R Notebook

Code

Hide

```
library(knitr)
library(readtext)
library(corpora)
library(quanteda)
library(dplyr)
```

Hide

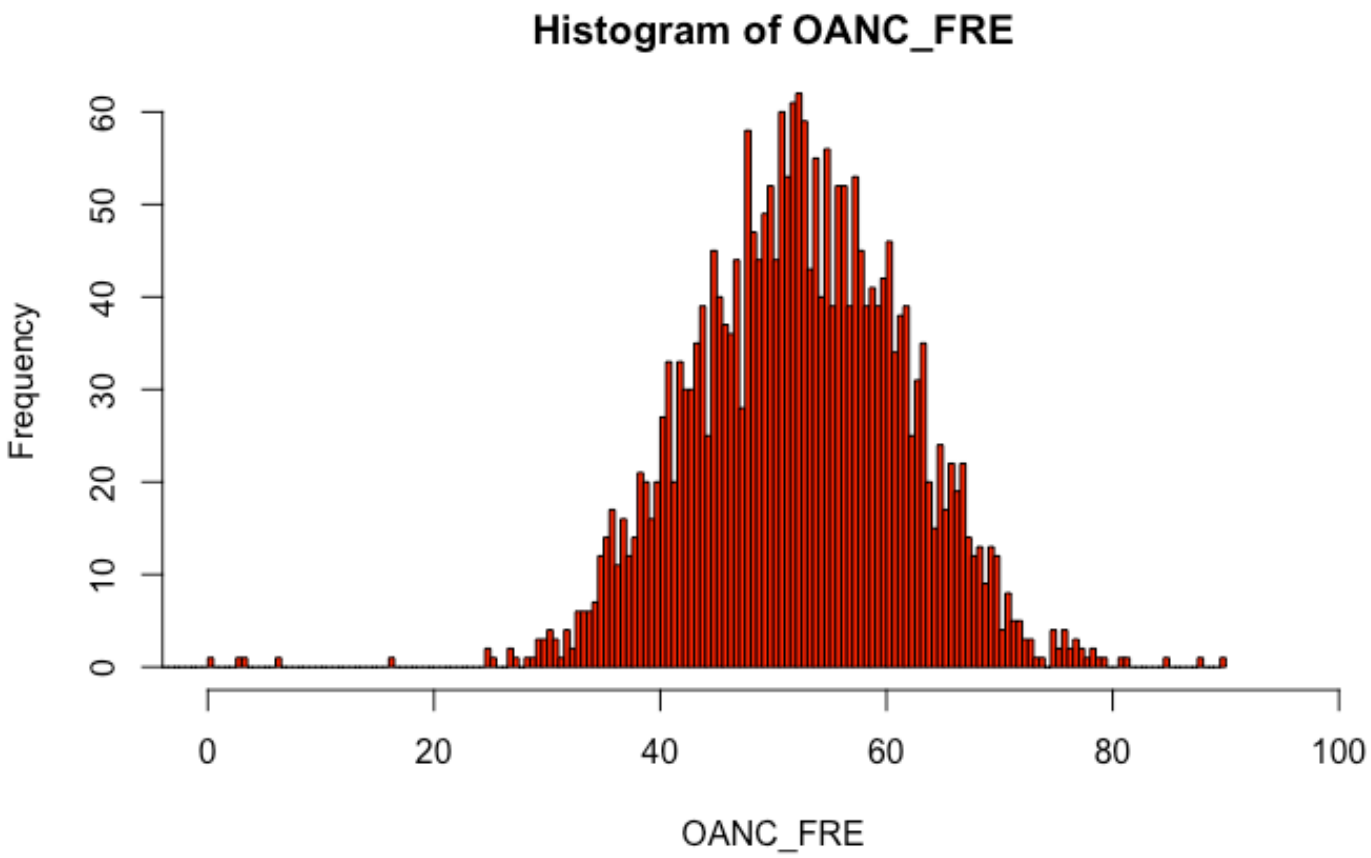
```
BNC_sub_corpus
```

```
Corpus consisting of 2,002 documents and 3 docvars.
```

Hide

```
OANC_corpus
```

```
Corpus consisting of 2,482 documents and 2 docvars.
```



Hide

```
OANC_FRE <- textstat_readability(texts(OANC_corpus, groups = "docvar2"), "Flesch")
hist(OANC_FRE, breaks=1000, col="red",xlim = c(0,100))
print(length((OANC_FRE)))
```

```
[1] 2480
```

Hide

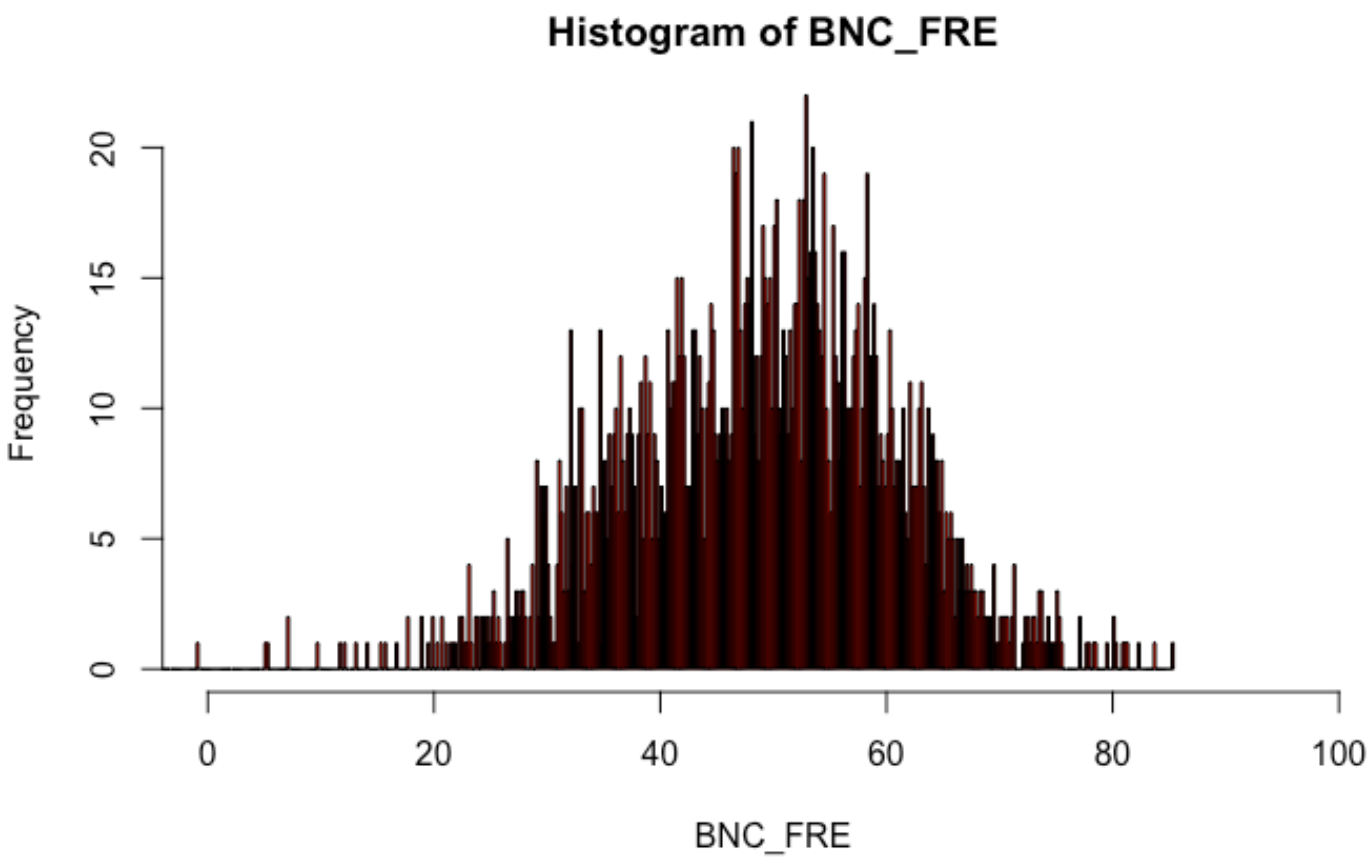
```
print(mean(OANC_FRE))
```

```
[1] 51.91524
```

Hide

```
print(var(OANC_FRE))
```

```
[1] 154.7285
```



Hide

```
t.test(BNC_FRE,OANC_FRE)
```

```
    Welch Two Sample t-test

data:  BNC_FRE and OANC_FRE
t = -8.4423, df = 4356.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.791732 -2.362555
sample estimates:
mean of x mean of y
 48.83810  51.91524
```

Hide

```
OANC_weight
```

```
      chatterbox            goes            soft              on         clinton
```

```
watching                a
    4.147260e-05     2.448858e-04     6.872602e-05     7.160146e-03     1.305399e-03
8.097032e-05     2.470108e-02
          c-span           replay             last            night               of
president        clinton's
    8.689497e-06     5.529680e-06     8.993630e-04     1.943288e-04     2.864019e-02
1.029705e-03     4.024817e-04
           press       conference            click             here               to
watch               it
    3.902374e-04     8.689497e-05     4.400046e-04     6.556621e-04     2.647690e-02
1.252078e-04     7.884139e-03
            read       transcript              was           struck               by
the            chief
    4.451393e-04     1.224429e-05     5.244507e-03     4.660730e-05     5.159192e-03
6.206553e-02     1.299475e-04
     executive's          comfort              and             ease          here's
sample         exchange
    1.579909e-06     3.752283e-05     2.359317e-02     2.804338e-05     7.781050e-05
3.120320e-05     8.610502e-05
               q          another         question            about     presidential
race            aside
    2.764840e-05     6.801507e-04     5.644224e-04     3.161792e-03     2.101278e-04
1.986735e-04     6.991096e-05
            from           asking           george                w             bush
come          forward
    3.962411e-03     8.492009e-05     2.887283e-04     1.425868e-04     3.258562e-04
4.510639e-04     6.951598e-05
            give        specifics           issues             that              you
mentioned            could
    3.839178e-04     4.344749e-06     2.152625e-04     1.430607e-02     4.062735e-03
4.937214e-05     1.121340e-03
            tell               us             what             find    objectionable
his           trying
    2.824087e-04     6.931849e-04     2.384477e-03     4.680479e-04     5.134703e-06
5.535210e-03     2.326415e-04
         present              new         moderate             face              for
party             just
    1.173082e-04     2.444909e-03     5.174201e-05     2.393562e-04     9.211657e-03
4.064315e-04     1.406119e-03
            like              did        democrats          whether           you're
worried               he
    1.803861e-03     8.638150e-04     2.476507e-04     4.708128e-04     2.934680e-04
5.016210e-05     5.470039e-03
            will           figure              out              how      republicans
can           occupy
    2.515215e-03     1.465365e-04     2.148676e-03     1.278541e-03     3.720685e-04
1.843358e-03     9.874429e-06
```

```
          center         american         politics               no               do
have              any
    1.283676e-04     8.187876e-04     2.737192e-04     2.267564e-03     1.544756e-03
4.351068e-03     1.048664e-03
           hints                i            don't            think           answer
those         questions
    1.895890e-05     5.038724e-03     9.538698e-04     7.788949e-04     4.498790e-04
9.475502e-04     2.725342e-04
        laughter              say             look              let               me
again         wouldn't
    1.105936e-05     8.661849e-04     4.107762e-04     2.749041e-04     9.834931e-04
3.566644e-04     1.520662e-04
            even            agree             with characterization             gave
my            first
    1.483139e-03     1.169132e-04     6.480390e-03     1.184931e-05     1.686552e-04
1.700772e-03     1.181377e-03
            when              ran               in             1991            thing
people          thought
    2.124977e-03     1.086187e-04     1.825624e-02     4.937214e-05     4.265753e-04
1.541201e-03     3.013676e-04
           going              our          country            would               if
remember             late
    4.467192e-04     1.095272e-03     3.637740e-04     2.372628e-03     2.596975e-03
1.216530e-04     2.263219e-04
         senator             paul         tsongas             were         actually
almost        ridiculed
    5.450685e-05     1.449566e-04     1.974886e-06     1.740664e-03     3.657488e-04
3.542945e-04     8.689497e-06
              at             time          because              we             both
put            these
    4.000724e-03     1.325938e-03     1.305794e-03     2.284153e-03     7.386073e-04
3.704886e-04     1.139114e-03
            very         detailed            plans               go             back
get             one
    5.853561e-04     3.120320e-05     1.105936e-04     5.766666e-04     6.694863e-04
9.669041e-04     2.870694e-03
             now              see        virtually       everything             said
except           things
    1.222454e-03     6.619817e-04     5.253196e-05     1.943288e-04     1.410858e-03
1.327123e-04     4.261803e-04
           tried         defeated          relaxed          posture           seemed
strikingly      reminiscent
    1.611507e-04     2.330365e-05     1.145434e-05     1.066438e-05     1.346872e-04
5.529680e-06     1.540411e-05
            john                f          kennedy            whose            image
has             been
    4.111712e-04     5.766666e-05     1.445616e-04     3.554794e-04     1.169132e-04
```

```
3.686322e-03     1.824399e-03
             all             over             tube              few             days
sad        aftermath
    2.627388e-03     1.090927e-03     1.145434e-05     5.869361e-04     3.712785e-04
4.147260e-05     1.066438e-05
           son's            death            who's              had          trouble
looking             bill
    1.224429e-05     3.570593e-04     8.926484e-05     2.080740e-03     1.074338e-04
1.575959e-04     5.075456e-04
              tv            least             year            found         himself
liking           what's
    2.144726e-04     4.155160e-04     6.517123e-04     3.294109e-04     3.965571e-04
3.949772e-06     2.045982e-04
      hypothesis               is           easier       popularity          ratings
down        according
    1.935388e-05     1.220163e-02     6.161644e-05     3.317808e-05     3.238813e-05
6.078698e-04     3.720685e-04
           roper          center's              job      performance             web
site          enjoyed
    1.579909e-06     1.974886e-06     2.828036e-04     1.177032e-04     3.570593e-04
3.092671e-04     3.752283e-05
    highest-ever          approval               73          percent           during
december          january
    3.949772e-07     4.976712e-05     9.479452e-06     6.272237e-04     3.926073e-04
2.922831e-05     4.463242e-05
        february             this           course      impeachment           senate
trial              bad
    2.646347e-05     4.415845e-03     4.719977e-04     2.448858e-04     2.725342e-04
1.789247e-04     3.558744e-04
        behavior       throughout          flytrap          scandal       especially
smug          refusal
    1.457466e-04     8.097032e-05     6.003653e-05     1.342922e-04     2.338265e-04
8.689497e-06     2.488356e-05
         confess          perjury      proceedings             made             him
easy          despise
    1.579909e-05     5.055708e-05     2.488356e-05     7.634908e-04     1.407304e-03
1.568059e-04     8.689497e-06
             yet            doubt          largely    congressional      overplaying
their             hand
    3.803630e-04     1.105936e-04     7.109589e-05     1.228379e-04     1.184931e-06
2.707963e-03     2.318516e-04
          public             more          support             than             ever
richard            nixon
    5.995753e-04     2.828826e-03     2.717443e-04     2.009249e-03     4.281552e-04
1.698402e-04     9.913927e-05
        personal             best               67               or          ronald
reagan               68
```

```
    2.614749e-04     4.759475e-04     9.479452e-06     3.616411e-03     4.502740e-05
1.101986e-04     8.294520e-06
            mass        adulation            being    retrospective       phenomenon
also              hit
    9.795433e-05     2.764840e-06     7.713904e-04     1.105936e-05     4.344749e-05
1.240623e-03     1.248128e-04
           early             1998            right            after            broke
instance         probably
    2.847785e-04     1.425868e-04     6.414429e-04     1.144249e-03     5.924657e-05
1.212580e-04     2.966278e-04
      expressing      disapproval               as            drama       dissipated
though         congress
    1.974886e-05     8.294520e-06     7.069696e-03     5.213698e-05     1.184931e-06
5.810114e-04     2.677945e-04
         started              lay              off           rating          dropped
since              may
    1.228379e-04     3.791781e-05     6.370982e-04     5.806164e-05     6.122146e-05
6.604018e-04     9.823082e-04
            it's          between               50               60            isn't
certainly         absolute
    1.610717e-03     6.860753e-04     1.145434e-04     8.255023e-05     3.384954e-04
1.591758e-04     2.448858e-05
          lowest               be             1993             1994          several
occasions               37
    1.974886e-05     4.875993e-03     7.346575e-05     9.163470e-05     2.839886e-04
2.053881e-05     1.540411e-05
   interestingly           period             felt     enthusiastic       presidency
still             vast
    9.874429e-06     1.003242e-04     1.153333e-04     1.777397e-05     6.003653e-05
7.208333e-04     6.754109e-05
           sweep          history     particularly            great            where
kennedy's              sat
    1.263927e-05     3.388904e-04     1.236278e-04     4.648881e-04     7.741552e-04
2.685845e-05     4.660730e-05
          higher         policies               an        situation             view
usually             much
    1.327123e-04     7.267580e-05     4.356598e-03     1.161233e-04     2.073630e-04
1.121735e-04     1.049059e-03
            hate      explanation      yesterday's            which              bit
reach              are
    7.662557e-05     8.886986e-05     4.186758e-05     2.351299e-03     1.745799e-04
6.477625e-05     4.773299e-03
            well              but             they             want           change
policy          dynamic
    7.038493e-04     5.488208e-03     3.206820e-03     6.118196e-04     2.010434e-04
3.325708e-04     1.579909e-05
        changing           should        continued         newfound            knows
```

```
thinking         reporter
    4.976712e-05     9.890228e-04     6.398630e-05     4.739726e-06     1.516712e-04
1.315274e-04     9.953424e-05
       reporters           always             take          popular          they're
trained       sympathize
    1.212580e-04     4.186758e-04     6.228790e-04     1.832694e-04     3.021575e-04
2.685845e-05     3.159817e-06
       underdogs          afflict      comfortable        afflicted            grant
some            crude
    2.369863e-06     2.369863e-06     3.831278e-05     1.066438e-05     4.463242e-05
1.477610e-03     2.685845e-05
      psychology             work             part        yesterday             less
salesman              apt
    2.211872e-05     6.679064e-04     4.321050e-04     8.413013e-05     5.770616e-04
1.658904e-05     1.263927e-05
      pissed-off             rude          willing     good-natured              way
confident              own
    2.764840e-06     1.856393e-05     1.074338e-04     4.739726e-06     1.026941e-03
2.646347e-05     8.112831e-04
    achievements          primary          message            spend          surplus
tax            cuts
    1.895890e-05     6.675114e-05     1.259977e-04     1.327123e-04     5.490182e-05
3.436301e-04     1.011142e-04
          didn't             seem           urgent              not            clear
house         majority
    4.380297e-04     2.966278e-04     1.540411e-05     4.627552e-03     2.223721e-04
7.567762e-04     1.279726e-04
        possibly          against           better         judgment            maybe
wasn't             such
    6.477625e-05     7.441370e-04     4.731826e-04     7.346575e-05     2.026233e-04
1.694452e-04     1.058934e-03
             guy            blame      blockbuster          earlier             week
members              los
    1.575959e-04     8.136529e-05     1.500913e-05     1.358721e-04     4.755525e-04
2.172374e-04     2.069680e-04
         angeles             film          critics      association        condemned
motion          picture
    1.990685e-04     3.799680e-04     2.721393e-04     8.215525e-05     2.685845e-05
3.594292e-05     1.516712e-04
         academy          america         refusing          stanley        kubrick's
eyes            wide
    3.080822e-05     3.207214e-04     2.962329e-05     3.475799e-05     7.504566e-06
1.477215e-04     6.043150e-05
            shut                r          without            bunch        ludicrous
digital      alterations
    5.964155e-05     3.949772e-05     5.241347e-04     4.068265e-05     1.421918e-05
3.357306e-05     1.184931e-06
```

```
      ritualized             orgy           yippee         occasion          bashing
board              saw
    1.579909e-06     9.479452e-06     3.949772e-07     3.831278e-05     8.294520e-06
9.005479e-05     1.335023e-04
          before              fig           leaves          applied         sequence
erotic           dental
    6.943698e-04     4.739726e-06     7.583561e-05     3.752283e-05     3.238813e-05
3.278310e-05     9.084475e-06
         surgery           that's          neither              nor            there
point          blaming
    3.554794e-05     4.716027e-04     1.216530e-04     1.536461e-04     1.680233e-03
4.968813e-04     1.856393e-05
           wrong         villains         scenario        anonymous              who
sit    sanctimonious
    2.749041e-04     1.342922e-05     2.764840e-05     3.317808e-05     3.138883e-03
5.608676e-05     2.764840e-06
      hypocrites          preside        country's          largest            video
media           chains
    2.369863e-06     4.739726e-06     6.793607e-05     5.055708e-05     9.716438e-05
3.088721e-04     1.974886e-05
           treat            rated          leprous       originally            meant
eliminate           stigma
    5.529680e-05     9.479452e-06     7.899543e-07     2.922831e-05     1.240228e-04
3.159817e-05     7.109589e-06
            came                x             same          refused            carry
ads            films
    2.717443e-04     3.475799e-05     6.469726e-04     6.043150e-05     6.082648e-05
1.145434e-04     9.874429e-05
          refuse             held          hostage               et              al
studio          million
    3.396804e-05     1.248128e-04     9.874429e-06     1.157283e-04     1.564110e-04
4.581735e-05     4.163059e-04
             100             line            feels        compelled             hack
movie             play
    1.220479e-04     2.903082e-04     8.650000e-05     1.500913e-05     2.053881e-05
5.193950e-04     2.164475e-04
         footsie            until           prides          itself           family
store           stroll
    3.159817e-06     3.333607e-04     3.554794e-06     2.812237e-04     3.752283e-04
7.623059e-05     4.739726e-06
         through           aisles            past         displays        wholesome
fare             meat
    5.861461e-04     3.159817e-06     3.215114e-04     1.816895e-05     8.294520e-06
9.479452e-06     3.357306e-05
         cleaver         massacre      innumerable        soft-core           flicks
wonder           caters
    1.974886e-06     3.317808e-05     5.924657e-06     3.159817e-06     5.924657e-06
```

```
1.074338e-04     3.949772e-07
           other          outfits             help            given       themselves
veneer    righteousness
    1.529352e-03     1.105936e-05     3.144018e-04     2.709543e-04     2.717443e-04
4.739726e-06     2.764840e-06
           while      trafficking             kind           sleaze            every
entertainment     conglomerate
    8.428812e-04     5.924657e-06     3.689087e-04     7.899543e-06     5.588927e-04
9.084475e-05     4.344749e-06
      capitalism           column            daily          variety           roger
ebert           argues
    4.384246e-05     1.722100e-04     2.223721e-04     6.517123e-05     4.147260e-05
3.041324e-05     9.953424e-05
       something            allow        audiences         original          movies
provide         category
    5.632374e-04     1.062489e-04     2.132877e-05     1.129635e-04     1.994635e-04
1.062489e-04     4.700228e-05
           adult       acceptable          theater           owners      advertising
outlets             shun
    4.621233e-05     2.922831e-05     8.413013e-05     4.937214e-05     9.005479e-05
1.777397e-05     3.554794e-06
            them        everybody      understands          concept      adults-only
remain         business
    1.454306e-03     8.886986e-05     2.330365e-05     4.858219e-05     3.949772e-07
8.768493e-05     4.147260e-04
      newspapers         stations             free     discriminate             wish
suggestion               so
    9.005479e-05     2.172374e-05     3.147968e-04     7.504566e-06     1.137534e-04
2.764840e-05     2.186594e-03
            sure          purpose            begin             does           really
trust      distinguish
    2.354064e-04     7.425571e-05     1.011142e-04     7.591461e-04     5.205799e-04
7.149086e-05     2.093379e-05
        artistic            merit            might       no-brainer         kubrick
imprimatur            south
    3.554794e-05     2.448858e-05     7.555913e-04     1.184931e-06     2.093379e-05
2.369863e-06     1.872192e-04
            park           bigger           longer            uncut            title
second          meaning
    6.872602e-05     5.924657e-05     1.643105e-04     4.739726e-06     8.334018e-05
3.088721e-04     1.062489e-04
       allegedly           passed         straight            heads              its
lines              sex
    3.357306e-05     9.163470e-05     7.149086e-05     6.556621e-05     2.152231e-03
1.149384e-04     2.903082e-04
        almighty             long              big        newspaper         announce
won't         carrying
```

```
    5.924657e-06     5.379589e-04     4.964863e-04     1.086187e-04     2.646347e-05
2.002534e-04     3.436301e-05
          larger            issue     distinctions          culture          release
angles            notes
    8.847488e-05     3.586393e-04     1.540411e-05     2.168425e-04     8.176027e-05
1.066438e-05     1.797146e-04
      disturbing           double         standard        continues         maintain
comes           sexual
    2.488356e-05     5.174201e-05     1.777397e-04     6.754109e-05     3.870776e-05
2.614749e-04     2.030183e-04
         violent          content             wild             west          example
opens         humorous
    4.976712e-05     9.044977e-05     6.991096e-05     1.591758e-04     3.298059e-04
3.199315e-05     6.714612e-06
    decapitation               pg             star            wars          episode
1          phantom
    1.184931e-06     1.579909e-06     2.500205e-04     7.267580e-05     5.806164e-05
5.032009e-04     2.014383e-05
          menace         contains         numerous           deaths             fond
citing           albert
    2.883333e-05     4.384246e-05     2.922831e-05     4.068265e-05     1.303425e-05
3.436301e-05     2.567352e-05
          brooks       experience             1985             lost       gorgeously
encapsulates            scene
    3.515297e-05     1.674703e-04     2.093379e-05     2.121027e-04     4.739726e-06
3.949772e-07     1.177032e-04
     protagonist             wife            julie          hagerty            wants
fuck             her
    2.448858e-05     2.101278e-04     1.303425e-05     3.949772e-07     2.022283e-04
1.382420e-05     2.297977e-03
          office             desk     unacceptable             used          context
pointed        character
    3.080822e-04     2.290867e-05     9.084475e-06     3.677237e-04     6.359132e-05
6.714612e-05     1.552260e-04
         instead           saying             gone    puritanically       intolerant
peculiar          notions
    3.317808e-04     3.230913e-04     1.402169e-04     3.949772e-07     4.344749e-06
2.330365e-05     1.856393e-05
        violence         director       frequently          tangled             told
instructive            story
    1.224429e-04     1.808995e-04     5.292694e-05     9.084475e-06     3.665388e-04
5.529680e-06     9.819132e-04
        research       fistfights       discovered              lot            punch
real             life
    1.528562e-04     7.899543e-07     5.727169e-05     2.595000e-04     1.974886e-05
4.475091e-04     7.804749e-04
        person's             nose             gets           broken          there's
```

```
blood             once
    2.369863e-05     4.423744e-05     2.219772e-04     5.569178e-05     3.712785e-04
9.558447e-05     3.965571e-04
         watched           friend            brief          remains             most
ghastly             i've
    2.804338e-05     2.057831e-04     5.450685e-05     1.015091e-04     1.482349e-03
5.529680e-06     1.808995e-04
            seen            let's          pretend            fight         purposes
making    realistically
    2.061781e-04     1.105936e-04     2.488356e-05     1.224429e-04     3.278310e-05
3.254612e-04     4.344749e-06
        horrible           anyone             sees            twice          getting
into            sound
    2.330365e-05     2.480457e-04     7.149086e-05     5.569178e-05     2.365913e-04
1.502493e-03     1.425868e-04
       cartilage         crunched             show           bright              red
pouring        someone's
    7.899543e-07     1.579909e-06     4.348698e-04     3.791781e-05     1.019041e-04
9.084475e-06     1.658904e-05
           close              up          fingers        beginning            swell
agony           likely
    1.951187e-04     1.805836e-03     2.053881e-05     9.992922e-05     5.134703e-06
7.899543e-06     2.322466e-04
       correctly             too         explicit         children           decide
restage             make
    2.251370e-05     8.638150e-04     3.199315e-05     3.562694e-04     7.228082e-05
3.949772e-07     9.641392e-04
        painless             ten          punches            forth             guys
walking             away
    3.949772e-06     3.357306e-05     8.294520e-06     3.633790e-05     7.465068e-05
4.265753e-05     3.262511e-04
     unblemished               g            folks           lesson             kids
funny        bloodless
    1.184931e-06     2.725342e-05     7.030593e-05     5.371689e-05     2.310616e-04
7.820548e-05     3.554794e-06
    consequences           matter          protect             next             step
ask        ourselves
    5.608676e-05     2.421210e-04     8.373516e-05     4.431644e-04     8.215525e-05
1.982785e-04     4.028767e-05
          cinema      exclusively           adults              why            can't
keep     stigmatizing
    2.093379e-05     1.856393e-05     4.976712e-05     7.520365e-04     4.451393e-04
2.768790e-04     7.899543e-07
       graveyard           spiral           whimsy             dear          cynthia
we're          talking
    2.369863e-06     7.109589e-06     4.739726e-06     3.689087e-04     4.344749e-06
2.077580e-04     1.378470e-04
```

```
          cyber         protocol           grouse        computers       constantly
internal       activities
    4.739726e-06     3.949772e-06     1.974886e-06     6.714612e-05     3.910274e-05
3.831278e-05     3.791781e-05
         anybody             know          signing            check           e-mail
gives       impression
    3.317808e-05     7.536164e-04     1.974886e-05     1.177032e-04     2.926781e-04
1.465365e-04     3.831278e-05
   inadvertently        launching          nuclear          missile          dialing
communicating           28,800
    9.084475e-06     1.421918e-05     1.082237e-04     5.687671e-05     5.134703e-06
5.924657e-06     1.579909e-06
        starting             ppp   authenticating           ending      inscrutable
status              box
    5.529680e-05     3.949772e-07     3.949772e-07     4.818721e-05     3.554794e-06
9.479452e-05     8.886986e-05
          graphs         blinking           lights          various   pronouncements
meanings            human
    2.764840e-06     3.949772e-06     2.567352e-05     1.354772e-04     9.479452e-06
6.714612e-06     3.298059e-04
       macarthur          crimson           clover           convoy            songs
battle          orleans
    3.554794e-06     2.369863e-06     7.899543e-07     3.554794e-06     4.700228e-05
8.452511e-05     6.319634e-06
             tie         kangaroo            sport           please           mister
custer           walter
    2.646347e-05     1.974886e-06     3.554794e-05     2.136826e-04     3.554794e-06
3.949772e-07     2.646347e-05
       brennan's    unforgettable        rendition             epic             ride
h            glenn
    7.899543e-07     2.764840e-06     3.159817e-06     2.843836e-05     3.317808e-05
3.436301e-05     1.856393e-05
            flip             side              old           rivers             pass
distance      discussions
    9.084475e-06     2.267169e-04     4.561986e-04     6.714612e-06     9.163470e-05
4.384246e-05     2.488356e-05
        suitably             high             note               jr        ephemeral
natures             soon
    1.974886e-06     3.649589e-04     2.030183e-04     7.978539e-05     4.344749e-06
1.974886e-06     1.753699e-04
             got             move       republican              cut          mexican
banking           crisis
    3.922123e-04     1.453516e-04     3.811530e-04     1.919589e-04     2.251370e-05
2.330365e-05     1.256027e-04
            fate           taiwan              use           phrase      unfortunate
only           effort
    4.147260e-05     3.357306e-05     4.431644e-04     7.307077e-05     2.172374e-05
```

```
1.553050e-03     1.145434e-04
         refrain        referring         favorite      publication             beer
frame          journal
    1.500913e-05     3.594292e-05     1.022991e-04     1.236278e-04     5.687671e-05
5.490182e-05     2.160525e-04
   inconspicuous      consumption          article      superiority          salute
products           edible
    1.579909e-06     2.922831e-05     3.752283e-04     1.105936e-05     7.899543e-06
8.965981e-05     7.899543e-07
       pillsbury         doughboy          charlie             tuna            slim
jim          finally
    7.899543e-07     1.974886e-06     2.804338e-05     3.159817e-06     8.689497e-06
6.714612e-05     1.789247e-04
      definitive          styptic           pencil        proclaims          editors
excellent             word
    1.461415e-05     7.899543e-07     6.714612e-06     9.084475e-06     1.568059e-04
5.687671e-05     2.997877e-04
             top  meaninglessness          welcome             try         meantime
stated           record
    2.429110e-04     1.184931e-06     5.253196e-05     1.943288e-04     9.084475e-06
2.883333e-05     1.414018e-04
            your             book         abortion          marvel        substance
necessarily           judged
    1.780162e-03     7.788949e-04     1.244178e-04     8.689497e-06     3.515297e-05
4.502740e-05     2.172374e-05
         lowbrow           online          company          you've          keeping
lard           pocket
    3.949772e-06     1.966986e-04     3.444201e-04     9.992922e-05     9.163470e-05
3.159817e-06     1.619406e-05
           steve               em        separated        difficult          imagine
working-class          radical
    8.452511e-05     4.621233e-05     1.263927e-05     1.078288e-04     1.101986e-04
8.689497e-06     5.411187e-05
          avowed          atheist          elected             th          century
sworn            guess
    3.554794e-06     1.974886e-06     6.161644e-05     2.456758e-04     2.721393e-04
1.303425e-05     7.307077e-05
           quite           modern       victorians         although        primitive
others             tom
    1.974886e-04     1.749749e-04     2.369863e-06     2.488356e-04     1.342922e-05
3.144018e-04     1.295525e-04
         delay's           recent           tirade         teaching        evolution
blamed        shootings
    7.899543e-06     3.404703e-04     2.764840e-06     3.001826e-05     3.475799e-05
3.238813e-05     1.303425e-05
       littleton          mention            month        mandating          posting
commandments          schools
```

```
    1.382420e-05     1.165183e-04     1.777397e-04     4.739726e-06     1.540411e-05
1.105936e-05     1.445616e-04
        recently       resolution          calling        americans            abase
god              140
    2.113128e-04     4.779224e-05     1.101986e-04     2.685845e-04     3.949772e-07
1.998584e-04     5.134703e-06
           voted          replace          capitol          revival             tent
dismiss          someone
    6.912100e-05     3.554794e-05     2.764840e-05     1.777397e-05     9.479452e-06
2.606849e-05     2.516004e-04
           delay        extremist             wake             many          pundits
politicians          disdain
    6.003653e-05     5.134703e-06     3.910274e-05     9.120022e-04     6.319634e-05
1.125685e-04     1.421918e-05
     creationism           joined         decrying         nihilism       supposedly
engendered          secular
    8.294520e-06     4.502740e-05     3.554794e-06     4.739726e-06     4.423744e-05
1.184931e-06     2.330365e-05
        disagree          alleged       skepticism       chattering          classes
perspective     professional
    3.515297e-05     6.872602e-05     2.014383e-05     2.764840e-06     2.448858e-05
3.870776e-05     7.662557e-05
      chatterers            today             tend      deferential         religion
occasionally           ponder
    3.949772e-07     3.819429e-04     7.860045e-05     2.369863e-06     5.450685e-05
4.147260e-05     5.924657e-06
     appropriate            place           rarely          express      irreverence
toward      established
    6.122146e-05     3.412603e-04     4.384246e-05     4.384246e-05     1.184931e-06
2.042032e-04     4.858219e-05
       religions             feel             mock              age       gladstones
scarce         pandering
    6.319634e-06     2.760890e-04     2.290867e-05     2.160525e-04     3.949772e-07
6.319634e-06     1.461415e-05
       preaching        loathsome           afraid          however        effective
comrades              far
    7.109589e-06     4.739726e-06     4.107762e-05     3.768082e-04     8.571004e-05
4.739726e-06     3.669338e-04
          aren't          playing       variegated    denominations            base
gore              aim
    1.331073e-04     1.173082e-04     1.579909e-06     2.369863e-06     4.542237e-05
2.922831e-04     2.251370e-05
        possible          traffic           mostly       ecumenical       banalities
virtue           values
    2.180274e-04     4.384246e-05     9.044977e-05     2.764840e-06     2.764840e-06
3.791781e-05     9.044977e-05
       godliness            talks    nonsensically          parents          struggle
```

```
confirms      religiosity
    1.184931e-06     7.425571e-05     3.949772e-07     1.907740e-04     5.450685e-05
8.294520e-06     1.974886e-06
       deploring           hollow       secularism         afflicts    unenlightened
fire
    2.764840e-06     1.421918e-05     2.764840e-06     3.554794e-06     7.899543e-07
9.242465e-05
 [ reached getOption("max.print") -- omitted 61116 entries ]
```

Choose the top 10000 features and sorted by difference ranking in American English and British English

	features	rank_in_american	rank_in_british	difference
	<fctr>	<dbl>	<dbl>	<dbl>
6575	percent	9854	336	9518
9379	u.s	9836	904	8932
2827	dole	9305	418	8887
7229	quiz	8780	171	8609
9647	vs	8986	427	8559
3972	globe	9177	823	8354
1553	cent	1560	9858	8298
8471	spin	9587	1574	8013
339	ads	9027	1021	8006
6841	posted	9324	1347	7977

1-10 of 12,603 rows Previous **1** 2 3 4 5 6 … 100 Next

look into where does these words pop up for words that doesn't show up:

Hide

```
print(setdiff(names(top_american)[100:300],names(top_british)[100:300]))
```

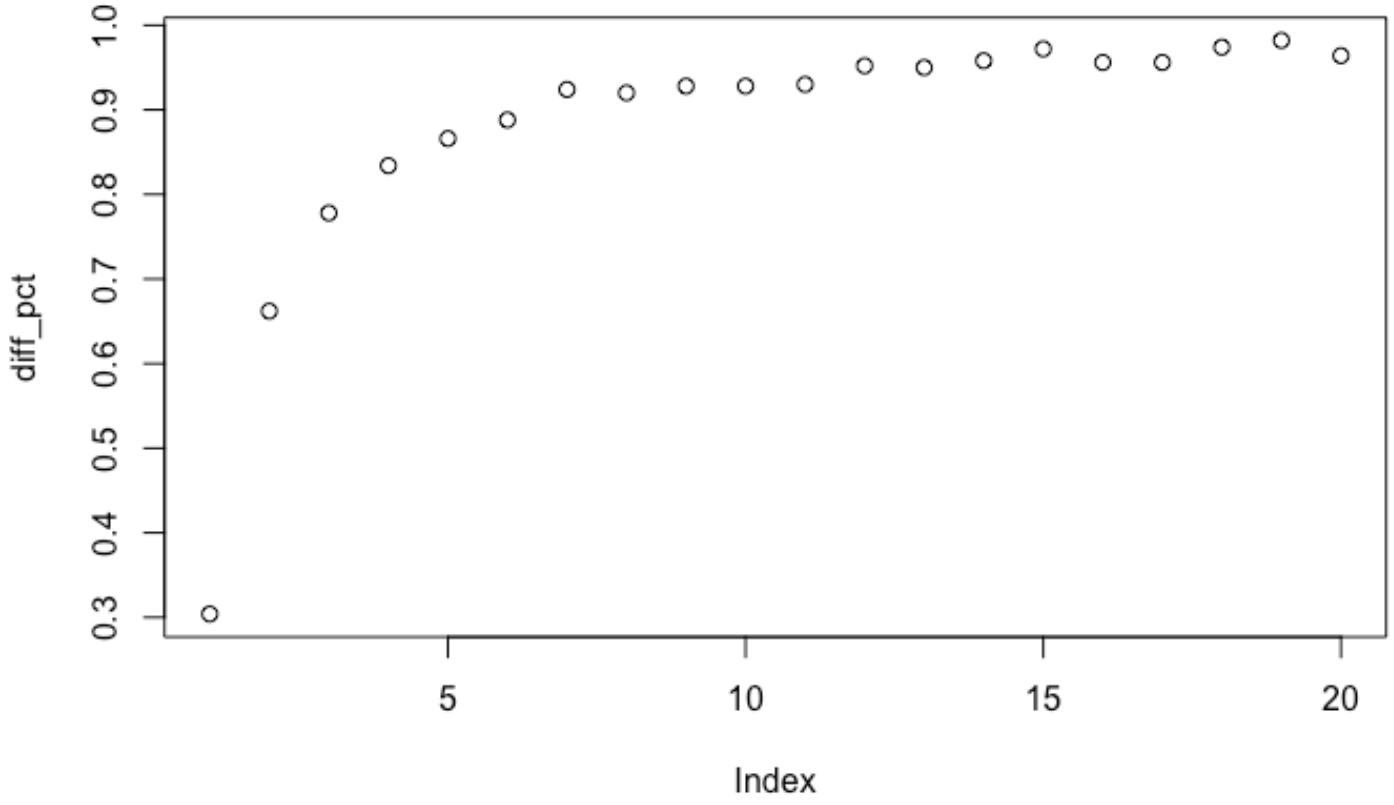
```
 [1] "those"       "then"        "many"        "york"        "american"    "book"
"where"       "being"
 [9] "made"        "why"         "slate"       "between"     "work"        "see"
"percent"     "white"
[17] "very"        "washington"  "u.s"         "question"    "something"   "states"
"news"        "really"
[25] "movie"       "doesn't"     "bill"        "called"      "big"         "paper"
"week"        "post"
[33] "better"      "that's"      "united"      "he's"        "seems"       "answer"
"campaign"    "real"
[41] "read"        "can't"       "click"       "didn't"      "show"        "ever"
"thing"       "things"
[49] "makes"       "clinton's"   "black"       "himself"     "got"         "press"
"today"       "republican"
[57] "film"        "name"        "among"       "article"     "republicans" "according
"   "love"        "there's"
[65] "kind"        "dear"        "enough"      "told"        "ago"         "actually"
"problem"     "internet"
[73] "issue"       "death"       "web"         "bad"         "whose"       "almost"
"perhaps"     "tax"
[81] "former"      "recent"      "history"     "isn't"       "piece"       "am"
"nothing"     "instead"
```

Hide

```
print(setdiff(names(top_british)[100:300],names(top_american)[100:300]))
```

```
 [1] "much"          "way"           "just"          "how"           "per"
"me"            "local"
 [8] "get"           "cent"          "within"        "british"       "although"
"it's"          "set"
[15] "different"     "general"       "system"        "second"        "group"
"given"         "important"
[22] "found"         "london"        "later"         "school"        "court"
"left"          "further"
[29] "order"         "example"       "four"          "information"   "members"
"market"        "early"
[36] "area"          "small"         "form"          "service"       "development"
"labour"        "council"
[43] "support"       "act"           "taken"         "help"          "education"
"large"         "major"
[50] "possible"      "police"        "minister"      "economic"      "3"
"interest"      "came"
[57] "change"        "p"             "following"     "already"       "seen"
"away"          "services"
[64] "control"       "took"          "research"      "five"          "particular"
"young"         "terms"
[71] "problems"      "available"     "working"       "held"          "d"
"period"        "society"
[78] "level"         "including"     "international" "community"     "english"
"went"          "full"
[85] "others"        "health"        "south"         "main"
```

compute the words difference in each language



Dale&Chall's Easy Word http://countwordsworth.com/blog/dale-chall-easy-word-list-text-file/ (http://countwordsworth.com/blog/dale-chall-easy-word-list-text-file/)

	features <fctr>	BNC_weight <dbl>	OANC_weight <dbl>	diff <dbl>
1698	of	5.475225e-02	4.404761e-02	1.070464e-02
2509	the	1.055940e-01	9.545461e-02	1.013935e-02
2507	that	1.518512e-02	2.200224e-02	6.817116e-03
1299	in	3.443451e-02	2.807746e-02	6.357055e-03
62	and	4.227725e-02	3.628547e-02	5.991782e-03
2700	was	1.262390e-02	8.065868e-03	4.558031e-03
171	be	1.082957e-02	7.499107e-03	3.330461e-03
2749	which	6.796766e-03	3.616216e-03	3.180550e-03
1221	his	5.870425e-03	8.512960e-03	2.642535e-03
1	a	3.545138e-02	3.798940e-02	2.538020e-03

1-10 of 2,848 rows Previous **1** 2 3 4 5 6 … 100 Next

is there any words outside the list is very frequent?

Hide

```
sum(dale_full_df_rm$diff)
```

```
[1] -1.152886e-17
```

boostrap

Hide

```
x
```

```
Document-feature matrix of: 2,002 documents, 40,436 features (90.8% sparse).
```