נופר יעקובי גורש, תז: 201033503

סיכום מאמר - dagibbs: A command for topic modeling in Stata using latent Dirichlet allocation

שמות המחברים:

Carlo Schwarz University of Warwick Coventry, UK.

The Stata Journal (2018) 18, Number 1, pp. 101–117

קישור למאמר:

https://med.mahidol.ac.th/ceb/sites/default/files/public/pdf/stata_journal/sj18-1.pdf#page=107

מהי הבעיה שהמאמר בא לפתור? מדוע היא מעניינת או חשובה?

אוסף של טקסט (Text Data) כמו למשל מסמכי טקסט מהווים מקור מידע עשיר לחוקרים. לדוגמא, ישנם המון DataSet (סטים רבים של מידע) אשר מורכבים מאוספים גדולים של Text data אשר אינם מסווגים. האוספים הללו (collections) מכונים "corpora" והם יכולים להכיל אלפי מסמכים בודדים עם מיליוני מלים ואף יותר.

לפיכך, ברוב המקרים, לא ניתן להשתמש בכל המידע הזה לניתוח סטטיסטי ללא עזרה בכלי אוטומטי שמנתח טקסט או מכונת למידה אוטומטית. וזה בדיוק מה שהמאמר מתייחס אליו. כמובן שהחשיבות של דבר זה נודע בהמון נישות בחיינו. בין אם מדובר בצורך לימודי או מדיני ואף צבאי. נושא מכונות למידה אוטומטית וניתוח טקטסים אוטומטים בפרט הינו דבר שעומד על הפרק בחברות גדולות באופן יום יומי ולכן המענה שניתן לו מעניין מאוד.

מה הם הפתרונות שהיו קיימים לפני המאמר הזה, ומדוע הם לא מספקים?

ב-Stata, בעוד הפקודה screening (2010) מאפשרת חיפוש טקסט על סמך מילת מפתח בודדות, והפקודה txttool (2014) מאפשרת הצגה של טקסט כקבוצת מילים , הפקודה Ldagibbs מספקת אפשרות חדשה של ניתוח תוכן עבור מסמכים גדולים.

מה הפתרון שהמאמר מציע - מה התרומה של המאמר מעבר למה שכבר נעשה?

LDA הינו מודל נושאים (Topic model) הפופלרי ביותר ומאפשר קיבוץ אוטומטי של כל סוג של מסמך טקסט למספר נבחר של קבוצות בעלי הקשר דומה כאשר המספר נבחר ע"י המשתמש, אלו נקראים topics.

Ldagibbs מרחיב את יכולת ניתוח טקסט בStata, מה שמאפשר השוואה בין מסמכים שלמים על סמך קווי דימיון.

ועל ידי אפשרות זאת, הפקודה Ldagibbs הופכת מידע שלא היה ניתן לשימוש קודם לכן לאפשרי עבור החוקרים.

Gibbs sampling היא Markov chain Monte Carlo אלגוריתם,אשר מבוסס על לקיחת מדגם שוב ושוב מהסתברות החלוקה (מדובר בכלי סטטיסטי). הפקודה WinBUGS מספקת framework ליישום Gibbs sampler ב-Stata.

צעד ראשון, ldagibbs מחלק את המסמך למילים בודדות – מה שמכונה word tokens.(איסמון).

מילים אלו מוקצות באופן אקראי לאחד מהנושאים עם הסתברות שווה. זה מספק הקצאה ראשונית של מילים ומכך מאפשר את תחילת תהליך הדגימה. לאחר מכן, ldagibbs מקצה topic חדש עבור כל word tokens.

במאמר מצויינות כל הנוסחאות המתמטיות ופירוט רב על העניין המתמטי של זיהוי מילה כ-Topic והחלוקה לקבוצות. בחרתי לא לציין זאת כאן בסיכום אך אתייחס לכך בהרחבה במצגת.

הפקודה ldagibbs דורשת רק את המשתנה שמכיל את מחרוזות הטקסט כקלט. האפשרויות האינדוודואליות של הפקודה מאפשרות למשתמש לשנות את אופן הפעולה של Gibbs sampler ולתת שמות עבור משתני הפלט. .

לנוחיות המשתמש, ldagibbs כוללת גם כמה יכולות בסיסיות לניקוי טקסט כדי לאפשר הסרה של מילות מפתח ומילים קצרות מהנתונים.

איזו עבודה נשארה לעתיד?

במאמר זה, אין פרק שמדבר על סיכום המאמר או דברים עתידיים שניתן לעשות. לכן אסיק בעצמי מהמאמר מה ניתן לשפר בעתיד על סמך נקודות התורפה של LDA.

בתיאוריה, LDA יכול להיות מיושם על כל סוג של טקסט (Text data) ללא תלות באורך האינדוודואלי של הטקסט. LDA עובד היטב עבור תקצירים של מאמרים מדעיים או כאשר המסמכים מכילים לפחות 50-100 מילים. מספר המילים הנדרש על מנת "לפרש" בקלות את הטקסט, משתנה בהתאם לסוג המסמך וגודל אוצר המילים. כאשר מסמכי הטקסט קצרים מאוד, LDA עשוי לייצר נושאים (Topics) פחות משמעותיים למשמעות הטקסט. לדוגמא, LDA לא יעבוד כמו שצריך במקרים של טקסט קצר כמו "ציוץ" בטוויטר, בגלל שבמקרה זה יש פחות מדי מילים בעלי מכנה משותף שניתן "לקבץ" אותם. ניתן להתגבר על בעיה זו בעזרת שילוב של מספר מסמכים קצרים יחד או באמצעות שימוש במודל נושאים (topic model) אחר השונה מLDA.

מה דעתכם על המאמר? האם יש בו יתרונות/חסרונות מעבר למה שצויין?

אני חושבת שהמאמר נגע בחיסרון בולט של המודל לניתוח טקסטים.

נכון שכרגע הבעיה המרכזית והעיקרית היא ניתוח טקסט גדול שLDA נותן את המענה העיקרי לבעיה זו וזה בעצם היתרון הגדול שלו.

אך מה בדבר ניתוח של טקסטים קצרים? כאן ישנה מחשבה לעתיד, כיצד נוכל להפוך את הכלי הזה לשמיש גם עבור טקסט המכיל מספר קטן של משפטים. נכון שמבחינה הגיונית אין מספיק מילים משותפות כדי לייצר "אשכולות", אך עדין, ניתן לחשוב על אפשרות אחרת לנתח את הטקסט. אני חושבת שהשימוש בשילוב של מספר טקסטים קצרים ביחד יתן קבוצות לא בעלות משמעות רלוונטית לטקסט. יש להשתמש כאן במודל המשתמש בנושאי טקסט בצורה אחרת ולא כפי שLDA משתמש.