**Please read carefully regarding assignment 2**

1. The assignment is now due on Friday 10/12 by 23:59.
2. Running lift with big\_city and small\_town

It came to our attention that the lift script produces an error if you use big\_city or small\_town (apparently the script gets thrown off by the underscore). Please use bigcity and smalltown as single words to avoid this problem.

1. If the MDS script does not run properly for candidates and issues, it is because you are not giving it a square matrix as input. That is, on both sides of the lift and dissimilarity matrix you need to put candidate1 candidate2 issue1 issue2 issue3 issue4. E.g., here is the lift matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Candidate1 | Candidate2 | Issue1 | Issue2 | Issue3 | Issue4 |
| Candiate1 |  | 2.1 | 1.3 | .8 | 1.6 | .6 |
| Candidate2 |  |  | .9 | 1.8 | 1.2 | 1.4 |
| Issue1 |  |  |  | .002 | .0001 | .003 |
| Issue2 |  |  |  |  | .005 | .03 |
| Issue3 |  |  |  |  |  | .02 |
| Issue4 |  |  |  |  |  |  |

The above matrix will let the MDS script run properly; however, the MDS map will not look right. It is because with 2 dimensions, it is impossible for the MDS algorithm to represent such extreme variation in values. E.g., the smallest value in the matrix is .0001, while the highest is 2.1. That is, the largest number 2.1 is 2.1/.0001 = 2100 times larger than the smallest number .0001. To avoid this situation, look at the smallest numbers (shown in yellow above), and artificially increase them, while keeping them smaller than the smallest of the reasonable lift values, E.g., we could do the following:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Candidate1 | Candidate2 | Issue1 | Issue2 | Issue3 | Issue4 |
| Candiate1 |  | 2.1 | 1.3 | .8 | 1.6 | .6 |
| Candidate2 |  |  | .9 | 1.8 | 1.2 | 1.4 |
| Issue1 |  |  |  | .3 | .3 | .3 |
| Issue2 |  |  |  |  | .3 | .3 |
| Issue3 |  |  |  |  |  | .3 |
| Issue4 |  |  |  |  |  |  |

Now the ratio has been reduced to 2.1/.3 = 7, which is much easier for MDS to represent.

1. For the parserforsentiment do not make any changes in the script other than the name of the input file. The script will prompt you for brand, attribute and the number of words when you run it.
2. In many cases the scripts are not running because the text is not in column 3. Similarly when you replace the location with bigcity or smalltown, the location must be made the 3rd column for the script to work.
3. For sentiment analysis, the last column of the output file (compound) is the final sentiment score of each row. **Ignore** the other three columns. There are two ways to report the sentiment score:
4. Take the average of the “compound” column – if the average is positive (negative), then the overall sentiment is positive (negative). But one problem is that since positive and negative numbers cancel each other, the average (either positive or negative) is small, unless the positives (negatives) overwhelmingly outnumber negatives (positives).
5. Report percentages (from the compound column) – E.g., 30% positive, 50% neutral and 20% negative. Which one to use is your choice.