תאריך: ‏03/04/2020

מסמך אפיון פרויקט

|  |  |  |
| --- | --- | --- |
| **הפרויקט** | | |
| שם הפרויקט: | זיהוי של בודק אנונימי של מאמר אקדמי | |
| מס' ב-LabAdmin: | 5614 | |
| סמסטר: | אביב 2020 | |
| חד/דו סמסטריאלי: | חד סמסטריאלי | |
| **הצוות** | | |
| שם המנחה: | פבל ליפשיץ | |
| שם סטודנט 1: | רועי החנוכי | מקצוע רישום: פרויקט ב' |
| שם סטודנט 2: | ליאור קיאסי | מקצוע רישום: פרויקט ב' |
| **חברה מלווה** | | |
| שם החברה: |  | |
| שם איש קשר: |  | |

**1. מטרת הפרויקט**

מטרת הפרויקט באופן כללי היא בניית אלגוריתמים ומודלים לזיהוי מחבר של טקסט – ובאופן פרטני יותר זיהוי של כותב Review על אודות מאמר אקדמי. מדובר בפרויקט שעוסק בחקר תחום ה-NLP ובפרט במשימת Authorship Attribution. הפרויקט בהתחלה יעסוק בחקר ובניית features מתאימים למשימה ובהמשך נבחן אלגוריתמי ML שונים (ייתכן תוך שימוש ב-DL) וטיוב ה-features הנ"ל תוך חק"ב מתאים.

**2. פירוט הנחות ודרישות**

אנו מניחים שקיימים Reviews רבים באינטרנט בעזרתם ניתן לייצר datasets אשר איתם ניתן לעבוד ולאמן את המודל. כמו כן, בשלבים הראשוניים של הפרויקט נעסוק בבניית מודל ראשוני בו נשתמש בטקסטים שונים הקיימים באינטרנט. הנחת המוצא של הפרויקט היא שקיים מספר מספק של Reviews איתם ניתן לעבוד, כולל מידע רלוונטי (כגון: מחבר הטקסט). נדרוש שהאלגוריתם יהיה מודולרי ויאפשר הוספת features או לחילופים שינוי המסווג באופן פשוט יחסית.

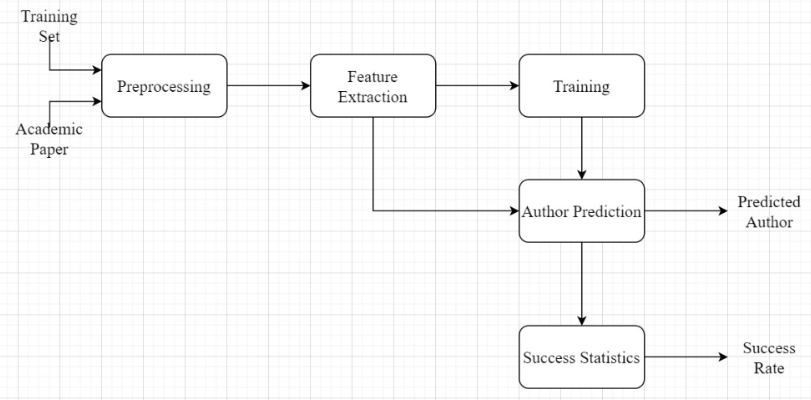
**3. פתרונות אפשריים וסיכום קצר של סקר הספרות**

בסקר הספרות הראשוני שבצענו למדנו חומר העוסק ב-NLP מקורס אונלייני של סטנפורד, קראנו ספציפית על בעיית Authorship Attribution במאמרים שונים וכן קראנו סיכומי תחרות בשם PAN שעוסקת בנושאים הרלוונטים לנו. למדנו בעיקר על ה-features השימושיים למשימה שלנו ואיזה מסווגים מניבים את התוצאות הטובות ביותר. כמו כן, למדנו על השיטות השונות לנתח תוצאות ולמדוד הצלחה.

פתרונות אפשריים שראינו עוסקים בניתוח לקסיקלי, סינטקטי וסמנטי של טקסטים ובחירת features שהניבו את התוצאות המדויקות ביותר (כגון: ngrams, "bag of words" וכו').

מקורות עיקריים:

* Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. "Overview of the Author Identiﬁcation Task at PAN 2018 - Cross-domain Authorship Attribution and Style Change Detection." Avignon, France, September 10-14, 2018.
* Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the Cross-Domain Authorship Attribution Task at PAN 2019. Lugano, Switzerland September 09-12,2019.
* Yunita Sari. "Neural and Non-neural Approaches to Authorship Attribution - A Thesis submitted to the University of Sheﬃeld for the degree of Doctor of Philosophy in the Faculty of Engineering". Department of Computer Science The University of Sheﬃeld, September 2018.
* <http://web.stanford.edu/class/cs224n/> - Stanford, CS224n: Natural Language Processing with Deep Learning. Instructor: Christopher Manning.
* Patrick Juola. Authorship Attribution. "In Foundations and Trends in Information Retrieval", Volume 1, Issue 3, March 2008.
* Moshe Koppel, Jonathan Schler, and Shlomo Argamon. "Computational Methods Authorship Attribution". Journal of the American Society for Information Science and Technology, Volume 60, Issue 1, pages 9-26, January 2009.
* Efstathios Stamatatos. A Survey of Modern Authorship Attribution Methods. "Journal of the American Society for Information Science and Technology", Volume 60, Issue 3, pages 538-556, March 2009.
* Bander Alsulami, Edwin Dauber, Richard Harang, Spiros Mancoridis and Rachel Greenstadt. "Source Code Authorship Attribution using LongShort-Term Memory Based Networks". Drexel University, Philadelphia PA 19104, USA and Invincea, Inc, Fairfax VA 22030, USA.

**4. תרשים מלבנים (block diagram) של הפתרון הנבחר או הנבדק**

**5. מודולים שנידרש לפתח**

- **Preprocessing** – מודול של עיבוד ראשוני של טקסט במטרה להתאימו למסווג שלנו (כגון: הקטנת אותיות ל-lowercase, הקטנת רווחים, מחיקת Unicode וכו').

- **Feature Extraction** – מודול להוצאת פיצ'רים באמצעותם נאמן ונבצע את הסיווג במודל (כגון: n-grams, "bag of words" וכו').

- **Classifier** – מסווג המודל אותו נאמן ובאמצעותו נבחן Data חדש.

- **Success Statistics** – מודול אשר מחשב ומוציא סטטיסטיקות אודות האימון והסיווג, שלמעשה ישקף מדדי הצלחה של הפרויקט.

**6. מודולים מוכנים שניתן להיעזר בהם**

- <https://www.nltk.org/> - ספריה עם כלים להוצאת פיצ'רים שונים ב-NLP לשימוש (כאשר ה-Datasets הינם טקסטים).

- PyTorch – ספריית DL של Python.

- <https://github.com/huggingface> - דף גיטהאב עם כלים שונים להתעסקות ב-NLP כולל חלוקת טקסט ל-Tokens.

- sklrean – ספריית ML של Python.

**7. סביבת עבודה וכלי פיתוח שיהיו בשימוש**

- מכונת לינוקס(?).

- Python 3.

- Pycharm(?).

- Anaconda/Google Collab(?).

**8. שיטת הבדיקה שתידרש בסיום הפרויקט**

לאחר אימון המודל וביצוע סיווג על test set, ניצור confusion matrix על סטטיסטיקות ההצלחה של המודל. נשתמש ב-confusion matrix לחישוב מדדי הצלחה כגון accuracy, precision, recall, F1 score. במהלך הפרויקט נבחן אילו מהמדדים מייצגים באופן הנכון ביותר את מטרת הפרויקט, ואף נשקול הוספת מדדים נוספים מהספרות במידת הצורך.

**9. רשימת משימות:**

|  |  |  |
| --- | --- | --- |
| מס' | שם המשימה | תיאור המשימה |
| 1 | למידה עצמית | למידת תחום ה-NLP ומשימת Authorship Attribution תוך ביצוע סקר ספרות |
| 2 | הקמת סביבת עבודה | בחירת סביבת פיתוח ומחקר המתאימה לבעיה שלנו |
| 3 | Toy Problem | בניית אלגוריתם פשוט למשימה קטנה יותר – כולל preprocessing, feature extraction, classification בסיסיים. |
| 4 | מצגת אמצע |  |
| 5 | מציאת והכנת datasets | מציאת ה-datasets שנעבוד איתם בפרויקט, וביצוע preprocessing. |
| 6 | פיתוח המודל | בחירת features, בנייה ובחירת מסווג מתאים, כתיבת מודול לניתוח תוצאות. |
| 7 | הרצת המודל ובחינת תוצאות | השמשת מודל הפרויקט. |
| 8 | ספר פרויקט ומצגת סוף |  |
| 9 | מאמר אקדמי (?) |  |

**10. תרשים גאנט (התקדמות הפרויקט):**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| מספר חודשים מתחילת הסמסטר | | | | | | | | | משימה |
| 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
|  |  |  |  |  |  |  |  |  | 1 |
|  |  |  |  |  |  |  |  |  | 2 |
|  |  |  |  |  |  |  |  |  | 3 |
|  |  |  |  |  |  |  |  |  | 4 |
|  |  |  |  |  |  |  |  |  | 5 |
|  |  |  |  |  |  |  |  |  | 6 |
|  |  |  |  |  |  |  |  |  | 7 |
|  |  |  |  |  |  |  |  |  | 8 |
|  |  |  |  |  |  |  |  |  | 9 |