Web Science: Assignment #9

Alexander Nwala

Mohd. Nauman Siddique

Thursday, April 25, 2019

Contents

Problem 1

Using the data from A7:
- Consider each row in the blog-term matrix as a 1000 dimension vector, corresponding to a blog.
- Use knnestimate() to compute the nearest neighbors for both: http://f-measure.blogspot.com/ http://ws-dl.blogspot.com/
for k=1,2,5,10,20.
Use cosine distance metric (chapter 8) not euclidean distance. So you have to implement numpredict.cosine() instead of using numpredict.euclidean() in: https://github.com/arthur-e/Programming-Collective-Intellig blob/master/chapter8/numpredict.py

SOLUTION
Not attempted

Problem 2

Re-download the 1000 TimeMaps from A2, Q2. Create a graph where the x-axis represents the 1000 TimeMaps. If a TimeMap has "shrunk", it will have a negative value below the x-axis corresponding to the size difference between the two TimeMaps. If it has stayed the same, it will have a "0" value. If it has grown, the value will be positive and correspond to the increase in size between the two TimeMaps.
As always, upload all the TimeMap data. If the A2 github has the original TimeMaps, then you can just point to where they are in the report.

SOLUTION
I have reused my memento urls from A2 assignment from https://github.com/naumansiddiqui4/anwala.github.io/blob/master/lectures/cs532-s19/assignments/A2/Url_timemaps.csv.
I downloaded the timemaps again for A9 resuing the code from A2 and created a histogram shown in figure 1 and a scatter plot shown in figure 2 to show the difference between memento count for all the urls.

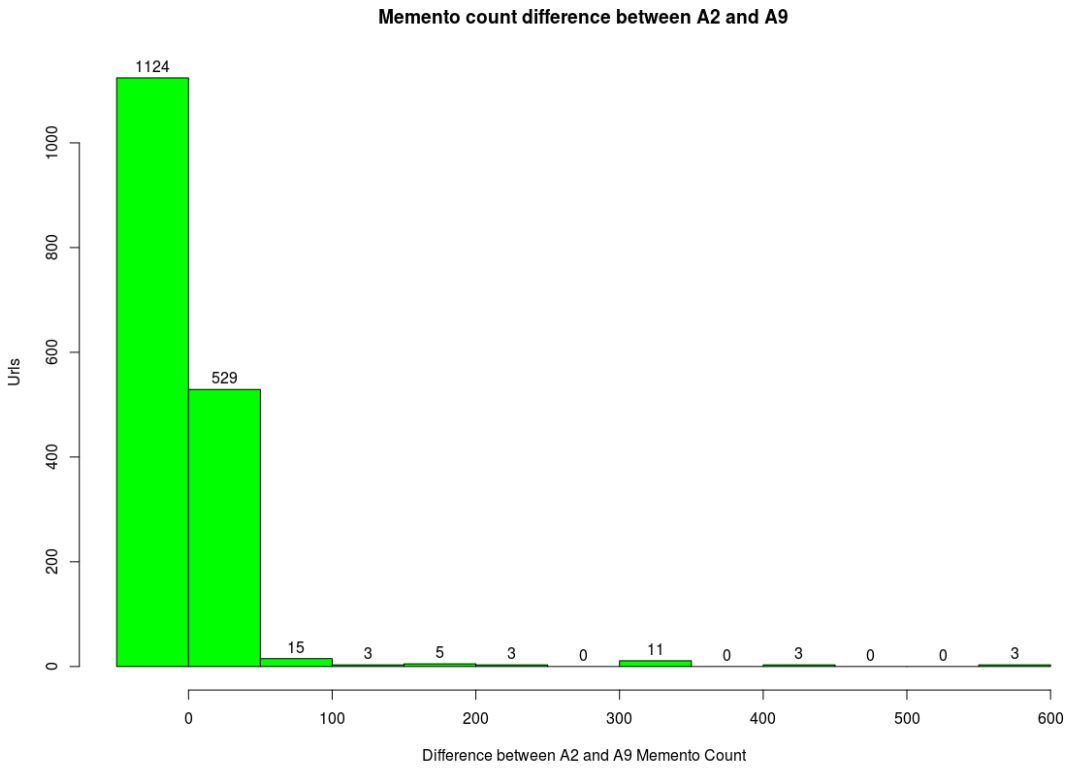

Figure 1: Histogram for memento cout difference between A2 and A9 mementos

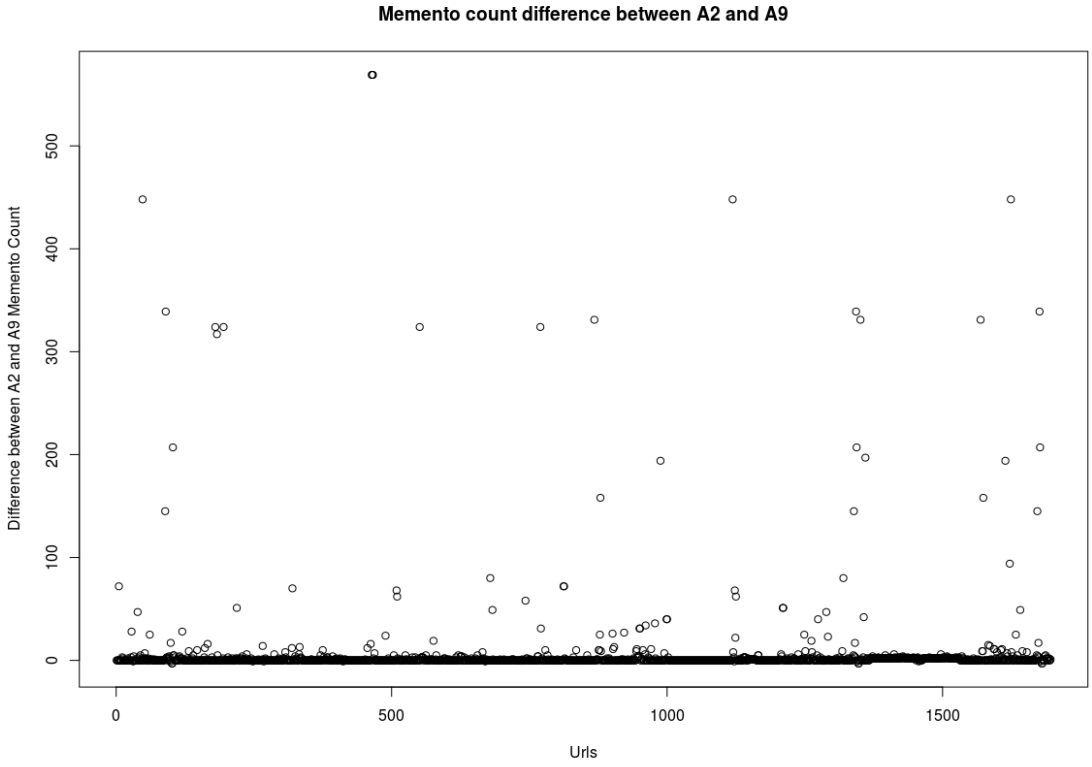

Figure 2: Scatterplot for memento cout difference between A2 and A9 mementos