

# דו"ח פרויקט — ניתוח נושאים בעברית

## Topic Modeling (Hebrew) | Wikipedia + UGC

מגיש: צור זיו

GitHub: TzurZiv1/hebrew-topic-modeling-text-mining

### תקציר

בפרויקט זה השווייתי בין גישות קלאסיות ומודרניות לניתוח נושאים (Topic Modeling) בטקסטים בעברית, בשני קורפוסים: Wikipedia (מסמכים ארוכים) ו-UGC (מסמכים קצרים ורועשים). הערכת הביצועים נעשתה באמצעות מדדים כמותיים ( $c_v$  coherence,  $topic\_diversity$  ו- $dominance\_penalty$ ), ענישה לדומיננטיות לפי  $largest\_topic\_share$  שמייצרת. בנוסף בוצעה בדיקה איכותנית (מילות טופיק מרכזיות ודוגמאות מסמכים). עקב מגבלות RAM, חלק מהריצות בוצעו ב-Colab וחלק ב-Kaggle. כל תוצאות הביניים נשמרו כ-artifacts לצורך השוואה מרוכזת.

### 1. הגדרת הבעיה ומטרות

- חילוץ טופיקים (Topics) אינטרפרטיביים בעברית והשוואת מודלים בין Wikipedia ל-UGC.
- השוואת השפעת עיבוד מקדים (Preprocessing) בין `clean_text` לבין `lemma_text`.
- תוצרים: מחברות, Stage 1–5 תיקיית `results_artifacts` ומחברת השוואה מסכמת.

### 2. הנתונים

- **Wikipedia**: טקסט אנציקלופדי ארוך, עקבי יחסית ורמת רעש נמוכה.
- **UGC (User-Generated Content)**: טקסט קצר ורועש (סלנג, שגיאות כתיב, קיצורים, אמוג'ים).
- לכל קורפוס הוגדרו תצורות נתונים כגון `Wiki:clean_text`, `Wiki:lemma_text`, `UGC:clean_text`, `UGC:lemma_text`.

### 3. עיבוד מקדים

- נרמול טקסט: ניקוי תווים לא רלוונטיים, סטנדרטיזציה של רווחים ופיסוק.
- Tokenization עקבי לאחר נרמול.
- `clean_text`: טקסט נקי ללא הפחתה מורפולוגית.
- `lemma_text`: למטיזציה (Lemmatization) לצמצום דלילות ולהתמודדות עם מורפולוגיה עברית.
- הנחה: למטיזציה יכולה לשפר יציבות, אך עלולה לטשטש הבחנות סמנטיות.

### 4. מודלים שנבדקו

- **Baselines**: מודלים קלאסיים מבוססי Bag-of-Words / פירוק מטריצות (LDA, LSA, NMF).
- **Embedding-based**: מודל Top2Vec גילוי טופיקים בעזרת Embeddings ו-Clustering.
- **מתקדמים**: CTM — C-Top2Vec שימוש ב-contextual embeddings לשיפור סמנטי.

### 5. תכנון ניסויים

- ריצות מרובות למודלים הקלאסיים עם ערכי K שונים.
- ריצות מצומצמות למודלים המתקדמים בגלל אילוצי RAM וזמן.
- שמירת כל המדדים והדוגמאות כקבצי CSV להשוואה סופית.

## 6. מדדי הערכה

- **coherence\_cv**: קוהרנטיות מילות הטופיק (גבוה יותר = טוב יותר).
- **topic\_diversity\_final**: שיעור מילים ייחודיות מתוך מילות הטופיק.
- **dominance\_penalty**: מחושב לפי 1 פחות) largest\_topic\_share שיעור המסמכים בטופיק הכי גדול.
- **score\_final**: מכפלת הקוהרנטיות בגיוון.
- **score\_penalized**: הסיון לאחר מכפלת הציון בענישת הדומיננטיות.

## 7. תוצאות מרכזיות

המודל המוביל לכל תצורה לפי הציון המשוקלל: (score\_penalized)

Dataset	Model	K	Coherence	Diversity	Score Penalized
ויקי (מלא)	CTM	20	0.749	0.921	0.645
ויקי (lemma)	NMF	20	0.687	0.875	0.601
ויקי (clean)	NMF	20	0.633	0.869	0.551
UGC (מלא)	CTM	12	0.610	0.819	0.437
UGC (lemma)	NMF	10	0.427	0.903	0.386

## 8. בדיקה איכותנית וניתוח שגיאות

- הצגת מילות מפתח (top terms) ודוגמאות מסמכים לכל טופיק נבחר.
- זיהוי "טופיק-על": (Topic dominance) "דומיננטיות גבוהה מדי פוגעת ביכולת לפרש את התוצאות.
- קורפוס ה UGC-נמצא כרגיש במיוחד לאי-יציבות בשל קוצר המסמכים.

## 9. מגבלות והמשך

- מחסור ב RAM-מנע בחלק מהמקרים את הרצת הפחתת הטופיקים ההיררכית.
- המשך מומלץ: הפקת נתוני דומיננטיות למודלי ה Baselines-ואיחוד זרעי הרצה (seeds) להפחתת שונות.