CS 418: Introduction to Data Science
Homework Assignment 05
Fall 2019

Instructions

This assignment is due Sunday, December 08, at 11:59PM (Central Time).

Answers for this assignment must be entered on *Blackboard*. You have 1 attempt to submit the assignment. This attempt is not timed. The assignment will close automatically at the due date and no late submissions will be accepted.

This assignment is individual. Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Part I: Text Analysis (80 pt.)

Problem 1 (5 pt.)

Suppose that a query returns 150 documents, 120 of which are relevant to the query. If there are 200 relevant documents in the collection, what is the precision and recall of the query?

Problem 2 (40 pt.)

Given the following collection of documents:

Doc 1	Jobs! Jobs in data science.
Doc 2	New data for data science.
Doc 3	New jobs in computer science.
Doc 4	Data science is new.

Suppose that you tokenize each document and process the tokens by removing punctuation, converting to lower case, and removing stop words "in," "for," and "is." After processing the tokens, answer the following questions:

a. (5 pts.) Complete the following term frequency matrix for the collection of documents.

Term	Doc 1	Doc 2	Doc 3	Doc 4
jobs				
data				
science				
new				
computer				

b. **(5 pts.) Complete the following inverse document frequency matrix for the collection of documents. Use a logarithm of base 10 to compute the inverse document frequency.**

Term	idf
jobs	
data	
science	
new	
computer	

c. **(10 pts.) Complete the following term frequency-inverse document frequency matrix for the collection of documents.**

Term	Doc 1	Doc 2	Doc 3	Doc 4
jobs				
data				
science				
new				
computer				

d. **(8 pts.) Compute the score of each document in the collection for the query "data science computer"?**

e. **(2 pts.) Which of the documents in the collection is the most relevant to the query "data science computer"?**

f. **(8 pts.) Represent each document in the collection as a vector with a component for each term in the dictionary and a weight for each component given by the tf-idf of the corresponding term in the document. Given these vectors, compute the cosine similarity between Doc 1 and every other document in the collection.**

g. **(2 pts.) Which of the documents in the collection is the most similar to Doc 1?**

Problem 3 (35 pt.)

Given the following collection of <u>training</u> documents:

Doc ID	Terms in Doc	Class
1	computer science jobs science	1
2	new data science computer	1
3	jobs computer algorithms	0
4	data science computer science	1
5	algorithms new jobs algorithms	0

And the following collection of <u>test</u> documents:

Doc ID	Terms in Doc	Class
6	jobs computer science	?
7	jobs science algorithms	?

Suppose that you use Naïve Bayes to classify the test documents and answer the following questions:

- **a. (3 pts.) What is the probability that a document belongs to class 1? What is the probability that a document belongs to class 0?**
- **b. (12 pts.) What is the probability that a document of class 1 contains the word "science"? Similarly, for the words "algorithms," "computer," and "jobs." Use Laplace smoothing to avoid zeros.**
- **c. (12 pts.) What is the probability that a document of class 0 contains the word "science"? Similarly, for the words "algorithms," "computer," and "jobs." Use Laplace smoothing to avoid zeros.**
- **d. (4 pts.) Using Naïve Bayes, what is the predicted class of Doc 6?**
- **e. (4 pts.) Using Naïve Bayes, what is the predicted class of Doc 7?**

Part II: Big Data (20 pt.)

Problem 4 (5 pt.)

Match each statement with the corresponding type of data.

- **Does not conform to a data schema.**
- **Has a faster growth rate.**
- **Includes text, images, audio, and video.**
- **Often stored in relational databases.**
- **Often stored in NoSQL databases.**

a. Structured

b. Unstructured

Problem 5 (5 pt.)

Match each statement with the corresponding replication method.

- **Provides fault tolerance.**
- **Writes can be inconsistent.**
- **Writes are always consistent.**
- **Reads can be inconsistent.**

a. Primary-replica (master-slave)

b. Peer-to-peer

c. Both

Problem 6 (5 pt.)

Match each statement with the corresponding data storage approach.

- **All data is stored.**
- **Some data may be discarded.**
- **Data must conform to a data schema.**
- **Data is available for analysis sooner.**
- **Decisions about future data use must be made in advance.**

a. Schema-on-write

b. Schema-on-read

Problem 7 (5 pt.)

Match each statement with the corresponding data processing approach.

- **Characterized by high-latency responses.**
- **Trades consistency for speed.**
- **Appropriate for MapReduce tasks.**
- **Appropriate for real-time processing of data streams.**

a. Batch processing

b. Transactional processing