BGMT404 Final Exam (1:30-3:30, 05/16) Spring 2016

Name: **UID:**

You have 2 hours to complete this exam (it should take much less time than that).

- You may use 1 page (one side, A4 paper) of handwritten notes.
- **You must turn in your handwritten notes along with your exam sheet. Without this, your final exam will not be graded.**
- You may not use any computers.
- Also, you may not communicate with any other person during this exam, (except the professor or the TA).
- Make sure you have answered all of the questions in the exam. It isn't over until it says "END EXAM".
- Do not forget to write your name and UID on this and your handwritten notes.

As strategies for completing the questions on this exam, please keep in mind the following:

- If you find a question ambiguous, please explain your confusion.
- You are more likely to get partial credit for a wrong answer if you show your work.
- However, be careful not to get carried away and run over the time limit. In other words, plan ahead.
- It is a good idea to work on questions with smaller points first (they are easier) and work harder questions later.

1. Please answer the following questions. (10 points)

1.1 The output after executing the following statements: (2 points)

```
str = "I love this world  " # two spaces at the end
print str.strip().upper()[2:6]
```

1.2 The output after executing the following statements: (2 points)

```
str = "I love programming. It is beautiful and helpful."
lst = str.split()
print lst[2][0:7]
```

1.3 what are key benefits of Tableau? (2 points)

1.4 In a python program using matplotlib, what's the layout of subplots if using the following statements? Please draw the layout of subplots. Each subplot is a rectangle. All subplots have the same size. (1 points)

```
subplot(1,2,1)
subplot(1,2,2)
```

1.5 The output of the following statements: (For the corresponding tag of each word, use TAG) (3 points)

```
import nltk
str = "I love this world"
ws = nltk.word_tokenize(str)
tags = nltk.pos_tag(ws)
for tag in tags:
        print tag
```

2. Please read the following program and write down the outputs. (10 points)

Suppose the content of file: test.txt is:

1. Hello world
2. Hello world world
3. I love this world
4. Life is good

```
import os
import sys

fh = open('test.txt','r')
for x in fh:
        lst = x.split('[\.\s]+')
        for element in lst:
                print element

fh.close()
```

3. Please read the following program and answer questions. (15 points)

```
html_doc = """
<html>
        <head>
                <title>The Dormouse's story</title>
        </head>
<body>
        <p class="title">
                <b>The Dormouse's story</b>
        </p>
        <p class="story">Once upon a time there were three little sisters; and their names were
                <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
                <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
                <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>
                ; and they lived at the bottom of a well.
        </p>
        <p class="story">...</p>
</body>
</html>
"""

from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc)

print soup.head
print soup.p[‘class’]
print soup.p.next_sibling
for link in soup.findAll(‘a’):
        href = link.get(‘href’)
        print href
```

4. Please read the following program and answer questions. (15 points)

```
ceos = """
{
        "President": "Alan Isaac",
        "CEO": "David Richardson",
        "Data": [
                {"country":"India",
                 "name":"Sachin Tendulkar" ,
                 "salary":"100"
                },
                {"country":"Srilanka",
                 "name":"Lasith Malinga",
                 "salary":"200"
                },
                {"country":"England",
                 "name":"Alastair Cook",
                 "salary":"300"
                }
        ]
}
"""
import json
json_obj = json.load(ceos)
print json_obj['CEO']
data = json_obj['Data']
for i in range(data):
        the_country  = data[i]
        print data[i]['name']
```

5. Please identify which following strings match the given regular expression. (15 points)

Define the following regular expression:

regex_str = "^I lost my:? (wallet|car|cell phone|marbles)"

"I lost my wallet"	(match)	(not match)
" I lost my wallets"	(match)	(not match)
"I lost my: car"	(match)	(not match)
"I lost my- car"	(match)	(not match)
"I lost my: cell"	(match)	(not match)
"I lost my: cell phone"	(match)	(not match)
"I lost my cell phone"	(match)	(not match)
"I lost my marbles"	(match)	(not match)

6. Please write the tf-idf matrix for the following 5 documents. Each row in the matrix represents a document and each column in the matrix represents a unique term. The value in the cell (i, j) means the tfidf value of the j^{th} term in the i^{th} document. (case in-sensitive) (15 points)

Hello world, I love you.
You have the blue sky.
You have the blue ocean.
You have the sun
I love the sun, the sky, the ocean, and you.

7. Please use regular expressions to write patterns to match the following strings. (20 points)

(1) A University of Maryland or Google email (e.g., username@umd.edu, username@gmail.com). The username must satisfy the following criteria: (a) at least 5 characters, (b) must start with letters or underscore, (c) must be combinations of letters, numbers, and underscores.

(2) A web page URL (e.g., http://www.google.com, https://www.facebook.com, http://www.umd.edu, etc.) consists of the following components: (a) start with http or https, (b) domain name starts with www and ends with .com, .edu, .org, or .net, (c) the middle part of the domain name (e.g., google, facebook, umd, etc.) is the word characters.

END EXAM