**CS5560 Knowledge Discovery Management**

**Lab Assignment 2**

**Lab Submission before September 29th 2018 (11:59 PM).**

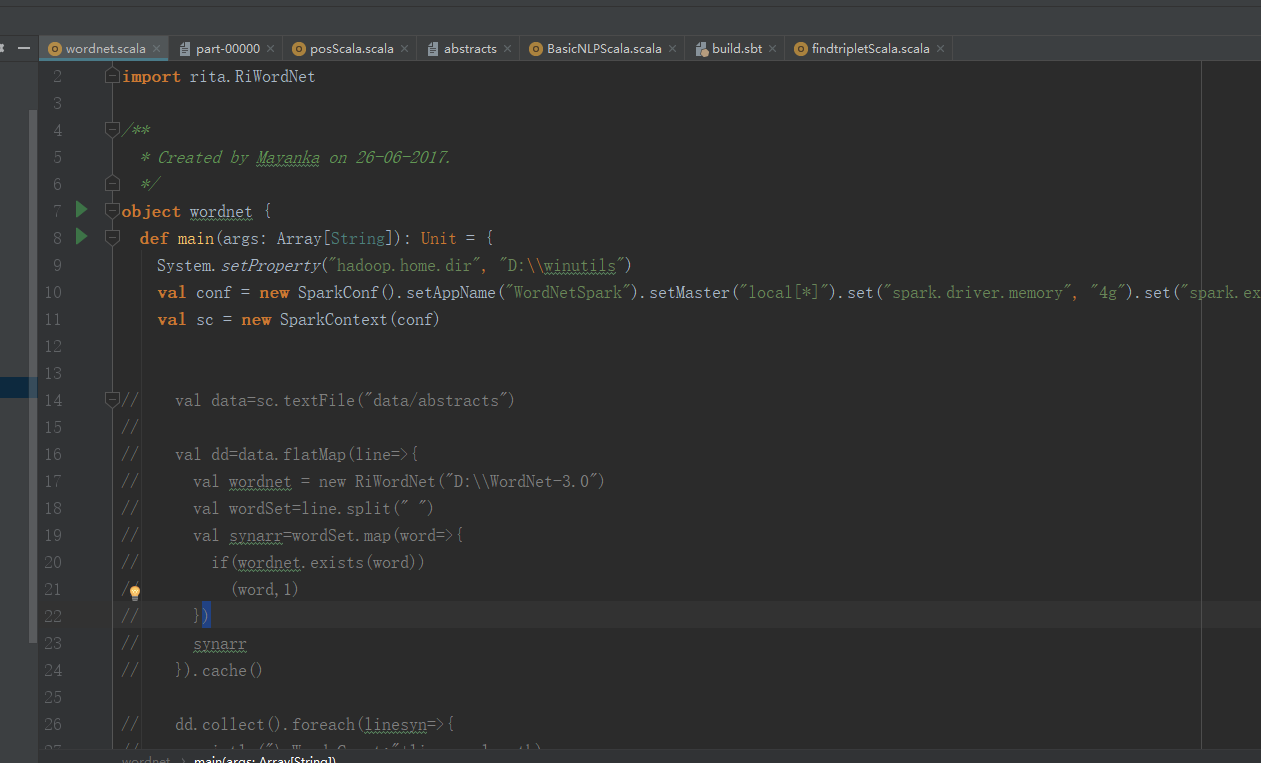
**YUXI MAO--6**

Submit a short lab report including screenshots and data description to Turnitin . Post your GitHub URL with Source Code (No wiki) through Lab 2 form at https://goo.gl/forms/MP2n7LQGmm2atsZo2

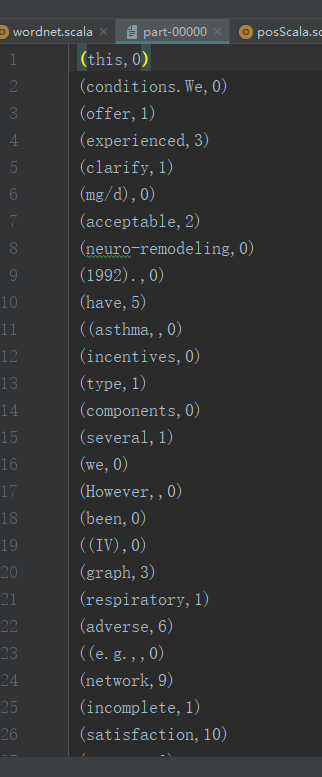
Consider 20 abstracts, write spark code for the following

1. ***Report Data Statistics as Lab Assignment 1 Based on Spark Code.***

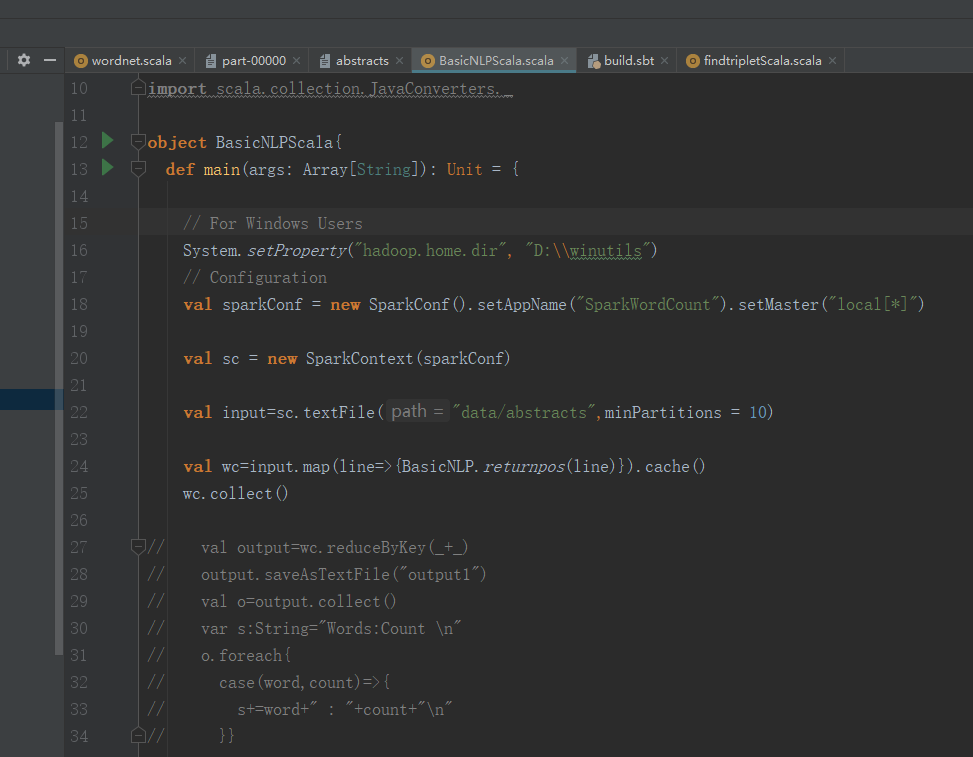
***Wordnet:***

******

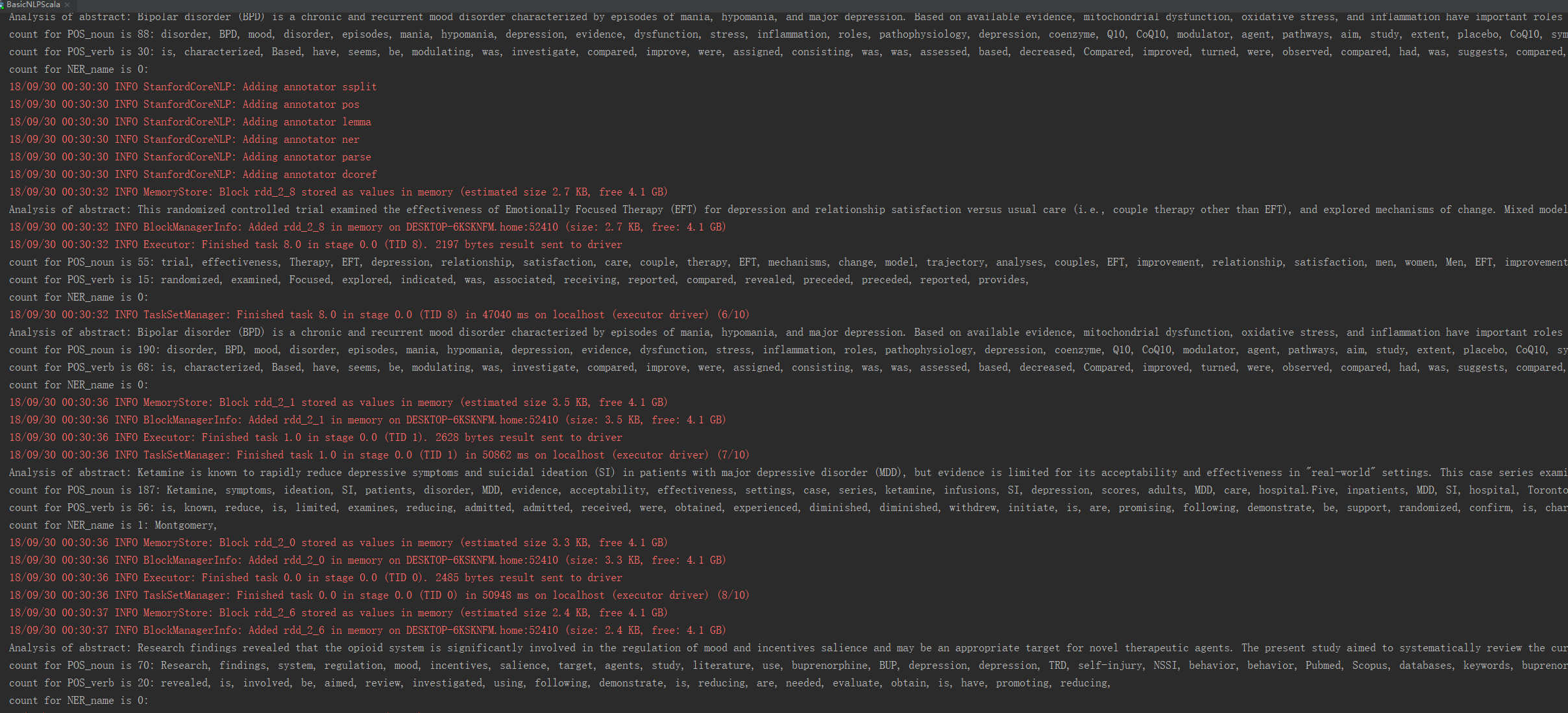
***Result:***

******

***basicNlp:***

******

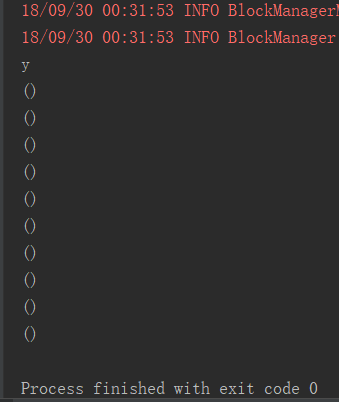
***Result:***

******

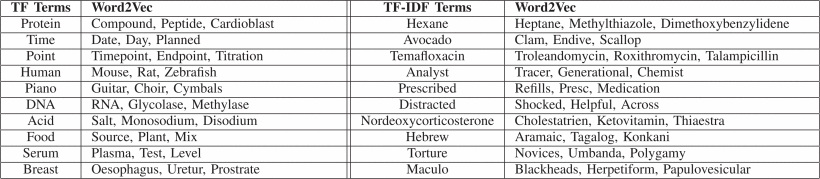
***findTripletScala:***

******

***Result:***

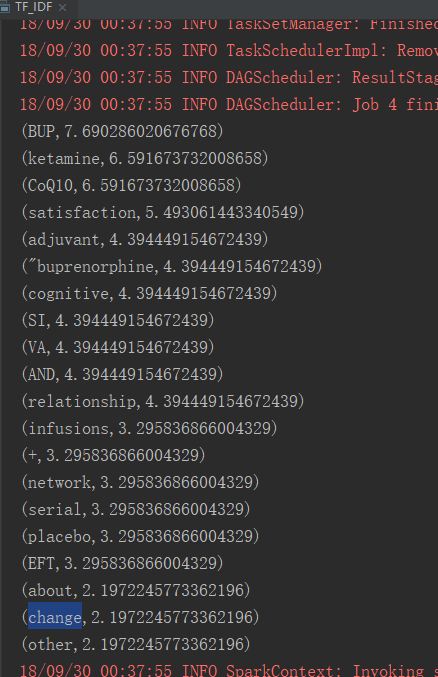
******

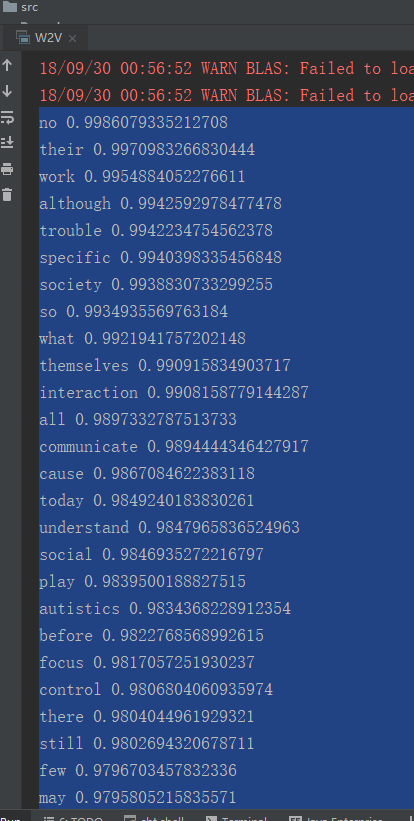
1. ***Word Statistics: Find out the top TF-IDF and corresponding Word2Vec words.***



|  |  |  |  |
| --- | --- | --- | --- |
| TF Terms | Woed2Vec | TF Terms | Woed2Vec |
| BUP |  | relationship |  |
| ketamine |  | infusions |  |
| CoQ10 |  | network |  |
| satisfaction |  | serial |  |
| adjuvant |  | placebo |  |
| buprenorphine |  | EFT |  |
| cognitive |  | about | local 0.9841563701629639  u 0.9839566946029663  c 0.9790564775466919  months 0.9778562784194946  man 0.9749913215637207  rise 0.9728242754936218  bc 0.9721618294715881  film 0.9655793905258179  important 0.9639584422111511  average 0.9589523673057556  with 0.9577916264533997  land 0.9557305574417114  name 0.9552657008171082  at 0.9543728232383728  environment 0.9539139270782471  island 0.9505839347839355  association 0.94980388879776  open 0.945336103439331  if 0.9445759654045105  difficulties 0.9405398964881897  form 0.9374629855155945  unrounded 0.9305189251899719  being 0.9281463027000427  lower 0.9203683137893677  from 0.9188121557235718  through 0.916932225227356  human 0.9125330448150635  particularly 0.9121606349945068  same 0.9105128049850464  examples 0.9090743660926819  during 0.9066547155380249  structures 0.9040712714195251  environmental 0.8981303572654724  however 0.8959760665893555  before 0.8907638788223267  control 0.8904466032981873  around 0.8903695344924927  little 0.88967365026474  even 0.88918137550354  few 0.888176441192627 |
| SI |  | change | system 0.9894139170646667  for 0.9887910485267639  according 0.9879485368728638  but 0.9868661761283875  part 0.9845648407936096  temperature 0.9842807054519653  regions 0.981249988079071  scale 0.9808996319770813  center 0.9800129532814026  typical 0.9798603057861328  because 0.9795851707458496  surface 0.9791500568389893  on 0.9788306355476379  light 0.978096604347229  color 0.9773216843605042  snow 0.9772832989692688  study 0.9749574065208435  size 0.972454309463501  trees 0.9716670513153076  than 0.9714263677597046  patterns 0.971346914768219  single 0.9712486863136292  direct 0.9708988070487976  as 0.9708071351051331  political 0.970360517501831  albedo 0.9701842665672302  earth 0.9701375961303711  also 0.9700176119804382  partly 0.9691537618637085  any 0.9682308435440063  known 0.9679285883903503  tropical 0.9670541286468506  when 0.9666322469711304  child 0.9665735960006714  covered 0.9660079479217529  significant 0.9643808007240295  community 0.9643326997756958  does 0.9609242677688599  public 0.9591552019119263  resulting 0.9590765833854675 |
| VA |  |  |  |
| AND | is 0.9666374325752258  became 0.966566801071167  to 0.951656699180603  greek 0.9511185884475708  capital 0.9306198358535767  condition 0.9282093644142151  word 0.9156255722045898  a 0.9127243757247925  effects 0.9084790349006653  universal 0.9039115309715271  back 0.9019659161567688  size 0.9017395377159119  postal 0.8946250677108765  effect 0.8912178874015808  later 0.8895605802536011  british 0.8820427060127258  set 0.8813002109527588  well 0.8805791139602661  part 0.8784922361373901  amount 0.8745626211166382  resulting 0.8730233907699585  radiation 0.8707491159439087  uae 0.8615657687187195  albedo 0.8594392538070679  in 0.8589170575141907  areas 0.8487327098846436  letter 0.8457925915718079  found 0.8456196784973145  of 0.8423497676849365  vowel 0.835583508014679  student 0.8344407081604004  symbol 0.8309782147407532  school 0.8304184079170227  surface 0.8299349546432495  global 0.8298646807670593  river 0.8265287280082703  low 0.8246257901191711  x 0.8239741325378418  dhabi 0.8215556740760803  area 0.8209086060523987 | other | no 0.9986079335212708  their 0.9970983266830444  work 0.9954884052276611  although 0.9942592978477478  trouble 0.9942234754562378  specific 0.9940398335456848  society 0.9938830733299255  so 0.9934935569763184  what 0.9921941757202148  themselves 0.990915834903717  interaction 0.9908158779144287  all 0.9897332787513733  communicate 0.9894444346427917  cause 0.9867084622383118  today 0.9849240183830261  understand 0.9847965836524963  social 0.9846935272216797  play 0.9839500188827515  autistics 0.9834368228912354  before 0.9822768568992615  focus 0.9817057251930237  control 0.9806804060935974  there 0.9804044961929321  still 0.9802694320678711  few 0.9796703457832336  may 0.9795805215835571  views 0.9793700575828552  daily 0.9792543649673462  activity 0.9785625338554382  non 0.9783878326416016  popular 0.9780572056770325  able 0.9772555232048035  however 0.9768030047416687  she 0.9759803414344788  interests 0.9753784537315369  behavior 0.974965512752533  stereotyped 0.9747211933135986  groups 0.9736477136611938  her 0.9736355543136597  uk 0.9730967283248901 |

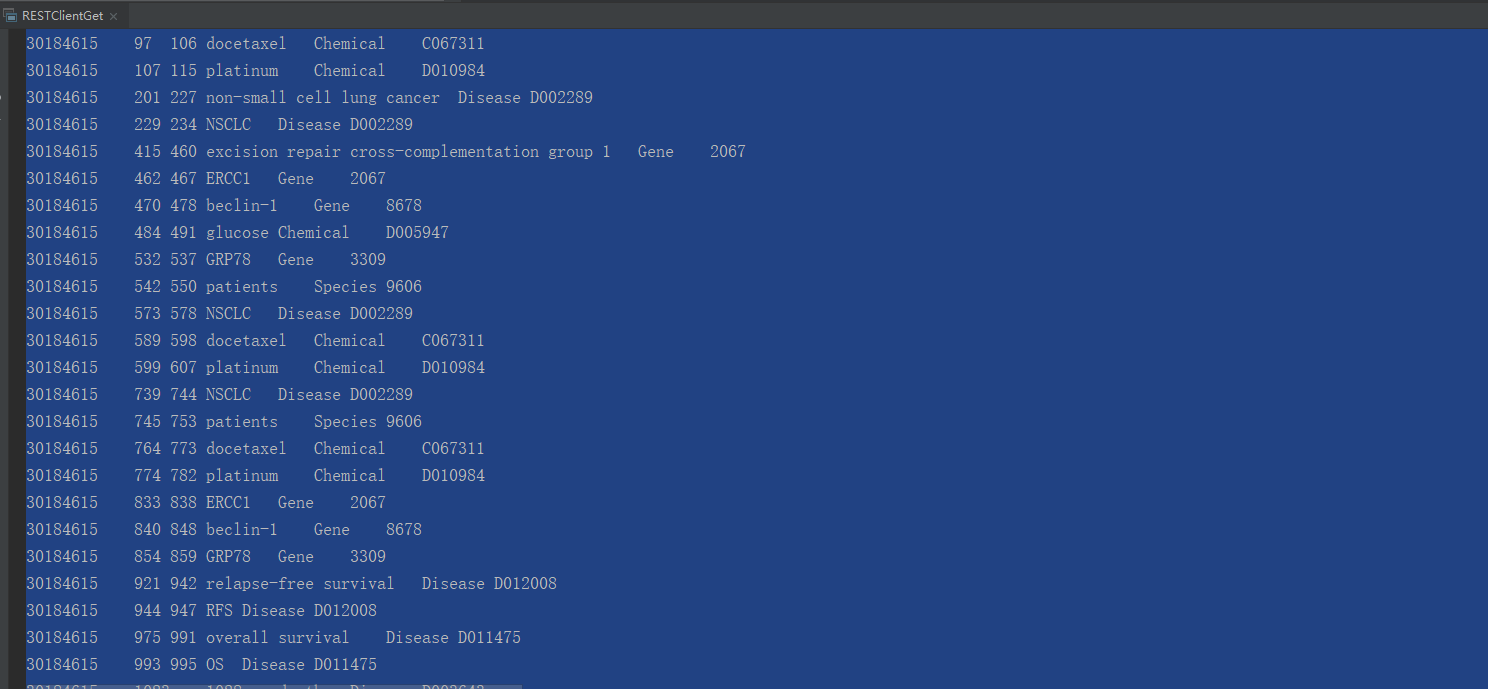
Screenshot for result:





1. ***Medical Word Statistics: Find out the top TF-IDF and corresponding Word2Vec words.***

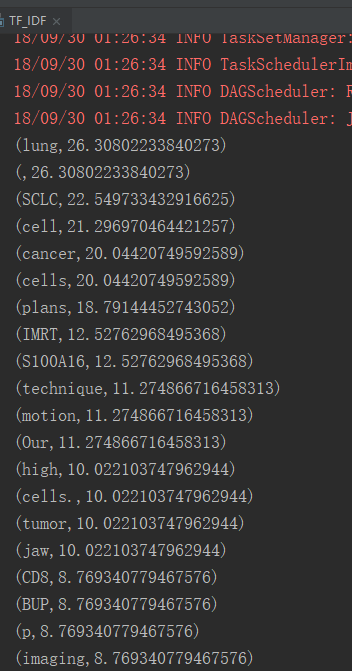
Get the medcial word:

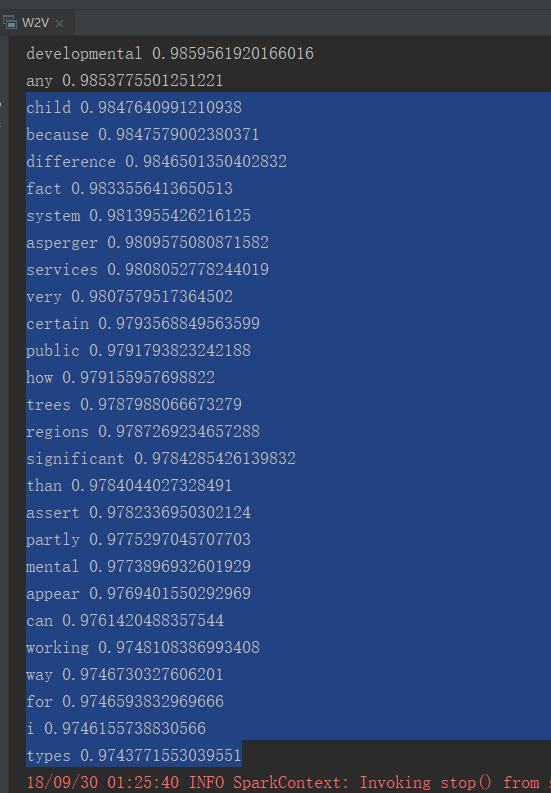


Then using those words for tfidf and w2v:

|  |  |  |  |
| --- | --- | --- | --- |
| TF-Terms | W2v | TF-Terms | W2v |
| lung |  | Our |  |
| SCLC |  | High | autism 0.9951085448265076  does 0.9950651526451111  com 0.9943599104881287  light 0.991661548614502  autistic 0.9901211857795715  single 0.9900636672973633  childhood 0.9896146059036255  scale 0.988215446472168  or 0.988158106803894  center 0.9880214333534241  tend 0.9873254299163818  body 0.9868731498718262  but 0.9864466786384583  developmental 0.9859561920166016  any 0.9853775501251221  child 0.9847640991210938  because 0.9847579002380371  difference 0.9846501350402832  fact 0.9833556413650513  system 0.9813955426216125  asperger 0.9809575080871582  services 0.9808052778244019  very 0.9807579517364502  certain 0.9793568849563599  public 0.9791793823242188  how 0.979155957698822  trees 0.9787988066673279  regions 0.9787269234657288  significant 0.9784285426139832  than 0.9784044027328491  assert 0.9782336950302124  partly 0.9775297045707703  mental 0.9773896932601929  appear 0.9769401550292969  can 0.9761420488357544  working 0.9748108386993408  way 0.9746730327606201  for 0.9746593832969666  i 0.9746155738830566  types 0.9743771553039551 |
| cell |  | Toumr |  |
| cancer |  | Jaw |  |
| cells |  | CD8 |  |
| plans |  | BUP |  |
| IMRT |  | P |  |
| S100A16 |  | imaging |  |
| technique |  | motion |  |

(maybe because of the bad data, most of them don’t have the w2v result)





1. ***Medical Word Statistics: Use Bio-portal API to report, the list of ontologies covered with corresponding count of the terms from the medical words in the abstracts.***

**Submission Guidelines (for all students):**

1. Submit your source code to GitHub.
2. Submit your documentation (submit your screenshots as well. The screenshot should have both the code and the output) to your Turnitin.
3. Comment your code appropriately.
4. Use the following google link to submit your assignment

(Lab Submission Link#) : https://goo.gl/forms/MP2n7LQGmm2atsZo2

***Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL:*** [*https://catalog.umkc.edu/special-notices/academic-honesty/*](https://catalog.umkc.edu/special-notices/academic-honesty/)