טו' בכסלו תשפ"א
1 בדצמבר, 2020

"מידע וטקסט"

עלון קבוצת עניין אחזור מידע וטקסט - SIGTRS
חוברת 2 כרך כז' – דצמבר 2020

חדשות מקבוצת העניין

שלום לכולם!

אנו נפגשים שוב בעלון חדש נוסף. עלון זה סוגר את כרך כז' של הקבוצה.
אני מאחל לכולנו המשך הנאה, עניין ופעילות ברוכה בתחום אחזור המידע.

ראוי לציין שנושאים רבים הנידונים בקבוצה נקבעים ע"י החברים עצמם בדפי המשוב בסיום המפגשים. אתם מוזמנים להמשיך ולהציע נושאים לדיון וכמובן להציע את עצמכם להרצאה.

אני מזכיר לכם את כתובת האתר של הקבוצה www.sigtrs.org. באתר חומרים רבים שנאספו במשך שנים רבות בתחום. אנא הפיצו בין חבריכם את כתובתנו ועודדו אותם להצטרף לרשימת התפוצה (הרשימה מפוקחת על ידי והתעבורה בה נועדה לעדכונים בלבד).

1. <u>קשר</u>

הקבוצה מקימת קשר עם חבריה באמצעות קבוצת דיוור אלקטרונית (Mailing-list) המאפשרת בצורה חופשית לכל אדם להירשם לקבוצה או למחוק עצמו ממנה. הדרכה כיצד להצטרף או לעדכן כתובת מופיעה באתר תחת התפריט "רשימת תפוצה של הקבוצה".

ספריות וגופים מרכזיים אחרים המעוניינים לקבל עותק מודפס של העלון צריכים לפנות בבקשה מיוחדת ליו"ר הקבוצה.

הרוצים להיות חברים בקבוצה צריכים להירשם לרשימת התפוצה האלקטרונית שלה באתר הקבוצה.

באתר הקבוצה מפה שנועדה להקל על ההגעה של החברים מחוץ לעיר. הדפיסו אותה לפני היציאה. אם אתם עושים שימוש בניווט לווייני (GPS) כוונו אותו לשע"ם, רחוב פועלי צדק 4, ירושלים. נסיעה טובה!

2. **<u>מפגשים</u>**

מאז הוצאת החוברת האחרונה (כרך כ״ז חוברת מס. 1) ביוני 2020 נפגשה הקבוצה פעמיים.

המפגש הראשון התקיים ביולי 2020

א. נושא ההרצאה: כתב יד, אחזור וקריאה בעברית ובערבית ע״י דרור בן דוד ועידן בר.
החברה פיתחה יישום OCR המבוסס על ארכיטקטורה של Encoder-Decoder וביניהם מנגנוני Attention ו-BILSTM לקידוד Sequence. פיתוח היישום נמשך למעלה משנה והוא כלל פיתוח יכולת ״לסנטז״ כתב בעברית ובערבית במגוון פונטים וסוגי הרעשות – ״תפור״ לסוגי ההרעשות שהלקוח פוגש. היכולת לאמן את המערכת נשענת על מחקרי עומק של חומרי הלקוח, כולל ניתוחים סטטיסטיים של תדירות הופעת מילים מסוימות, התמודדות עם כותרות, טבלאות, הערות שוליים, משפטים באנגלית שהשתרבבו לטקסט וכיוצא באלו. היישום שולב במערכות הלקוח, כולל התאמת הפתרון למערכות ההפעלה של הלקוח ויכולותיו בתחום תמיכה בתשתיות. המעבר לסביבת הלקוח כלל ביצוע בדיקות על חומרים שלו, ביצוע Fine Tuning וכיוצא באלו. על בסיס ממלאי התפקיד אצל הלקוח – המערכת נותנת ביצועים טובים יותר מכל מערכת אחרת שהלקוח הצליח למצוא בעולם (סד״ג 93% סיכויי הצלחה בפענוח מילה בערבית).

המפגש השני התקיים באוקטובר 2020

ב. נושא ההרצאה: חיפוש בטקסט חופשי גדול (BIG DATA) בעברית ע״י ד״ר אברהם מידן.
בהרצאה נבדקה מהי הדרך היעילה לחפש קטעים רלוונטיים (לדוגמה, תשובות לשאלות) בטקסט חופשי בעברית. הדרך המקובלת ו״הפשוטה״ ביותר היא: רושמים מילה, והתוכנה מחפשת את המחרוזת. השיטה עובדת לא רע באנגלית, אבל בעברית היא יוצרת הרבה החטאות ואזעקות שווא, עקב הגורמים הבאים: ניקוד: לדוגמה, ״שבת״ במובן ״יום שבת״, לעומת ״שבת״ במובן ״השתתף בשביתה״. הטיות לשוניות: לדוגמה הם מכים, הוא הכה (רק האות ״כ״ משותפת) אותיות שבשפות אחרות הן מילות יחס עצמאיות מהוות חלק מהמחרוזת: לדוגמה, ״מועד״ במובן חג, לעומת ״מועד העובדים״. אין כתיב אחיד, לדוגמה שיפור לעומת שפור.

3. **<u>מאמרים ועבודות</u>**

א. פרופ׳ יעקב שויקה הלך לעולמו (1936-2020), עפר דרורי
כולל פרסום חוזר של מאמרו של פרופ׳ שויקה על המורכבות בפיתוח מנועי חיפוש בעברית משנת 2005

ב. על כתפי ענקים: פרידה מפרופ׳ שויקה ז״ל, חיים סבתו

ג. על פרופ׳ שויקה בעקבות ההספד של הרב סבתו, עפר דרורי

ד. אקמול לקומה ג׳ נוחת עכשיו, דרור בן דוד

קריאה מהנה.

4. **<u>חסות</u>**

נכון להיום הקבוצה ללא חסות.
נשמח לקבל הצעות לחסות מארגונים בתחום העיסוק של הקבוצה.

5. **<u>כללי</u>**

קבוצת הענין "אחזור מידע וטקסט" (SIGTRS) מהווה פורום לאנשי מקצוע העוסקים בתחום אחזור טקסט, אחזור מידע וטכנולוגיות קשורות. אנשי המקצוע הם מפתחים (מנתחי מערכות ותוכניתנים) ו/או משתמשים.
הקבוצה הוקמה בשנת 1994 ע"י החתום מטה ומאז פועלת ברציפות הן במפגשי הרצאות והן בהפצת מידע בין היתר באמצעות עלון הקבוצה.
הקבוצה נפגשת ארבע פעמים בשנה להרצאות ולהחלפת רעיונות.
עלון הקבוצה יוצא פעמים בשנה בקביעות משנת 1994. כל חומר העלון בטקסט מלא נמצא באתר הקבוצה.
הצעות להרצאות, רעיונות או חומר כתוב אחר ניתן לשלוח ליו"ר הקבוצה:

עפר דרורי
שע"ם
offerd@gmail.com
טל: 02-5688439

ב ב ר כ ה

עפר דרורי
יו"ר הקבוצה

בקרו אותנו באתר הבית של הקבוצה: http://www.sigtrs.org