Scoring English using Quad Grams

How can we make a computer recognize how “English” a piece of text is?

Initialization

Computation of Quadgram List and Scoring

To begin with, we must create our own scoring for how English a piece of text is by using data of English text. Due to the limitations of the processing power of my computer, we will be using only the first 15000 characters of Pride and Prejudice as our data set. Of course, a legetimate scoring would require a much larger data set to cover all possibilites, including newer english, colloquial english, etc. To start, we create a function for making a text into a string of english letters without any special characters or spaces or numbers or capitals etc.

```
stringCleaner[mesg_] := ToLowerCase@
   StringJoin[StringSplit[mesg, {" ", ",", ".", "!", "?", ":", "'", "\"", ";", "-",
      "_", "(", ")", "*", "1", "2", "3", "4", "5", "6", "7", "8", "9", "0"}]];
```

```
cleanPride = stringCleaner[pride]
```

```
chapteritisatruthuniversallyacknowledgedthatasinglemaninpossessionofagoodfortunem⋱
   ustbeinwantofawifehoweverlittleknownthefeelingsorviewsofsuchamanmaybeonhisfir⋱
   stenteringaneighbourhoodthistruthissowellfixedinthemindsofthesurroundingfamil⋱
   iesthatheisconsideredtherightfulpropertyofsomeoneorotheroftheirdaughtersmydea⋱
   rmrbennetsaidhisladytohimonedayhaveyouheardthatnetherfieldparkisletatlastmrbe⋱
   nnetrepliedthathehadnotbutitisreturnedsheformrslonghasjustbeenhereandshetoldm⋱
   eallaboutitmrbennetmadenoanswerdoyounotwanttoknowwhohastakenitcriedhiswifeimp⋱
   atientlyyouwanttotellmeandihavenoobjectiontohearingitthiswasinvitationenoughw⋱
   hymydearyoumustknowmrslongsaysthatnetherfieldistakenbyayoungmanoflargefortune⋱
   fromthenorthofenglandthathecamedownonmondayinachaiseandfourtoseetheplaceandwa⋱
   ssomuchdelightedwithitthatheagreedwithmrmorrisimmediatelythatheistotakeposses⋱
   sionbeforemichaelmasandsomeofhisservantsaretobeinthehousebytheendofnextweekwh⋱
   atishisnamebingleyishemarriedorsingleohsinglemydeartobesureasinglemanoflargef⋱
   ortunefourorfivethousandayearwhatafinethingforourgirlshowsohowcanitaffectthem⋱
   mydearmrbennetrepliedhiswifehowcanyoubesotiresomeyoumustknowthatiamthinkingof⋱
   hismarryingoneofthemisthathisdesigninsettlingheredesignnonsensehowcanyoutalks⋱
   obutitisverylikelythathemayfallinlovewithoneofthemandthereforeyoumustvisithim⋱
   assoonashecomesiseenooccasionforthatyouandthegirlsmaygooryoumaysendthembythem⋱
   selveswhichperhapswillbestillbetterforasyouareashandsomeasanyofthemmrbingleym⋱
   aylikeyouthebestofthepartymydearyouflattermeicertainlyhavehadmyshareofbeautyb⋱
   utidonotpretendtobeanythingextraordinarynowwhenawomanhasfivegrownupdaughterss⋱
   heoughttogiveoverthinkingofherownbeautyinsuchcasesawomanhasnotoftenmuchbeauty⋱
   tothinkofbutmydearyoumustindeedgoandseemrbingleywhenhecomesintotheneighbourho⋱
   oditismorethaniengageforiassureyoubutconsideryourdaughtersonlythinkwhatanesta⋱
   blishmentitwouldbeforoneofthemsirwilliamandladylucasaredeterminedtogomerelyon⋱
   thataccountforingeneralyouknowtheyvisitnonewcomersindeedyoumustgoforitwillbei⋱
   mpossibleforustovisithimifyoudonotyouareoverscrupuloussurelyidaresaymrbingley⋱
```

```
willbeverygladtoseeyouandiwillsendafewlinesbyyoutoassurehimofmyheartyconsentt⋱
ohismarryingwhicheverhechoosesofthegirlsthoughimustthrowinagoodwordformylittl⋱
elizzyidesireyouwilldonosuchthinglizzyisnotabitbetterthantheothersandiamsures⋱
heisnothalfsohandsomeasjanenorhalfsogoodhumouredaslydiabutyouarealwaysgivingh⋱
erthepreferencetheyhavenoneofthemmuchtorecommendthemrepliedhetheyareallsillya⋱
ndignorantlikeothergirlsbutlizzyhassomethingmoreofquicknessthanhersistersmrbe⋱
nnethowcanyouabuseyourownchildreninsuchawayyoutakedelightinvexingmeyouhavenoc⋱
ompassionformypoornervesyoumistakememydearihaveahighrespectforyournervestheya⋱
remyoldfriendsihaveheardyoumentionthemwithconsiderationtheselasttwentyyearsat⋱
leastahyoudonotknowwhatisufferbutihopeyouwillgetoveritandlivetoseemanyyoungme⋱
noffourthousandayearcomeintotheneighbourhooditwillbenousetousiftwentysuchshou⋱
ldcomesinceyouwillnotvisitthemdependuponitmydearthatwhentherearetwentyiwillvi⋱
sitthemallmrbennetwassooddamixtureofquickpartssarcastichumourreserveandcapric⋱
ethattheexperienceofthreeandtwentyyearshadbeeninsufficienttomakehiswifeunders⋱
tandhischaracterhermindwaslessdifficulttodevelopshewasawomanofmeanunderstandi⋱
nglittleinformationanduncertaintemperwhenshewasdiscontentedshefanciedherselfn⋱
ervousthebusinessofherlifewastogetherdaughtersmarrieditssolacewasvisitingandn⋱
ewschaptermrbennetwasamongtheearliestofthosewhowaitedonmrbingleyhehadalwaysin⋱
tendedtovisithimthoughtothelastalwaysassuringhiswifethatheshouldnotgoandtillt⋱
heeveningafterthevisitwaspaidshehadnoknowledgeofititwasthendisclosedinthefoll⋱
owingmannerobservinghisseconddaughteremployedintrimmingahathesuddenlyaddresse⋱
dherwithihopemrbingleywilllikeitlizzywearenotinawaytoknowwhatmrbingleylikessa⋱
idhermotherresentfullysincewearenottovisitbutyouforgetmammasaidelizabeththatw⋱
eshallmeethimattheassembliesandthatmrslongpromisedtointroducehimidonotbelieve⋱
mrslongwilldoanysuchthingshehastwoniecesofherownsheisaselfishhypocriticalwoma⋱
nandihavenoopinionofhernomorehaveisaidmrbennetandiamgladtofindthatyoudonotdep⋱
endonherservingyoumrsbennetdeignednottomakeanyreplybutunabletocontainherselfb⋱
eganscoldingoneofherdaughtersdontkeepcoughingsokittyforheavenssakehavealittle⋱
compassiononmynervesyoutearthemtopieceskittyhasnodiscretioninhercoughssaidher⋱
fathershetimesthemillidonotcoughformyownamusementrepliedkittyfretfullywhenisy⋱
ournextballtobelizzytomorrowfortnightayesoitiscriedhermotherandmrslongdoesnot⋱
comebacktillthedaybeforesoitwillbeimpossibleforhertointroducehimforshewillnot⋱
knowhimherselfthenmydearyoumayhavetheadvantageofyourfriendandintroducemrbingl⋱
eytoherimpossiblemrbennetimpossiblewheniamnotacquaintedwithhimmyselfhowcanyou⋱
besoteasingihonouryourcircumspectionafortnightsacquaintanceiscertainlyverylit⋱
tleonecannotknowwhatamanreallyisbytheendofafortnightbutifwedonotventuresomebo⋱
dyelsewillandafterallmrslongandherdaughtersmuststandtheirchanceandthereforeas⋱
shewillthinkitanactofkindnessifyoudeclinetheofficeiwilltakeitonmyselfthegirls⋱
staredattheirfathermrsbennetsaidonlynonsensenonsensewhatcanbethemeaningofthat⋱
emphaticexclamationcriedhedoyouconsidertheformsofintroductionandthestressthat⋱
islaidonthemasnonsenseicannotquiteagreewithyoutherewhatsayyoumaryforyouareayo⋱
ungladyofdeepreflectioniknowandreadgreatbooksandmakeextractsmarywishedtosayso⋱
methingsensiblebutknewnothowwhilemaryisadjustingherideashecontinuedletusretur⋱
ntomrbingleyiamsickofmrbingleycriedhiswifeiamsorrytohearthatbutwhydidnotyoute⋱
llmethatbeforeifihadknownasmuchthismorningicertainlywouldnothavecalledonhimit⋱
isveryunluckybutasihaveactuallypaidthevisitwecannotescapetheacquaintancenowth⋱
eastonishmentoftheladieswasjustwhathewishedthatofmrsbennetperhapssurpassingth⋱
eresttthoughwhenthefirsttumultofjoywasovershebegantodeclarethatitwaswhatshehad⋱
expectedallthewhilehowgooditwasinyoumydearmrbennetbutiknewishouldpersuadeyoua⋱
tlastiwassureyoulovedyourgirlstoowelltoneglectsuchanacquaintancewellhowplease⋱
diamanditissuchagoodjoketoothatyoushouldhavegonethismorningandneversaidaworda⋱
boutittillnownowkittyyoumaycoughasmuchasyouchoosesaidmrbennetandashespokehele⋱
```

```
fttheroomfatiguedwiththerapturesofhiswifewhatanexcellentfatheryouhavegirlssai⋱
dshewhenthedoorwasshutidonotknowhowyouwillevermakehimamendsforhiskindnessorme⋱
eitherforthatmatteratourtimeoflifeitisnotsopleasanticantellyoutobemakingnewac⋱
quaintanceseverydaybutforyoursakeswewoulddoanythinglydiamylovethoughyouarethe⋱
youngestidaresaymrbingleywilldancewithyouatthenextballohsaidlydiastoutlyiamno⋱
tafraidforthoughiamtheyoungestimthetallesttherestoftheeveningwasspentinconjec⋱
turinghowsoonhewouldreturnmrbennetsvisitanddeterminingwhentheyshouldaskhimtod⋱
innerchapternotallthatmrsbennethoweverwiththeassistanceofherfivedaughterscoul⋱
daskonthesubjectwassufficienttodrawfromherhusbandanysatisfactorydescriptionof⋱
mrbingleytheyattackedhiminvariouswayswithbarefacedquestionsingenioussuppositi⋱
onsanddistantsurmisesbuteheludedtheskillofthemallandtheywereatlastobligedtoac⋱
ceptthesecondhandintelligenceoftheirneighbourladylucasherreportwashighlyfavou⋱
rablesirwilliamhadbeendelightedwithhimhewasquiteyoungwonderfullyhandsomeextre⋱
melyagreeableandtocrownthewholehemeanttobeatthenextassemblywithalargepartynot⋱
hingcouldbemoredelightfultobefondofdancingwasacertainsteptowardsfallinginlove⋱
andverylivelyhopesofmrbingleysheartwereentertainedificanbutseeoneofmydaughter⋱
shappilysettledatnetherfieldsaidmrsbennettoherhusbandandalltheothersequallywe⋱
llmarriedishallhavenothingtowishforinafewdaysmrbingleyreturnedmrbennetsvisita⋱
ndsataboutenminuteswithhiminhislibraryhehadentertainedhopesofbeingadmittedto⋱
asightoftheyoungladiesofwhosebeautyhehadheardmuchbuthesawonlythefathertheladi⋱
esweresomewhatmorefortunatefortheyhadtheadvantageofascertainingfromanupperwin⋱
dowthatheworeabluecoatandrodeablackhorseaninvitationtodinnerwassoonafterwards⋱
dispatchedandalreadyhadmrsbennetplannedthecoursesthatweretodocredittoherhouse⋱
keepingwhenananswerarrivedwhichdeferreditallmrbingleywasobligedtobeintownthef⋱
ollowingdayandconsequentlyunabletoacceptthehonouroftheirinvitationetcmrsbenne⋱
twasquitedisconcertedshecouldnotimaginewhatbusinesshecouldhaveintownsosoonaft⋱
erhisarrivalinhertfordshireandshebegantofearthathemightbealwaysflyingaboutfro⋱
moneplacetoanotherandneversettledatnetherfieldasheoughttobeladylucasquietedhe⋱
rfearsalittlebystartingtheideaofhisbeinggonetolondononlytogetalargepartyforth⋱
eballandareportsoonfollowedthatmrbingleywastobringtwelveladiesandsevengentlem⋱
enwithhimtotheassemblythegirlsgrievedoversuchanumberofladiesbutwerecomfortedt⋱
hedaybeforetheballbyhearingthatinsteadoftwelvehebroughtonlysixwithhimfromlond⋱
onhisfivesistersandacousinandwhenthepartyenteredtheassemblyroomitconsistedofo⋱
nlyfivealtogethermrbingleyhistwosistersthehusbandoftheeldestandanotheryoungma⋱
nmrbingleywasgoodlookingandgentlemanlikehehadapleasantcountenanceandeasyunaff⋱
ectedmannershissisterswerefinewomenwithanairofdecidedfashionhisbrotherinlawmr⋱
hurstmerelylookedthegentlemanbuthisfriendmrdarcysoondrewtheattentionoftheroom⋱
byhisfinetallpersonhandsomefeaturesnoblemienandthereportwhichwasingeneralcirc⋱
ulationwithinfiveminutesafterhisentranceofhishavingtenthousandayearthegentlem⋱
enpronouncedhimtobeafinefigureofamantheladiesdeclaredhewasmuchhandsomerthanmr⋱
bingleyandhewaslookedatwithgreatadmirationforabouthalftheeveningtillhismanner⋱
sgaveadisgustwhichturnedthetideofhispopularityforhewasdiscoveredtobeproudtobe⋱
abovehiscompanyandabovebeingpleasedandnotallhislargeestateinderbyshirecouldth⋱
ensavehimfromhavingamostforbiddingdisagreeablecountenanceandbeingunworthytobe⋱
comparedwithhisfriendmrbingleyhadsoonmadehimselfacquaintedwithalltheprincipal⋱
peopleintheroomhewaslivelyandunreserveddancedeverydancewasangrythattheballclo⋱
sedsoearlyandtalkedofgivingonehimselfatnetherfieldsuchamiablequalitiesmustspe⋱
akforthemselveswhatacontrastbetweenhimandhisfriendmrdarcydancedonlyoncewithmr⋱
shurstandoncewithmissbingleydeclinedbeingintroducedtoanyotherladyandspentther⋱
estoftheeveninginwalkingabouttheroomspeakingoccasionallytooneofhisownpartyhis⋱
characterwasdecidedhewastheproudestmostdisagreeablemanintheworldandeverybodyh⋱
opedthathewouldnevercomethereagainamongstthemostviolentagainsthimwasmrsbennet⋱
```

```
whosedislikeofhisgeneralbehaviourwassharpenedintoparticularresentmentbyhishav
ingslightedoneofherdaughterselizabethbennethadbeenobligedbythescarcityofgentl
ementositdownfortwodancesandduringpartofthattimemrdarcyhadbeenstandingneareno
ughforhertohearaconversationbetweenhimandmrbingleywhocamefromthedanceforafewm
inutestopresshisfriendtojoinitcomedarcysaidheimusthaveyoudanceihatetoseeyoust
andingaboutbyyourselfinthisstupidmanneryouhadmuchbetterdanceicertainlyshallno
tyouknowhowidetestitunlessiamparticularlyacquaintedwithmypartneratsuchanassem
blyasthisitwouldbeinsupportableyoursistersareengagedandthereisnotanotherwoman
intheroomwhomitwouldnotbeapunishmenttometostandupwithiwouldnotbesofastidiousa
syouarecriedmrbingleyforakingdomuponmyhonourinevermetwithsomanypleasantgirlsi
nmylifeasihavethiseveningandtherearesev eralofthemyouseeuncommonlyprettyyouare
dancingwiththeonlyhandsomegirlintheroomsaidmrdarcylookingattheeldestmissbenne
tohsheisthemostbeautifulcreatureieverbeheldbutthereisoneofhersisterssittingdo
wnjustbehindyouwhoisveryprettyandidaresayveryagreeabledoletmeaskmypartnertoin
troduceyouwhichdoyoumeanandturningroundhelookedforamomentatelizabethtillcatch
inghereyehewithdrewhisownandcoldlysaidsheistolerablebutnothandsomeenoughtotem
ptmeiaminnohumouratpresenttogiveconsequencetoyoungladieswhoareslightedbyother
menyouhadbetterreturntoyourpartnerandenjoyhersmilesforyouarewastingyourtimewi
thmemrbingleyfollowedhisadvicemrdarcywalkedoffandelizabethremainedwithnoveryc
ordialfeelingstoward
```

The next step is to extract every quadgram that exists in the string we have created. An example of the quadgrams of “chapter” is “chap”, “hapt”, “apte”, “pter”. Quadgrams however extend to the following words as well, since what we have is not individual words, but a long string of characters.

```
getQuadGram[mesg_] := StringTake[stringCleaner[mesg],
   Table[{i, i + 3}, {i, StringLength[stringCleaner[mesg]] - 3}]];
prideQuadGrams = getQuadGram[cleanPride]
```

```
{chap, hapt, apte, pter, teri, erit, riti, itis, tisa, isat, satr, atru, trut, ruth,
 uthu, thun, huni, univ, nive, iver, vers, ersa, rsal, sall, ally, llya, lyac, yack,
 ackn, ckno, know, nowl, owle, wled, ledg, edge, dged, gedt, edth, dtha, that,
 hata, atas, tasi, asin, sing, … 11 480 …, eliz, liza, izab, zabe, abet, beth,
 ethr, thre, hrem, rema, emai, main, aine, ined, nedw, edwi, dwit, with, ithn,
 thno, hnov, nove, over, very, eryc, ryco, ycor, cord, ordi, rdia, dial, ialf,
 alfe, lfee, feel, eeli, elin, ling, ings, ngst, gsto, stow, towa, owar, ward}

large output | show less | show more | show all | set size limit...
```

Now that we have all the quadgrams in those first 15000 character of Pride and Prejudice, we must create a score that each quadgram is “English” like. Our score this time will be the log base 10 of the number of times the quadgram appears over the total number of quadgrams. For example: $\log_{10}$[count(chap)/count(quadgrams)]

```
prideQuadGramsFreq =
 Reverse[Sort[Tally[getQuadGram[cleanPride]], #1[[2]] < #2[[2]] &]]
```

```
{{ther, 44}, {that, 36}, {ingl, 31}, {with, 30}, {ngle, 28}, {them, 24}, {bing, 24},
 {gley, 24}, {ofth, 23}, {benn, 23}, {enne, 23}, {nnet, 23}, {fthe, 22}, {mrbi, 22},
 {rbin, 22}, {nthe, 21}, {dthe, 21}, {eyou, 21}, {will, 20}, {have, 19}, {tion, 19},
 {tthe, 18}, {ould, 18}, {ethe, 17}, {efor, 17}, {ance, 17}, {athe, 16},
 ... 5801 ..., {cywa, 1}, {ywal, 1}, {doff, 1}, {offa, 1}, {ffan, 1}, {fand, 1},
 {ethr, 1}, {hrem, 1}, {rema, 1}, {emai, 1}, {main, 1}, {nedw, 1}, {ithn, 1},
 {thno, 1}, {hnov, 1}, {nove, 1}, {eryc, 1}, {ryco, 1}, {ycor, 1}, {cord, 1},
 {rdia, 1}, {dial, 1}, {ialf, 1}, {alfe, 1}, {lfee, 1}, {gsto, 1}, {stow, 1}}
```

large output | show less | show more | show all | set size limit...

```
probCalc[freqList_] := N[Log[10, freqList[[2]] / Length[prideQuadGrams]]]
probCalc[prideQuadGramsFreq[[1]]]
```

```
-2.41992
```

```
logProbCalc[freqList_] := Map[probCalc, freqList];
```

```
prideLogProb = logProbCalc[prideQuadGramsFreq]
```

```
{-2.41992, -2.50707, -2.57201, -2.58625, -2.61621, -2.68316,
 -2.68316, -2.68316, -2.70164, -2.70164, -2.70164, -2.70164,
 -2.72095, -2.72095, -2.72095, -2.74115, ... 5824 ..., -4.06337,
 -4.06337, -4.06337, -4.06337, -4.06337, -4.06337, -4.06337, -4.06337,
 -4.06337, -4.06337, -4.06337, -4.06337, -4.06337, -4.06337, -4.06337}
```

large output | show less | show more | show all | set size limit...

Now to make it a list of rules using Dispatch

```
ruleCreator[freqList_, logProb_] :=
 Dispatch[Thread[Flatten[Drop[#, {2}] & /@ freqList] → logProb]]
prideProb = ruleCreator[prideQuadGramsFreq, prideLogProb]
```

```
Dispatch[ Length: 5855
          Rules:
          ther → −2.42
          that → −2.51
          ingl → −2.57
          with → −2.59
          ngle → −2.62
            ⋮            ]
```

Our final score for a piece of text will be using the formula log[AB]=log[A]+log[B]. The score of a length of text will be the sum of the log probability of each quadgram divided by the length of the text. The higher the score, the more English it is. Feel free to try out some strings.

```
scoreEnglishProb[text_] := (score = 0;
  scoreList = "";
  scoreList = ReplaceAll[getQuadGram[text], prideProb];
  For[i = 1, i < Length[scoreList], i++, If[MatchQ[scoreList[[i]], _String],
    score = score + (-8.0), score = score + scoreList[[i]]]];
  score / Length[scoreList])
scoreEnglishProb["It will be no use to us, if
   twenty such should come, since you will not visit them"]
scoreEnglishProb[
 "ITWILLBENOUSETOUSIFTWENTYSUCHSHOULDCOMESINCEYOUWILLNOTVISITTHEM"]
scoreEnglishProb["Hello there, friend"]
scoreEnglishProb["Hola Amigo Como estas"]
```

```
-3.56499

-3.56499

-5.09663

-6.90997
```

As described before, our sample data size is severely compromised due to the limitations in computational power of my laptop. As such, many quadgrams that may be found in colloquial language or, in all honesty, any other text besides the first 10000+ characters of Pride and Prejudice, will not have a probability associated with it in our list of rules. As such, a common phrase such as “hello friend” does not even appear in our data, and receives a low score.