Department of Computer and Information Science and Engineering

University of Florida

CAP4770/CAP5771 Fall 2015

Midterm I Sample Test

Instructor: Dr. Daisy Zhe Wang

THIS IS A SAMPLE TEST

Name:

UFID:

I. [0 points] Data Modeling and Similarity Metrics.

Given the following text fragments extracted from Wikipedia:

"Connery's breakthrough came in the role of secret agent James Bond. He was reluctant to commit to a film series, but understood that if the films succeeded his career would greatly benefit. He played the character in the first five Bond films: Dr. No, From Russia with Love, Goldfinger, Thunderball, and You Only Live Twice then appeared again as Bond in Diamonds Are Forever and Never Say Never Again. All seven films were commercially successful."

"The James Bond film series is a British series of spy films based on the fictional character of MI6 agent James Bond, 007, who originally appeared in a series of books by Ian Fleming. It is one of the longest continually-running film series in history, having been in on-going production from 1962 to the present (with a six-year hiatus between 1989 and 1995). "

(1) Model the documents using

- Bag of Words
- 3-grams

(**2**) Given the query "James Bond film":

- Which similarity metric would you use to retrieve the most relevant document?
- Apply such similarity metric to each document.

II. [0 points] Natural Language Processing.

Consider the following context-free grammar

- $S- > NP\ VP$
- $NP- > NP\ PP$
- $NP- > Det\ N$
- $VP- > VP\ PP$
- $VP- > V\ NP$
- $PP- > Prep\ NP$
- $Det- > the$
- $Det- > a$
- $Prep- > with$
- $V- > kissed$
- $N- > man$
- $N- > woman$
- $N- > dog$

(**1**) Show a parse tree for the sentence “The man kissed the woman with a dog”.

(**2**) Write a regular expression to extract named entities from the following sentence "Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt."

III. [0 points] Map Reduce.

Consider the following pandas code,

```
pd.merge(Orders, LineItem, on=’order_id’)
```

Where Orders and LineItem are two DataFrames that share a common attribute *order_id*. Consider *order_id* to be the key of the Orders DataFrame (No two orders share the same *order_id*). We want to join them together to get a complete view.

(**1**) Your Map and Reduce functions should produce the same result as the pandas command. The two input DataFrames are concatenated in a single file that will be processed.

Map Input: each input is a list of strings representing a tuple in the DataFrame. Each list element corresponds to a different attribute of the table. The first item on each record (Index 0) has two possible values $\{line\_item|order\}$ the second element on each record is the *order_id*. LineItem records have 17 attributes including the identifier string. Order records have 10 elements including the identifier string.

Input example:
[order, 1, ..., 230]
[order, 2, ..., 242]
[line_item, 1, ..., asdc]
[line_item, 3, ..., ates]

Reduce output: the reduce output should be a joined record: a list of lenght 27 that contains the attributes from the *order* record followed by the fields from the *LineItem* record.

Output Example:
[order, 1, ..., 230, line_item, 1, ..., asdc]

IV. [0 points] Exploratory Data Analysis.

Consider a file with one day worth of ads shown and clicks recorded on the New York Times home page. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged-in.

Note: An impression (in the context of online advertising) is when an ad is fetched from its source, and is countable. Clicking or not is not taken into account. Each time an ad is fetched it is counted as one impression

(**1**) [**points**] What steps would you take to create the following data visualizations:

- The distribution of number impressions
- The distribution of click-throught-rate ($CTR = \#clicks/\#impressions$)

for six age categories: < 18, $18 - 24$, $35 - 34$, $35 - 44$, $45 - 54$, 54+

(*This page is intentionally left blank*)