CSE 4/587 - Final Exam

12/14/22 @ 8AM, 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: ___________________________

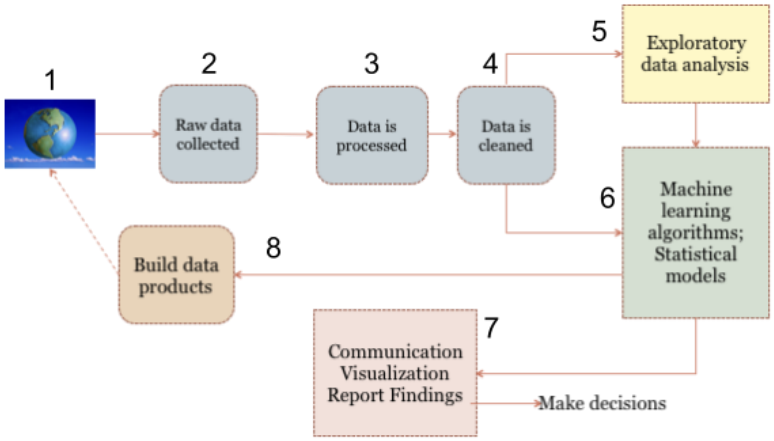
Instructions

1. This exam contains **7 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 3 hours 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	Q6	Total
20	30	40	15	35	10	150

Question 1 - Ethics in Data Science [20 Points]


1
2
Raw data collected
3
Data is processed
4
Data is cleaned
5
Exploratory data analysis
6
Machine learning algorithms; Statistical models
8
Build data products
7
Communication Visualization Report Findings
Make decisions

The diagram to the right shows the data science pipeline we've referred to over the course of the semester. Each stage of the pipeline has been labeled with a number, which you can use in the following questions related to ethics in data science.

a) Name and define (**in at most one sentence**) 4 of the 6 types of colloquial biases discussed in class. [12 points]

b) For each of the 4 types of bias mentioned above, name one of the stages in the data science pipeline where this type of bias may show up. [4 points]

c) Choose ONE of the 4 types of bias mentioned above and briefly describe a solution to correct for that type of bias, and which stage(s) of the pipeline this fix would involve. [4 points]

Question 2 - Distributed Systems [30 Points]

a. List 2 distinct benefits that Spark has over MapReduce [6 points]

b. Briefly explain how HIVE improves programmer productivity when compared to MapReduce [2 points]

c. In **one sentence** explain the primary way fault-tolerance is achieved in:

   i. HDFS [4 points]

   ii. Spark [4 points]

d. Explain the difference between a transformation and an action in Spark [4 points]

e. Name one transformation that results in a ***narrow*** dependency and draw a DAG showing how this transformation acts on the RDDs and their partitions [5 points]

f. Name one transformation that results in a ***wide*** dependency and draw a DAG showing how this transformation acts on the RDDs and their partitions [5 points]

Question 3 - Spark [40 Points]

For Question 3, refer to the following Spark code. Assume that the number of files in the **`files`** list is 2, and that **`fullOuterJoin`** is a transformation resulting in a narrow dependency.

```
result = None
for file in files:
    start = sc.textFile(file)
    middle = start.flatMap(lambda x: x.split(' ')) \
                  .map(lambda x: (x, 1)) \
                  .reduceByKey(lambda a,b: a + b)

    if result == None:
        result = middle
    else:
        result = result.fullOuterJoin(middle)
result.collect()
```

a. Draw the lineage graph for the RDD **`result`** right before the call to **`result.collect()`**. Include nodes for all intermediate RDDs, even if they are unnamed. [10 points]

b. How many stages will your DAG be broken up into for execution? [10 points]
Draw the stage boundaries in your diagram.

c. Identify in the above code (**excluding line 11**) one instance of the following:

   i. A transformation that results in a wide dependency [4 points]

   ii. A transformation that results in a narrow dependency [4 points]

   iii. An action [4 points]

d. How many "jobs" will the above code run? [4 points]

e. What algorithm is the above code an implementation of? [4 points]

Question 4 - Classifiers [15 Points]

a) Name the three classification algorithms we described in lecture. For each algorithm state whether it is a structural classifier, or statistical classifier. [6 points]

b) In the second half of the class we used two motivating examples when discussing classifiers: spam email classification, and ad click probability. State what was used as the features for each example, and give one reason why k-NN may not be effective for these two use cases. [4 points]

c) Naive bayes is an exceedingly simple algorithm. Name one reason why it can still be effective for something like spam classification. [2 points]

d) How can we adapt binary classifiers to multi-class problems? [3 points]

Question 6 - Misc [10 Points]

a) What are the four Vs of data intensive computing? [2 points]

b) Name two sources of uncertainty in data. [2 points]

c) Briefly explain why fault tolerance in distributed systems is important. [2 points]

d) Name one concrete example of a data intensive application that you interact with on a frequent basis. [2 points]

e) Name two solutions you could employ if the dataset you are working on becomes too big to store/process in a timely manner. [2 points]

Scrap Paper

Scrap Paper

Scrap Paper