

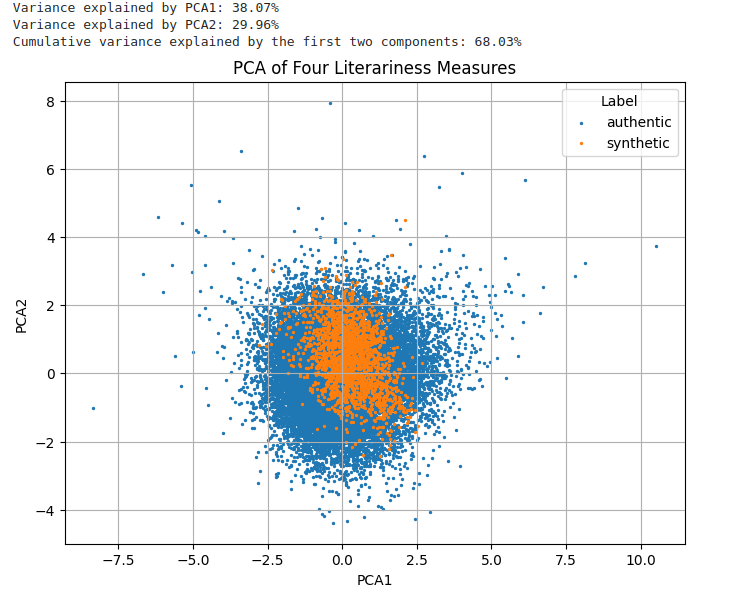
List of texts mislabeled in above K-means clustering below.

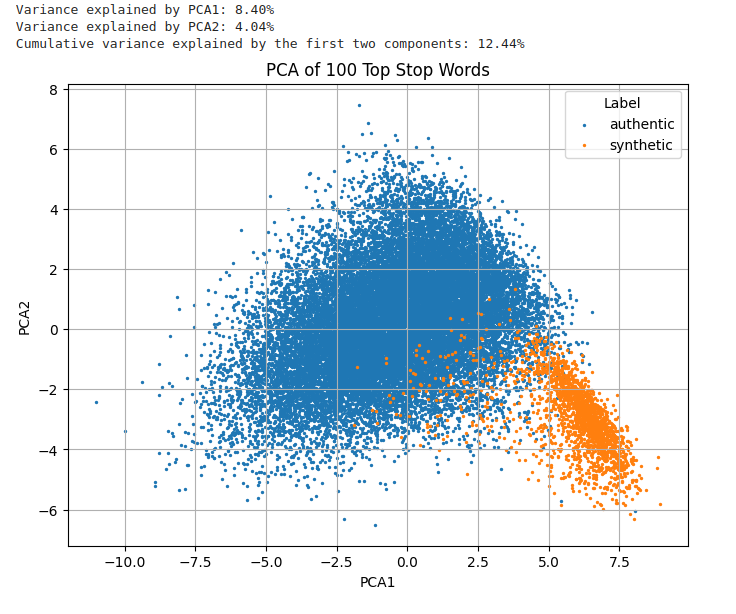
Synthetic labeled authentic (62 Twains, 8 Austens):

['AUSTEN\_synthetic\_combined\_156.txt', 'AUSTEN\_synthetic\_combined\_157.txt', 'AUSTEN\_synthetic\_combined\_158.txt', 'AUSTEN\_synthetic\_combined\_160.txt', 'AUSTEN\_synthetic\_combined\_165.txt', 'AUSTEN\_synthetic\_combined\_166.txt', 'AUSTEN\_synthetic\_combined\_177.txt', 'AUSTEN\_synthetic\_combined\_180.txt', 'TWAIN\_synthetic\_combined\_10.txt', 'TWAIN\_synthetic\_combined\_100.txt', 'TWAIN\_synthetic\_combined\_101.txt', 'TWAIN\_synthetic\_combined\_104.txt', 'TWAIN\_synthetic\_combined\_105.txt', 'TWAIN\_synthetic\_combined\_108.txt', 'TWAIN\_synthetic\_combined\_11.txt', 'TWAIN\_synthetic\_combined\_113.txt', 'TWAIN\_synthetic\_combined\_116.txt', 'TWAIN\_synthetic\_combined\_12.txt', 'TWAIN\_synthetic\_combined\_121.txt', 'TWAIN\_synthetic\_combined\_122.txt', 'TWAIN\_synthetic\_combined\_123.txt', 'TWAIN\_synthetic\_combined\_127.txt', 'TWAIN\_synthetic\_combined\_15.txt', 'TWAIN\_synthetic\_combined\_16.txt', 'TWAIN\_synthetic\_combined\_163.txt', 'TWAIN\_synthetic\_combined\_165.txt', 'TWAIN\_synthetic\_combined\_167.txt', 'TWAIN\_synthetic\_combined\_169.txt', 'TWAIN\_synthetic\_combined\_17.txt', 'TWAIN\_synthetic\_combined\_171.txt', 'TWAIN\_synthetic\_combined\_174.txt', 'TWAIN\_synthetic\_combined\_177.txt', 'TWAIN\_synthetic\_combined\_179.txt', 'TWAIN\_synthetic\_combined\_182.txt', 'TWAIN\_synthetic\_combined\_184.txt', 'TWAIN\_synthetic\_combined\_185.txt', 'TWAIN\_synthetic\_combined\_189.txt', 'TWAIN\_synthetic\_combined\_19.txt', 'TWAIN\_synthetic\_combined\_20.txt', 'TWAIN\_synthetic\_combined\_23.txt', 'TWAIN\_synthetic\_combined\_25.txt', 'TWAIN\_synthetic\_combined\_27.txt', 'TWAIN\_synthetic\_combined\_31.txt', 'TWAIN\_synthetic\_combined\_33.txt', 'TWAIN\_synthetic\_combined\_34.txt', 'TWAIN\_synthetic\_combined\_36.txt', 'TWAIN\_synthetic\_combined\_39.txt', 'TWAIN\_synthetic\_combined\_53.txt', 'TWAIN\_synthetic\_combined\_54.txt', 'TWAIN\_synthetic\_combined\_55.txt', 'TWAIN\_synthetic\_combined\_6.txt', 'TWAIN\_synthetic\_combined\_65.txt', 'TWAIN\_synthetic\_combined\_66.txt', 'TWAIN\_synthetic\_combined\_67.txt', 'TWAIN\_synthetic\_combined\_68.txt', 'TWAIN\_synthetic\_combined\_71.txt', 'TWAIN\_synthetic\_combined\_72.txt', 'TWAIN\_synthetic\_combined\_73.txt', 'TWAIN\_synthetic\_combined\_74.txt', 'TWAIN\_synthetic\_combined\_75.txt', 'TWAIN\_synthetic\_combined\_80.txt', 'TWAIN\_synthetic\_combined\_84.txt', 'TWAIN\_synthetic\_combined\_85.txt', 'TWAIN\_synthetic\_combined\_86.txt', 'TWAIN\_synthetic\_combined\_88.txt', 'TWAIN\_synthetic\_combined\_9.txt', 'TWAIN\_synthetic\_combined\_92.txt', 'TWAIN\_synthetic\_combined\_94.txt', 'TWAIN\_synthetic\_combined\_95.txt', 'TWAIN\_synthetic\_combined\_99.txt']

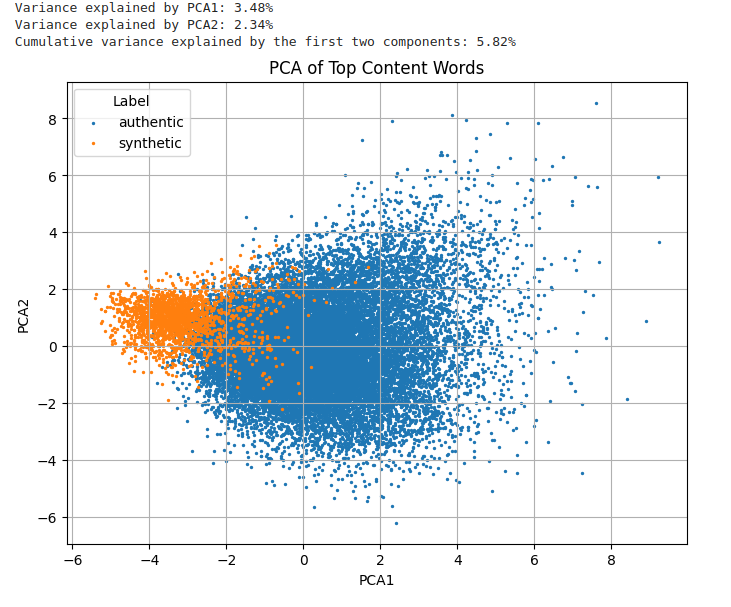
Authentic labeled synthetic (17 Hopkins, 11 Griggs, 3 Chesnutt, 2 Twain, 1 each for Austen, Bronte, Alcott):

['Alcott\_2029.txt', 'Austen\_1080.txt', 'Bronte\_1374.txt', 'Chesnutt\_166.txt', 'Chesnutt\_312.txt', 'Chesnutt\_634.txt', 'Griggs\_2.txt', 'Griggs\_213.txt', 'Griggs\_258.txt', 'Griggs\_309.txt', 'Griggs\_310.txt', 'Griggs\_393.txt', 'Griggs\_395.txt', 'Griggs\_401.txt', 'Griggs\_428.txt', 'Griggs\_92.txt', 'Griggs\_96.txt', 'Hopkins\_1.txt', 'Hopkins\_108.txt', 'Hopkins\_117.txt', 'Hopkins\_118.txt', 'Hopkins\_157.txt', 'Hopkins\_162.txt', 'Hopkins\_177.txt', 'Hopkins\_186.txt', 'Hopkins\_188.txt', 'Hopkins\_189.txt', 'Hopkins\_194.txt', 'Hopkins\_2.txt', 'Hopkins\_230.txt', 'Hopkins\_61.txt', 'Hopkins\_80.txt', 'Hopkins\_81.txt', 'Hopkins\_82.txt', 'Twain\_2189.txt', 'Twain\_2273.txt']

The four literariness measures are mean sentence length, lexical density, type-token ratio and concreteness.



The top stop words are the relative frequency for each chunk of the 100 most common stop words in the corpus.



The top content words are the relative frequency for each chunk of the 100 most common non-stop words in the corpus.