North Carolina Agricultural and Technical State University
College of Engineering
CSE 817 Fundamentals of Big Data Analysis
Spring 2026
Assignement-2 (100 Points)

(Word Count, SQL, Classification, Clustering, Tokenization/Feature Vector)

Deadline: March 10th, 2026 (11:59 pm)

Midterm Timeline: March 12th

1. Multi-Class Classification:

For the multiclass classification problem, there were six different datasets. Some of the datasets contain missing values. For example, TrainData1, TestData1 and TrainData3 contain some missing values. Therefore, the first approach needs to handle the missing values for selecting the features. Then compare the accuracy on train dataset to find out which classifier gives best result for each dataset with cross validation to verify the accuracy based on test dataset. Implementation needs to be done using PySpark with Google Colab.

Dataset

Dataset	Training (Features)	Training (Samples)	Testing (Features)	Testing (Samples)	No of Classes
Dataset 1	3312 (NAN Values)	150	3312	53	5

Hint:

- Missing Value Estimation
  - ✓ (KNN method for imputation of the missing values)
- Dimensionality Reduction
- Use Several Classifiers/ Ensemble Method
  - ✓ Logistic Regression (with different c values)
  - ✓ Random Forest (with different estimator values)
  - ✓ SVM (with different kernels)
  - ✓ KNN (with k = 1,2,5,10,20)
  - ✓ K (3,5,10) Fold Cross Validation
- Performance Comparison
  - ✓ Classification Accuracy, Precision, Recall (Sensitivity), Specificity, F1 Score
  - ✓ AUC, ROC Curve

2. Word Count MapReduce:

For this problem, the first step of is to create a Spark context. Next, you will need to read the target file into an RDD. You now have an RDD filled with strings, one per line of the file. Next you will want to split the lines into individual words

Next, you will want to replace each word with a tuple of that word and the number 1. Now, to get a count of the number of instances of each word, you need only group the elements of the RDD by key (word) and add up their values. Finally, you can show the results and stop the context

- Follow the above steps to write a code in Google Colab using PySpark for this word counting problem with MapReduce operation based on **wc.txt** dataset to find the results.

3. Clustering:

The examined group comprised kernels belonging to three different varieties of **wheat: Kama, Rosa and Canadian**, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The data set can be used for the tasks of classification and cluster analysis.

Attribute Information:

To construct the data, seven geometric parameters of wheat kernels were measured:
1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

All of these parameters were real-valued continuous.

- **K-means clustering** using PySpark in Google Colab based on the **seed.txt** dataset
- Apply **Gaussian Mixture Model** with the same **seed.txt** dataset

4. Spark SQL Problem:

Perform the following operations in Google Colab using PySPark using the **mcar.txt** dataset

a) Create and display Spark DataFrames
b) Filter the DataFrame to only retain rows with mpg less than 18
c) Compute the average weight of cars by their cylinders using group by or aggregation
d) Select gear of the car for the cylinder values greater or equal to 4 and less or equal than 9

5. NLP Feature Extraction:

a) Apply **HashingTF and IDF** in Google Colab using PySPark using the **shakespeare.txt** dataset. *(calculate DF, IDF, TF-IDF, search for specific keyword in the document)*

b) Apply **Word2Vec** in Google Colab using PySPark using the **shakespeare.txt** *(get word vectors and find similarities)* dataset.

6. Sentiment Analysis Twitter Dataset:

This problem is for sentiment analysis of Tweets related to the Covid-19 pandemic, which is a multi-label text classification task. Since the outbreak of coronavirus, it has affected more than 180 countries where massive losses in the economy and jobs globally and confining about 58% of the global population are caused. The research on people's feelings is essential for keeping mental health and informed about Covid-19.

The training data contains 5000 labeled tweets while the validation data have 2500 pieces of unlabeled tweets. The training data have 3 columns, containing Tweet ID, Tweet text, and labels. Note that the orders are shown as Optimistic (0), Thankful (1), Empathetic (2), Pessimistic (3), Anxious (4), Sad (5), Annoyed (6), Denial (7), Surprise (8), Official report (9), Joking (10). For example, if the labels are 3 and 6, it means that this piece of the tweet is labeled as Pessimistic and Annoyed. The prediction needs to be done on the validation dataset (test data will be considered to justify your work).

- Perform Covid-19 sentiment analysis using Twitter dataset to classify Covid-19 spread sentiments based on the **training.csv** and **validation.csv** file.