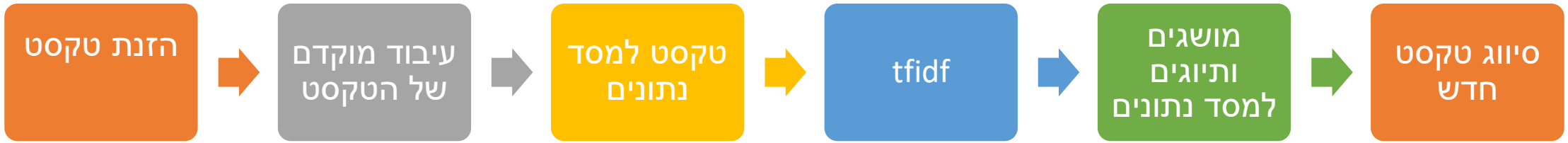
עבודת גמר: מסווג נושאים

יישום tf-idf

מבט מלמעלה


הזנת טקסט
עיבוד מוקדם של הטקסט
טקסט למסד נתונים
tfidf
מושגים ותיוגים למסד נתונים
סיווג טקסט חדש

צוותים

- צוות סריקת מידע - 2 או 3 חברים
- צוות עיבוד מקדים - 3 או 4 אנשים
- צוות מסד נתונים - 6 או 7 אנשים
  - לשקול חלוקה ל: (1) צוות שאילתות (2) צוות חיבור בפייתון
- צוות גרעין המערכת - 5 או 6
  - לשקול חלוקה ל: (1) צוות TFIDF (2) צוות סיווג טקסט חדש

הזנת טקסט

input: url address, label

output: text (not html) , label

צוות: סריקת מידע

בוחרים מאמר מויקיפדיה באנגלית ומעתיקים ידנית את הכתובת

מקבלים גם סיווג כמו "כדורסל"

השתמשו בחבילות "html2text" ו "requests" תחלצו רק טקסט בלבד! שישמר במשתנה

המודול מחזיר צמד של טקסט והסיווג שלו

עיבוד מוקדם של טקסט

input: raw text

output: list of documents

צוות: עיבוד מקדים

להפטר ממידע מיותר כמו הפניות להערות, כותרות וכדומה

הכינו רשימה של מילים מובחנות מתוך הטקסט, תסכלו עליה ותחשבו אם אפשר לאחד מילים שונות למושגים .קחו זמן להתבונן זה לא פשוט

אם מצאתם משהו כזה תייצרו טקסט מקביל שמורכב מהמושגים במקום מילים.

אבל אל תמחקו את הטקסט המקורי

השתמשו ב-regex

לחלק את הטקסט לפסקאות.

מעתה המסמכים שלנו אינם המאמר אלא כל פיסקה היא מסמך

החזירו את כל המסמכים, גם עם המילים גם עם המושגים

טקסט למסד נתונים

input: list of documents (terms & words) + labels

output: get method in python that selects documents by label

צוות: מסד נתונים

לשמור מסמכים עם מושגים בטבלה אחת. לשמור מסמכים עם מילים בטבלה נוספת. לכל מסמך צריך להיות מספר מזהה נפרד שמופיע בשתי הטבלאות.

תוסיפו עמודה לטבלת טקסט המושגים עם הסיווג של הטקסט

מעטפת בפייתון לכל השאילתות במסד הנתונים במשהו כמו psycopg2

תכתבו פונקציה בפייתון שתעשה שמירה כמו שמתואר בשלבים הקודמים (השאילתות בתוך הפונקציה, המשתמש לא צריך להיות חשוף ל-SQL)

תכתבו פונקציות שמביאה נתונים מתוך הטבלאות

i.e. get("basketball") returns a list of terms documents

or

get() -> gets all documents

בניית המידע המזויף: מסמך מושגים הוא מסמך שיש בו הרבה טקסט ולא רק רשימת מושגים, רק הוא מחליף מילים במושגים מסמך מילים זה אותו דבר אבל עם המילים המקוריות

tfidf

input: all the documents

output: list of dict {terms, labels & scores}

צוות: גרעין המערכת

תשמש במודול של שכבת מסד הנתונים כדי להביא את כל הנתונים. להביא את כל הטבלה או רק לפי סיווג מסוים

יישם idf עבור כל המסמכים במערכת שירוץ על כל מושג

כלומר

len(מסמכים)

---

count(מסמכים עם המושג)

ישם תדירות מושג (tf) שירוץ על כל מושג

כלומר

sum(המושג במסמך)

---

max(sum(כל מושג במסמך)

צור רשימה של מושגים וציון ה- tfidf עבור סיווגים שונים

השתמשו במודול מסד נתונים לשמור שהארגומנט לפונקציה כך:

list of dictionaries: {label,term,score}

אתם לא צריכים לדאוג למודולים של שכבת מסד הנתונים ההנחה היא שאתם מקבלים אותם מצוות אחר

טבלאות מבד נתונים ל tfidf

input: list of {term, label, tfidf_score}

output: save_tfidf([{t,l,score}]), get_tfidf(**label)

צוות: מסד נתונים

מושג, סיווג וציון ישמר לטבלה במסד הנתונים

לשאילתא הזו תכתבו מודול בפייתון

save_tfidf

מקבל רשימה של dictionaries

{label,term,score}

אם המושג והסיווג כבר קיים בטבלה צריך לעדכן את הציון כך:

N = מספר המסמכים בסיווג הזה.

תחשבו את הממוצע החדש של הציון בהתחשב בציון החדש

מודול שמביא מידע מהטבלה הזו, או הכל או לפי סיווג

get_tfidf() -> מביא הכל

get_tfidf

(label="basketball")

→ מביא את כל המושגים והציונים בסיווג כדורסל

the list of dict is none of your concern they will be implemented by core team

סיווג של טקסט חדש

input: url address

output: assumed label

צוות: סריקת מידע וגרעין המערכת

תביא את הטקסט מהאינטרנט באופן המתואר בהבאת הטקסטים הקודמים

חשב את ציוני tfidfעבור הטקסט הנבחן והשווה את הציונים של כל מושג לציונים של אותו מושג ברשימות הסיווגים.

כדי לערוך את ההשוואה השתמש במכפלה סקלרית ב-numpy
(אם יש צורך אסביר יותר אבל תנסו אולי לפתור את זה)

תחליטו על ערכי סף של דימיון כדי להחליט אם לשייך לסיווג מסוים סווגו את המסמך כסיווג1, סיווג2 או לא ידוע
(אפשר להשתמש בשמות הסיווגים זו רק הדגמה...)

להדפיס את התוצאות יפה בבקשה

אינטגרציה

- כל חלק שתיארנו כאן ישמר כחבילה (כדאי אולי לשמור את כל שכבת המידע כחבילה אחת עם כמה מודולים)
- חברי צוות האינטגרציה הם ראשי הצוותים של שאר הצוותים ובראשם ראש צוות אינטגרציה (שגם הוא יכול להיות חלק מצוות אחר). תעשו ישיבה מקדימה לפני שמתחילים לפתח כדי להבין טוב מה כל צוות צריך להוציא בסוף ולראות שכולם מבינים אותו דבר, ואחרי זה תתכנסו עוד פעם אחת באמצע העבודה לדווח זה לזה מה קורה וכשמסיימים לבנות את המרכיבים תתחילו לתפור.
- הרעיון הוא להפטר מכל הזיופים במרכיבים השונים וליצור חיבורים אמיתיים לזרימת הנתונים

בהצלחה