שאלה 1

(a) בהינתן מודל שפה bigram למשפט S: START, $W_1, W_2, \dots, W_n$, STOP

אנו רואים כי שם הסתברויות המעברים מושגות היא (נסמנת ב- 1)

וכל ההסתברויות קיימת הסתברות גדולה מאפס להגיע ל- STOP

אזי סכום ההסתברויות מעל כל הרצפים האפשרים הוא 1.

הוכחה:

נתון $\forall w \quad P(STOP|w) > 0$

נסמן ע $\exists P > 0$ כך ש $\forall w \quad P(STOP|w) \geq P$

$$A_i = \left\{ \begin{matrix} \text{המשפט יגמר} \\ \text{במילה ה- } i \end{matrix} \right\} = \{ X_i = STOP \mid X_{i-1} \neq STOP \} \geq P$$

$$A = \left\{ \begin{matrix} \text{כל המשפטים} \\ \text{שנגדירם} \end{matrix} \right\} = \bigcup_{i=1}^{\infty} \left\{ \begin{matrix} \text{המשפט יגמר} \\ \text{במילה ה- } i \end{matrix} \right\} = \bigcup_{i=1}^{\infty} A_i$$

$$A^c = \left\{ \begin{matrix} \text{כל המשפטים} \\ \text{שלא נגדירם} \end{matrix} \right\} = \bigcap_{i=1}^{\infty} \{ X_i \neq STOP \mid X_{i-1} \neq STOP \}$$

$$P(A^c) = P\left(\bigcap_{i=1}^{\infty} \{ X_i \neq STOP \mid X_{i-1} \neq STOP \} \right) =$$

$$\overset{\text{נובע מאי תלות}}{=} \prod_{i=1}^{\infty} P(\{ X_i \neq STOP \mid X_{i-1} \neq STOP \})$$

$$= \sum_{i=1}^{\infty} (1-P)$$

$$\overset{1+x \leq e^x}{\leq} \exp\left(-\sum_{i=1}^{\infty} P\right)$$

$$= \frac{1}{\underbrace{\exp\left(\sum_{i=1}^{\infty} P\right)}_{\text{טור הסדרה הגבולה מתבדר}}} \xrightarrow{i \to \infty} \infty \longrightarrow \frac{1}{\infty} \longrightarrow 0 \qquad \Rightarrow \boxed{P(A) = 1}$$

ב) בחלק א' הראנו כי:

שאלה 1

(a) עלינו להראות שסכום ההסתברויות של כל המשפטים הסופיים הוא 1.
נשתמש בדרך ורדה כאשר ע"י האורך המשלים.

יהי משפט אינסופי $S = \{START, W_1, \dots, W_n \dots\}$.

סוף יצליח $\mathbb{P}(STOP \mid W) > 0 \quad \forall W$

לכן $\mathbb{P}(W' \mid W) < 1 \quad \forall W, \forall W' \neq STOP$

לפיכך:

$$\mathbb{P}(S) = \mathbb{P}(START, W_1, \dots, W_n, \dots) \overset{\text{bigram def.}}{=} \underbrace{\mathbb{P}(W_1 \mid START)}_{C \in \mathbb{R}} \cdot \underbrace{\prod_{i=2}^{\infty} \mathbb{P}(W_i \mid W_{i-1})}_{<1} \xrightarrow{i \to \infty} 0$$

כה נובע מכך שאר אינסופי של איברים קטנים מ- 1, מתכנס ל- 0.

מאחר והדבר יהיה נכון לכל משפט אינסופי, נסיק שסכום ההסתברויות של כל המשפטים האינסופיים יהיה 0.

לפיכך מאחר וראינו כי סכום ההסתברויות למשפט אינסופי (לא STOP) הוא אפס.
נסיק שסכום ההסתברויות למשפטים הסופיים (יתר האפשרויות בפלוס) הוא 1 כנדרש.

$\boxtimes$

(b) ורדה לטאן שהתוכחה מסעיף א' לא תעבוד במקרה למודלים שאינם מרקוביים.

נתבונן במרחב הגלים $C = \{W\}$

ונשים

$$\mathbb{P}(W_1 = W \mid W_0 = START) = e^{-1}$$

$$\mathbb{P}(W_n = W \mid W_1 \cdots W_n = \underbrace{W \cdot \ldots \cdot W}_{n}) = \prod_{i=1}^{n} e^{-\frac{1}{i^2}} = e^{-\sum \frac{1}{n^2}} = e^{-\frac{\pi^2}{6}} = \frac{1}{e^{\frac{\pi^2}{6}}} < 1$$

$$\mathbb{P}(W_n = STOP \mid W_1 \cdots W_n = W \cdot \ldots \cdot W) = 1 - e^{-\sum \frac{1}{n^2}}$$

נראה שההסתברויות בין 0 ל- 1
לכן חיקות ובנוסף אינן נסכמות ל- 1

$$e^{-\sum \frac{1}{n^2}} + 1 - e^{-\sum \frac{1}{n^2}} = 1$$

בנוסף נראה כי ההסתברות לכך שתמיד נלך אינסופית היא

$$\lim_{n \to \infty} \mathbb{P}(W_n = W \mid W_1 \cdots W_n = \underbrace{W \cdot \ldots \cdot W}_{n}) \longrightarrow \frac{1}{e^{\frac{\pi^2}{6}}}$$

קיבלנו שגם עבור $n \to \infty$ יש הבדלים
גורר קיומן הסתברות חיוביות וחיוביות למאוד.
לכן הוא אפשרי.
בשונה מהמקרה עם המודל המרקובי.

שאלה ב

אלגו רתשר Unigram.

$\forall w \in C \quad \mathbb{P}(w) = \#w$

(a) תאור ה- Unigram:

בהינתן קורפוס בגודל N מילים ומשפט $S = (w_1, \dots, w_n)$ נגדיר את הסתברות המשפט באופן הבא:

$$\mathbb{P}(S) = \prod_{i=1}^{n} \mathbb{P}(w_i) = \prod_{i=1}^{n} \frac{Count(w_i)}{N}$$

כלומר, כל מילה מקבלת הסתברות בהתאם לשכיחותה היחסית מבין סך כל המילים בקורפוס.

- נניח במשפט לשים where

$$\mathbb{P}(where) = \frac{count(where)}{N} > \frac{count(were)}{N} = \mathbb{P}(were)$$

- נניח במשפט לשים were

$$\mathbb{P}(where) = \frac{count(where)}{N} < \frac{count(were)}{N} = \mathbb{P}(were)$$

כלומר בשני המקרים, נבחרת המילה שיש לה שכיחות גבוהה יותר, למרות שגם count גבוה יותר בקורפוס.

מבנה המשפט שלנו לא יתכן שנבחר נכונה באי המופעים כי נמצאים מאחרי ולכן יתכן כי $\mathbb{P}(where) < \mathbb{P}(were)$ למרות $\mathbb{P}(where) > \mathbb{P}(were)$.

סעיף ד'

מודל ה-bigram יוגדר באופן הבא:

עבור קורפוס בגודל N מילים, נחבר את הסתברות המילים Where, were בהינתן מילה W המופיעה לפניהן

$$P(Where|W) = \frac{Count(Where, W)}{Count(W)}$$

כלומר את שכיחות מופע המילה Where לאחר W ביחס לסך מופעי המילה W.
נעשה אותו דבר עבור were ופעם מילה אחרת כאופן המילים.

- מודל ה bigram יכל להיות טוב יותר לכיוון שעבור הקיומה שלנו "יין ויכתר את המילה נמצאת בשני המשפטים. באות כאשר יתקיים:

$$P(Where|went) = \frac{Count(Where, went)}{Count(went)} > \frac{Count(were, went)}{Count(went)} = P(were|went)$$

וגם:

$$P(Where|there) = \frac{Count(Where, there)}{Count(there)} > \frac{Count(were, there)}{Count(there)} = P(were|there)$$

- "יין ומשפט יקבל הסתברות גבוה, מסתיק שבמנד מילים אחד במשפט לא יופיע כלל בקורפוס עלה לגבה את הסתברות המשפט כולו.

נאר מאחר והסתברות משפט S עפ"י המודל היא: $P(S) = \prod_{i=1}^{n} P(w_i|w_{i-1})$

לכן אם יש צמד מילים w_{i-1}, w_i אשר לא הופיעה (בסיכוי גבוה) בקורפוס האימון שלנו $P(w_i, w_{i-1}) = 0$.

נאר בעיה נוספת שמשפט הגיוני ותקין שכן עדיו יקבל הסתברות 0 מאחר ולא נמצא קשר בין מילה אחת לקבוצה

שאלה 3:

a) נראה שמתקיים השוויון:

N_c ההופעות של

ההלקה של מילה שמופיעה N_c פעמים

$$\sum_{c=1}^{C_{max}} \left(\frac{(c+1)N_{c+1}}{N_c \cdot N} \cdot N_c \right) = 1 - P_{unseen}$$

$$\Leftrightarrow \quad \frac{1}{N} \sum_{c=1}^{C_{max}} (c+1)N_{c+1} = \frac{N - N_1}{N}$$

$$\Leftrightarrow \quad \sum_{c=1}^{C_{max}} (c+1)N_{c+1} = N - N_1$$

$$N_1 + 2N_2 + 3N_3 + \dots + C_{max}N_{C_{max}} + C_{max+1} \cdot \overset{0}{N_{C_{max+1}}} = N$$

$$N_1 + 2N_2 + 3N_3 + \dots + C_{max}N_{C_{max}} = N$$

ומכן השוויון נכון מהגדרת N ∎

b) עלינו לרשום את המשוואה של עומר add-one של מילה המופיעה c פעמים.

מספר המילה w מופיעה c פעמים

$$q_{add-1}(w) = \frac{c(w)+1}{\sum_{w' \in V} (c(w')+1)} = \frac{c(w)+1}{c() + |V|} = \frac{c+1}{N + \sum_{i=0}^{C_{max}} N_i}$$

מספר הקורפוס מאז N

מספר המילים בשפה.

$$\frac{c+1}{N+|V|} > \overset{MLE}{\frac{c}{N}} \Leftrightarrow \frac{c+1}{c} > \frac{N+|V|}{N} \Leftrightarrow 1 + \frac{1}{c} > 1 + \frac{|V|}{N} \Leftarrow \boxed{\frac{N}{|V|} > c}$$

עומר את מספר ההופעות של מילה קטן $\frac{N}{|V|}$ אז נקבל הלקה מעל ה MLE

ואם ——— " ——— גדול ——— " ——— מתחת ל MLE.

עומר $\frac{N}{|V|} = \mu$ מאמר את הגרים. ∎

c) עלינו להראות שהתכונה מעל ב' נכונה להלקה מעל א'.

נראה בעזרת ל"ד דוגמה נגדית:

נניח $C_{max} = 5$ ולכן c בין 1 ל-5 נניח $N_c = c$.
אזי עבור מילה w המופיעה c פעמים בקורפוס יתקיים:

$$\frac{(c+1) \cdot N_{c+1}}{N \cdot N_c} = \frac{(c+1)^2}{c \cdot N} = \frac{c^2+2c+1}{c \cdot N} > \frac{c(c+2)+1}{c \cdot N} > \frac{c(c+2)}{c \cdot N} = \frac{c+2}{N} > \frac{c}{N} = MLE$$

עומר עם שיטת ההלקה גו, והדוגמה הנ"ל, תמיד אומדן ההלקה יהיה גדול יותר מה MLE

ולכן סתירה לתכונה מעל ב.

שאלה 4:

(a) המשוואה למודל triagram היא:

$$\forall w \quad \mathbb{P}(x_1,\dots,x_{n-1}) = \prod_{i=1}^{n} \mathbb{P}(x_i \mid x_{i-1}, x_{i-2})$$

ההנחות שמודל זה נשען עליו הן שמילה תלויה רק בשתי המילים שהיו לפניה ולא בכל אחת מאחור.

(b) עבור: המילים נובכים על הכדור. The dogs barks on the ball

המילה "נובכים" נמצאת למטה לכן המודל ילמד אותה.

(c) שגיאה: הכלבים רצו במהירות אל האגם היפה ונבך שלו.
0 1 2 3 4 5

המילה "נבח" תלויה בצמד "האגם היפה". ואינה תלויה במילים כלבים או רצו לכן המודל לא יזיה את נבחו למרות שברור שזה הקרוי יותר.

לו המודל היה 5-גרם כנראה שהמודל היה מבין שזה "נבחו".

The dogs ran to the beautiful lake, and bark on it.
6 5 4 3 2 1

המילה bark הוא נטוה מ- barked, אך עדיין מודל כזה רק 3 מילים לאחור הוא לא יכיר שזה אנגלית שפה.
לו המודל היה 6-גרם הוא היה מבין שאחרי הטוען לפני העיקר בהשוואה.

שאלה 5

הכלבים רצו במהירות אל האגם היפה ונבך שלו.

לעומת: הכלב רץ במהירות אל האגם היפה ונבך שלו.

נמצא כי כל שהמודל המרקובי בעל n גדול יותר הסיכוי שלו להכיר משפט לא תקין קטן.

פתרונות של שאלות תכנות:

```
*** Question 2 ***
The predicted sentence using bigram model is:
I have a house in the

*** Question 3 *** , using the bigram model:
Question 3 (a) - The probability of the following two sentences:
Probability for question 3a:  -inf
Probability for question 3b:  -29.667
Question 3 (b) - The perplexity of both the following two sentences:
Perplexity for question 3b:  inf

*** Question 4 *** - Estimate a new model using linear interpolation smooth
between the bigram model and unigram model:
The probability of 1th sentence is:  -36.192
The probability of 2nd sentence is:  -30.993
The perplexity of sentences 1 and 2 is:  270.076

Process finished with exit code 0
```