תרגיל רשות (ללא ציון) – Clustering

אשכולות מסמכים, או אשכולות טקסטים, הוא יישום פופולרי מאוד של אלגוריתמי אשכולות (clustering). מנוע חיפוש באינטרנט, כמו גוגל, מחזיר לעיתים קרובות אלפי תוצאות לשאילתה פשוטה. לדוגמא, אם תקליד את מונח החיפוש "יגואר" בגוגל, סביב 200 מיליון תוצאות יוחזרו. זה מקשה מאוד על גלישה או חיפוש מידע רלוונטי, במיוחד אם למונח החיפוש יש משמעויות מרובות. אם אנו מחפשים "יגואר", אנו עשויים לחפש מידע על החיה, המכונית או קבוצת הכדורגל Jacksonville Jaguars.

ניתן להשתמש בשיטות אשכולות כדי לקבץ תוצאות חיפוש באופן אוטומטי לקטגוריות, וכך יהיה קל יותר למצוא תוצאות רלוונטיות. שיטה זו משמשת במנועי החיפוש PolyMeta ו- Helioid, וכן ב- FirstGov.gov, פורטל האינטרנט הרשמי של ממשלת ארה"ב. שני האלגוריתמים הנפוצים ביותר המשמשים לאשכול מסמכים הם Hierarchical ו- k-means.

בבעיה זו אנו נקבץ מאמרים המתפרסמים בDaily Kos , בלוג פוליטי אמריקאי המפרסם מאמרי חדשות ודעה שנכתבו מנקודת מבט מתקדמת. דיילי קוס הוקם על ידי מרקוס מוליצאס בשנת 2002, ונכון לספטמבר 2014, באתר הייתה תנועת ימי חול בממוצע של מאות אלפי ביקורים.

הקובץ dailykos (CSV - 10.1MB) מכיל נתונים על 3,430 כתבות חדשות או בלוגים שהתפרסמו ב- Daily Kos. מאמרים אלה פורסמו בשנת 2004, והובילו לבחירות לנשיאות ארצות הברית. המועמדים המובילים היו הנשיא המכהן ג'ורג 'וו. בוש (רפובליקני) וג'ון קרי (דמוקרטי). מדיניות חוץ הייתה נושא דומיננטי בבחירות, ובמיוחד הפלישה לעירק בשנת 2003.

כל אחד מהמשתנים במערך הנתונים הוא מילה שהופיעה בלפחות 50 מאמרים שונים (1,545 מילים בסך הכל). מערך המילים גוזם על פי חלק מהטכניקות הקיימות בניתוח טקסטים (פיסוק הוסר והוסרו מילות עצירה). עבור כל מסמך, ערכי המשתנה הם מספר הפעמים שהמילה הופיעה במסמך.

משימות:

א. עליכם לחלק את הדאטה לאשכולות על פי שני אלגוריתמים – האלגוריתם ההיררכי (AgglomerativeClustering), ואלגוריתם k-means. **שימו לב, כבר ממומשים בספריה sklearn ולכן אין צורך לממשם בצורה עצמאית.**

עבור k-means:

ב. חלקו את הדאטה ל2,3,7,8 אשכולות והציגו את איכות החלוקה על פי הקריטריונים שנלמדו בהרצאה בכיתה (SSE,silhouette).

ג. בחרו את מספר האשכולות האופטימלי מבחינתכם, הסבירו למה בחרתם ערך זה

ד. לאיזה אשכול יש את המספר המקסימלי של דוגמאות ולאיזה יש את המספר המינימלי (מהו המספר)?

ה. צרו קובץ CSV המכיל את הנתונים השייכים לכל אשכול בנפרד

ו. הסתכלו על הקבצים השונים ונסו להבין לאיזה נושא כל אחד מתייחס, תעדו בתשובה מה הערכים שגרמו לכם להגיע למסקנה

עבור האלגוריתם ההיררכי:

ז., בצעו סיווג על פי ארבעת ערכי הlinkage בספרייה. השוו את הדנדוגרמים שאתם מקבלים.

עליכם להגיש דוח מדעי המסכם את כל המשימות, התוצאות וההסברים. בנוסף לקובץ הקוד המשחזר את **כל** העבודה שעשיתם לאורך התרגיל. ממליץ לצרף את הפרויקט לגיט שלכם על מנת לעבות את תיק העבודות שאתם מציגים בו.