Name of Course – Machine Learning for Trading

Project Name - 2014Fall7646 Project 1B

Student Name – Aishvarya Krishnan

Approach to solve the problem statement

To compute the tf-idf values for each word in variable number of documents I followed the steps –

1. Read each document and store the following –
    a. Words in a list which will be used as a 'Bag of words'
    b. Append list of words from each document to a list - docwords
2. Calculate the number of occurrences of each word in the documents - count
3. Using Counter for each list of words in docwords –
    a. calculate the number of occurrences of the each word in every document - count_t
    b. calculate the word that occurs maximum times in each document – max_w
4. Using values of count_t and max_w calculate term frequency 'tf'
5. Using count and number of documents calculate idf
6. Calculate tf-idf as per formula and print output

Output of tfidf.py

term	ballmer01.txt	ballmer02.txt	ballmer03.txt	ballmer04.txt	ballmer05.txt	ballmer06.txt	ballmer07.txt
the	0.559616	0	0.559616	0.559616	0	0.559616	0
worlds	1.94591	0	0	0	0	0	0
largest	1.94591	0	0	0	0	0	0
software	1.94591	0	0	0	0	0	0
maker	1.94591	0	0	0	0	0	0
microsoft	1.94591	0	0	0	0	0	0
ceo	0	0.847298	0.423649	0.847298	0	0	0
steve	0	1.94591	0	0	0	0	0
ballmer	0	0.847298	0.423649	0	0.847298	0	0
resigned	0	1.94591	0	0	0	0	0
from	0	1.252763	0.626381	0	0	0	0
its	0	1.94591	0	0	0	0	0
board	0	1.94591	0	0	0	0	0
yearold	0	0	0.626381	1.252763	0	0	0
retired	0	0	0.972955	0	0	0	0
his	0	0	0.972955	0	0	0	0
position	0	0	0.972955	0	0	0	0
as	0	0	0.972955	0	0	0	0

satya	0	0	0	1.94591	0	0	0
nadella	0	0	0	1.94591	0	0	0
was	0	0	0	1.94591	0	0	0
named	0	0	0	1.94591	0	0	0
new	0	0	0	1.94591	0	0	0
said	0	0	0	0	1.94591	0	0
it	0	0	0	0	1.94591	0	0
would	0	0	0	0	1.94591	0	0
give	0	0	0	0	1.94591	0	0
him	0	0	0	0	1.94591	0	0
more	0	0	0	0	1.94591	0	0
time	0	0	0	0	1.94591	0	0
he	0	0	0	0	0	1.94591	0
recently	0	0	0	0	0	1.94591	0
acquired	0	0	0	0	0	1.94591	0
a	0	0	0	0	0	1.94591	0
basketball	0	0	0	0	0	1.94591	0
team	0	0	0	0	0	1.94591	0
clippers	0	0	0	0	0	1.94591	0
for	0	0	0	0	0	1.94591	0
billion	0	0	0	0	0	1.94591	0
revenues	0	0	0	0	0	0	1.94591
increased	0	0	0	0	0	0	1.94591
threefold	0	0	0	0	0	0	1.94591
during	0	0	0	0	0	0	1.94591
ballmers	0	0	0	0	0	0	1.94591
tenure	0	0	0	0	0	0	1.94591