**

COMP814 – Text Mining

Vectorising Lab

# Objective

To be able to make your own “Similarity Calculator” and compare it with some of the ones from Python library.

# Task

1. Download the example code from lectures and ensure that it runs error free.
2. Download 4, approximately 1 page news articles from NZ Herald or elsewhere.
3. Read them into a variable and vectorise them using the two types of vectorisers as given in the lecture code, They are :
   1. Count Vectoriser
   2. TFIDF Vectoriser
   3. Use the following formula to compute the distance between any two of those 4 articles using both Count and the TFIDF vectorisers.
   4. That is, the difference between the two vectors, squared and the elements summed up. Use the following code example for subtraction of two vectors.
4. **import** numpy **as** NP
5. A = **[**4, 8, 7**]**
6. B = **[**5, -4, 8**]**
7. print**(**"The input arrays are :\n","A:",A ,"\n","B:",B**)**
8. Res1= NP.add**(**A,B**)**
9. Res2= NP.subtract**(**A,B**)**
10. print**(**"Result of Addition is :",Res1,"\n","Result of Subtraction is:",Res2**)**
    1. Compare with the distances computed with nltk library functions (from lecture code) and see if your one is **proportionally** similar to any or some of them. Report this.