Text Mining

Assignment 1

Kevin Zollicoffer

# 1a)

Please see attached python project. Output of code is:

/usr/local/Cellar/python/2.7.9/Frameworks/Python.framework/Versions/2.7/bin/python2.7 /Users/zollie/PASS/TextMining/Week1/assign1.py

TMPY> Tokenizing documents

TMPY> Generating dictionary

vocab has length of 354

all,dollar,results,month,four,manager,higher,talks,issues,note:,to,program,th,under,rise,fall,level,did,notes,companies,p,div,says,japanese,likely,further,estimated,investment,shares,current,capital,international,net,public,told,exchange,commercial,reported,canadian,k,credit,products,analysts,sugar,market,shareholders,from,would,two,next,francs,markets,today,more,dlrs,share,company,marks,it,high,this,nine,can,december,tax,states,currency,reserves,united,end,six,earnings,revs,economy,plans,may,earlier,such,a,banks,operations,reuter,over,expects,years,ended,including,committee,japan,still,its,before,industries,group,pct,lt,production,policy,non,finance,they,half,not,now,day,bank,february,january,each,quarter,bond,development,financial,energy,ec,trading,year,agriculture,out,canada,profit,since,increase,issue,announced,after,",york,ministry,could,american,ltd,south,first,oper,major,owned,one,president,management,system,their,rates,stock,statement,that,part,than,stg,rate,were,acquisition,and,investors,outstanding,imports,have,any,sell,also,chairman,which,added,price,who,most,eight,plan,dealers,average,sale,surplus,gold,agreed,spokesman,based,merger,should,only,treasury,exports,his,bonds,new,report,during,trade,shr,countries,common,through,set,qtr,national,around,officials,are,tender,brazil,said,federal,dlr,state,between,crude,we,terms,sold,however,dividend,convertible,debt,both,last,country,annual,against,foreign,april,s,figures,co,loan,period,basis,union,west,due,been,commission,interest,expected,meeting,firm,gas,minister,demand,prices,an,cts,mln,cash,will,while,seven,is,coffee,unit,vs,in,avg,if,prior,pay,make,same,grain,european,week,oil,yesterday,recent,lower,well,the,proposed,yen,being,money,domestic,sources,cut,rose,had,board,has,take,march,increased,government,term,early,five,world,loss,securities,payments,output,because,some,economic,export,"we,total,for,per,deficit,t,be,wheat,business,agreement,corp,loans,by,on,about,central,debentures,of,industry,months,or,into,three,down,shrs,there,long,was,billion,buy,offering,offer,but,gain,subsidiary,with,he,made,official,up,lead,record,growth,sales,general,as,at,inc,compared,no,stake,when,other,department,includes,fell,ago,assets,u,time,tonnes,"the

# 1b)

Please see attached python project. Output of code is:

TMPY> Removing stopwords

TMPY> Stemming tokens

TMPY> Generating dictionary

vocab has length of 239

operations,stake,surplus,proposed,yen,being,money,agreed,fall,dollar,results,years,ended,manager,including,committee,talks,follow,still,issues,industries,spokesman,cut,group,gold,merger,show,pct,sugar,economy,lt,production,board,treasury,policy,government,loan,earnings,non,finance,bonds,domestic,rise,term,expects,early,half,report,month,day,bank,common,loss,shr,level,notes,countries,companies,securities,p,payments,output,div,quarter,added,bond,development,financial,grain,national,japanese,ec,likely,economic,trading,"we,japan,tender,wheat,estimated,agriculture,investment,canada,rate,profit,lead,shares,current,dlr,state,deficit,capital,crude,international,net,issue,public,announced,t,terms,business,exchange,commercial,agreement,canadian,however,reported,market,base,corp,york,chairman,convertible,debt,loans,dlrs,central,debentures,country,industry,months,foreign,prior,american,figures,analysts,ltd,francs,south,oper,major,around,period,owned,shrs,dividend,qtr,total,unit,shareholders,",basis,union,west,outstanding,management,system,long,commission,program,growth,stock,interest,imports,rates,markets,energy,firm,ago,buy,statement,offering,offer,tax,company,marks,gas,sales,part,gain,demand,prices,subsidiary,assets,states,stg,meeting,official,mln,cash,ministry,record,officials,nine,export,k,acquisition,seven,annual,investors,give,december,share,coffee,high,currency,reserves,further,avg,trade,inc,sell,compared,credit,pay,european,take,department,products,revs,higher,week,brazil,cts,oil,plans,price,earlier,yesterday,united,eight,plan,president,recent,billion,note:,lower,"the,exports,dealers,average,sale,federal,sources,time,banks,tonnes,increase,minister

# 1c)

Please see attached python project. Output of code is:

TMPY> Tokenizing documents

TMPY> Removing stopwords

TMPY> Stemming tokens

TMPY> Generating dictionary

vocab has length of 303

operations,sector,stores,yen,consolidated,fall,dollar,major,years,ended,including,higher,restated,go,cut,exchange,previous,industries,rd,yr,group,earned,show,royal,pct,better,lt,production,board,equity,quarterly,",include,resources,take,real,division,non,good,return,around,provisions,domestic,rise,term,preferred,results,early,b,tax,half,report,month,provision,bank,common,executive,loss,shr,estate,name,level,companies,l,securities,banking,continue,t,qtly,div,quarter,reduced,added,outstanding,declared,financial,billion,p,energy,reduce,pacific,second,cost,trading,result,payable,growth,further,balance,expects,investment,declined,plans,brazil,trust,business,rate,communications,profit,spokesman,diluted,shares,forecast,current,dlr,market,dividends,international,net,issue,public,announced,payments,full,power,excludes,respectively,agreement,however,reported,restructuring,base,plc,dividend,debt,loans,distribution,change,payout,dlrs,revenues,mths,cts,k,months,united,foreign,prior,american,figures,analysts,ltd,operating,holders,dec,discontinued,oper,raised,rev,parent,revenue,previously,annual,loan,period,owned,capital,included,shrs,raises,national,approval,qtr,extraordinary,total,unit,ct,writedown,city,shareholders,strong,additional,service,basis,holdings,adjusted,initial,petroleum,charge,long,payment,commission,estimated,interest,francs,pretax,final,systems,acquisitions,buy,north,statement,offering,estimates,company,marks,gas,fund,decline,corp,prices,subsidiary,analyst,assets,steel,stg,improved,f,meeting,canadian,mln,cash,value,record,future,nine,making,ending,property,industry,acquisition,"we,pre,seven,losses,sees,ag,give,cents,december,share,sales,expenses,high,chairman,reserves,properties,periods,exclude,accounting,avg,orders,inc,sell,compared,credit,related,pay,write,stake,nd,savings,crowns,insurance,note,split,fourth,income,products,revs,gains,stock,development,canada,oil,fiscal,nil,earnings,price,earlier,effect,reflect,class,regular,eight,plan,services,president,data,approved,ago,note:,lower,"the,off,performance,third,sets,charges,flow,sale,st,federal,chief,costs,gain,time,banks,management,credits,increase,profits,turnover

# 2)

Please see attached python project.

TMPY> Vectorizing documents

# 3)

The tokenizer.

# 4)

## 1.

If fully constructed as a spreadsheet in a Text Mining problem, it would likely be a very sparse matrix

## 2.

In Text Mining the values of the spreadsheet cells would be non-negative.

## 3.

For Text Mining, missing values are a non issue.