Final Project - NLP

הפרויקט ניתן לביצוע ביחידים / זוגות. יש לבחור את אחד הנושאים המתוארים להלן ולבצע עבורו את כל האנליזות הנדרשות. בכל אחד מהפרויקטים ישנו סעיף אחד מובלט והוא הסעיף המרכזי באותו הפרויקט ולכן יש לשים את הדגש עליו. אם יש לכם רעיון לפרויקט אחר – אנא פנו אלי לקבלת אישור לביצועו.

כל אחד מהפרויקטים כולל כריית נתונים טקסטואליים, עיבודם והצגת התוצאות. תוצרי הפרויקט:

* GITHUB הכולל את כל קבצי הפייתון + תיעוד שלהם
* קובץ WORD המתאר את המתודולוגיה + התוצאות
* הצעה לשינוי של אלגוריתם / מדד איכות קיים והסבר על הראציונל
* השוואה של איכות התוצאות המתקבלות עבור אחד מהאלגוריתמים, כאשר משנים את אחד מהפרמטרים / ההיפרפרמטרים, לכמה ערכים שונים.

פרויקט 1:

ייתכן ונקבל טקסטים אנונימיים בעברית של סיכומים רפואיים שנכתבו ע"י רופאים המחלקה הפנימית בביה"ח. מטרת הפרויקט היא להפוך טקסט UNSTRUCTED ל STRUCTED לפי אינדיקטורים שונים שיימסרו לנו מהסגל הרפואי.

1.) כתוב תכנית בפייתון לקליטת התמלילים מתוך מערכת שנמצאת בביה"ח ואשר תנתן לנו הרשאה לגשת אליה.
2.) בצע עיבודי pre processing הכוללים:

- Tokenization
- Lemmatization
- Stop words removal

3.) יש למצוא את המילים הנפוצות ביותר באמצעות tf-idf ולהציגן באמצעות:

- Bar chart
- Word cloud

4.) השתמש באלגוריתם word2vec למציאת המילים הנפוצות ביותר
5.) השתמש ב AUTOENCODER לצורך מציאת המילים החשובות ביותר
6.) השווה את התוצאות של סעיף 3, 4 , ו 5.
7.) הפעל אלגוריתם למציאת NER
8.) בצע EDA על הנתונים (למשל, מספר פניות ביום / השעות או הימים בשבוע בהן יש שכיחות פניות גבוהה יותר)
9.) יש לבצע חיפוש של מילות מפתח ומושגים נבחרים לפי חוקיות שתמסר לנו מהצוות הרפואי. יש להפוך את הטקסטים הללו לדאטבייס מובנה לפי השדות שבעניין.
10.) יש לבצע summarization לכל אחת מהטקסטים.
11.) הפעל מודל GPT למציאת השדות שבעניין באמצעות ANSWERING QUESTION
12.) השווה את התוצאות שקיבלת בסעיף 10 לאלו של סעיף 11.
13.) יש לכתוב מסמך המסכם את התוצאות כולל הדאטבייס המובנה.

פרויקט 2 :

1.) כתוב תכנית בפייתון להורדת ציוצי טוויטר משני אתרי חדשות שונים (לבחירתך). ציוצים שפורסמו בתוך טווח תאריכים כרצונך (אותו פרק הזמן לשני האתרים). היקף הנתונים הרצוי הוא לפחות 5000 ציוצים.

רצוי לעבוד על אתר חדשות באנגלית (New York Times, CNN) וכו'. אך ניתן גם בעברית.
רצוי להיעזר בספריה TWEETPY
ניתן להיעזר בשני הקישורים הבאים:
https://www.jcchouinard.com/twitter-api/
https://developer.twitter.com/en/docs/tutorials/step-by-step-guide-to-making-your-first-request-to-the-twitter-api-v2

2.) בצע עיבודי pre processing הכוללים:

- Tokenization
- Lemmatization
- Stop words removal

3.) יש למצוא את המילים הנפוצות ביותר בכל אחד מהאתרים ולהשוות ביניהם. יש להציג זאת באמצעות:

- Bar chart
- Word cloud

את המילים הנפוצות ביותר יש למצוא באמצעות tf-idf

4.) השתמש באלגוריתם word2vec לצורך מציאת המילים הנפוצות ביותר
5.) השתמש ב AUTOENCODER לצורך מציאת המילים החשובות ביותר
6.) השווה את התוצאות של סעיף 3, 4 , ו 5.
7.) הפעל אלגוריתם למציאת NER
8.) בצע EDA על הנתונים (למשל, מספר ציוצים ליום / שעה) ולהציג מגמות
9.) יש לבצע sentiment analysis עבור כל אחד מהאתרים בנפרד ולהשוות ביניהם ולהציג מגמות
10.) יש לבצע summarization לכל הציוצים מכל אחד מהאתרים בנפרד.
11.) הפעל אלגוריתם RNN ליצירת טקסט.
12.) הפעל מודל GPT ליצירת טקסט
13.) השווה את הטקסט שנוצר בסעיף 11 לזה שנוצר בסעיף 12.

פרויקט 3:

בצע סקירת ספרות מתוך אתר SCOPUS שמהווה דאטהבייס למאמרים אקדמיים.

1.) כתוב תכנית בפייתון להתממשקות ל API של אתר SCOPUS והורדת הדאטהבייס הכולל של שם המאמר, הכותבים, הג'ורנל בו התפרסם, מילות מפתח, שנת פרסום ו ABSTRACT. יש להוריד את כל המסמכים בהם מופיע המושג NLP או השם המאמר, או במילות המפתח או ב ABSTRACT. יש להוריד קליטת התמלילים מתוך מערכת שאת הרשומות הרלוונטיות מ 10 השנים האחרונות.
2.) בצע עיבודי pre processing הכוללים:

- Tokenization
- Lemmatization
- Stop words removal

3.) יש למצוא את המילים הנפוצות ביותר באמצעות tf-idf ולהציגן באמצעות:

- Bar chart
- Word cloud

4.) השתמש באלגוריתם word2vec למציאת המילים הנפוצות ביותר
5.) השתמש ב AUTOENCODER לצורך מציאת המילים החשובות ביותר

6.) השווה את התוצאות של סעיף 3, 4 , ו 5.
7.) הפעל אלגוריתם למציאת NER
8.) בצע EDA על הנתונים (למשל, מספר מאמרים שפורסמו בכל שנה, מיהם המחברים שפרסמו הכי הרבה וכו')
9.) יש למצוא את המילים / המושגים הנפוצים ביותר ולגבי כל אחד מהם להראות מגמה לאורך 10 השנים. יש למצוא מילים / מושגים שיש ביניהם תלות הדדית גבוהה.
10.) יש לבצע summarization לכל אחת מה ABSTRACTS.
11.) הפעל מודל GPT למציאת הנושאים הנפוצים ביותר באמצעות ANSWERING QUESTION
12.) השווה את התוצאות שקיבלת בסעיף 9 לאלו של סעיף 11.
13.) יש לכתוב מאמר REVIEW המסכם מגמות בפרסומים אקדמיים בנושא NLP.

פרויקט 4:

ישנו דאטהבייס בו נמצאים כל פסקי הדין של בית המשפט העליון. ניתן להוריד את פסקי הדין הרצויים מתוך אתר בית המשפט העליון:
https://supreme.court.gov.il/pages/fullsearch.aspx
מטרת הפרויקט היא להוריד כ 5000 פסקי דין / החלטות של בית המשפט העליון ולהציג ניתוח של הטקסטים המשפטיים.

1.) כתוב תכנית בפייתון להורדת פסקי הדין בהם תבחר מתוך אתר בית המשפט העליון. החלופה לכך היא הורדה ידנית. חלופה נוספת היא שימוש בבסיס נתונים קיים הנמצא באתרים של אוני בר אילן / אוני העברית / אוני' רייכמן.
2.) בצע עיבודי pre processing הכוללים:

- Tokenization
- Lemmatization
- Stop words removal

3.) יש למצוא את המילים הנפוצות ביותר באמצעות tf-idf ולהציגן באמצעות:

- Bar chart
- Word cloud

4.) השתמש באלגוריתם word2vec למציאת המילים הנפוצות ביותר
5.) השתמש ב AUTOENCODER לצורך מציאת המילים החשובות ביותר
6.) השווה את התוצאות של סעיף 3, 4 , ו 5.
7.) הפעל אלגוריתם למציאת NER: שם השופט, התובע והנתבע.
8.) בצע EDA על הנתונים (למשל, מספר פסקי דין לכל שופט, מגמה במספר פסקי הדין בכל שנה וכו').
9.) יש לבצע sentiment analysis לפי : חיובי / שלילי / נייטראלי. ולהציג סטטיסטיקה לכל שופט וסטטיסטיקה לכל אחת ממילות המפתח הנפוצות כפי שמצאת בסעיפים 3 / 4 / 5.
10.) יש לבצע summarization לכל פסק דין.
11.) הפעל אלגוריתם RNN ליצירת טקסט של החלטת בית המשפט. (הפסקה האחרונה)
12.) הפעל מודל GPT לצורך החלטת בית המשפט (פסקה אחרונה)
13.) השווה את סעיפים 11 ו 12.