BINF-734 Homework -7 Report **Monica Devi Manam**
George Mason University

Abstract:
This paper describes about whether a given hypothesis is true or false. We arrive at this conclusion by following the protocol or the steps asked to perform in the homework.

Hypothesis:
Topical documents can be separated by indicative words.

Steps Followed:

- Firstly two diverse topics are chosen to perform this assignment. The topics chosen were a)Plant BioInformatics b) skin care and beauty.
  12 documents were collected and converted to text files, which are attached along with this.
  Gathering the data:
  The 2 functions are preformed which have to be input :
  >>> l1,l2=mh.Names('BIO_Text','BEAU_Text')
  >>> d1,d2=mh.Texts(l1,l2)
  The files are named in this way : ['BIO_Text/1.txt']
  The 2nd function extracts all the text from these files.
- **Creating the Dictionaries:**
  a) MakeDictionaries : By passing the texts lists we get the dictionaries
  >>> list1=mh.MakeDictionaries(d1,d2)
  >>> len(list1)
  24
  b)Making good words: Get the good words by passing the list of dictionaries
  >>> gwords = mh.GoodWords(list1)
  >>> len(gwords)
  8080
- **Making the Matrices**: It should have D rows and T columns were D is the number of documents and T is the number of terms (unique words).
  >>> m1,m2,m3 = mh.MakeMatrices(list1, gwords)
  >>> m1.shape
  (24, 8080)
  checking length :
  >>> a= m2 [:,1].sum()
  >>> 1.0
- **Pruning and Indicative Words: Creates a valid list of words, which should appear in the N docs**
  >>> p = mh.Prune(m1, gwords,mindoc=7)
  >>> len(p)
  779
  Running MakeMatrices again:

```
>>> c1,c2,c3 = mh.MakeMatrices(list1,p)
>>> c1.shape
(24, 779)
```

Indicatives: After running the MakeMatrices again, the miner's function IndicWords is called for removing words that do not help in discrimination of docs.

```
>>>pdoc=range(0,12)  # Input the pdoc range as 1-12, since my positive documents are in this range
>>>ndoc=range(12,24) # Input the ndoc range as 12-24, since my negative documents are in this range

>>> gwords2 = mh.Indicatives(c1,c2,c3,p,pdoc,ndoc)
>>> len(gwords2)
137
>>> f1,f2,f3 = mh.MakeMatrices(list1, gwords2)
>>> f2.shape
(24, 137)
```

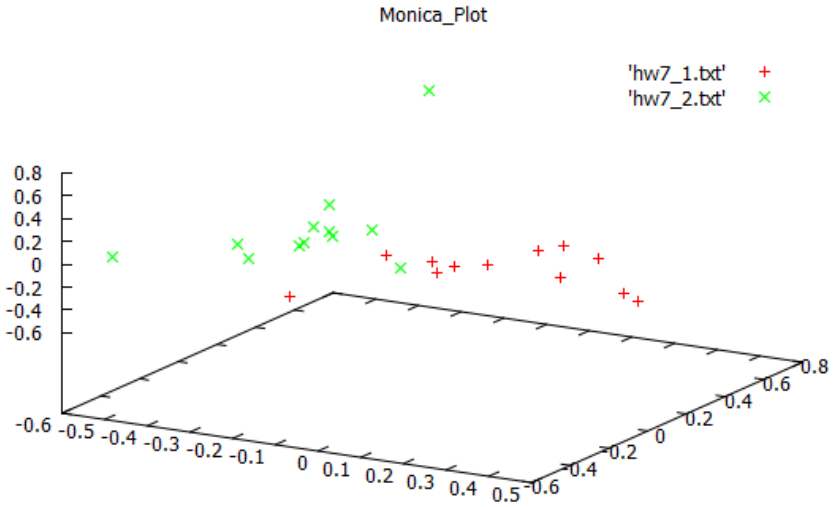
- **PCA functions:**

  Now, the matrix obtained from the above function should be plotted. For this import the given files "pca.py" and gnuPlot

  Then perform the following steps:

```
>>> import pca
>>> evals,evecs = pca.NewCoords(f2)
>>> len(evecs)
137
>>> cffs=pca.DataInNewSpace(f2,evecs[:,:3])
>>> cffs.shape
(24, 3)
>>> gnu.Save('hw7_1.txt', cffs[:12])
>>> gnu.Save('hw7_2.txt', cffs[12:])
```

PLOT:



Conclusion:

The Plot above shows 2 different documents plots. This indicated that the hypothesis is true. Although a 10% overlap is observed, this is probably due to the relevance of biology in Skin Care and beauty. But, otherwise I think the plot shows that the 2 documents can be separated based on the indicative words that are generated.