עיבוד שפה טבעית ש3:
סיווג מסמכים
Document Classification

פרקים: SLP 4-5, Eisenstein 2-4

מיני-הודעות

- להגיד שם כשמשתתפיות בשיעור
- קורס מבוא לבלשנות - מיד אחרינו באותה כיתה בימי רביעי...

פתרון הבוחן

שאריות - טוקניזציה עם NLTK

ביקורות על מחשבים ניידים

חוות דעת

מחשב מצויין וחזק !!

★★★★★

חוות דעת

מחשב מושלם לכל סוג עבודה. מסך מעולה, חיי סוללה מדהימים, ספיקרים איכותיים, בנוי טוב, קל וחזק, לא מתחמם, דק וקל לנשיאה.

★★★★★

חוות דעת

מוצר טוב מאוד ומוצלץ לכל אהובי אפל

★★★★★

חוות דעת

מחשב מעולה. מהיר מאד, נוח לתפעול, מסך רמקולים ומיקרופון מצוינים, המעבד החדש שלהם עובד מצויין

★★★★★

ביקורות על מחשבים ניידים

חוות דעת

כמה ימים אחרי הקנייה הפסיק להיתחבר לאינטרנט וכל מה שניסיתי לעשות לא עזר
עד שהגיעו מהחברה והחליפו את כרטיס הרשת האלחוטי.

חוות דעת

מחשב בסיסי בעל יחס עלות/תועלת מצויין.
מתאים לגלישה/זום/אפיס ושאר שימושים בייתיים ועסקיים קלים ולסטודנטים (לא למדעי המחשב).
זה לא מחשב גיימינג וככזה יש להתייחס אליו.
המקלדת קצת קטנה, אין כניסת רשת, אבל מי צריך רשת כשיש WIFI 5GHZ.
לי המחשב התחמם (הגיע ל-100C), הוחלף ולא חזרה התקלה.

חוות דעת

למחשב יש בעיית רעש נוראית, שריקות וצפצופים, גם לאחר שחזר מהמעבדה של HP. לא קיבלתי מענה
משירות לקוחות ונשארתי עם מחשב יקר ומרעיש. תחסכו לעצמכם את עגמת הנפש.

חוות דעת

מחשב מצוין, יעיל וזריז! נראה טוב.

זיהוי סנטימנט (sentiment) מתוך אתרי ביקורת

- פעולה שאינה תמיד פשוטה גם לבני אדם
- "תיוג" אנושי, אבל סובייקטיבי
  - תיוג אפשרי לאחר עיבוד מקדים (*-***** ⇐ +/-)
- פתרון באמצעות חוקים (rule-based) לפי מאפיינים לשוניים - אולי ייתן קירוב טוב
  - מילים מסוימות?
  - מילים בעלות אופי מסוים?
  - אורך המסמך?
  - סימני פיסוק?
- השיטות שנלמד יהיו גישות **לימוד מבוקר** (Supervised Learning)

עוד סוגים של סיווג מסמכים

- זיהוי דיעה, זיהוי עמדה (opinion mining, stance detection)
- סוגת המסמך (genre)
  - מתוך אתר חדשות, נתייג מאמרים לפי המדור ממנו הגיעו
  - (איזה "אופי" של הבדל בין מילים נחפש כאן, אל מול סנטימנט?)
- זיהוי ספאם
- זיהוי מחבר (authorship)

סיווג מסמכים - הגדרה פורמלית

- קלט: מסמך d, שהוא רצף תווים
- קלט: אוסף תגים $C = \{c_1, \ldots , c_k\}$
- פלט: תג רצוי $c \in C$

הערכת סיווג (evaluation)

- מטריצת בלבול confusion matrix

אמיתי /// חזוי	+	-
+	60	20
-	10	10

הערכת סיווג (evaluation)

- מטריצת בלבול confusion matrix

אמיתי /// חזוי	+	-
+	60	20
-	10	10

הערכת סיווג (evaluation)

- מטריצת בלבול confusion matrix

מספרים לפי הריצה

אמיתי /// חזוי	+	-
+	60 TP	20 FN
-	10 FP	10 TN

הערכת סיווג (evaluation)

אמיתי /// חזוי	+	-
+	60 TP	20 FN
-	10 FP	10 TN

- דיוק accuracy = אמיתי חלקי הכל
  - (TP+TN)/(TP+TN+FP+FN)
- מתי מאוד לא כדאי להשתמש במטריקת דיוק?
- נניח שמעניין אותנו התג החיובי (+)
  - Precision = TP/(TP+FP) ← כמה המודל פוגע
  - Recall = TP/(TP+FN) ← כמה המודל תופס
- שתי מטריקות שמתארות רצונות סותרים, ולכן נרצה מערכת ש"די טובה" בשתיהן

מטריקת F

אמיתי /// חזוי	+	-
+	60 TP	20 FN
-	10 FP	10 TN

Precision = TP/(TP+FP) ← כמה המודל פוגע

Recall = TP/(TP+FN) ← כמה המודל תופס

- **ממוצע הרמוני** בין P ל-R
- פרמטר β קובע כמה כל חלק "חשוב"
  - תיקון שגיאות - תפיסה זה חשוב, אבל פגיעה הרבה יותר
  - (לא נל"פ) איתור וירוס קורונה בבדיקה מהירה - תפיסה יותר חשובה
  - כמעט תמיד נשתמש במטריקה המאוזנת בכל-זאת.
- הנוסחה:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- מה הערך המאזן?
- איך מחזקים את R?

מה קורה כשיש יותר משני תגים?

/// חזוי אמיתי ///	**א**	**ב**	**ג**	**ד**
א	120	20	5	5
ב	60	60	30	20
ג	5	0	5	20
ד	10	10	10	10

הערכת סיווג מרובה-תגים

/// חזוי אמיתי ///	א	ב	ג	ד
א	120	20	5	5
ב	60	60	30	20
ג	5	0	5	20
ד	10	10	10	10

- עבור כל תג אפשר לחשב בנפרד P, R, F
- אנחנו רוצים לדווח **מספר אחד** למערכת שלנו
- אפשרות א': למצע F בין התגים
  - מיצוע מאקרו (macro-averaging)
  - למי הוא נותן יותר משקל?
- אפשרות ב': לחשב P, R על הכל ולחשב F אחד
  - מיצוע מיקרו (micro-averaging)
  - מי מקבל כאן יותר משקל?

לימוד מכונה

- אמצעי ללמוד **מודל** (model) מתוך נתונים
  - המודל מכיל **פרמטרים** (parameters) המתעדכנים/נלמדים תוך כדי תהליך **אימון** (training)
- מחלקים את הדאטא לאימון (training set) ומבחן (test set)
  - שלב המבחן לפעמים ייקרא הסקה (inference)
  - לרוב נוסיף סט שלישי בשם פיתוח (development, dev) או אישוש (validation)
- לאופן הייצוג של הנתונים נקרא **פיצ'רים** (features)
- חומר עזר במודל

פיצ'רים במערכת סיווג

- ספירת מילים (Bag of words)
- לקסיקונים (Lexicons)
  - רשימות מילים מקוטלגות כבעלות "אופי" או "תכונה" - ניתן לספור הופעות
- "צורת המסמך" (surface features)
  - מספר מילים
  - מספר types
  - סימני פיסוק
  - מבנים תחביריים
- פיצ'רים חוץ-טקסטואליים
  - (לביקורות) שם המשתמש
  - זמן הפרסום

בייז התם (Naive Bayes) - לוח