קורס R 52414 - מעבדה 2 - הסברים

זהו קובץ עזר המסביר מה נדרש בשאלות המעבדה בעברית. בכל, מקרה, הנוסח המחייב הוא בקובץ ה-Rmd של המעבדה ועל קובץ הפתרון להשתמש בו ולהתייחס אליו.

1. בשאלה זו עליכם לקרוא קובץ html של הספר Moby-Dick.
   a. יש לקרוא את קובץ ה-html, להוציא מתוכו את הטקסט, ולהדפיס את השורה הראשונה בטקסט
   b. יש לחלק את הטקסט למילים, בעזרת המפרידים: נקודה, רווח, פסיק ושני תוי ירידת שורה \r, \n. אח"כ יש לחשב ולהראות את התפלגות אורכי המילים, תוך ציון הממוצע, החציון, השכיח והערך המקסימלי של התפלגות זו.
   c. יש לחשב את השכיחות (מספר ההופעות) של כל מילה בטקסט, ולהדפיס את 10 המילים בעלות השכיחות המקסימלית. יש להתייחס למילים יחידניות, כלומר אם מילה מופיעה מספר פעמים בטקסט בדיוק באותו איות יש להתייחס לכך כמספר הופעות של אותה המילה.

2. a. יש לחלק את הטקסט לפרקים, בעזרת זיהוי המחרוזות המפרידות בין פרקים. מומלץ להשתמש ב-regular expressions, וכן להסתכל היטב על הקובץ כדי לזהות אילו מחרוזות מפרידות בין הפרקים. יש להתייחס לחלקי הטקסט שלפני הפרק הראשון (אטימולוגיה ותמציות) וכן לאפילוג כפרקים נפרדים.
   b. יש לכתוב פונקציה המחשבת את השכיחות היחסית של מילה מבין כל המילים באותו פרק, להריצה על מילים נבחרות, ולהציג את השכיחות כפונקציה של הפרק, על פי סדר הפרקים.

3. a. יש לחשב את ההסתברות ששני אנשים הבוחרים מילה אקראית באופן אחיד מהספר יבחרו את אותה המילה בשתי דרכים:
   (i) לכתוב נוסחא בה על פי שכיחויות המילים, מקבלים את ההסתברות לבחירה של אותה מילה ע"י שני האנשים.
   (ii) לעשות סימולציה בה אנו בוחרים באופן אקראי מספר רב של פעמים שתי מילים, וסופרים את השכיחות היחסית של הסימולציות בהן קיבלנו את אותה המילה פעמיים.

b. נניח שכעת בוחרים באופן אקראי מרשימת המילים היחידניות. מה ההסתברות שבמקרה זה שניים יבחרו באותה המילה? האם היא גבוהה או נמוכה ביחס להסתברות הקודמת, ומדוע?

4. a. יש למצוא את המילים בנות חמש אותיות, לאחר הסרת מילים המכילות גם תוים שאינם אותיות באנגלית, ולאחר מעבר לאותיות קטנות. יש להציג את 10 המילים בנות חמש אותיות הנפוצות ביותר.

b. יש לחשב ולהציג את טבלת השכיחויות של כל אות אנגלית בכל אחד מ-5 המקומות במילה.

c. יש לחשב את הנראות של כל מילה בעזרת טבלת הסתברויות ומודל הסתברויות של אותיות בודדות - כלומר הנראות מוגדרת כמכפלת ההסתברויות של 5 האותיות במילה, כאשר ההסתברות של כל אות נקבעת על פי הטור המתאים בטבלה.
יש להריץ את הפונקציה על כל המילים בנות 5 אותיות, כאשר משתמשים בטבלת השכיחות מהסעיף הקודם, אחרי שמנרמלים אותה לטבלת הסתברויות, ולהראות את 10 המילים בעלות הנראות הגבוהה ביותר.

5. a. יש להוריד טבלת מילים בנות חמש אותיות מהקישור המצורף, לחשב ולהציג את טבלת ההסתברויות של אותיות בודדות בהתאם למיקום, בדומה לשאלה 4b , ולהשוות לטבלה הקודמת.

b. יש לכתוב פונקציה המקבלת רשימת מילים (מילון) וכן תוצאות של ניחושי wordle, ומחזירה את המילים החבויות האפשריות לאחר תוצאות אלו. עבור כל ניחוש אנו מקבלים מידע לגבי המקומות בהן יש התאמה נכונה בין האות בניחוש למילה החבויה, המקומות בהם האות משותפת אך המיקום שגוי, והמקומות בהם האות של הניחוש לא מופיעה כלל במילה החבויה. יש להחזיר רק את המילים עבורן המידע המתקבל מתאים לכל הניחושים.

6. בשאלה זו אנו בוחנים אסטרטגיה פשוטה לפתרון wordle, המתחשבת רק בפגיעות מדוייקות (כלומר אותיות בניחוש הזהות לאותיות במילה החבויה ובאותו מקום).

a. יש לממש אסטרטגיה זו ולהריץ אותה כל המקרה בו המילה החבויה היא "mouse", תוך הדפסת הניחושים לאורך הדרך עד שמזהים נכונה את המילה.

b. יש לכתוב נוסחה להתפלגות המצטברת של מספר הניחושים עד לזיהוי המילה החבויה. לנוסחא יש קשר להתפלגות הגיאומטרית, אך ההתפלגות המתקבלת אינה ההתפלגות הגיאומטרית אלא התפלגות אחרת. יש לחשב מתוך נוסחא זו באופן נומרי את התוחלת, בעזרת מעבר על כל הערכים האפשריים (עד ערך גדול כ-10,000 תוך הזנחת ערכים גדולים יותר) וחישוב הסתברותם.

c. יש לחשב באופן אמפירי את ההתפלגות ע"י סימולציה בה אני מגרילים מילים חבויות ומנחשים אותן בעזרת האסטרטגיה, וכך בעזרת 100 חזרות מקבלים התפלגות אמפירית של מספר התורות הדרושים לזיהוי המילה החבויה. לבסוף יש להשוות את ההתפלגות המצטברת והתוחלת המתקבלות באופן אמפירי לחישוב מהסעיף הקודם.

7. בשאלה זו אנו בוחנים 2 אסטרטגיות אחרות לפתרון wordle. באסטרטגיה 2, בכל שלב אנו מוצאים את המילים שעדיין אפשריות בהינתן תוצאות הניחושים הקודמים, ומנחשים את המילה הממקסמת את הנראות מבין מילים אלו. באסטרטגיה 3 אנו בוחרים מילה אקראית מבין מילים אלו.

a. יש לממש שתי אסטרטגיות אלו ולהריץ אותן על המקרה בו המילה החבויה היא "mouse", תוך הדפסת הניחושים לאורך הדרך עד שמזהים נכונה את המילה.
b. יש לחשב באופן אמפירי את ההתפלגות ותוחלת מספר הניחושים ע"י סימולציה כמו בשאלה 6b
c. (בונוס) יש לפתח אסטרטגיה חדשה יעילה יותר, ולהראות שהיא מביאה למספר ניחושים ממוצע נמוך יותר מהאסטרטגיות הקודמות.