מעבדה בעיבוד שפות טבעיות

סמסטר ב' 2017

פרופ' וינטנר שולי

מגישים:

בן-יעקוב נדב

אפראימוב אורן

סביבת עבודה:

התוכנה פותחה במערכת הפעלה Ubuntu 16.04LTS ופועלת גם במערכת הפעלה Windows10. התוכנה נכתבה בשפה פייתון 3.5.4 גרסת אנקודה ועל מנת להריץ את התוכנה יש להתקין גרסה זו או להתקין את החבילות pandas ו-skleran דרך pip.

סביבת עבודה:

הוספנו לתוכנה קובץ קונפיגורציה (config.ini) דרכה ניתן לשנות שמות תיקיות, להריץ 'מוד בדיקות' ולאפשר הרצה במקביל בחלקים ב' ו-ג'.

* על מנת להריץ עבודה במקביל, יש צורך להגביל את מספר ה-cores בהתאם לכמות הראם שיש במחשב. בהתאם לכמות הchunks שיש בידנו, במחשב של 16 גיגה אפשר לעבוד על 2 ליבות במקביל.

חלק א':

המשימה בחלק זה הייתה:

"הכינוי את הקורפוס ותייגו אותו עם מטה-אינפורמציה כפי שתואר לעיל. ניתן להעזר [בספריית פייתון scrapy](https://scrapy.org/) לצורך משימה זאת."

בחלק זה, הכינו את הקורפוס מ-4 מקורות מידע:

* 'עם עובד' – סיפורת מקור וסיפורת תרגום ([קישור לסיפורים](http://www.am-oved.co.il/page_23289)(
* 'דורטומה' – כולל סיפרות מקור ותרגום ([קישור לסיפורים](http://www.dortome.com/153719/%D7%A4%D7%A8%D7%A7%D7%99%D7%9D-%D7%A0%D7%91%D7%97%D7%A8%D7%99%D7%9D)(
* 'הקיבוץ המאוחד' - כולל סיפרות מקור ותרגום ([קישור לסיפורים](http://www.kibutz-poalim.co.il/%D7%A4%D7%A8%D7%A7%D7%99%D7%9D_%D7%A8%D7%90%D7%A9%D7%95%D7%A0%D7%99%D7%9D))
* 'פרויקט הסיפור הקצר - כולל סיפרות מקור ותרגום ([קישור לסיפורים](http://www.shortstoryproject.com/he/%D7%A1%D7%A4%D7%A8%D7%99%D7%99%D7%94/))

מלבד המידע שקצרנו מ-'פרויקט הסיפור הקצר', הקצירה הייתה מאוד קשה!

כל אתר בנוי בפורמט שלו, אין חוקיות בין הסיפורים באותו אתר, התבנית בין סיפור לסיפור באותו אתר שונה זו מזו ואפילו בחלק מן המקרים הסיפורים היו תמונות, pdf או הובילו לאתר אחר.

בנוסף, שמות הקבצים היו ברוב המקרים לא תקנים: אורך גדול מידי, מכילים תווים לא חוקיים בווינדוס, מכילים תווים charmap וכד.

על מנת לתמוך גם במערכת ווינדוס וגם בלינוקס, בנינו מערכת שיודעת להתמודד עם בעיות אלו, להודיע בהתאם ולשנות את שמות הקבצים.

הבעיה הכי קשה בה נתקלנו בקצירה, היא שלכמעט חצי מסיפורים שקצרנו, היה חסר שפת המקור. לכן, על מנת להתמודד עם בעיה זהו, הגדרנו מחלקה אשר מקבלת קובץ טקסט מפורמט הבא "שם הסיפור המקורי (שם הסיפור לשינוי במידה ויש) : שם המחבר : שם המתרגם במידה ויש : שפת מקור.

לדוגמא:

* אדיוס, המינגוויי (אדיוס, המינגווי) : ליאונרדו פדורה : טל ניצן : ספרדית

הסוגריים בחלק הראשון אינם חובה ובמידה וקיים אנו משנים את שם הקובץ לפי ערך זה.

* אבן, נייר : תומר גרדי : None : עברית

עבור חלק הקדשנו קרוב ל-200-250 שעות עבודה, על מנת להכין מערכות שיודעת לקרות בצורה אוטומטית (למעט הוספת מידע נוסף) ולהוריד סיפורים חדשים מכול אחד מן 4 המקורות.

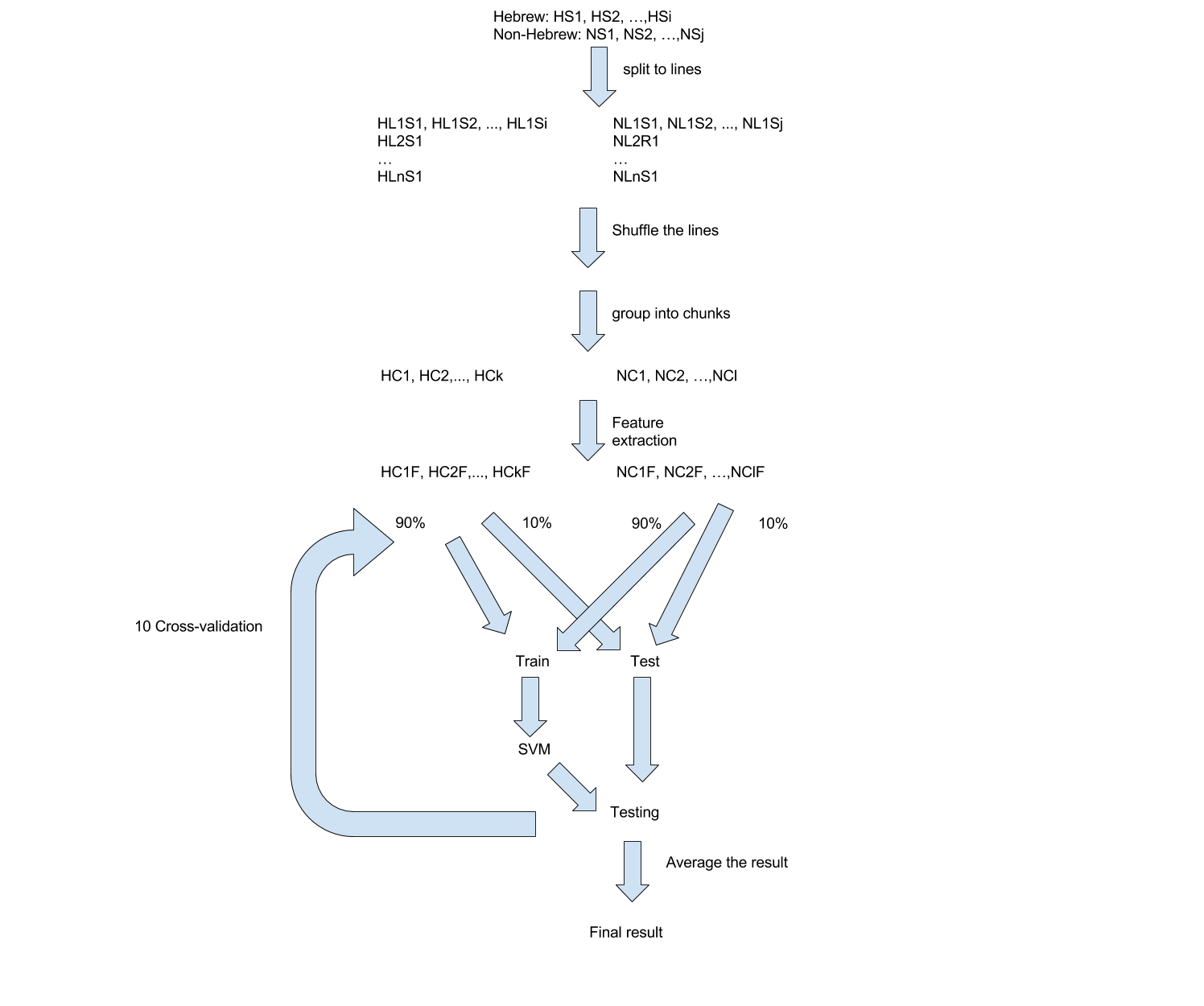
חלק ב':

המשימה בחלק זה הייתה:

" הראו כי שני סוגי הטקסטים (עברית מקור ותרגום) ניתנים להפרדה **בינארית**. השתמשו ב character-ngrams ו POS-ngrams (צירופים של חלקי דיבר) כתכונות לסיווג. ניתן להעזר ולקבל השראה מעבודה שנעשתה על טקסטים בעברית  - [זיהוי תרגומית בעברית](https://academic.oup.com/dsh/article/31/1/30/2605458/Identifying-translationese-at-the-word-and-sub). הגדרת תכונות (feature extraction) צריכה להתבסס על יחידות טקסט בגודל משתנה: בין 500 ל 2000 מילים. הקוד שלכם צריך לתמוך בקלות בשינויים בגודל יחידות הטקסט. העריכו את איכות התוצאות בשיטת ten-fold cross-validation. לאור עבודות קודמות, הציפיה היא לדיוק גבוה מאוד באיכות ההפרדה."

בחלק זה הוכחנו כי 2 סוגי קורפוסים, אשר השגנו בחלק א', עברית מקור ועברית תרגום ניתנים להפרדה לינארית.

לשם כך, פעלנו באופן הבא:



תחילה, פיצלנו כל סיפור לשורות. לאחר מכאן, ערבלנו את כל שורות הסיפורים עבור קורפוס בנפרד. בשלב הבא, איחדנו את כל השורות ל-chunks (אשר ניתן לעריכה מתוך קובץ הקונפיגורציה) בעלי 5000 תווים (פחות או יותר).

בשלב זה חילצנו את ה-features מכך chunk. ה-features בהם בחרנו להשתמש הם מניית מופעי כל תו, זוג תווים ושלשות תווים בשורות ה-chunk.

בשלב האחרון, יש בידינו את כל feature vectors. על מנת להוכיח כי ניתן לסווג את המידע, בחרנו לבצע 10-fold cross-validation (אשר גם את משתנה זה ניתן לערוך דרך קובץ הקונפיגורציה). כלומר, חילקנו את ה-feature vectors לקבוצת לימוד (90%) וקבוצת אימון (10%).

אימנו 4 מסווגים, אשר נלקחו מספרייה skleran:

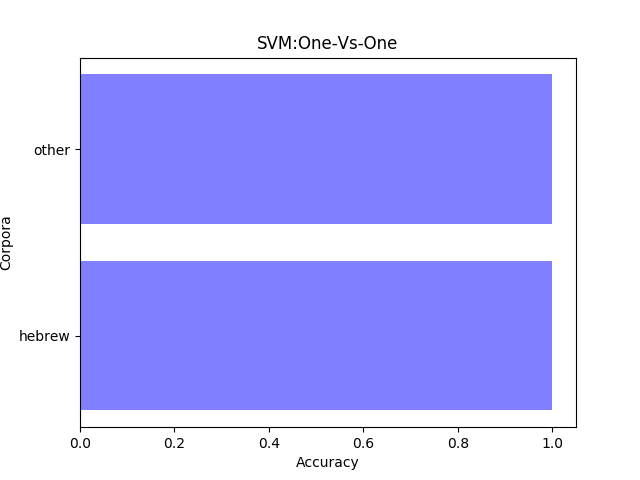
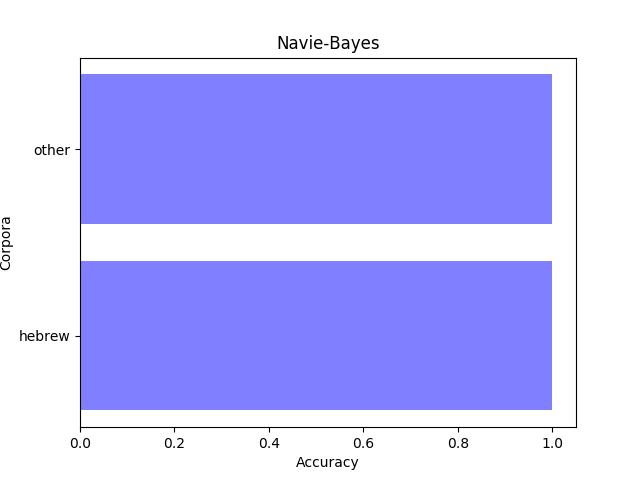
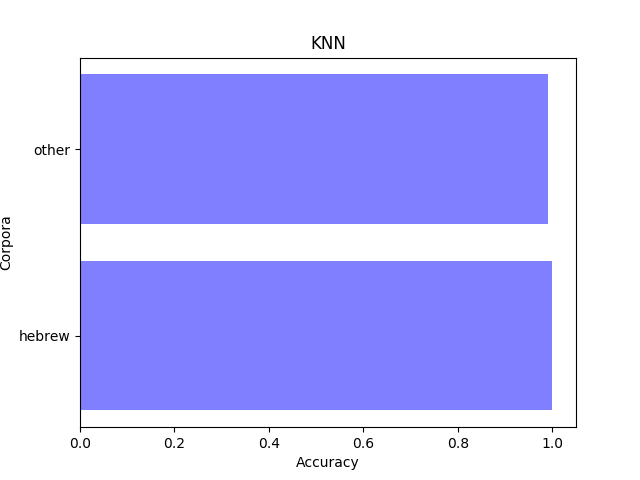
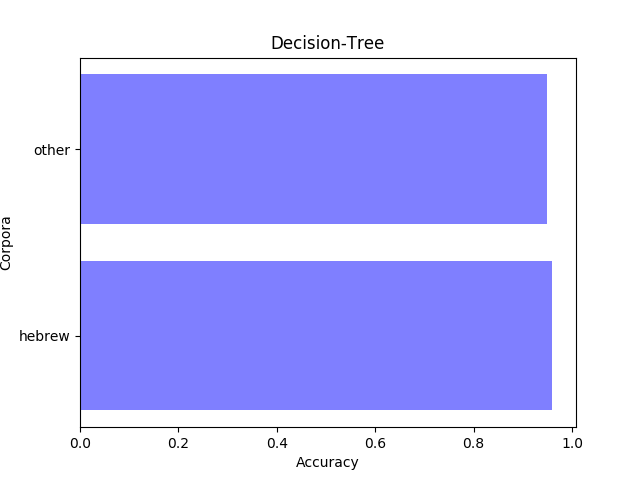
* SVM – One Vs One.
* Navie-Bayes.
* Decision-Tree.
* KNN.

אימנו כל אחד מן המסווגים על אותו קבוצת לימוד ואימון, על מנת להשוות את איכות המסווג ואת ההבדלים בניהם. על התהליך חזרנו 10 פעם כמספר kfold, כאשר כל פעם השתמשנו בקבוצות אימון ובדיקה שונות (פונקצית ספריה של sklearn), כאשר כל קבוצות הבדיקה זרות זו לזו.

100 הפיצ'רים המשמעותיים ביותר

בקובץ במיקום NLP-LAB\Meta Data\Part\_B\best\_features ניתן לראות את 100 הפיצ'רים המשמעותיים ביותר. כל פיצר תחום ב-' (בשתי גרשים על מנת לתפוס תווים מיוחדים, שבירת שורה, הזחה, רווח וכד').

עבור כל מסווג הכינו גרף אשר מראה תוצאות המסווג עבור כל קורפוס שנבדק. רצינו לבחון את טיב המסווג, בגלל שיש כמות לא מאוזנת של דגימות מכול שפה, דבר זה עוזר לנו להבין טוב יותר את איכות הסיווג.



תוצאות המסווג:

SVM:One-Vs-One - avg 0.9984

Navie-Bayes - avg 0.9994

Decision-Tree - avg 0.9966

KNN - avg 0.9847

בנוסף ניתן להסתכל בקובץ classifier\_analysis.txt אשר מנתח עבור כל קורפוס את הזיהוי של כל מסווג.

הקובץ נראה ככה:

Kfold number '0', 'KNN' classifier name, 'hebrew' corpora name:

The number of the vectors learned '1204'

The number of the vectors tested '130'

The number of vectors identified correctly '128'

The list of the wrong languages has been identified '['other']'

השורה הראשונה מציינת את מספר kfold בו אנו נמצאים, לאחריו סוג המסווג בו אנו עושים שימוש, ועל איזה קורפוס אנחנו בודקים.

השורה האחרונה מציינת את רשימת השפות הלא נכונות שמסווג סיווג את הווקטורים. בחלק זה, בגלל שיש 2 שפות אין משמעות לשורה זו!

**לסיכום, ניתן לראות בקלות, כי שתי סוגי הקורפוסים ניתנים להפרדה לינארית.**

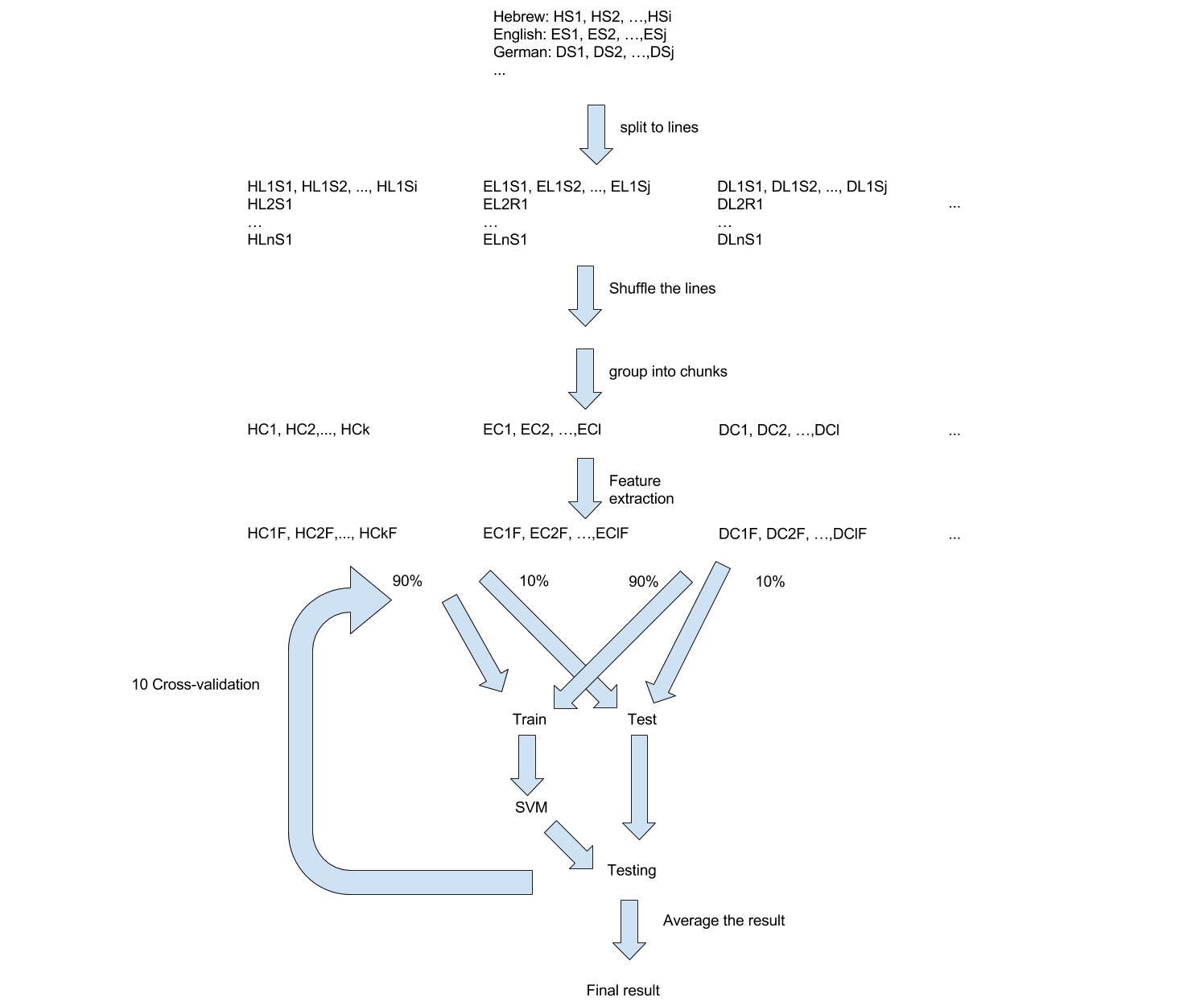
חלק ג':

המשימה בחלק זה הייתה:

"כעת הפרידו בין תרגומים לעברית משפות מקור שונות (בנוסף לעברית מקורית). כאן משימת הסיווג כבר אינה בינארית: מספר המחלקות הוא כמספר שפות המקור. לפיכך, צפוי הדיוק לרדת (שוב, העריכו את התוצאות ב"-cross-validation).

בחלק זה הוכחנו את יכולת הסיווג של טקסטים, אשר השגנו בחלק א', בין השפות העיקריות: איטלקית, אנגלית, גרמנית, הונגרית, יוונית, יידיש, נורווגית, ספרדית, עברית, ערבית, פולנית, פורטוגזית, צ'כית, צרפתית, רוסית ושוודית.

* ניתן בחלק זה לקרוא נתונים מתוך הקובץ corpora\_names\_to\_classifier.txt או לקרוא את כל השפות שיש להם לפחות כמות chunks גדולה או שווה ל-k ב-k-cross validation).



אותו אלגוריתם כמו בחלק הקודם.

השוני היחידי, שכן אימנו עם 2 סוגי SVM, One-VS-One ו- One Vs All.

רצינו לראות את ההבדלים באיכות הסיווג ועד כמה הם משמעותיים.

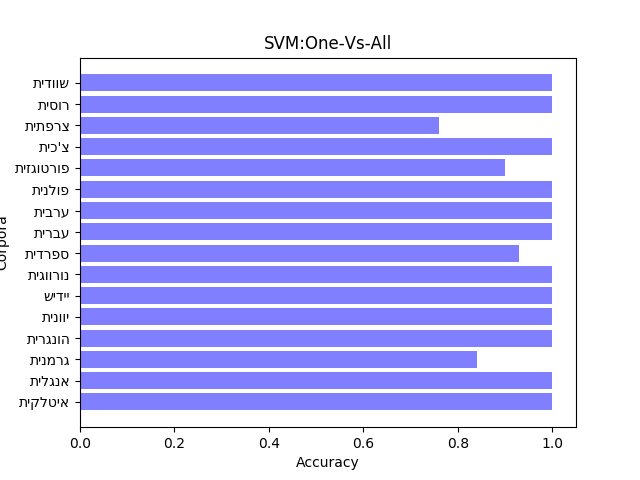
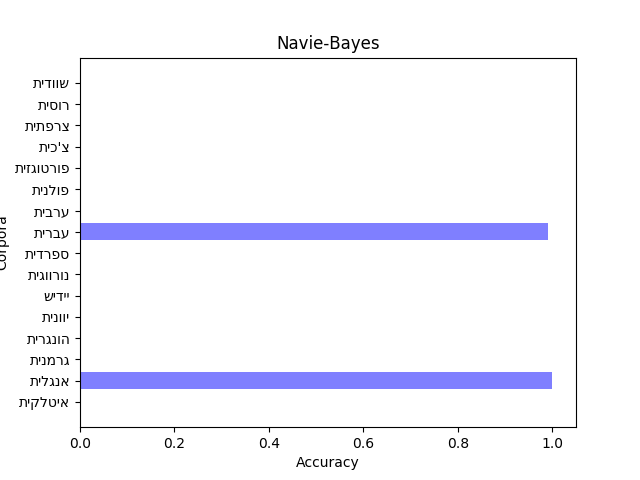
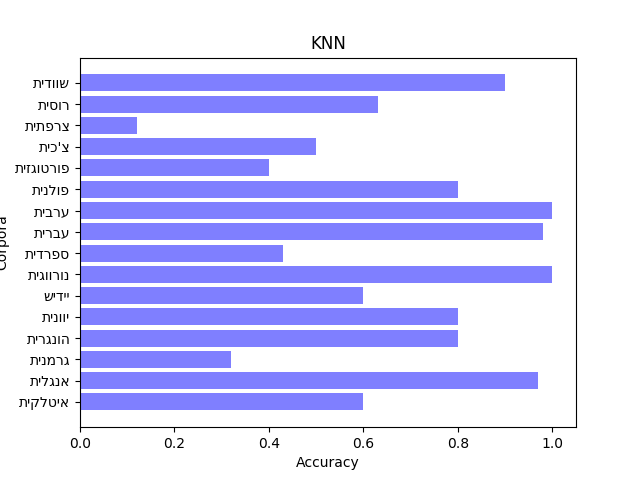
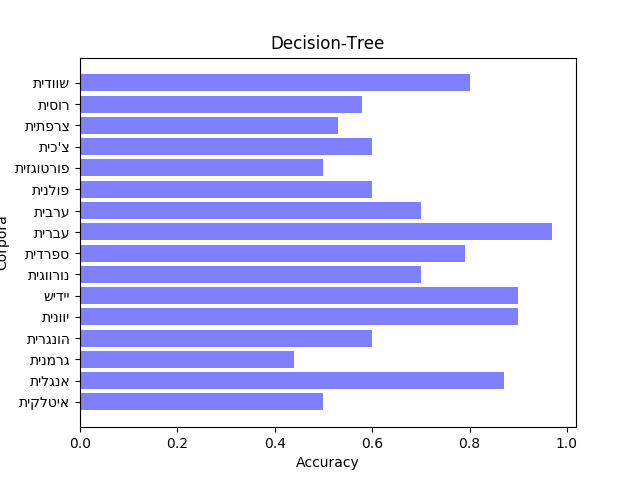
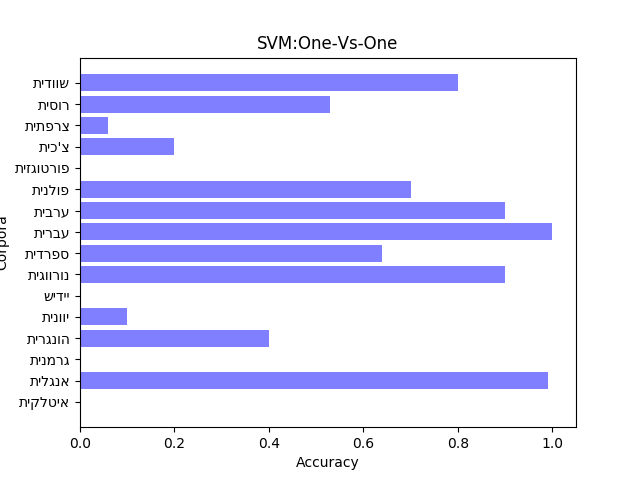
אימנו 5 מסווגים, אשר נלקחו מספרייה skleran:

* SVM – One Vs One.
* SVM – One Vs All
* Navie-Bayes.
* Decision-Tree.
* KNN.

100 הפיצ'רים המשמעותיים ביותר

בקובץ במיקום NLP-LAB\Meta Data\Part\_C\best\_features ניתן לראות את 100 הפיצ'רים המשמעותיים ביותר. כל פיצר תחום ב-' (בשתי גרשים על מנת לתפוס תווים מיוחדים, שבירת שורה, הזחה, רווח וכד').

על מנת לבחון את טיב המסווג, עבור כל מסווג יצרנו גרף אשר את אחוז הדיוק של המסווג. למרות שאחוזי הזיהוי מאוד גבוהים בכל אחד מן המסווגים, אנו רואים פערים משמעותיים ובחלק מן המסווגים עבור קורפוסים מסוימים אנו רואים אחוזי דיוק נמוכים ואפילו קורפוסים שבהם המסווג לא סיווג דגימה!



לאחר הרצה של כמה ניסויים שונים, עם מספר שונה של קורפוסים, התוצאות היו מעל 90 אחוז למסווגים ובחלק מן המקרים אף מעל 96 אחוז!!!

ברוב המקרים SVM – One Vs One היה מעל 95 אחוז ואילו SVM – One Vs All היה מעל 97%. למרות זאת, רק עבור המסווג SVM – One Vs All אנו מקבלים סיווג טוב ואיכותי שיודע להבדיל בין כל הקורפוסים ובנוסף לזהות בדיוק די גדול את השפה שלה קורפוס!

ניתן להסתכל בתיקייה Images - Analysis Classifier, יש שם עבור כל קורפוס את הסיווג שלו מול כל אחד מן המסווגים.

* שורה שאין בה כלום, מעיד שמסווג לא זיהה כהלכה את הווקטור (בדרך כלל במקרים בהם יש רק מספר קטן של בדיקות לסיווג).

תוצאות המסווג:

SVM:One-Vs-One - avg 0.9643

SVM:One-Vs-All - avg 0.9911

Navie-Bayes - avg 0.9323

Decision-Tree - avg 0.9091

KNN - avg 0.9068

ניתן לראות, רק על ידי המסווג SVM: One Vs All ניתן לסווג את הקורפוסים בצורה טובה ובאחוזי זיהוי מאוד גבוהים עובר כל אחד מקורפוסים.

חלק ד':

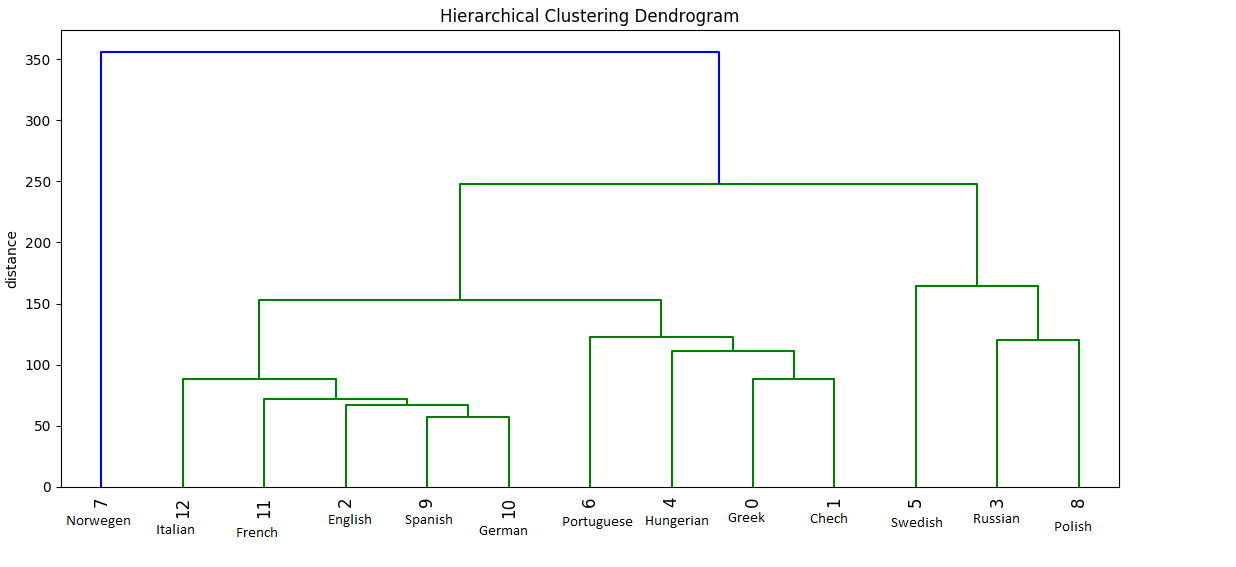
המשימה בחלק זה הייתה:

" בשלב זה תיישמו שיטות לא-מונחות של למידה ממוכנת. מטרתכם לבצע hierarchical clustering בינארי על תרגומים משפות שונות על מנת לראות האם תופעת ה interference חזקה מספיק כדי לשחזר עץ שפות פילוגנטי המשקף את ההתפתחות האבולוציונית של שפות המקור. תחילה בצעו clustering על תרגומים משפות מקור אשר משתייכות לשלוש המשפחות העיקריות של שפות ההודו-אירופאיות: גרמניות, רומניות וסלביות. לאחר מכן צרפו את שאר התרגומים, כמו גם עברית מקור, ובדקו כיצד משתנה התוצאה."

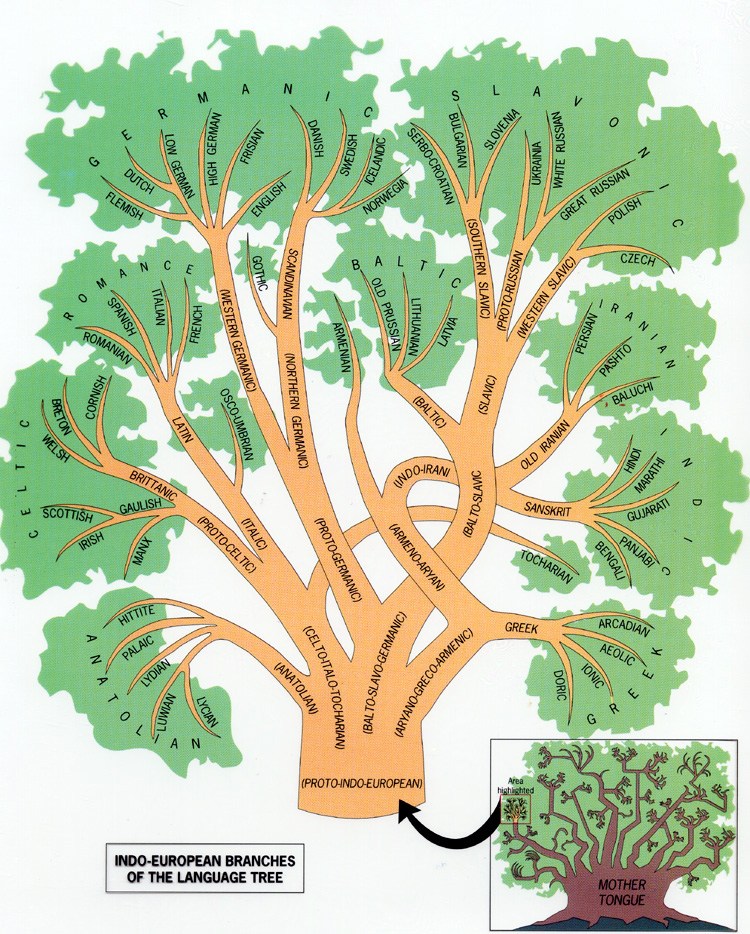
בחלק זה השתמשנו בסיפורים מן השפות איטלקית, אנגלית, גרמנית, הונגרית, יוונית, נורווגית, ספרדית, פולנית, פורטוגזית, צ'כית, צרפתית, רוסית ושוודית. פיצלנו כל סיפור לשורות. לאחר מכאן, ערבלנו את כל שורות הסיפורים עבור קורפוס בנפרד. בשלב הבא, איחדנו את כל השורות ל-chunks (אשר ניתן לעריכה מתוך קובץ הקונפיגורציה) בעלי 5000 תווים (פחות או יותר). בשלב זה חילצנו את ה-features מכך chunk. ה-features בהם בחרנו להשתמש הם מניית מופעי כל תו, זוג תווים ושלשות תווים בשורות ה-chunk.

בשלב האחרון, יש בידינו את כל feature vectors. חישבנו בעבור כל שפה את ממוצע הוקטורים, ובו השתמשנו כמרכז צביר השפה.

כאשר הרצנו את האלגוריתם, התקבל העץ הבא:



ניתן בקלות להבחין בקרבתן של שפות דוגמת הרוסית והפולנית, כמו גם קרבתן של האנגלית אל הגרמנית. עם זאת, לא נראה כי זוהתה הפרדה בין השפות הגרמניות לרומניות.

ניתן לראות כי קיימים מספר קווי דימיון בין העץ שהתקבל לבין עץ השפות האינדו-אירופאיות

כאשר הוספנו את השפות עברית, ערבית , ויידיש התקבל העץ הבא:

