CSE 4/587 - Midterm Exam

10/19/22 @ 9AM, NSC 201

Name: ______________________________ Person #: ________________________

UBIT: ______________________________ Seat #: ________________________

Academic Integrity

My signature on this cover sheet indicates that I agree to abide by the academic integrity policies of this course, the department, and university, and that this exam is my own work.

Signature: _____________________________________ Date: ___________________________

Instructions

1. This exam contains 9 total pages (including this cover sheet). Be sure you have all the pages before you begin.
2. Clearly write your name, UBIT name, person number, and seat number above. **Additionally, write your UBIT name at the top of every page now.**
3. You have 1 hour and 20 minutes to complete this exam. Show all work where appropriate, but keep your answers concise and to the point.
4. After completing the exam, sign the academic integrity statement above. Be prepared to present your UB card upon submission of the exam paper.
5. You must turn in all of your work. No part of this exam booklet may leave the classroom.

DO NOT WRITE BELOW

Q1	Q2	Q3	Q4	Q5	Total
20	20	20	20	20	100

Question 1 - Modeling and Algorithms [20 Points]

a. How can we help ensure that our supervised learning models are not overfit? [4 points]

[4 points] Split data into training and test. Fit against training, compare against test. Full credit for anything related to the above answer, or other answers that are correct. Partial (2 points) for mentioning training set but not saying how it is used.

b. In linear regression, what does the error term, ε, capture? What about R^2? [4 points]

[2 points] E captures [variance, noise, effects from features not included in our model]

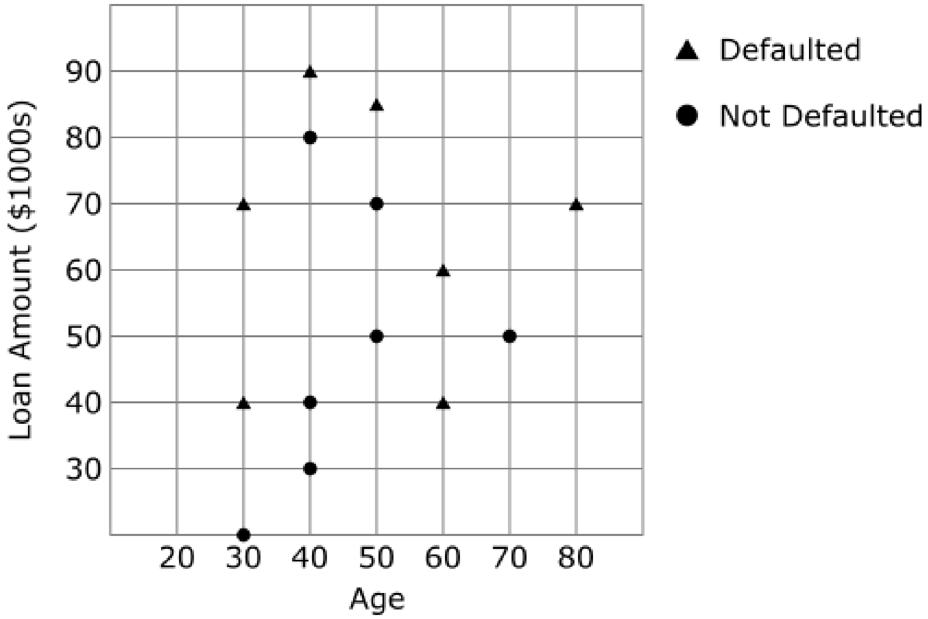
[2 points] R2 captures [the effectiveness/accuracy of our model, the amount of the variance our model captures]

Give 1 point for either of the above if they give just a formula for how to calculate them but don't explain what they mean.

c. Name the four parameters we need to define in order to fully specify and evaluate a K-NN model for a given dataset. [4 points]

[1 point] Similarity / distance metric to define how similar / close two points are
[1 point] scaling done on the features
[1 point] evaluation metric (ie accuracy, precision, recall, etc)
[1 point] k

Below is a plot containing a number of data points with a known classification.
The x-axis represents a person's age, and the y-axis represents a loan amount in thousands of dollars. Each point is classified as ● or ▲. Points classified as ▲ represent people that have defaulted on their loan (did not pay it back). Points classified as ● represent people who have not.



d. If we assume euclidean distance determines the similarity of two points, does K-NN predict that a 40 year old client with a loan of $70k will default for k=3? What about for k=5? [4 points]

[2 points] k=3 predicts Not Default
[2 points] k=5 predicts Default

e. Explain what would happen if we give the loan amount in terms of dollars instead of thousands of dollars. Based on that explanation, what would our model predict for the same client from part (d) with k=3? [4 points]

[2 points] Explain that changing the scaling would result in the loan amount becoming the dominant feature in determining distance between two points (or some similar answer explaining that the scaling will change the results)

[2 points] With the rescaled data, k=3 now predicts Default

Question 2 - Hadoop and HDFS [20 Points]

a. List two major differences between Hadoop1.x and Hadoop2.x versions. [4 points]
[2 points] per difference listed

Possibilities:

- **Hadoop 1.0 had MR as resource manager/Hadoop 2.0 has YARN as a resource manager**
- **Hadoop 2.0 supports more than just MR**
- **Hadoop 2.0 supports larger clusters**
- **Hadoop 2.0 supports multiple name nodes**
- **Hadoop 2.0 supports windows**
- **Other correct answers possible**

b. How is an HDFS block replicated? Where are map and reduce tasks executed? [4 points]
[2 points] A block is replicated 3 times: one on local node, one on remote rack, one on different node of that same remote rack. (only award one point if they only specify replication factor and not where the replicas are placed)

[2 points] MR tasks are executed on the same node as the data they are operating on (only award one point if they specify that the tasks execute on data nodes but don't mention locality to the data)

c. In HDFS, what is a (i) heartbeat (ii) BlockReport? Explain. [4 points]
[2 points] Heartbeat is sent from DataNode to NameNode to inform that the datanode is still alive

[2 points] Block report sent from DN to NN with information on what blocks exist on that DN, replication factor, etc.

d. List two functions of a NameNode. List two functions of a DataNode. [4 points]
[1 point] per NN function up to a max of 2
[1 point] per DN function up to a max of 2

Acceptable NN functions: Manage namespace, manage metadata, manage filesystem, master node, manage edit log, handle client requests, any other correct function
Acceptable DN functions: Store blocks of data, give data to clients, report to NN with blockreport/heartbeat, execute MR tasks

e. What is the primary data type of the MapReduce model? Why are Maps able to run in parallel over the data? [4 points]
[2 points] <key, value> pairs
[2 points] Data is WORM/write once, read many, read only, etc.

Question 3 - MapReduce [20 Points]

SETI@home is a long-running project searching for extra-terrestrial life by analyzing radio frequency signals recorded by various telescopes. The radio signal intensity was scaled and "printed" (stored) as integers from 0-35 inclusive, with digits from 0-9, and a-z representing 10-35. We want to configure a Hadoop-MapReduce infrastructure to analyze this voluminous repository for any significant contact from extraterrestrials. Consider a Hadoop-MapReduce configuration as given below:.

- We use the word count algorithm from class. Here a "word" is a digit or character representing the SETI signal. Our reducer class is also used as our combiner class.
- Assume that the input has a total of **G = 40Tbyte data**. (1T = 10^{12} bytes, 1M= 10^{6} bytes)
- Input corpus is split equally into **S** sites, each running a MR cluster.
- Assume you plan to configure **M = 200** mappers per site. There are **R** reducers.

Answer the following questions for the configuration listed above:

a) What is the: [6 points]

(i) Input keyspace of the mappers?

[1 point] keyspace includes {0-9}

[1 point] keyspace includes {a-z}

[1 point] They only mention that there are 36 keys (or 35 if off by one)

(Max of 2 points)

(ii) Size of the input processed by each *site*?

[2 point] 40TB / S

(iii) Workload of each mapper in bytes?

[2 point] 40TB / S / 200 = 40 * 10^{12} bytes / S / 200 = 2 * 10^{11} / S bytes

(Still award full credit if they have the right expression but not in terms of bytes)

b) Assume that mappers suppress the range of values (0-15) and emit only the radio signals of values (16-35) inclusive. Assume that all combiners will run right before the shuffle and sort step. [6 points]

(i) What is the maximum number of <key,value> pairs that will be shuffled and sorted?

[2 points] 20 keys * 200 mappers * S sites = 4000 * S key value pairs

(ii) How many distinct keys will each reducer have to reduce?

[2 points] 20 / R

(iii) How many <key,value> pairs will be in the final output?

[2 points] 20

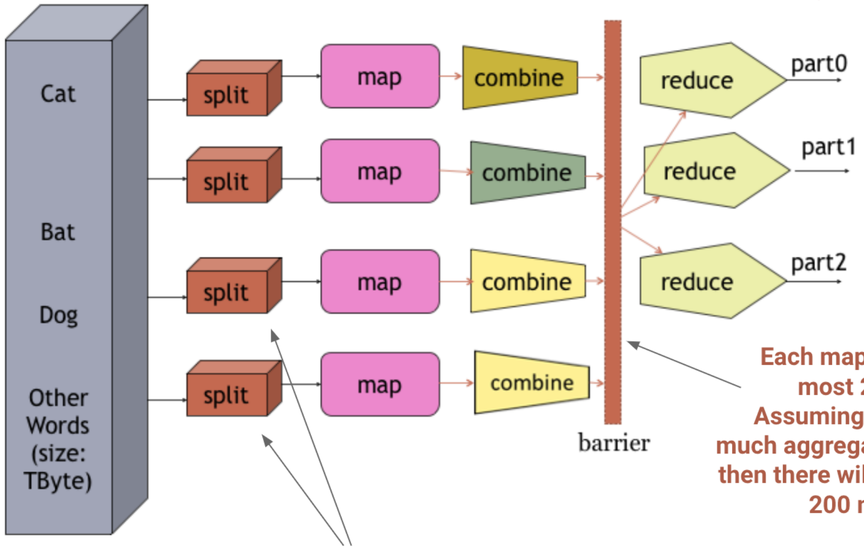
d) Draw a diagram that shows how the data flows in your MapReduce application, starting from input and resulting in output. Include mappers, reducers, and combiners. Label your diagram with the expressions you have derived above. [8 points]

Award points for each of the following (up to a max of 8):

- **[2] Shows the data split to multiple mappers**
- **[1] Shows that each mapper has a combiner**
- **[2] Barrier between mappers/combiners and reducers**
- **[1] Shows multiple reducers and num reducers independent from num mappers**
- **[1] point per correct label from parts b and c**

Input keyspace to mappers: {0-9}∪{a-z}

20 different keys / R reducers = 20 / R keys per reducer



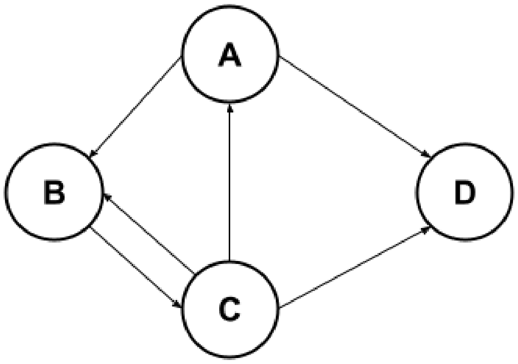
20 kv pairs in final output

Each mapper will output at most 20 different keys. Assuming combiners do as much aggregation as possible, then there will be 20 kv pairs x 200 mappers x S sites

(40TB / S) per site

40TB / S / 200 = 2×10^{11} / S bytes per mapper

Question 4 - GraphProcessing [20 Points]



a) Given the above graph, write down the adjacency matrix used to compute the PageRank of the graph. (Use the naive formulation without using teleportation) [4 points]

[2 points] Has the right entries in the matrix filled in

[2 points] Entries have the correct weights

Subtract one point for minor errors

b) State the initial condition $\boldsymbol{r}_0$ for power iteration. Perform 3 iterations of power iteration to find $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$. [4 points]

[1 point] Has the right initial condition

[1 point] for each iteration that is correct (still give credit if answer is correct given an incorrect adjacent matrix)

Subtract one point if most entries are correct but there are some minor mistakes

c) Will the power iteration solution for the above graph converge to what we want? Why or why not? If not, explain how to implement a fix. [6 points]

[2 points] No. (take a away a point if they claim it will not converge - it does converge)

[2 points] Recognize there is a dead end, D. (there is no spider trap)

[2 points] Fix is teleportation.

d) Describe at least 2 differences in the MapReduce implementation of PageRank. [6 points]

[3 points] Adjacency list instead of adjacency matrix (or graph must be split across many nodes/can't use matrix/represented in KV pairs, etc)

[3 points] Need a separate step to redistribute rank from spider traps and dead end nodes

[2 points] The computation is split over mappers and reducers to run in parallel.

Question 5 - Word Co-Occurrence [20 Points]

Computing word Co-Occurrence is an important problem in a number of different domains that involves counting the number of times one word appears in the same context as another. Sequentially, this can be accomplished by creating an ***N*** x ***N*** matrix ***M***, where ***N*** is the number of words in our vocabulary, and M_{ij} is the number of times that word w_i appears in the same context as word w_j.

a) Write pseudocode for a mapper and a reducer to compute word co-occurrence using the pairs approach. You can assume that the function ***Neighbors*(*w*)** is already defined for you, and returns a list of words in the same context as ***w***. [12 points]

```
Mapper
  map(docid, doc d)
    for word w in d
      for word u in neighbors(w)
        emit((w,u), 1)
```

```
Reducer
  reduce(pair p, counts[] c)
    sum ← 0
    for count in c
      sum ← sum + c
    emit(p, sum)
```

[2 points] Mapper loops over words in document
[2 points] Mapper loops over words in Neighbors of the loop index
[2 points] Mapper emits a pair of two words as key, and a count of 1 as value

[2 points] Reduce takes a single pair of words, and a list of counts as input
[2 points] Reduce loops over input counts to come up with a total sum
[2 points] Reduce emits the same pair as it was given as input, and the sum it computed

b) The other approach for computing word co-occurrence is using stripes. How do pairs and stripes relate to our original sequential formulation using matrix ***M***. [4 points]

[2 points] Pairs computes a single entry in M/computes M_{ij}/creates NxN keys
[2 points] Stripes computes a single row in M/computes M_i/creates N keys

c) Describe one advantage and one disadvantage that the stripes approach has compared to the pairs approach. [4 points]
[2 points] For identifying an advantage
[2 points] For identifying a disadvantage

Possible advantages: Faster, fewer keys, more dense data, easier to aggregate data, less shuffle/sort overhead
Possible disadvantages: More complex implementation, memory usage does not scale

Scrap Paper