תרגיל 1 קורס: עיבוד שפה טבעית מגישים: מאיה כהן, נריה בן דוד

הלק תיאורטי

1) א. נגדיר $W=\{w_1,\dots,w_m\}$ אוסף המילים ללא start, stop.

נתון: $\forall_{1\le i\le m} \sum_{j=1}^{m} P(w_j|w_i)+P(stop|w_i)=1$

$\forall_{1\le i\le m} \exists \varepsilon_i>0 \quad P(stop|w_i)>\varepsilon_i$

נגדיר X_k משתנה מקרי המייצג את המילה ה-k במשפט.

$A=\{\exists k>0 \ X_k=stop\}$. ל.ז.ש. $P(A)=1$.

נראה כי $P(A^c)=0$ $\quad A^c=\{\forall_{k>0} \ X_k\neq stop\}$

מכך שישנו מספר סופי של מילים $0<\varepsilon=\min\{\varepsilon_i \mid 1\le i\le m\}$ כך שלכל $1\le i\le n$

$$P(stop|w_i)>\varepsilon>0$$

$$P(A^c)=\prod_{k=1}^{\infty} P(X_k=stop \mid X_{k-1}\neq stop)\le \prod_{i=1}^{\infty}(1-\varepsilon)\le \prod_{i=1}^{\infty} e^{-\varepsilon}=e^{-\sum_{i=1}^{\infty}\varepsilon}=0$$

נוסח מתוקן ה- bigram

הומוגניות של השרשרת מרקוב

ב. לא, נראה דוגמה נגדית.

נבנה מודל שבו ההסתברות לעבור ל-stop קטנה באופן המשפט עד רגע המעבר.

כלומר: $P(X_k=stop \mid X_1,\dots,X_{k-1}\in W)=e^{-\frac{1}{2^i}}$

$$P(A^c)=\prod_{k=1}^{\infty} P(X_k=stop \mid \underbrace{X_1,\dots,X_{k-1}\in W}_{stop\notin W})=\prod_{k=1}^{\infty} e^{-\frac{1}{2^i}}=e^{-\sum_{k=1}^{\infty}\frac{1}{2^i}}=e^{-1}>0$$

לכן $P(A)<1$,

כלומר קיימת מילה אינסופית שאינה מגיעה ל- stop.

ב. א. נגדיר מודל unigram באופן הבא:

בהינתן משפט, המודל מחזיר את ההסתברות של המשפט

על פי כמות ההופעות של המילים בסט אימון.

לומר: $P(w_i) = \frac{c(w_i)}{\sum_{w \in W} c(w)}$ עבור כל מילה.

ועבור כל משפט: $P(w_1, \dots, w_n) = \prod_{i=1}^{n} P(w_i)$.

בהינתן המשפט המבוקש,

במאורח הראשון אני נהיה צודקים במקרה כי $c(where) > c(were)$.

במאורח השני אני נהיה צודקים במקרה כי $c(where) < c(were)$.

לסומת זאת לא נצדוק בשני המאורחות מכיוון שהתנאים

$c(where) > c(were)$, $c(where) < c(were)$ אינם יכולים להתקיים

במקביל.

ב. נגדיר מודל bigram באופן הבא:

בהינתן משפט המודל מחזיר את ההסתברות של המשפט

על פי כמות ההופעות של המילים בסט אימון תוך תלות במילה

הקודמת להן. לומר: $P(w_j | w_i) = \frac{c(w_i, w_j)}{\sum_{k} c(w_i, w_k)}$ עבור כל מילה.

ועבור כל משפט: $P(w_1, \dots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1})$.

המודל יכול להגדיל תוצאה יותר טובה מהמודל בסעיף א'

משום שבניגוד למודל הקודם, מודל זה מתייחס גם להקשר.

ולכן יכול להבדיל בין מקרים בהם הקדמות בין יותר את

where לעומת were.

- במודל זה משפטים יכולים לקבל הסתברות 0, כיוון שלמרות שכלל המילים מופיעות לפחות פעם אחת.
ייתכן כי צמדים של מילים כלשהן לא יופיעו כלל בסט אימון ומכאן כאשר נחשב את הסתברות המשפט הצמד לא יופיע ולכן נקבל הסתברות אפס.
למשל: סט אימון: the dog ate ומכאן כי $P(\text{the ate})=0$.
- נוכל לפיר בעיה כיוון שמשפט שאמור לקבל הסתברות גבוהה מהמודל יכול להתאפס כיוון שהיו בו שני מילים סמוכות שלא הופיעו, ומאופן חישוב ה-bigram איפסו את כל הסתברות המשפט.

3. ט.

לפי כך $A_c = \{w \mid$ w מופיעה c פעמים בקורפוס $\}$

$$* \sum_{c=1}^{c_{max}} \sum_{w \in A_c} \frac{(c+1)N_{c+1}}{N_c \cdot N} = \frac{1}{N} \sum_{c=1}^{c_{max}} \sum_{w \in A_c} \frac{(c+1)N_{c+1}}{N_c} = \frac{1}{N} \sum_{c=1}^{c_{max}} N_c \cdot \frac{(c+1)\cdot N_{c+1}}{N_c} = \frac{1}{N} \sum_{c=1}^{c_{max}} (c+1)\cdot N_{c+1} =$$

$$\frac{1}{N}\left(2N_2 + 3N_3 + \ldots + c_{max}N_{c_{max}} + 0\right) = \frac{1}{N}\left(N - N_1\right) = 1 - \frac{N_1}{N} = 1 - P_{unseen}$$

* מתל $c_{max}+1$ - N לא מופיעים מילים באורכים אלו.

ב. $q_{ML}(w) = \frac{c}{N}$, $q_{add-1}(w) = \frac{c+1}{\sum_{w'}(c(w')+1)} = \frac{c+1}{N+|V|}$.

V- קבוצת המילים בקורפוס.

עבור $\mu = \frac{N}{|V|}$

$$* \; q_{add-1}(w) = \frac{c+1}{N+|V|} \le \frac{c}{N} = q_{ML}(w) \iff (c+1)N \le c(N+|V|) \iff$$

$$\iff cN + N \le cN + |V|c \iff N \le |V|\cdot c \iff \mu := \frac{N}{|V|} \le c$$

לכן עבור $c \ge \mu$: $q_{add-1}(w) \le q_{ML}$.

$$* \; q_{add-1}(w) = \frac{c+1}{N+|V|} \ge \frac{c}{N} = q_{ML}(w) \iff (c+1)N \ge c(N+|V|) \iff$$

$$\iff cN + N \ge cN + |V|c \iff N \ge |V|\cdot c \iff \mu := \frac{N}{|V|} \ge c$$

לכן עבור $c \le \mu$: $q_{add-1}(w) \ge q_{ML}$.

ד.

$$q_{GT}(w) \leq q_{ML}(w) \Leftrightarrow \frac{(c+1)N_{c+1}}{N_c \cdot N} \leq \frac{c}{N} \Leftrightarrow (c+1)N \cdot N_{c+1} \leq c \cdot N_c \cdot N \Leftrightarrow$$

$$(c+1)N_{c+1} \leq c \cdot N_c \Leftrightarrow cN_{c+1} + N_{c+1} \leq c \cdot N_c \Leftrightarrow N_{c+1} \leq cN_c - cN_{c+1} \Leftrightarrow$$

אם $N_c - N_{c+1} > 0$

$$\frac{N_{c+1}}{N_c - N_{c+1}} \leq c$$

אם $N_c - N_{c+1} = 0$

$N_{c+1} \leq 0$

אם $N_c - N_{c+1} < 0$

$$\frac{N_{c+1}}{N_c - N_{c+1}} \geq c$$

מכאן כי לא קיים threshold עם קבוצה עבורו מתקיימים אי השוויונות הרצויים. (כל N_c יכול להיות שונה עבור כל c ולכן לא יהיה חסם קבוע)

4.

ד. הנוסחה עבור מודל trigram:

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

ההנחה במודל זה על התלות בין מילים היא שלא תלויה אך ורק בשני המילים המופיעות לפניה.

ב. הם אוכלים ארוחה.

במקרה זה המודל יתחשב בצורת הפועל אוכלים בהם כיוון שלפניו באה המילה הם, ומודל trigram מתחשב בשתי המילים האחרונות ובמקרה זה בפרט במילה האחרונה.

She always sings.

במקרה זה המודל יתחשב בהטיה הנכונה של sings משום שלפניו באות המילים she, always ומודל trigram מתחשב ב-2 המילים האחרונות.

ג. The book that I bought yesterday is very interesting.

מודל שמבחין בין שימוש בין is ל- are בצורה נכונה במשפט צריך להתחשב במילה book שהיא המילה המופיעה 5 מילים לפני המילה is.

מודל n-gram שיבחין בכך יהיה לכל הפחות 6-gram.

. "קנינו ספרים כדי שנוכל לקרוא אותם."

מודל שמבחין בצורה אותם צריך להתחשב במילה ספרים.

המילה ספרים מופיעה 4 מילים לפני מילה אותם ולכן מודל ram שמתחשב בכך יהיה לכל הפחות 5-gram.

5. היא פתרה את הלוח שבולה

ב שתי מילים שוקבר נטיות תחביריות שך המשפט שינו תקין תחבירי.

שמשון שיבור שבול נפל

ב שלושה מילים שוקבר נטיות תחביריות שך המשפט שינו תקין תחבירי.

אתם ומאיה מחר בבוקר הולכים לבית-ספר.

ב 4 מילים שוקבר נטיות תחביריות שך המשפט שינו תקין תחבירי.

מכאן שניתן לראות ש מתן שאולי מרקוב יהיו טובים כמודל שפה בהינתן משפט באורך n נדבה להשתמש במודל n-gram, לומר בשביל שאולי מרקוב יהיו מדוייקים הם צריכים להיות בעלי סיבוכיות גבוהה.

חלק פרקטי

2. the

3. ב. משפט 1: $-\infty$. משפט 2: -29.666.

ג. perplexity $= \infty$.

4. ב. משפט 1: -36.19 משפט 2: -30.99

ג. 270.076