

OLD DOMINION UNIVERSITY

CS 495: INTRODUCTION TO WEB SCIENCE
INSTRUCTOR: MICHAEL L. NELSON, PH.D
FALL 2014 4:20PM - 7:10PM R, ECSB 2120

Assignment # 11

GEORGE C. MICROS UIN: 00757376

Honor Pledge

I pledge to support the Honor System of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community it is my responsibility to turn in all suspected violations of the Honor Code. I will report to a hearing if summoned.

Signed _____
December 11, 2014

George C. Micros

Written Assignment 11

Fall 2014

CS 495: Introduction to Web Science

Dr. Michael Nelson

December 11, 2014

Contents

1	Written Assignment 11	5
1.1	Question 1	6
1.1.1	The Question	6
1.1.2	The Answer	6

Chapter 1
Written Assignment 11

1.1 Question 1

1.1.1 The Question

Using the data from A9:

- Consider each row in the blog-term matrix as a 500 dimension vector, corresponding to a blog.
- From chapter 8, replace `numpredict.euclidean()` with cosine as the distance metric. In other words, you'll be computing the cosine between vectors of 500 dimensions.
- Use `knestimate()` to compute the nearest neighbors for both:
<http://f-measure.blogspot.com/> <http://ws-dl.blogspot.com/>
 for `k=1,2,5,10,20`.

1.1.2 The Answer

A function for the cosine similarity was made based off the definition of the dot product. This was used with the `nnestimate()` function

```

1  #!/usr/bin/python
3  import math
5  def cos_sim(v1, v2):
        sumxx, sumyy, sumxy = 0, 0, 0
7      for i in range(len(v1)):
            x = v1[i]; y = v2[i]
9            sumxx += float(x)*float(x)
            sumyy += float(y)*float(y)
11           sumxy += float(x)*float(y)
        return sumxy/math.sqrt(sumxx*sumyy)
13
14  def getdistances(data, vec1):
        distancelist=[]
15
16      # Loop over every item in the dataset
17      for i in range(len(data)):
18              vec2=data[i]
19
20          try:
21                  distancelist.append((cos_sim(vec1, vec2), i))
22          except:
23                  pass
24
25      # Sort by distance
26          distancelist.sort()
27          return distancelist
28
29  def knestimate(data, vec1, k=5):
30      # Get sorted distances
31          dlist=getdistances(data, vec1)
32          avg=0.0
33          return dlist
34
35  vecs = {}
36
37  f = open("blogdata2.txt", "r")
38
39  for line in f:
40          a = line.strip('\n').split('\t');
41          b = a.pop(0)
42          vecs[b] = a
43
44  print len(vecs)
45
46  fm = 'F-Measure'
47  ws = 'Web Science and Digital Libraries Research Group'
48
49  a = vecs[fm]
50  temp = vecs.values()
51  temp.pop(vecs.keys().index(fm))

```

```
53 a = knnestimate(temp,a,k=5)
55
57 k = [1, 2, 5, 10, 20]
57 print "-----F-Measure kNN-----"
57 for i in k:
59     print "----k = "+str(i)
59     for j in range(i):
61         b = a[j][1]
61         print vecs.keys()[b]
63
63 a = vecs[ws]
65 temp = vecs.values()
65 temp.pop(vecs.keys().index(ws))
67
67 a = knnestimate(temp,a,k=5)
69
71 k = [1, 2, 5, 10, 20]
71 print "-----WS-DL kNN-----"
73 for i in k:
73     print "----k = "+str(i)
75     for j in range(i):
75         b = a[j][1]
77         print vecs.keys()[b]
```

Listing 1.1: Python script that computes kNN based on cosine similarity

```

george@george-K55VD: ~/Dropbox/FALL14/CS495/hw/HW11/q1
george@george-K55VD: ~/Dropbox/FALL14/CS495/hw/HW11/q1 119x59
george@george-K55VD:~/Dropbox/FALL14/CS495/hw/HW11/q1$ ./q1.py
120
-----F-Measure KNN-----
---k = 1
Faces / Gesichter
---k = 2
Faces / Gesichter
Octopus Grigori
---k = 5
Faces / Gesichter
Octopus Grigori
Wee Kitchen
Japan Farmers Markets
If There's One Thing I've Learned...
---k = 10
Faces / Gesichter
Octopus Grigori
Wee Kitchen
Japan Farmers Markets
If There's One Thing I've Learned...
Ever Changing Streams
Tea Obsession
KikiMin
DustysDinners
Essdras M Suarez - Photographer - Blog
---k = 20
Faces / Gesichter
Octopus Grigori
Wee Kitchen
Japan Farmers Markets
If There's One Thing I've Learned...
Ever Changing Streams
Tea Obsession
KikiMin
DustysDinners
Essdras M Suarez - Photographer - Blog
My Little Slice of Pie
Bombay Boy
Downtown Elgin
Baker's Cakes
yours deliciously
Passey Family
somewhere in time
My Name is June. I Like To Cook
Carpe Diem Acreage
The Wineauxs
george@george-K55VD:~/Dropbox/FALL14/CS495/hw/HW11/q1$ █

```

Fig. 1.1: Clustings for the F-Measure blog


```

george@george-K55VD: ~/Dropbox/FALL14/CS495/hw/HW11/q1
george@george-K55VD: ~/Dropbox/FALL14/CS495/hw/HW11/q1 119x59
george@george-K55VD:~/Dropbox/FALL14/CS495/hw/HW11/q1$ ./q1.py
120
-----WS-DL kNN-----
---k = 1
This Fabulous Life
---k = 2
This Fabulous Life
neoscribe
---k = 5
This Fabulous Life
neoscribe
The Louisville-St. Louis Connection
Octopus Grigori
makarios: blessed
---k = 10
This Fabulous Life
neoscribe
The Louisville-St. Louis Connection
Octopus Grigori
makarios: blessed
striving to live each day HIS way
Winton Families & More
My Name is June. I Like To Cook
life with lily
How To: Mobile Phones, Joomla, SEO...
---k = 20
This Fabulous Life
neoscribe
The Louisville-St. Louis Connection
Octopus Grigori
makarios: blessed
striving to live each day HIS way
Winton Families & More
My Name is June. I Like To Cook
life with lily
How To: Mobile Phones, Joomla, SEO...
The Erratic Homemaker
Japan Farmers Markets
The FDC Report
Practically Magic
Wee Kitchen
A Truth From www.emmetssentials.com
The Jenn and Zui Kim Ohana
Vinson Boys
Burp! Recipes
Bella Terra
george@george-K55VD:~/Dropbox/FALL14/CS495/hw/HW11/q1$ █

```

Fig. 1.2: CLusterings for the WS-DL blogs