from pyspark.sql import SparkSession

from pyspark.sql.functions import col, udf, when

from pyspark.sql.types import DoubleType

from fuzzywuzzy import fuzz

# ✅ Initialize Spark Session

spark = SparkSession.builder.appName("FuzzyMatching").getOrCreate()

# ✅ Fuzzy Matching Functions

def fuzz\_ratio(str1, str2):

return fuzz.ratio(str1, str2) / 100.0 if str1 and str2 else 0.0

def fuzz\_partial\_ratio(str1, str2):

return fuzz.partial\_ratio(str1, str2) / 100.0 if str1 and str2 else 0.0

def fuzz\_token\_sort\_ratio(str1, str2):

return fuzz.token\_sort\_ratio(str1, str2) / 100.0 if str1 and str2 else 0.0

def fuzz\_token\_set\_ratio(str1, str2):

return fuzz.token\_set\_ratio(str1, str2) / 100.0 if str1 and str2 else 0.0

# ✅ Register UDFs

fuzz\_ratio\_udf = udf(fuzz\_ratio, DoubleType())

fuzz\_partial\_ratio\_udf = udf(fuzz\_partial\_ratio, DoubleType())

fuzz\_token\_sort\_ratio\_udf = udf(fuzz\_token\_sort\_ratio, DoubleType())

fuzz\_token\_set\_ratio\_udf = udf(fuzz\_token\_set\_ratio, DoubleType())

# ✅ Read CSV File (Assumes `Name\_Class` & `Variant\_Class` are already classified)

df = spark.read.csv("your\_file.csv", header=True, inferSchema=True)

# ✅ Apply All Fuzzy Matching in a \*\*Single Step\*\*

df = df.withColumns({

"Fuzz\_Ratio": when(

(col("Name\_Class").isin(["Alphanumeric", "Character Only"])) &

(col("Variant\_Class").isin(["Alphanumeric", "Character Only"])),

fuzz\_ratio\_udf(col("Name"), col("Name\_Variant"))

).otherwise(0.0),

"Fuzz\_Partial\_Ratio": when(

(col("Name\_Class").isin(["Alphanumeric", "Character Only"])) &

(col("Variant\_Class").isin(["Alphanumeric", "Character Only"])),

fuzz\_partial\_ratio\_udf(col("Name"), col("Name\_Variant"))

).otherwise(0.0),

"Fuzz\_Token\_Sort\_Ratio": when(

(col("Name\_Class").isin(["Alphanumeric", "Character Only"])) &

(col("Variant\_Class").isin(["Alphanumeric", "Character Only"])),

fuzz\_token\_sort\_ratio\_udf(col("Name"), col("Name\_Variant"))

).otherwise(0.0),

"Fuzz\_Token\_Set\_Ratio": when(

(col("Name\_Class").isin(["Alphanumeric", "Character Only"])) &

(col("Variant\_Class").isin(["Alphanumeric", "Character Only"])),

fuzz\_token\_set\_ratio\_udf(col("Name"), col("Name\_Variant"))

).otherwise(0.0)

})

# ✅ Show Final Results

df.select("Name", "Name\_Variant", "Name\_Class", "Variant\_Class",

"Fuzz\_Ratio", "Fuzz\_Partial\_Ratio",

"Fuzz\_Token\_Sort\_Ratio", "Fuzz\_Token\_Set\_Ratio").show(truncate=False)