms/edit?usp=sharing>
   * הזינו את תעודות הזהות של חברי הצמד המגיש לצד המשימה הנבחרת, לא לשכוח לשמור שינויים.
2. עליכם לאמן מסווג על בסיס הרצפים שאספתם, אשר בהינתן רצף טקסטואלי ידע לסווג באיזה סוג רצף מדובר (בכל משימה מוגדרים רצפים טקסטואליים שונים וסוגי רצפים שונים).
3. עליכם לאמן מודל שפה על בסיס הרצפים הטקסטואליים שנאספו, ולייצר רצפים חדשים בעזרת המודל שאומן.
4. כדי לבדוק את איכות המודל שלמדתם, עליכם לבחון את הצלחת המסווג שלמדתם בשלב 2 על הרצפים הטקסטואליים שייצרתם.

**שלב א- איסוף הנתונים:**

1. להלן רשימת המשימות שעליכם לבחור מתוכן:
2. עליכם לבחור בכמה זמרים, ללמוד מודל לפי שיריו של כל זמר, ולייצר לו שיר/ים חדשים. (הרצף: מילותיו של שיר, סוג הרצף: הזמר של אותו שיר)
3. עליכם לבחור בכמה פוליטיקאים, ללמוד מודל לפי נאומיו של כל פוליטיקאי, ולייצר לו נאומים חדשים. (הרצף: נאום, סוג הרצף: הפוליטיקאי הנואם)
4. עליכם לבחור בכמה ידוענים, ללמוד מודל לפי ציוצי טוויטר של כל אחד, ולייצר לו ציוצים חדשים. (הרצף: ציוץ בטוויטר, סוג הרצף: הידוען המצייץ)
5. עליכם לבחור בכמה חברי פייסבוק (אם זה חשבונות פרטיים אז נדרש אישורם), ללמוד מודל לפי פוסטים של כל אחד, ולייצר לו פוסטים חדשים. (הרצף: פוסט בפייסבוק, סוג הרצף: מפרסם הפוסט)
6. עליכם לבחור בכמה חברי ווטסאפ שלכם (נדרש אישורם), ללמוד מודל לפי הודעות כל שולח, ולייצר הודעות חדשות ממנו. (הרצף: תוכן הודעת ווטסאפ, סוג הרצף: השולח)
7. עליכם לבחור בכמה מכותבי דוא"ל שלכם, ללמוד מודל לפי תכתובות כל אחד, ולייצר תכתובות חדשות ממנו (הרצף: תוכן הודעת דוא"ל, סוג הרצף: השולח).
8. עליכם לבחור בכמה ספרי קריאה, ללמוד מודל לכל אחד, ולייצר ספר מכל סוג (הרצף: משפט/פיסקה בספר, סוג הרצף: שם הספר).
9. לבחור בתמלילים של מספר סרטים, ללמוד מודל לכל סרט, ולייצר תמליל לכל סרט (הרצף: משפט/פיסקה בתמליל, סוג הרצף: שם הסרט).
10. לבחור מספר סוגי תוכן ברשת: למשל ערכי ויקיפדיה, מתכונים מתוך אתר בישול ואתר חדשות, ללמוד מודל לכל אחד, ולייצר דפי תוכן חדשים מכל סוג (הרצף: טקסט עם תוכן דף הHTML, (סוג הרצף: סוג התוכן (ויקיפדיה ,מתכון או חדשות למשל)).
11. עליכם לבחור מאמרים ממספר תחומים: למשל מחשבים, פסיכולוגיה ורפואה, ללמוד מודל לכל תחום, ולייצר מאמרים חדשים מכל תחום. (הרצף: תוכן המאמר, סוג הרצף: התחום (למשל מחשבים/ פסיכולוגיה/ רפואה))
12. עליכם להגדיר אוכלוסיה תחומה, ואז למדל בנפרד טוויטים של גברים וטוויטים של נשים בתוכה, ולייצר טוויטים נשיים וטוויטים גבריים. (הרצף: תוכן הציוץ, סוג הרצף: זכר או נקבה).
13. עליכם להגדיר אוכלוסיה תחומה, ואז למדל בנפרד פוסטים בפייסבוק של גברים ופוסטים של נשים בתוכה, ולייצר פוסטים נשיים וגבריים. (הרצף: תוכן הפוסט, סוג הרצף: זכר או נקבה).
14. עליכם לבחור בכמה חברי ווטסאפ שלכם (נדרש אישורם), ללמוד מודל לפי הודעות ששלחו גברים, ולפי הודעות ששלחו נשים, ולייצר הודעות חדשות לנשים ולגברים. (הרצף: תוכן הודעה, סוג הרצף: זכר או נקבה).
15. עליכם לבחור בכמה מכותבי דוא"ל שלכם, ללמוד מודל לפי הודעות ששלחו גברים, ולפי הודעות ששלחו נשים, ולייצר הודעות חדשות מנשים ומגברים. (הרצף: תוכן הדוא"ל, סוג הרצף: זכר או נקבה).
16. עליכם לבחור ברצפים גנטיים מכמה מוצאים של בני אדם, ללמוד מודל לכל מוצא, ולייצר רצפים גנטיים לכל מוצא (<http://www.internationalgenome.org/data>). (הרצף: רצף גנטי, סוג הרצף: מוצא האדם הנדגם).
17. לבחור ברצפים גנטיים של זכר ונקבה, ללמוד מודל לכל מין, ולייצר רצפים גנטיים לכל מין (<http://www.internationalgenome.org/data>). (הרצף: רצף גנטי, סוג הרצף: זכר או נקבה).
18. אספו את הנתונים הנדרשים לטובת המשימה שבחרתם. ככל שתאספו יותר נתונים המודל יהיה איכותי יותר, אך זמני הריצה יהיו ארוכים יותר.
    1. במשימות 1-2, 9-10 נדרשים לפחות 50 רצפים עבור לפחות שלושה סוגים.
    2. במשימות 3-6 נדרשים לפחות 100 רצפים עבור לפחות חמישה סוגים.
    3. במשימות 7-8 נדרשת למידה לפי ספר/תמליל סרט מלא. וודאו שהספר/ תמליל מכיל לפחות 1000 משפטים. נדרשים לפחות שלושה סוגים.
    4. במשימות 11-14 נדרשים לפחות 1000 רצפים עבור שני סוגים (זכר ונקבה).
    5. במשימות 15-16 הרצף מורכב מאותיות ולא ממילים. נדרשים קצת הסברים נוספים לגבי קריאת הנתונים, הפרויקטים הרלוונטיים יקבלו הדרכה.

* מומלץ לאסוף יותר רצפים מהנדרש, בהמשך תוכלו לדגום רק את כמות הרצפים המינימאלית מתוך מה שאספתם במידה ותתקלו בקשיים עם זמני הריצה.

1. תעדו את שלב איסוף הנתונים בדוח מסכם הכתוב בפורמט MD, הממוקם בGitHub שלכם וכולל את קטעי הקוד שבהם השתמשתם. שימו לב לכך שהקבצים שאתם מגישים לא מכילים סיסמאות ששימשו אתכם לטובת שימוש בAPIs השונים.
   1. השתמשו בפקודות במקום בהורדה ידנית של נתונים, כדי לאפשר את שחזור התהליך.
   2. תארו את הנתונים, כמה רצפים הורדתם ומאלו סוגים, השתמשו בשיטות שהכרנו כדי להראות שבחנתם את הנתונים והגעתם לתובנות בסיסיות לגביהם.

**שלב ב- בניית מסווג לנתוני טקסט:**

1. השתמשו בטכניקות מתוך אוסף השיטות לעיבוד טקסט שהכרתם בכדי לייצג את הטקסטים שאספתם כמאגר מובנה: כל רצף יהיה רשומה במאגר, וכל רשומה תיוצג ע"י n תכונות (n הוא גודל המילון שתגדירו).

* שימו לב לגודל המילון שאתו אתם עובדים, אפשר לכוונן פרמטר זה, ו/או לספק הסבר מדוע בחרתם בגודל זה.

1. בחרו אלגוריתם סיווג ואמנו אותו על נתוני הטקסטים המעובדים, בחנו את תוצאות הסיווג על סט מבחן (שהמודל לא אומן עליו) ודווחו על התוצאות (דיוק, טבלת הצלבה וכו').

* במידה ובחנתם מספר אלגוריתמים, תארו את תהליך בחינת האלגוריתמים, השתמשו בגרפים כדי להציג השוואה ברורה של תוצאות האלגוריתמים השונים.
* משתנה המטרה שתרצו לחזות הוא סוג הרצף: למשל במשימה שבה אספתם טקסטים של שלושה כותבים שונים, משתנה המטרה הוא הכותב.
* תעדו שלב זה בדוח המסכם, שלבו קטעי קוד רלוונטיים (ניתן להפנות לקוד מלא בקבצי python שיושבים גם הם בפרוייקט הgithub).

**שלב ג- חילול רצפי טקסט:**

1. השתמשו בחבילה keras כדי להגדיר רשת עמוקה לטובת למידת מודל שפה עבור הרצפים שאספתם. אתם יכולים להיעזר בארכיטקטורות שיוצגו בהרצאה, או לנסות ארכיטקטורות נוספות. **לימדו מודל נפרד לכל סוג רצף**.
   1. מי שמסתבך עם התקנת החבילה יכול לעבוד דרך VMware במעבדות, שם מותקן keras.
   2. במידה וניסיתם כמה ארכיטקטורות, תארו אותן, והשוו את תוצאותיהן.
   3. דווחו על דיוק הרשת מול נתוני test.
2. צרו רצפי טקסט חדשים באמצעות המודל שנלמד לכל סוג. כמות הרצפים שיש ליצור היא כ30% מכמות הרצפים שעליהם נלמד המודל.
3. תעדו שלב זה בדוח המסכם באופן דומה לשלבים הקודמים.

**שלב ד – סיווג רצפי טקסט מחוללים:**

1. חזו את סוג הרצף, בעזרת המסווג שאימנתם בשלב ב', עבור כל אחת מהרשומות שייצרתם. השוו את סוג הרצף שחזיתם לסוג הרצף שלפיו נלמד המודל ששימש ליצירת הרצפים
   1. חשבו את דיוק התוצאה על הרצפים המחוללים.
   2. הציגו מטריצת הצלבות המפרטת את סוגי הטעויות.
2. תעדו שלב זה בדוח המסכם באופן דומה לשלבים הקודמים.

**הגשה:**

יש להגיש את התרגיל בזוגות עד 25/2/2018.

אנא וודאו כי פרויקט הGitHub מכיל את כל המידע הרלוונטי לטובת שחזור התהליך:

* קובץ/קבצים עם רצפי הטקסט שאספתם
* מודלים שנבנו
* קובץ/קבצים עם רצפי הטקסט שחוללו
* קוד שבו השתמשתם (יש לטשטש מפתחות רגישים בהם השתמשתם)
* דוח מסכם המתאר את כל שלבי הפרויקט (בשאיפה שכל המידע הרלוונטי לבדיקת הפרויקט יתואר בדוח)
* כמו בכל משימה במדעי הנתונים, נסו להסביר את התוצאות, להסיק מתוכן, ולציין מה יכול לשפר בעתיד.

תרגילים שיוגשו באיחור "יקנסו" באופן הבא: 5 נקודות יורדו מציון התרגיל לכל יום איחור.

ההגשה תתבצע בתיקיה המתאימה באתר הקורס באופן הבא:

1. על כל אחד מבני הזוג להגיש את העבודה.

2. יש להגיש קובץ ובו קישור למאגר ה Githubשל הפרויקט המסכם וכן שמות השותפים לתרגיל. שם הקובץ יהיה מהסוג ID1\_ID2.txt (תעודות הזהות של המגישים מופרדות בקו תחתון).

בהצלחה ובהנאה!