בס"ד

Alt-Max_EM_methods

במטלה נתבקשנו לקחת אוסף של מאמרים בעלי תשעה נושאים שאיננו מכירים, לבצע את האלגוריתם EM, ולסווג את המאמרים לפי התוצאות לתשעה נושאים נתונים.

<u>ישנם **פרמטרים רבים** איתם ניתן לנתח את תוצאות ריצת האלגוריתם שלנו:</u>

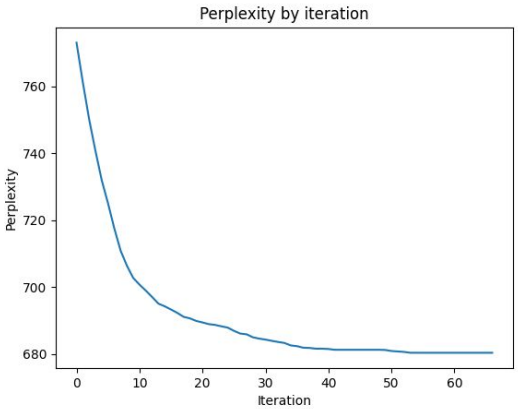
1. **Threshold to stop the EM iterations**

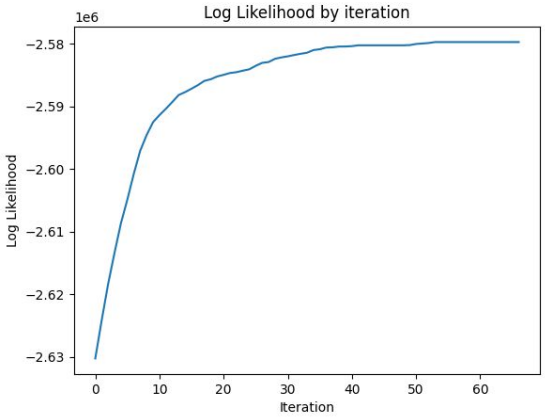
בחרנו epsilon = 0.01 עבורו נעצור את ריצת האלגוריתם. בסוף כל איטרציה של הרצה של EM ושיפור הפרמטרים, נבדוק האם ה- likelihood החדש קרוב לקודמו על כדי אפסילון, אם כן, נחדול את הריצה כיוון שהתכנסנו מספיק, אחרת, נמשיך לאיטרציה נוספת.

2. **Log likelihood & Perplexity graphs**

יצרנו שני גרפים המתארים את עליית ה- likelihood וירידת ה- perplexity ביחס לכמות האיטרציות, ומראות את השינוי מאיטרציה אחת לבאה אחריה, כאשר ציר ה- X מסמל את כמות האיטרציות וציר ה- Y מסמל את ln(likelihood) ואת ה - perplexity בהתאמה.

3. **Confusion matrix M of 9X9**



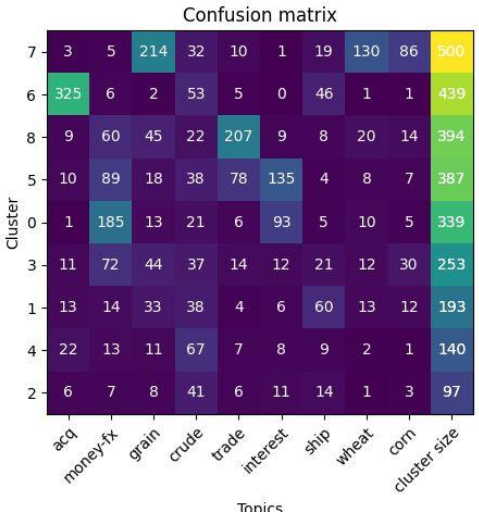


מטריצה המציגה את היחס בין כמות הסיווגים שמצאנו בהרצת המודל לבין אלו האמיתיים, כך שמתקיים:

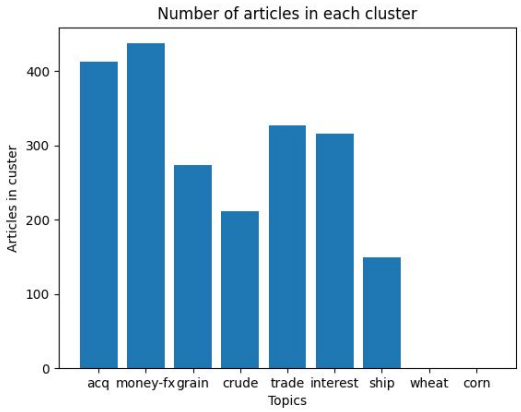
$$M_{ij} = \text{ the number of articles from the } j^{th} \text{ topic in the } i^{th} \text{ cluster}$$

בנוסף, סדר השורות הינו על פי כמות הסיווגים ל- cluster מסוים.

4. **Histograms graph for the confusion matrix**



היסטוגרמה בה ציר ה- X מייצג את תשעת הנושאים על פי הסדר במטריצה M וציר ה- Y מייצג את כמות המאמרים אשר סווגו לכל נושא על פי המודל שלנו.



5. **Accuracy**

מידת הדיוק במודל שלנו הינו מספר המסמל את יחס בין כמות הסיווגים הנכונים אותם מצא המודל שלנו לבין כלל הסיווגים שמצאנו:

$$accuracy = \frac{\text{num of our correct assignments}}{\text{total num of our assignments}} = \frac{1306}{2124} \approx 0.615 = 61.5\%$$

6. **More hyperparameters, and some parameters**

- **vocabulary size after filtering** –

גודל המילון לאחר הפלטור (הסרת כל מילה המופיעה שלוש פעמים ומטה) אותו קיבלנו הינו – 6800.

- λ –

לאחר ניסוי ובחירה של ערכים שונים, בחרנו ב- λ=0.023 עבורו קיבלנו תוצאות טובות. בשלב ה- M אנחנו מבצעים החלקה למילים ע"י אלגוריתם lidstone smoothing, כיוון שיש לנו יותר מילים במילון (300,000) מאשר בקובץ ממנו אנו לומדים (6800 – לאחר פלטור), לכן, כאשר אנו מחפשים את ההסתברות למילה בנושא מסוים $P(w|x)$, עלינו למנוע הסתברות השווה לאפס עבור מילה (הרי תמיד קיימת הסתברות כלשהי עבור מילה מסוימת), ו- lidstone נותן למילה זו הסתברות של λ.

- k –

במהלך החישובים הקורים בריצת האלגוריתם אנו עלולים להגיע ל- underflow, על מנת לטפל בכך אנו עושים מספר דברים, בניהם חישוב על ידי לוגים והפיכת מכפלות לסכימה, הסרת החזקה הגבוהה ביותר ועוד.
בין היתר אנו משתמשים ב- k על מנת לנרמל את הערכים, דבר הגורם לערכים להיות פחות שליליים – כלומר, פחות רחוקים מהאפס, ובכך מסייע למניעת ה- underflow.