# עיבוד שפות טבעיות סמסטר א' תשפ"א
## תרגיל בית מספר 2: **מודלי שפה**
**מועד הגשת התרגיל: 28.11.2020 בשעה 23:59**

בתרגיל זה עליכם לממש **מודלי שפה** (language models). הקלט לתרגיל יורכב מאוסף קבצים שמבוססים על ערכי ויקיפדיה ב 3 שפות שונות (אנגלית – en, ספרדית – es, ואנגלית פשוטה – simple). הקבצים יהיו בפורמט שהוגדר בתרגיל הראשון – הן מבחינת שם הקובץ (קוד השפה, קו תחתון ושם הערך – סיומת txt.), והן מבחינת התוכן (הטקסט שמופיע בדף ויקיפדיה, מחולק לטוקנים ומופרד למשפטים). הספריה שבה נמצאים הקבצים תועבר כפרמטר לתכנית. נתייחס לאוסף הקבצים שמכילים טקסט בשפה מסויימת כאל קורפוס נפרד (כלומר, נעבוד עם 3 קורפוסים שונים בהתאם לשפות שפורטו לעיל). אוסף קבצי הקלט יופיע באתר בסמוך להוראות התרגיל.

**שלב א:**

**הכנות**

לכל אחד מהקורפוסים חשבו טבלה הקובעת את אורכי המשפטים בקורפוס, וכמה משפטים יש בקורפוס מכל אורך. הגדירו פונקציה המגרילה אורך באקראי על פי התפלגות האורכים. כעת, הוסיפו לכל משפט בקורפוס סימן מיוחד לתחילת משפט וסימן נפרד לסוף משפט.

**Unigrams**

כתבו קוד שמחשב את הסתברות ההופעה של כל token ב כל אחד מהקורפוסים שיצרתם. הוסיפו סימן לציון סוף משפט. ההסתברות להופעת כל טוקן במודל ה- unigrams היא מספר המופעים של הטוקן חלקי אורך הקורפוס.

א. דגמו באקראי מתוך 3 הקורפוסים 4 צירופי מילים (טוקנים) באורכים 1,2,3,5. חשבו על 4 צירופים נוספים משלכם באותם אורכים (סך הכל 8 צירופים – 2 צירופים מכל אחד מהאורכים, כשאחד נדגם אקראית מתוך הקורפוס והשני הוא צירוף שחשבתם עליו בעצמכם). עבור כל אחד מצירופי המילים קבעו, על סמך מודל ה unigrams, לאיזו שפה מתוך ה-3 שעבדתם איתן הוא שייך (פרטו את התוצאות שקיבלתם בדו"ח שאתם מגישים).

ב. על סמך מודל ה- unigrams שבניתם, צרו משפטים רנדומליים, בהם כל מילה "נבחרת" לפי ההסתברות שלה. תחילה הגרילו את אורך המשפט על פי התפלגות האורכים בקורפוס, ואז את המילים. הדפיסו 3 דוגמאות למשפטים שיצרתם בדרך זו, עבור כל אחת משפות העבודה.

**Bigrams**

כתבו קוד שמחשב את הסתברות ההופעה של כל צמד tokens בקורפוס. השתמשו ב- Laplace smoothing לקביעת הסתברויות לצירופים שאינם קיימים בקורפוס (Out of Vocabulary).

א. דגמו באקראי מתוך 3 הקורפוסים 4 צירופי מילים באורכים 2,3,4,5. חשבו על 4 צירופים נוספים משלכם באותם אורכים (סך הכל 8 צירופים – 2 צירופים מכל אחד מהאורכים , כשאחד נדגם אקראית מתוך הקורפוס והשני הוא צירוף שחשבתם עליו בעצמכם) עבור כל אחד מצירופי המילים קבעו, על סמך מודל ה bigrams, לאיזו שפה הוא שייך (פרטו את התוצאות שקיבלתם בדו"ח שאתם מגישים).

ב. צרו משפטים רנדומליים המבוססים על מודל ה Bigrams (הסתברות להופעת המילה, בהינתן הופעת המילה הקודמת – כפי שהוסבר בהרצאה). גם כאן, הגרילו תחילה את אורך המשפט. הדפיסו 3 דוגמאות למשפטים שיצרתם בדרך זו, עבור כל אחת משפות העבודה.

**Trigrams**

הרחיבו את מודל ה bigrams למודל Trigrams. חשבו את ההסתברות לכל שלישית טוקנים. השתמשו ב Backoff עם אינטרפולציה לינארית פשוטה לקביעת הסתברויות לצירופים שאינם קיימים בקורפוס (Out of Vocabulary). בחרו את המקדמים בדרך שנראית לכם נכונה, והסבירו את פרטי המודל שיצרתם בדו"ח.

א. דגמו באקראי מתוך 3 הקורפוסים 4 צירופי מילים באורכים 3,4,5,7 חשבו על 4 צירופים נוספים משלכם באותם אורכים (סך הכל 8 צירופים – 2 צירופים מכל אחד מהאורכים , כשאחד נדגם אקראית מתוך הקורפוס והשני הוא צירוף שחשבתם עליו בעצמכם) עבור כל אחד מצירופי המילים קבעו, על סמך מודל ה trigrams, לאיזו שפה הוא שייך (פרטו את התוצאות שקיבלתם בדו"ח שאתם מגישים).

ב. צרו משפטים רנדומליים המבוססים על מודל ה Trigrams כך שהופעה של כל מילה תיקח בחשבון את שתי המילים שלפניה. חשבו איך כדאי לבחור את המילה הראשונה במשפט ופרטו את ההחלטה שקיבלתם בדוח. הדפיסו 3 דוגמאות למשפטים שיצרתם בדרך זו, עבור כל אחת משפות העבודה.

## שלב ב:

בעת הכינו מודלי Bigram ו Trigram שמבוססים על הטקסט של כל 3 השפות, כלומר אחדו את שלושת הקורפוסים לקורפוס אחד. צרו משפטים רנדומליים ע"ב שני המודלים, והדפיסו 5 דוגמאות לכל מודל (בסך הכל 10 משפטים).

קובץ ההגשה:
עליכם לכתוב קוד שמקבל כקלט את רשימת הערכים בקובץ ואת רשימת קודי השפות בקובץ, בדומה לאופן שבו קיבלתם את הקלט בתרגיל מס' 1. קובץ הפלט יכלול:
3 דוגמאות למשפטים מ 3 השפות ממודל ה Unigram - סה"כ 9 משפטים
3 דוגמאות למשפטים מ 3 השפות ממודל ה Bigrams - סה"כ 9 משפטים
3 דוגמאות למשפטים מ 3 השפות ממודל ה Trigrams - סה"כ 9 משפטים
5 דוגמאות למשפטים ממודל ה Bigrams שמבוסס על כל שפות העבודה
5 דוגמאות למשפטים ממודל ה Trigrams שמבוסס על כל שפות העבודה

הפלט צריך להיות בפורמט הבא:

Unigrams model based on complete dataset (English):
<sentence 1>
<sentence 2>
<sentence 3>

Unigrams model based on complete dataset (Spanish):
<sentence 1>
<sentence 2>
<sentence 3>

Unigrams model based on complete dataset (simple English):
<sentence 1>
<sentence 2>
<sentence 3>

Bigrams model based on complete dataset (English, Spanish, Simple English):
<sentence 1>
<sentence 2>
<sentence 3>
<sentence 4>
<sentence 5>

Trigrams model based on complete dataset (English, Spanish, Simple English):
<sentence 1>
<sentence 2>
<sentence 3>
<sentence 4>
<sentence 5>

**שאלות לדו"ח:**

1. השוו את המשפטים שקיבלתם בכל אחד משלושת המודלים שהתבססו על כל שפה בנפרד. האם יש הבדלים באיכות השפה? האם אתם מזהים הבדלים באיכות המשפטים בין השפות השונות? אם כן, נסו להציע סיבות שיסבירו את ההבדלים הללו (איכות של משפט בהקשר זה היא הסבירות שהמשפט שהפקתם אוטומטית יופק ע"י אדם באופן טבעי).
2. האם אתם יכולים לזהות הבדלים בין המשפטים שנוצרו על בסיס כל השפות לבין משפטים שנוצרו על בסיס שפה אחת? ממה לדעתכם נובעים ההבדלים, אם יש כאלה?
3. פרטו עבור כל סעיף האם לדעתכם ישפר את איכות המשפטים שיופקו:
   - הרחבת הקורפוס של שפה מסויימת – הוספת ערכים נוספים מאותה שפה, שימוש במודל ה trigrams
   - הגדלת הקורפוס – הוספת טקסט ממקורות אחרים באותן שפות, שימוש במודל ה trigrams
   - שימוש ב 5-grams - על הקורפוס הקיים.

עליכם להכין שני קבצים:

א. קובץ פייתון בשם **hw2.py** שיבצע את הנ"ל. הקובץ יקרא ע"י הפקודה:

```
python hw1.py <input_dir_path> <output_file_path>
```

באשר:

- **<input_dir_path>** הוא הנתיב המלא לתיקייה שבה נמצאים קבצי הקלט שעליהם יתבסס הקורפוס. הקבצים יהיו בפורמט של תרגיל 1 ותצטרכו להרכיב את הקורפוס המתאים מהטקסט שבקבצים על סמך המבנה שלהם.
- **<output_file_path>** הוא השם והנתיב המלא של קובץ הפלט שהתכנית תכתוב.

ב. דו"ח על התרגיל בפורמט PDF . על הדו"ח לכלול את **כל** הדוגמאות למשפטים שייצרתם (ולא רק את אלה שהדפסתם לפלט), וכן את התשובות לשאלות. הוסיפו לדו"ח את שמכם ומס' ת"ז. יש להגיש את שני הקבצים המפורטים לעיל - קוד פייתון ודו"ח (לא ב zip).

**הערות**:

- מותר להשתמש בספריות הסטנדרטיות של פייתון.
- **אין** להשתמש בספרייה NLTK
- יש להקפיד על עבודה עצמאית. צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
- תאריך הגשה: 28.11.2020 , עד השעה 23:59.
- שאלות על התרגיל נא להפנות לפורום .

**בהצלחה!!**