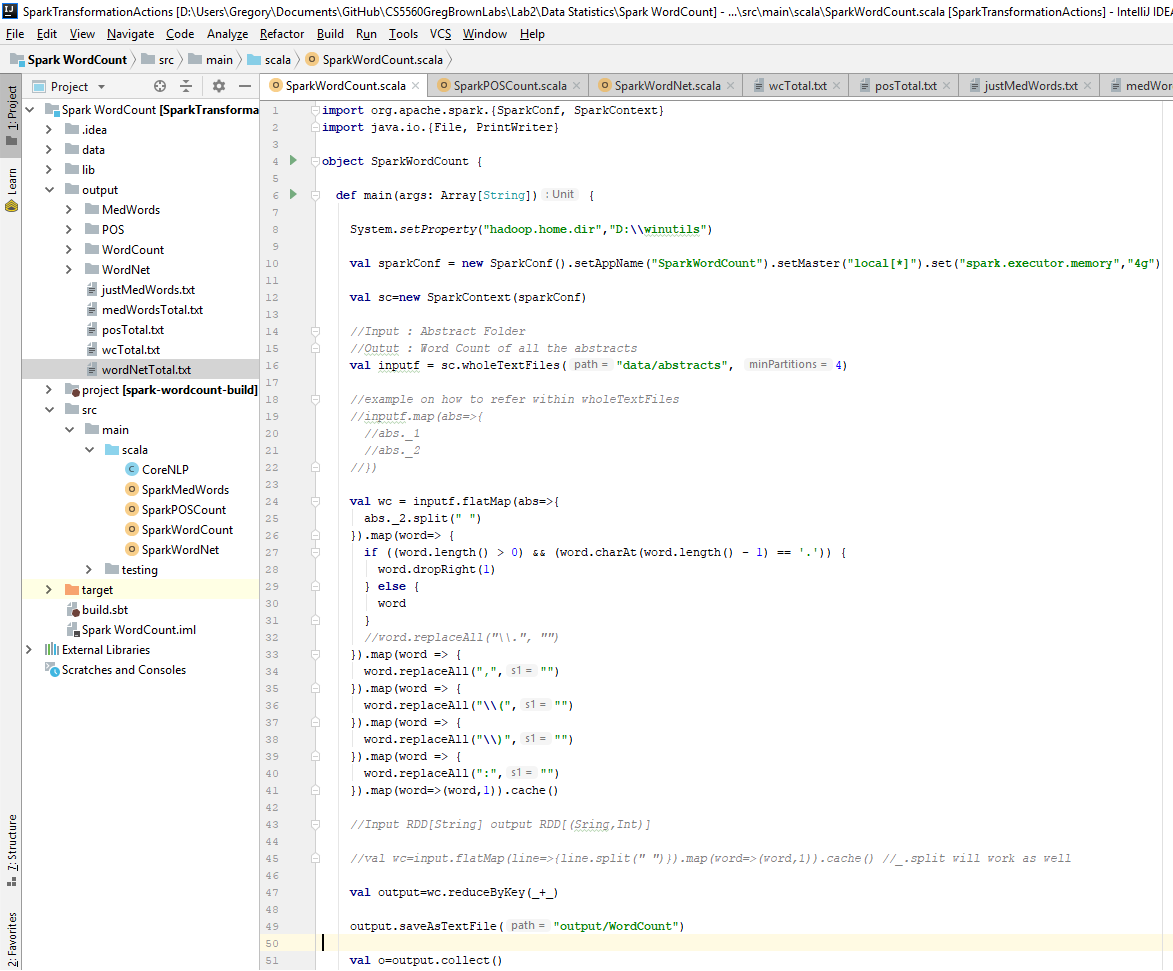
Greg Brown

Class ID: 1

Please find the discussed code and output text files at: <https://github.com/toadSTL/CS5560GregBrownLabs/tree/master/Lab2>

Please note that reports for ICP-4 and ICP-5 provide details on Spark code to report data statistics and to find top words via TF-IDF and synonyms for those words using Word2Vec. The explanation of these programs in this Lab write-up are abbreviated to avoid redundancy.

1. Below is a screenshot of the code for simple word counting:



As shown in this code, the first step to using Spark to process the data is to initialize the Spark context and take in the input. In the transition between ‘inputf’ and ‘wc’ pre-processing of the data is performed--including separation of words and removal of adjacent punctuation--and the words are mapped: ‘word’ -> ‘(word, 1)’. This value is the accumulated using the reduceByKey() function, to get the final count for each word. In the process of printing these word counts to a text file, the total word count is added-up.

The data statistics for Parts-of-Speech, WordNet Words, and Medical Words (via BioNLP) are collected in a similar way, varying in ways discussed in the ICP-4 Write-up. Below is a table of the 20 abstracts on which these data statistics programs were run, followed by a table showing the overall data statistics:

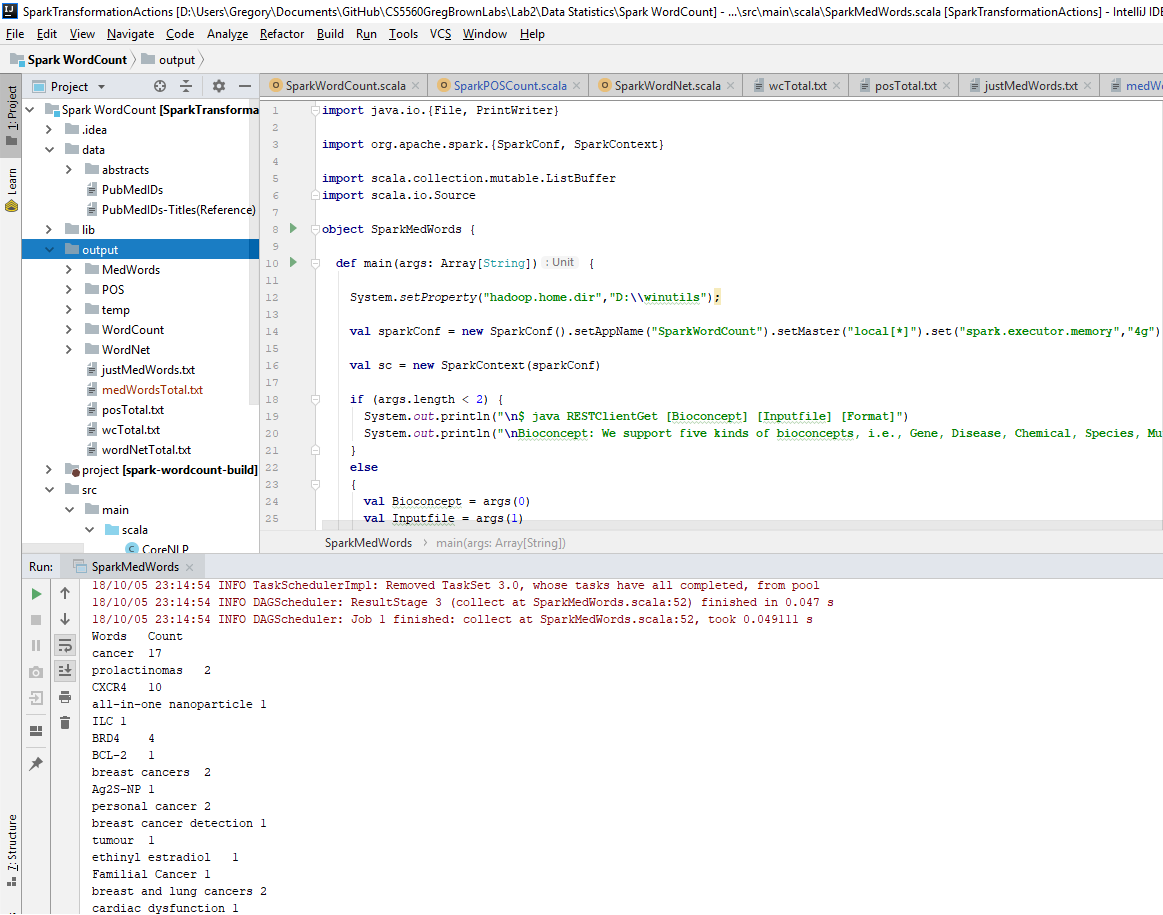
|  |  |  |
| --- | --- | --- |
| No. | PubMed ID | Title |
| 1 | 30196424 | Familial Cancer Clustering in Patients with Prolactinoma. |
| 2 | 30196358 | Seroma in breast surgery: all the surgeons fault? |
| 3 | 30194749 | Abbreviated MRI of the Breast: Does It Provide Value? |
| 4 | 30193687 | Hormonal contraception and breast cancer. |
| 5 | 30191351 | PET imaging of chemokine receptor CXCR4 in patients with primary and recurrent breast carcinoma. |
| 6 | 30191237 | An all-in-one nanoparticle (AION) contrast agent for breast cancer screening with DEM-CT-MRI-NIRF imaging. |
| 7 | 30190194 | Vitamin A and Breast Cancer Survival: A Systematic Review and Meta-analysis. |
| 8 | 30184043 | Breast cancer prognosis signature: linking risk stratification to disease subtypes. |
| 9 | 30181939 | Traumatic neuroma as a rare cause of intractable neuropathic breast pain following cancer surgery: Management and review of the literature. |
| 10 | 30180725 | Radiotherapy for patients with unresected locally advanced breast cancer. |
| 11 | 30262052 | A barcode mode based on glycosylation sites of membrane type mannose receptor as a new potential diagnostic marker for breast cancer. |
| 12 | 30261243 | Low-dose aspirin use and risk of contralateral breast cancer: A Danish nationwide cohort study. |
| 13 | 30260254 | Osteopontin-targeted probe detects orthotopic breast cancers using optoacoustic imaging. |
| 14 | 30259975 | BET proteins regulate homologous recombination-mediated DNA repair: BRCAness and implications for cancer therapy. |
| 15 | 30259416 | Emerging ways to treat breast cancer: will promises be met? |
| 16 | 30259342 | Risk perception and screening behavior of Filipino women at risk for breast cancer: implications for cancer genetic counseling. |
| 17 | 30258918 | Anti-HER-2 therapy following severe trastuzumab-induced cardiac toxicity. |
| 18 | 30257646 | Addition of triple negativity of breast cancer as an indicator for germline mutations in predisposing genes increases sensitivity of clinical selection criteria. |
| 19 | 30257411 | Troponin as a cardiotoxicity marker in breast cancer patients receiving anthracycline-based chemotherapy: A narrative review. |
| 20 | 30256567 | Aiding the Digital Mammogram for Detecting the Breast Cancer Using Shearlet Transform and Neural Network |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Words | Unique Words | Nouns | Verbs | Other (POS) | WordNet Words | Unique Wordnet Words | Medical Words | Unique Medical Words |
| 4536 | 1515 | 940 | 346 | 1120 | 2352 | 841 | 424 | 135 |

As the lists of Words and Wordnet Words would be very long, below is the list of medical words found corresponding to these articles. The lists of Words and WordNet recognized words can be found in the output folder of the ‘DataStatistics/SparkWordCount’ directory on github, or reproduced in the top level ‘output’ folder.

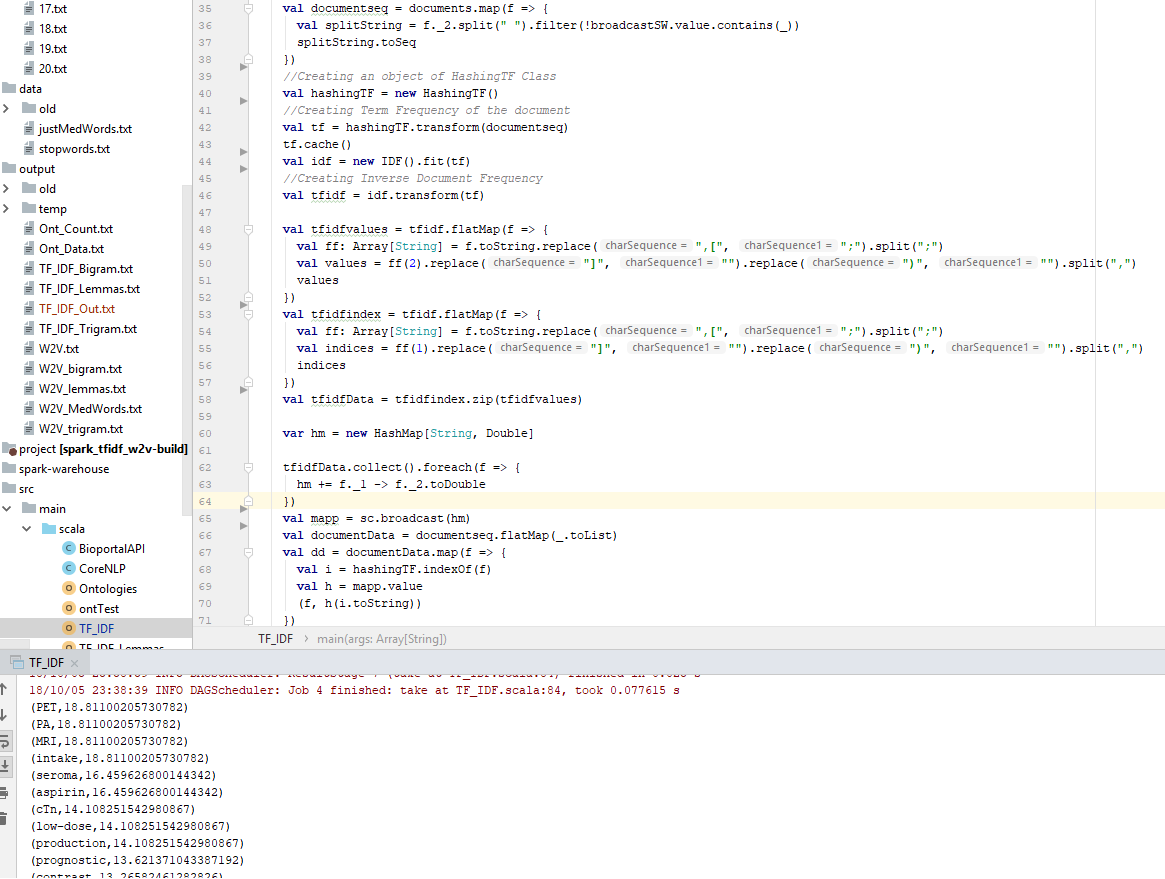
|  |
| --- |
| Words |
| cancer, prolactinomas, CXCR4, all-in-one nanoparticle, ILC, BRD4, BCL-2, breast cancers, Ag2S-NP, personal cancer, breast cancer detection, tumour, ethinyl estradiol, Familial Cancer, breast and lung cancers, cardiac dysfunction, colon cancer, PA, BC, RAD51, disease and fatal disease next to lung cancer, familial cancer, contralateral breast cancer, PI3K, epidermal growth factor receptor 2, BET, pituitary adenomas, BAX, colon cancers, Akt, increased total cancer, RAD51C, PR, metastasis, Triple Negative Breast Cancer, mTOR, TP53, disproportionate, MCL-1, CDH1, Chemical, anthracyclines, Colorectal cancer, ray, HER-2, triple negative breast cancer, ATM, NBN, malignancy, Breast cancer, breast cancer, silver sulfide, RAD51D, Troponin, MEK, invasive lobular carcinoma, placed in breast and axilla, Breast Cancer, levonorgestrel, BRCA1, PALB2, VEGFA, Seroma, Prolactinoma, 18F-FDG, prolactinomavcardiotoxicities, TNBC, capecitabine, aspirin, lapatinib, HER1, Osteopontinvcardiac toxicity, osteopontin, platinum, VEGFB, triple-negative breast cancers, BRCA2, human epidermal growth factor receptor 2, aid, tumors, breast tumor, Disease, breast cancer diagnosis, CBC, breast carcinoma, Species, invasive carcinoma, ErbB, progesterone, CDDP, chest wall disease, cardiotoxicity, BCL-W, DiR, estrogen, VEGFC, mannose, MR, Gene, HER2, patient, human, ERK, acromegaly, p53, HER3, patients, DCE, Patients, cancer cell-related components, BRCA1/2, c.1813dupA, relapsed disease, women, VEGF, Lung cancer, seroma production, lung cancer, nude mice, BCL-X, chest wall lesions, tumor, CHEK2, MAPK, visual Fatigue, People, HER4, AION, PGF, GSK525762A, deaths, Raf, toxicity |

Finally the following screenshot shows the Spark script for Medical Words alongside the start of the output:

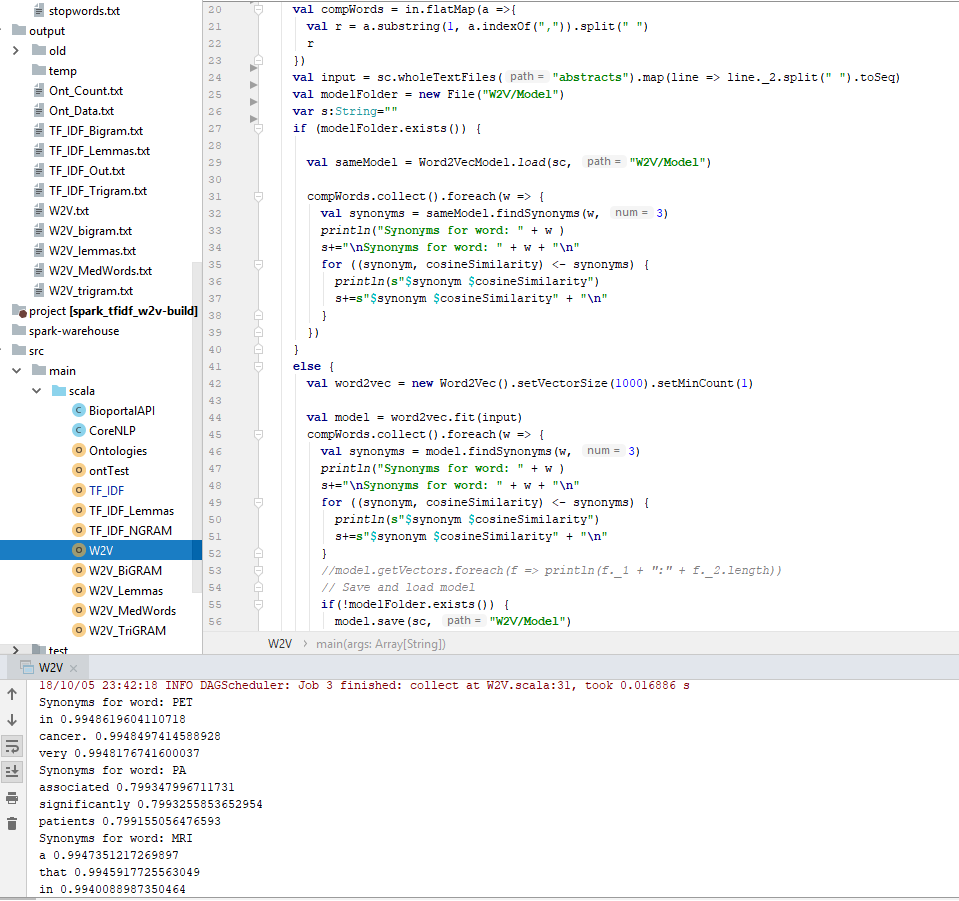


2) Word Statistics: Top TF-IDF word and corresponding Synonyms via Word2Vec

Similar to the Spark scripts for the data statistics, the TF-IDF and Word2Vec programs begin with initialization of a Spark context and reading of the input. As discussed in the write-up for ICP-5, next Term Frequency is created for the input, then Inverse Document Frequency is created using that Term Frequency, and all of this is done using library functions from Spark. Then the value and index of the Term Frequency and Inverse Document Frequency are added to the TF-IDF hashmap. Finally the top TF-IDF words are collected and output to a text file. The screenshot on the subsequent page shows the code with output for the TF-IDF

The full output for TF-IDF for the input abstracts can be found in the ‘output’ folder of the ‘Spark\_TF-IDF\_W2V’ folder or in the top level folder ‘output’

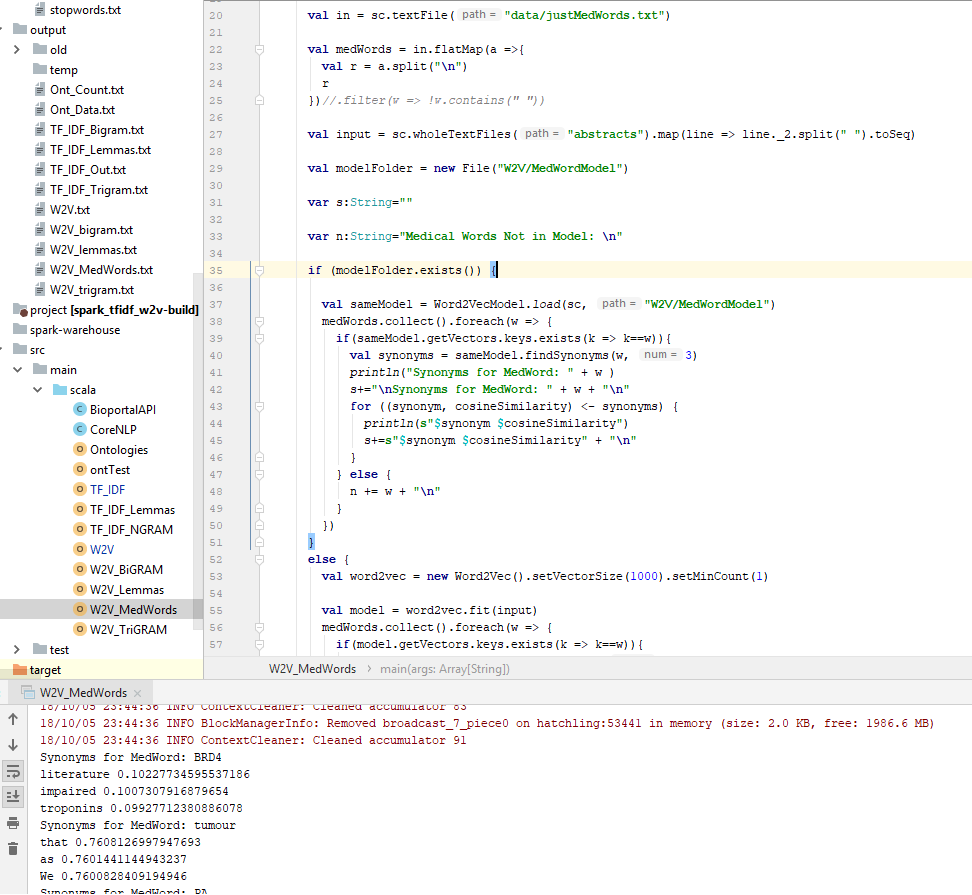
Next, using the output from the TF-IDF we perform Word2Vec. As with TF-IDF, more detailed description can be found in the write-up of ICP-5. The screenshot on the next page shows the code and output of Word2Vec. First the input is split into individual words, and then a Word2Vec model is trained (if one did not already exist) using the input. For each of the top TF-IDF words, which are read into the compWords variable, the top 3 synonyms are returned and output. Still the synonyms found using word2vec, do not appear to be very accurate, however, in discussion with other student, I learned that it may work better for words which are further down the list of TF-IDF terms, so in the future I will attempt to do Word2Vec, with a different set of TF-IDF terms.



The full output of the Word2Vec can be found in the same folders as mentioned for TF-IDF of the github repository for this lab.

3) Medical Word Statistics: Top TF-IDF medical words and Synonyms via Word2Vec

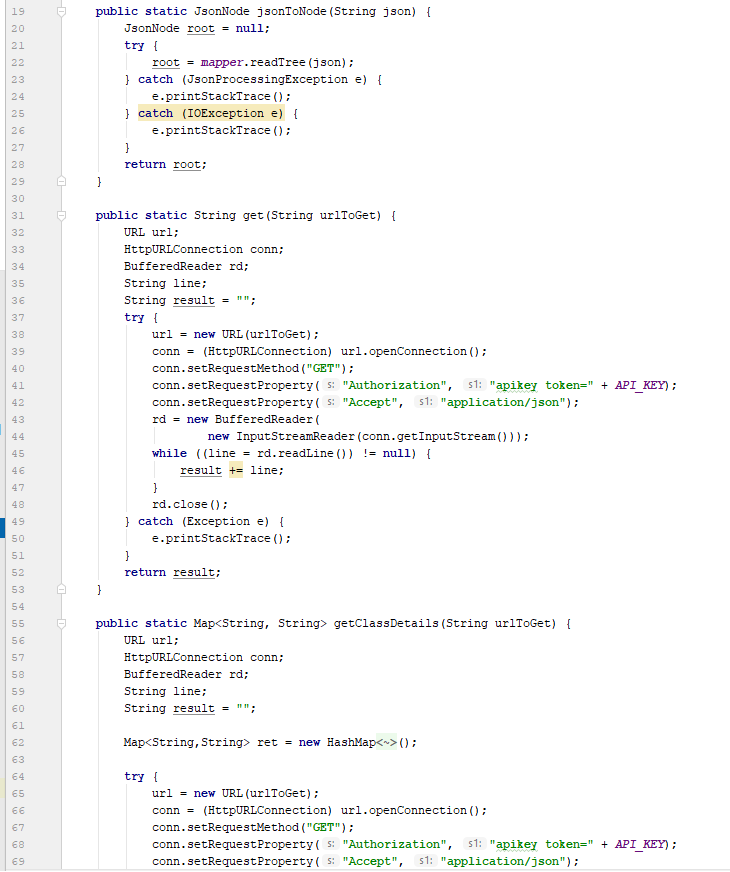
On the next page there is the is the code for the Word2Vec for medical words, along with the output of that program. The major difference between this program and the previously discussed word2vec program, is that, since the medical words are may not appear exactly as they appear in the text, I needed to check if the the medical word being compared is in the vocabulary. Thus the checks on lines 39 and 57 of the code were added, which explicitly check if the medical word for which we are seeking synonyms is in the vocabulary of the Word2Vec model.



Other than that, and the fact that we are comparing with medical terms, instead of the top TF-IDF terms this program is very similar to the previous Word2Vec program. The full output of this program can be found in the same folders as mentioned for TF-IDF and Word2Vec in the linked github repository.

4) Ontologies

Lastly I find the corresponding ontologies for the medical terms found from the document. For this program we will use the same input as for we did in the Medical Word Word2Vec program. However, this program calls the bioportalAPI and to facilitate this I used helper functions defined in java. The following page shows a screenshot of those helper functions: jsonToNode(), get() and getClassDetails().



These functions are called in the ontologies scala program which takes the medical words found from the abstracts and manipulates them to include them in the url for the api call. The get() call returns a json object which is then converted to a node via jsonToNode(). For each term the class details are obtained via getClassDetails(). These class details, generally listing the ontology, the preferred term-name within the ontology, and the url of the ontology, are made into a RDD of ‘(preferred term-name, ontology url)’ which is output to a text file.



The output of this program can be found, in the same directories as listed for the TF-IDF, Word2Vec, and Medical Words Word2Vec.