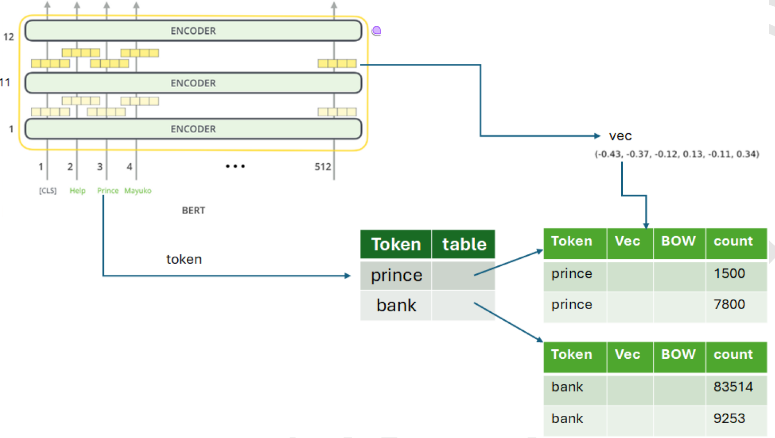
**יצירת word2vec המתמודד עם מילים הומוגרפיות וגם עם OOV**

לכל מילה ולכל חלק-של-מילה נתונים לנו במילון של Bert2Vec הווקטור המייצג, ה- BOW וגם מספר העדכונים.



קישור לדאטה-פריים המתואר בתרשים:

<https://drive.google.com/drive/folders/1zdt_Y0skocOyge4QN9Da7mjUCyE9vAib>

אנו מעוניינים לעבור על הטקסט וליצור מחרוזות ייחודיות לכל המשמעויות של המילים ההומוגרפיות, למשל, נניח שלמילה bank ישנן רק שתי משמעויות, האחת בהקשר של מוסד פיננסי והשניה בהקשר של גדה של נהר, נניח שמבחינת אינדקסים אצלנו במילון הראשונה היא משמעות פיננסית (bank1) והשניה משמעות של גדה של נהר (bank2).

מקבלים משפט

מריצים עליו את הטוקנייזר של BERT

"רואים" את המילה “bank”

דרך סביבת המילים (BOW) המופיעה ב- Bert2Vec ושימוש ב- BM25 יודעים באיזו משמעות המילה, אם במשמעות מוסד פיננסי נחליף את המילה bank במילה bank1, אם במשמעות גדה של נהר נחליף את המילה bank במילה bank2.

נשמור מילון חדש שש מימדי (bank1, Vec, NormVec, BOW, BOW\_B2V, count), המילה, הווקטור המייצג אותה ב- Bert2Vec, ה- BOW שלה במשפט המקורי (לפני טוקניזציה), ה- BOW\_B2V שלה לאחר טוקניזציה (כולל חלקי מילים) וכמה פעמים פנינו לכניסה/לווקטור זה. (והיה והגענו לעוד מילה בעלת משמעות זהה, נחליף את המילה בטקסט, נעדכן את ה- BOW, את ה- BOW\_B2V ונעלה את ה- count ב- 1)

כך נעבור על כל הטקסט.

נניח כי קיבלנו את המשפט הבא: This bank is **unprofessional** at all today.

לאחר שימוש בטוקנייזר נקבל את המשפט בצורה הבאה: This bank is **un ##pro ##fe ##ssion ##al** at all today .

עקרונית לכל חלק של מילה יש מספר כניסות (ולכן גם מספר ווקטורים ומספר קבוצות של BOW\_B2V).

אנו נרצה ליצור מילה אחת (**unprofessional**) בעלת ווקטור אחד; היצירה של הווקטור עצמו היא די פשוטה, היא תיעשה בשתי דרכים (1) סוכמים את כל המימדים (2) סוכמים את כל המימדים ואז מחלקים במספר חלקי המילים :

ל- (1)

(ערך מימד 0 של ווקטור התוצאה הוא: מימד 0 של **un** + מימד 0 של **##pro** + מימד 0 של **##fe** + מימד 0 של **##ssion** + מימד 0 של **##al**)

(ערך מימד 1 של ווקטור התוצאה הוא: מימד 1 של **un** + מימד 1 של **##pro** + מימד 1 של **##fe** + מימד 1 של **##ssion** + מימד 1 של **##al**)

וכן הלאה.

ל- (2) עושים את מה שעשינו ב-(1) ואז מחלקים במספר חלקי המילים, בדוגמא מחלקים ב- 5

בדרך זו נוצרו לנו לכל מילה שני ווקטורים שונים [שההבדל ביניהם הוא מכפלה בסקאלר] כמובן כל ווקטור יהיה שמור במקום אחר על מנת שנדע להבדיל ביניהם. יהיה סט של ווקטורים של הסכימה ויהיה סט לש ווקטורים של הסכימה המנורמלת.

לכל טוקן שהוא תת-מילה יכולים להיות כמה כניסות (הקשרים שונים) האתגר הוא לבחור את הכניסה/ה- BOW\_B2V המתאים ב- Bert2Vec.

This bank is **un ##pro ##fe ##ssion ##al** at all today .

אסביר ע"י דוגמה של תת-המילה **##pro**: נניח כי יש לה מספר כניסות, ניקח את הסביבה של **##pro** במשפט (כל הטוקנים המסומנים במארקר צהוב, עד 5 טוקנים מכל צד), נפעיל את הפונקציה BM25, הכניסה/הווקטור הכי קרוב, הוא יהיה הווקטור של הטוקן **##pro**; כך לכל תת-המילים האחרות. כך מצאנו את הווקטורים של כל תתי-המילה הבונים את המילה, סוכמים, ובכך יצרנו מילה עם ווקטור מתאים. לצורך השלמת המילון החמש-מימדי חסר לנו את ה- BOW, את ה- BOW\_B2V ואת ה- count, לצורך BOW נוסיף ל- BOW 5 מילים מימין ומשמאל למילה **unprofessional** המופיעות במשפט המקורי (לפני הטוקניזציה), לצורך ה- BOW\_B2V, נוסיף ל-BOW\_B2V 5 טוקנים מימין ומשמאל למילה **unprofessional** ואת ה- count נעלה ב-1.

בסוף נקבל את השישיה הבאה [כמובן עם האינדקס]: (**unprofessional**1, Vec, NormVec, BOW, BOW\_B2V, count)

נניח שהגענו למשפט שבו המילה **unprofessional** מופיעה במשמעות אחרת (נניח) אז נעשה את אותו תהליך וניתן אינדקס עוקב, **unprofessional**2, וכמובן נחליף בטקסט את המילה המקורית במילה **unprofessional**2.

איך נדע האם מדובר במשמעות שונה, נעשה דימיון קוסינוס בין הווקטור שנמצא אצלינו בחמישייה ולבין הווקטור שחישבנו ע"י הסכימה. אם הם דומים יותר מ- 0.68 אז מדובר באותה משמעות, אם הם דומים פחות מ- 0.62 אז מדובר במשמעות שונה, ניצור רשומה חדשה בטבלה, אם מדובר בטווח בין לבין, בטווח "האפור", רק נחליף את המילה בטקסט המקורי במילה @@**unprofessional**@@, לא נעדכן את הטבלה.

נניח שקיבלנו משפט חדש שבו המשמעות היא של **unprofessional**1 אבל סביבת המילים מעט שונה:

His behavior during the unprofessional meeting upset both clients and colleagues.

לאחר טוקניזציה התקבל המשפט במבנה הבא:

His behavior during the **un ##pro ##fe ##ssion ##al** meeting up ##set both clients and colleagues .

אופן בניית הווקטור תהיה זהה אבל בחירת הכניסות של הטוקן **##pro** תיתכן שונה, כי הסביבה שונה. התהליך יהיה אותו תהליך סביבת המילים תהיה מה שכעת במארקר ירוק. דבר זה יגרום, קרוב לוודאי שהסכימה של הווקטורים של תתי-המילים תהיה שונה, משמע, הווקטור יהיה שונה. אנו נרצה להחליף את המילה unprofessional במשפט האחרון במילה **unprofessional**1 כי זו אותה משמעות אבל זה אינו אותו הווקטור. במקרה כזה אם הווקטור שנוצר בסכימה הוא קרוב מבחינת cosin-similarity לווקטור של **unprofessional**1 שבטבלה, ז"א cosin-similarity גדול מ- 0.68 אז נעשה ממוצע משוכלל של שני הווקטורים בהתאם ל- count, נאחד את ה- BOW, את ה-BOW\_B2V ונעלה את ה- count ב-1.

(**unprofessional**1, Vec, NormVec, BOW, BOW\_B2V, count)

כל פעם שנתקלים במילה שקיימת אצלינו בטבלה באותה משמעות, מעדכנים את הטבלה באופן הבא:

1. Vec מתעדכן ע"י ממוצע משוכלל, אם count=5 🡨 Vec=Vec\*5+New-Vec (כמובן ברמה של כל מימד בפני עצמו)
2. בדומה גם NormVec מתעדכן ע"י ממוצע משוכלל
3. BOW מתעדכן ע"י הוספת 5 המילים מימין ומשמאל למילה הרלוונטית
4. BOW\_B2V מתעדכן ע"י הוספת 5 הטוקנים מימין ומשמאל למילה הרלוונטית
5. count=count+1

חשוב לציין, BOW ו- BOW\_B2V יהיו בנויים במבנה של זוגות, (word,num) ל- BOW, (token,num) ל- BOW\_B2V. המשמעות של num היא כמה פעמים המילה/הטוקן היו בסביבת המילה.

למשל, בהנתן המשפט הבא:

His behavior during the unprofessional meeting upset both clients and colleagues.

לאחר טוקניזציה:

His behavior during the **un ##pro ##fe ##ssion ##al** meeting up ##set both clients and colleagues .

נניח כי הכניסה/הרשומה הבאה צריכה להתעדכן: (**unprofessional**1, Vec, NormVec, BOW, BOW\_B2V, count)

נניח כי לפני העדכון תוכן BOW הוא: [(the,7),(His,3),(during,1), (maintains,1),(both,5)]

כעת: [(the,8),(His,4),(during,2), (maintains,1),(both,6),(behavior,1),(meeting,1), (upset,1),(clients,1),(and,1)]

נניח כי לפני העדכון תוכן BOW\_B2V הוא: [(the,7),(His,3),(during,1), (main,4),(##tains,1),(both,5)]

כעת: [(the,8),(His,4),(during,2),(main,4),(##tains,1),(both,6),(behavior,1),(meeting,1),(up,1),(##set,1),(clients,1)]

לאחר שעברנו על כל הקובץ (שינינו את כל המילים למילים עם אינדקס) מריצים את האלגוריתם Word2Vec **ואז עושים אנליזה על משימות מוכרות וידועות (לא חייב על הכל, זה דוגמאות חשובות, מוכרות ומקובלות) כמו:**

YELP, Question Answering (QA) and Natural Language Inference (NLI)**,** Quora Question Pairs (QQP), Question Natural Language Inference (QNLI), Stanford Sentiment Treebank (SST-2), Corpus of Linguistic Acceptability (CoLA), Semantic Textual Similarity Benchmark (STS-B), Microsoft Research Paraphrase Corpus (MRPC), Recognizing Textual Entailment (RTE), Winograd NLI (WNLI). אנלוגיות והמילים הקרובות ביותר

הדאטה סט של הניסויים דלעיל מופיע ומתואר במאמרים הבאים:

GLUE from https://gluebenchmark.com/ leaderboard and https://blog. openai.com/language-unsupervised. The GLUE originally summarized in Wang et al. (2018a)

<https://arxiv.org/pdf/1810.04805>

<https://arxiv.org/pdf/1804.07461>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Bidirectional encoder representations from transformers. *arXiv preprint arXiv:1810.04805*, 15.‏

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.‏

קישור לאתר הרשמי:

GLUE Benchmark

https://gluebenchmark.com/tasks/

hugging face:

DATASET:

https://huggingface.co/datasets/nyu-mll/glue

code example:

https://huggingface.co/docs/datasets/v1.16.1/quickstart.html

evaluate:

https://huggingface.co/spaces/evaluate-metric/glue

חשוב לראות ולהשוות בין התוצאות DyVat של Word2Vec של גוגל לבין Word2Vec שלנו, להערכתי התוצאות שלנו יהיה טובות יותר ולו רק בגלל שיש פחות רעש בגלל ההומוגרפיות.

בנוסף:

על כל אחד מהווקטורים של Brert2Vec להפעיל: (1) PCA, (2) TNSE, (3) UMAP; להוריד מימדים, מגודל של 768 לגודל של 300 מימדים, הגודל הסטנדרטי של Word2Vev & GloVe. וגם לגודל של 700 מימדים, המוטיבציה היא, אם יש רעש בווקטורים שהורדת המימדים תצמצם את הרעש ועדיין תשמור על רוב המידע החיוני.

ואז יש לנו מבנה נתונים של ווקטורים למילים בצורה הבאה:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **word** | **Word2Vec** | **Bert2Vec** | **W2V-BOW** | **B2V-BOW** | **B2V-PCA** | **B2V-TNSE** | **B2V-UMAM** |
| bank1 | <300-dim> | <768-dim> | … | … | <300-dim> | <300-dim> | <300-dim> |
| bank2 |  |  |  |  |  |  |  |
| happy1 |  |  |  |  |  |  |  |
| present1 |  |  |  |  |  |  |  |
| present2 |  |  |  |  |  |  |  |
| present3 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |