** SCRIPT START in TRAIN mode **

READING IN DATA

READING EXCEL DATA
XLS_DF:

	TITLE	AUTHORS	DOI	...	CAT_5_1	CAT_6	CAT_7
0	The English language in the Philippines	Yule, E.S	10.2307/452557	...	1.0	0.0	0.0
1	Creole and "Cajan"	William A. Read	https://www.jstor.org/stable/452755	...	0.0	0.0	0.0
2	Aesop in Negro Dialect	Addison Hibbard	10.2307/452758	...	0.0	0.0	0.0
3	The Chinook Jargon	Douglas Leechman	10.2307/452146	...	0.0	0.0	0.0
4	Indian Songs and English Verse	Hartley Alexander	10.2307/452122	...	0.0	0.0	0.0
..	...	...	...	...	...	...	...
268	Continuity versus English Influence in the Wes...	Thomas B. Klein	10.1215/00031283-4201998	...	0.0	0.0	0.0
269	Unstressed been: Past and Present in African A...	Arthur K. Spears	10.1215/00031283-4202009	...	0.0	0.0	0.0
270	“Put the Groceries Up”: Comparing Black and Wh...	Martha Austen	10.1215/00031283-43120064	...	0.0	0.0	0.0
271	“My Presiden(t) and Firs(t) Lady Were Black”:S...	Nicole Holliday	10.1215/00031283-6903954	...	0.0	0.0	0.0
272	“You’re So Not Going to Believe This”:The Use ...	Ulrike Stange	10.1215/00031283-4395168	...	0.0	0.0	0.0

[273 rows x 25 columns]

> XLS_ROWS PREVIEW:

>> {'TITLE': 'The English language in the Philippines', 'AUTHORS': 'Yule, E.S', 'DOI': '10.2307/452557', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': 'Y', 'MACHINE_READABLE': 'Y', 'PDF_NAME': '10-2307_452557', 'VENUE': 'American Speech', 'YEAR': 1925, 'MONTH': 'NOV', 'VOLUME': 1, 'ISSUE': 2, 'UNKNOWN': 0.0, 'CAT_0': 0.0, 'CAT_1': 0.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': 0.0, 'CAT_4': 0.0, 'CAT_4_1': 0.0, 'CAT_5': 0.0, 'CAT_5_1': 1.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> {'TITLE': 'Creole and "Cajan"', 'AUTHORS': 'William A. Read', 'DOI': 'https://www.jstor.org/stable/452755', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': '?', 'MACHINE_READABLE': '?', 'PDF_NAME': nan, 'VENUE': 'American Speech', 'YEAR': 1926, 'MONTH': 'JUN', 'VOLUME': 1, 'ISSUE': 9, 'UNKNOWN': 1.0, 'CAT_0': 0.0, 'CAT_1': 0.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': 0.0, 'CAT_4': 0.0, 'CAT_4_1': 0.0, 'CAT_5': 0.0, 'CAT_5_1': 0.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> ..

>> {'TITLE': '“My Presiden(t) and Firs(t) Lady Were Black”:Style, Context, and Coronal Stop Deletion in the Speech of Barack and Michelle Obama', 'AUTHORS': 'Nicole Holliday', 'DOI': '10.1215/00031283-6903954', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': nan, 'MACHINE_READABLE': nan, 'PDF_NAME': '10-1215_00031283-6903954', 'VENUE': 'American Speech', 'YEAR': 2017, 'MONTH': 'Winter', 'VOLUME': 92, 'ISSUE': 4, 'UNKNOWN': 0.0, 'CAT_0': 0.0, 'CAT_1': 1.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': 0.0, 'CAT_4': 0.0, 'CAT_4_1': 0.0, 'CAT_5': 0.0, 'CAT_5_1': 0.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> {'TITLE': '“You’re So Not Going to Believe This”:The Use of Genx so in Constructions with Future going to in American English', 'AUTHORS': 'Ulrike Stange', 'DOI': '10.1215/00031283-4395168', 'INFO_COMPLETE': 'Y', 'PDF_DOWNLOADED': 'Y', 'MACHINE_READABLE': 'Y', 'PDF_NAME': '10-1215_00031283-4395168', 'VENUE': 'American Speech', 'YEAR': 2017, 'MONTH': 'Winter', 'VOLUME': 92, 'ISSUE': 4, 'UNKNOWN': 0.0, 'CAT_0': 0.0, 'CAT_1': 1.0, 'CAT_2': 0.0, 'CAT_2_1': 0.0, 'CAT_3': 0.0, 'CAT_3_1': 0.0, 'CAT_4': 0.0, 'CAT_4_1': 0.0, 'CAT_5': 0.0, 'CAT_5_1': 0.0, 'CAT_6': 0.0, 'CAT_7': 0.0}

>> found and read XLSX file (american-speech-dataset-complete-rows.xlsx - 273 rows)

READING PDF DATA
>> found 390 PDF files -- 1728798 total words (avg: 4433)

> PDF TEXT PREVIEW (after preprocessing):
everybody likes pizza, doesn't he or she?
author(s): george jochnowitz
source: american speech , autumn, 1982 , vol. 57, no. 3 (autumn, 1982), pp. 198-203
published by: duke university press
stable url: https://www.jstor.org/stable/454860

 everybody likes pizza, doesn't he or she?

george jochnowitz
college of staten island

w hile reading margaret mitchell's gone with the wind, my younger daughter found what she called a grammatical error in the text of the 1940 edition: "everyone was very polite and kind to her because he felt sorry for her.. ." (p. 36). when i decided to write an essay arguing that they is grammatical with indefinite antecedents, i checked the text of the original (1936) edition. this time, the passage read "everyone was very polite and kind to her because they felt sorry for her..." (p. 94). apparently between 1936 and 1940 some anonymous copy editor decided to correct mitchell's prose, and in doing so created the jarring phrase that had offended my daughter.

wha ..

CONSTRUCTING DOCUMENT OBJECTS
>> Completed making 195 document objects - two quick preview documents:

PRINT_OUT for document <10-2307_454860.pdf>, titled <Everybody Likes Pizza, Doesn't He or She?> (authors: George Jochnowitz)
DOI 10.2307/454860 | YEAR 1982 | MONTH Autumn | VOLUME 57 | ISSUE 3
CATEGORY LIST: ['CAT_7']

TEXT: everybody likes pizza, doesn't he or she?
author(s): george jochnowitz
source: american speech , autumn, 1982 , vol. 57, no. 3 (autumn, 1982), pp. 198-203
published by: duke university press
s ..

X_VECTOR: []
Y_GOLDS: []
Y_PREDS: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1]

PRINT_OUT for document <10-2307_455246.pdf>, titled <Grammar of the English-Based Jamaican Proverb> (authors: David Lawton)
DOI 10.2307/455246 | YEAR 1984 | MONTH Summer | VOLUME 59 | ISSUE 2
CATEGORY LIST: ['CAT_2_1']

TEXT: grammar of the english-based jamaican proverb
author(s): david lawton
source: american speech , summer, 1984 , vol. 59, no. 2 (summer, 1984), pp. 123-130
published by: duke university press
st ..

X_VECTOR: []
Y_GOLDS: []
Y_PREDS: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1]

MAKING and FIT-TRANSFORMING TF-IDF VECTORISER
>> X_RAW has been made (not displaying all texts, as that would be too long)

>> X_ALL is currently:
  (0, 42684)   0.014000053961212856
  (0, 68344)   0.009121181920301901
  (0, 104144)   0.004195391920941417
  (0, 23815)   0.014823380308575626
  (0, 52810)   0.011377692772811882
  (0, 48367)   0.013361432431983296
  (0, 95531)   0.012839640649404211
  (0, 83803)   0.005768036024719076 4
  (0, 96989)   0.006659344935963278
  (0, 96081)   0.003320186242980694 3
  (0, 111314)   0.010414732324128524
  (0, 97691)   0.005977028949321841
  (0, 45265)   0.00620249441490 35
  (0, 52751)   0.006829410687044194
  (0, 68558)   0.008993964026606368
  (0, 67705)   0.009695487678424151
  (0, 19463)   0.011679227337595565
  (0, 94297)   0.011679227337595565
  (0, 61226)   0.012839640649404211
  (0, 74834)   0.014823380308575626
  (0, 98432)   0.009254319012319 88
  (0, 96867)   0.011377692772811882
  (0, 18081)   0.014823380308575626
  (0, 96123)   0.010414732324128524
  (0, 80267)   0.014823380308575626
  :   :
  (9, 70125)   0.03445450989388633
  (9, 17067)   0.0018226709568640944
  (9, 14264)   0.05077506721204301
  (9, 71847)   0.021872051482369134
  (9, 55254)   0.029311909541367066
  (9, 111695)   0.0018226709568640944
  (9, 47966)   0.020049380525505 04
  (9, 104558)   0.0018319943463354416
  (9, 88689)   0.003626790515145929 3
  (9, 76825)   0.003626790515145929 3
  (9, 104128)   0.005440185772718894
  (9, 33283)   0.005468012870592 2836
  (9, 21591)   0.043521486181751 1154
  (9, 78022)   0.005495983039006 3245
  (9, 76155)   0.014655954770683533
  (9, 65245)   0.027200928863594 47
  (9, 106383)   0.0018133952575729647
  (9, 88108)   0.012693766803010752
  (9, 9834)   0.021760743090875577
  (9, 87338)   0.0018133952575729647
  (9, 16915)   0.0018226709568640944
  (9, 84514)   0.007155313298889479
  (9, 71190)   0.039894695666605 22
  (9, 45871)   0.085438610635 2638
  (9, 58096)   0.006356904109438025 ..

ENCODING LABELS with MLB
>> Y_RAW is currently:
[['CAT_7'], ['CAT_1'], ['CAT_2_1'], ['CAT_1', 'CAT_2'], ['CAT_1', 'CAT_2'], ['CAT_4'], ['CAT_2_1'], ['CAT_4'], ['CAT_3_1'], ['CAT_1']] ..

>> Y_ALL is currently:
[[0 0 0 0 0 0 0 0 0 0 0 0 1]
 [0 0 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 0 0 0 0]
 [0 0 1 1 0 0 0 0 0 0 0 0 0]
 [0 0 1 1 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 0 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 0 0 0 0 0]
 [0 0 0 0 0 0 1 0 0 0 0 0 0]
 [0 0 1 0 0 0 0 0 0 0 0 0 0]] ..

SPLITTING DATA (TRAIN-TEST SPLIT)
>> After splitting, there are 156 training documents/rows and 39 test documents/rows -- 195 documents/rows total

TRAINING CLASSIFIER
X_TRAIN is:

  (0, 107735)   0.006052120147080106
  (0, 60093)   0.006052120147080106
  (0, 56516)   0.006052120147080106
  (0, 76815)   0.006052120147080106
  (0, 86830)   0.006052120147080106
  (0, 26190)   0.006052120147080106
  (0, 66389)   0.006052120147080106
  (0, 109904)   0.006052120147080106
  (0, 66497)   0.006052120147080106
  (0, 75492)   0.006052120147080106
  (0, 95167)   0.006052120147080106
  (0, 64994)   0.006052120147080106
  (0, 91820)   0.006052120147080106
  (0, 42133)   0.006052120147080106
  (0, 44591)   0.006052120147080106
  (0, 14427)   0.006052120147080106
  (0, 7930)   0.006052120147080106
  (0, 86306)   0.006052120147080106
  (0, 10240)   0.006052120147080106
  (0, 53279)   0.006052120147080106
  (0, 81471)   0.006052120147080106
  (0, 78383)   0.006052120147080106
  (0, 95692)   0.006052120147080106
  (0, 72900)   0.006052120147080106
  (0, 97715)   0.006052120147080106
  :   :
  (155, 39583) 0.08652558102774713
  (155, 70125) 0.03580368870113674 6
  (155, 17067) 0.005997804695514129
  (155, 14264) 0.07160737740227349
  (155, 71847) 0.0299890234775706 46
  (155, 55254) 0.042199394113666924
  (155, 111695) 0.0029989023477570644
  (155, 47966) 0.026990121129813 58
  (155, 104558) 0.003014242436690495
  (155, 88689) 0.005967281450189457

```
  (155, 76825)  0.005967281450189457
  (155, 104128) 0.011934562900378915
  (155, 33283)  0.005997804695514129
  (155, 21591)  0.05370553305170512
  (155, 78022)  0.006028484873380 99
  (155, 76155)  0.003014242436690495
  (155, 65245)  0.005967281450189457
  (155, 5678)   0.011314490896863184
  (155, 106383) 0.002983640725 09047287
  (155, 88108)  0.011934562900378915
  (155, 9834)   0.011934562900378915
  (155, 87338)  0.002983640725 09047287
  (155, 16915)  0.002998902347757 0644
  (155, 71190)  0.020885485075663 1
  (155, 45871)  0.003123889867453084

Y_TRAIN is:

[[0 0 0 ... 1 0 0]
 [0 0 0 ... 0 0 1]
 [0 0 0 ... 0 0 1]
 ...
 [0 0 1 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 [0 0 0 ... 0 1 0]]


/home/michaelinwords/.local/lib/python3.8/site-packages/sklearn/multiclass.py:84: UserWarning: Label not 0 is present in all training examples.
  warnings.warn(
/home/michaelinwords/.local/lib/python3.8/site-packages/sklearn/multiclass.py:84: UserWarning: Label not 1 is present in all training examples.
  warnings.warn(

 REVIEWING PERFORMANCE
-- This training was on 156 documents (out of 195 total documents)

SUBSET ACCURACY: 0.00%
CLASSIFICATION REPORT:
              precision    recall  f1-score   support

     UNKNOWN       1.00      0.00      0.00         1
       CAT_0       1.00      1.00      1.00         0
       CAT_1       1.00      0.00      0.00        18
       CAT_2       1.00      0.00      0.00         5
     CAT_2_1       1.00      0.00      0.00         3
       CAT_3       1.00      0.00      0.00         2
     CAT_3_1       1.00      1.00      1.00         0
       CAT_4       1.00      0.00      0.00         3
     CAT_4_1       1.00      1.00      1.00         0
       CAT_5       1.00      0.00      0.00         2
     CAT_5_1       1.00      0.00      0.00         6
       CAT_6       1.00      0.00      0.00         3
       CAT_7       1.00      0.00      0.00         1

   micro avg       1.00      0.00      0.00        44
   macro avg       1.00      0.23      0.23        44
weighted avg       1.00      0.00      0.00        44
 samples avg       1.00      0.00      0.00        44

per_label_accuracy: 0.91
```

SAVING MODEL
>> Successfully saved classifier as model.joblib and vectoriser as vectoriser.joblib

SCRIPT (train mode) COMPLETED SUCCESSFULLY ✅