

דו"ח פרויקט — ניתוח נושאים בעברית

Topic Modeling (Hebrew | Wikipedia + UGC)

מגיש: צור זיו

GitHub: [TzurZiv1/hebrew-topic-modeling-text-mining](https://github.com/TzurZiv1/hebrew-topic-modeling-text-mining)

תקציר

בפרויקט זה השוויתי בין גישות קלאסיות ומודרניות לנתח נושאים (Topic Modeling) בטקסטים בעברית, בשני קורפוסים Wikipedia (מסמכים ארוכים) ו-UGC (מסמכים קצרים ורוכשים).

הערכת הביצועים נעשתה באמצעות מדדים מסוימים (v_c) ו- $coherence_{cv}$: $topic_diversity$, $largest_topic_share$, $dominance_penalty$. ענישה לדומיננטיות לפי $share$ של נושא בדף בדיקה אינטנסיבית (מיilot topic מרכזיות ודוגמאות מסמכים). עקב מוגבלות, RAM חלק מהריצות בוצעו ב Colab; חלק ב-Kaggle. כל תוצאות הבניינים נשמרו כ-artifacts-לצורך השוואה מרכזת.

1. הגדרת הבעיה ומטרות

- חילוץ טופיים (Topics) אינטראקטיביים בעברית והשוואת מודלים בין Wikipedia ל-UGC.
- השוואת השפעת עיבוד מקדים (Preprocessing) בין `clean_text` ל-`lemma_text`.
- תוצאות: מחברות Stage 1–5 תקיית `results_artifacts` ומחברת השוואה מסכמת.

2. הנתונים

- **Wikipedia** : טקסט אנציקלופדי ארוך, עקי יחסית ורמת רעש גבוהה.
- **UGC (User-Generated Content)** : טקסט קצר ורועל (סלג', שגיאות כתיב, קיצורים, אמוג'ים).
- לכל קורפוס הוגדרו תכורות נתונים כגן, `Wiki:clean_text`, `Wiki:lemma_text`, `UGC:clean_text`, `UGC:lemma_text`.

3. עיבוד מקדים

- נרמול טקסט: ניקוי תווים לא רלוונטיים, סטנדרטיזציה של רוחחים ופיסוק.
- Tokenization עקי לאחר נרמול.
- `clean_text`: טקסט נקי ללא הפקחה מורפולוגית.
- `lemma_text`: למתייציה (Lemmatization) לצמצום דليلות ולהתמודדות עם מורפולוגיה עברית.
- הנחה: למתייציה 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 Clustering**: מודל Embedding-based — Top2Vec גליי טופיים בעזרת Embeddings ו-Clustering.
- **מתקדמים CTM**: 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.
- שבירת כל המددים והדוגמאות לקבצי CSV להשוואה סופית.

6. מדדי הערכה

- **coherence_cv:** קוהרנטיות מילוֹת הטופִיך (גבוה יותר = טוב יותר).
- **topic_diversity_final:** שיעור מילים ייחודיות מתוך מילוֹת הטופִיך.
- **largest_topic_share dominance_penalty:** מחושב לפי 1 פחות) שיעור המסמכים בטופִיך הכי גדוֹל).
- **score_final:** מכפלת הקוהרנטיות בגיאן.
- **score_penalized:** הסופי 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 אינטנסיבית וניתוח שגיאות

- הצגת מילוֹת מפתח (top terms) ודוגמאות מסמכים לכל טופִיך נבחר.
- דיזי "טופִיך-על": (Topic dominance) "דומיננטיות גבוהה מדי פוגעת ביכולת לפרש את התוצאות.
- קורפוֹם ה-UGC-נמצא כרגע במיוחד לא-יציבות בשל קוצר המסמכים.

9. מגבלות והמשר

- מחסור ב RAM-מנע בחלוקת מהמקרים את הרצת הפחתת הטופִיים ההיררכית.
- המשך מומלץ: הפקת נתוני דומיננטיות למודלי Baselines-ואיחוד זרעי הרצת (seeds) להפחיתה שונות.