Maciej Medyk – COT6930 – Natural Language Processing – Homework 02

Question 1 – [10.00pt] – Write a report including the screenshots of generating document-word matrix, input data files for training and testing, command lines you use to train and test, and the output results by libsvm.

After running svm-train.exe with parameter -t0 for linear kernel and -s0 for C-SVC the results were as follows:

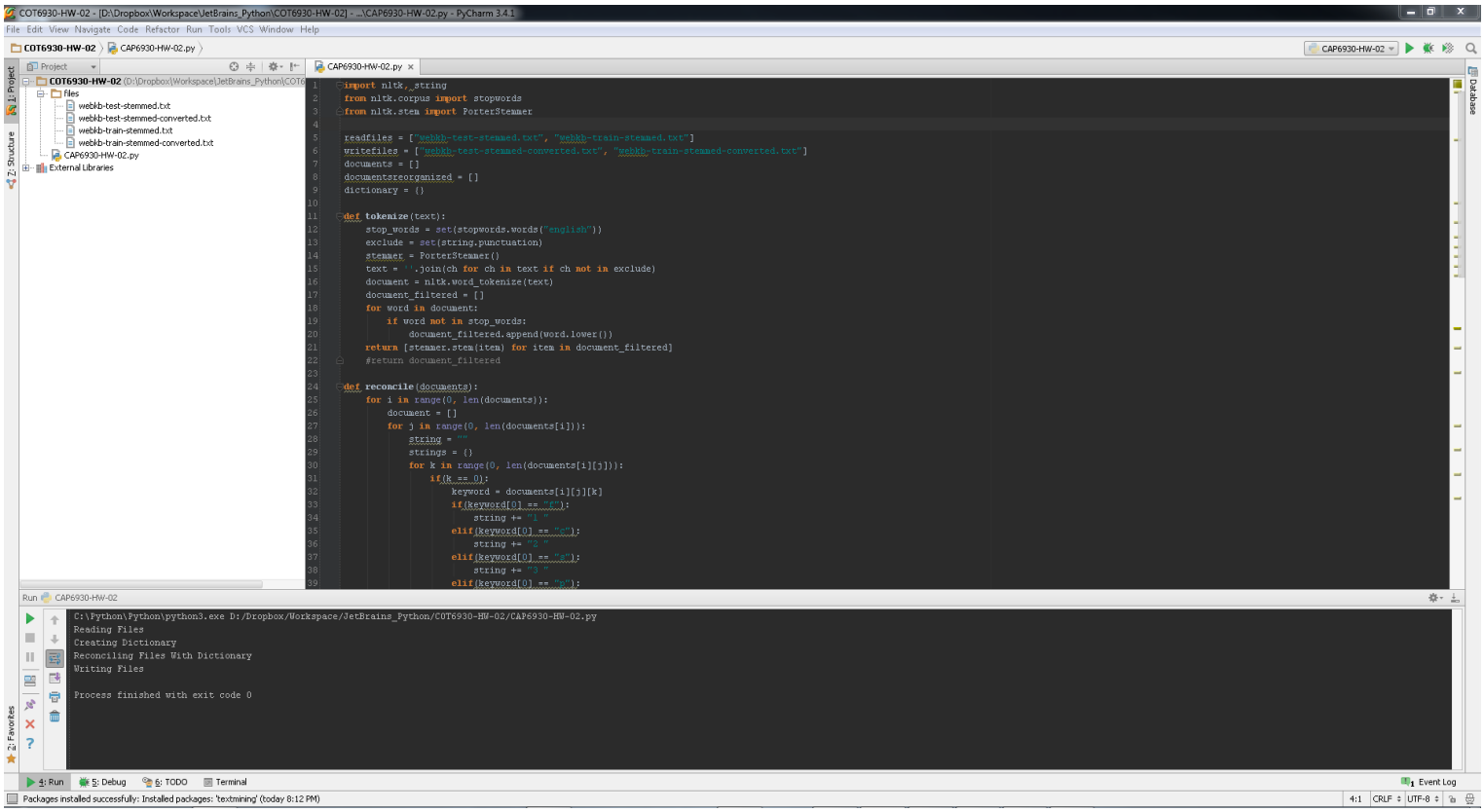
```
C:\Windows\System32\cmd.exe

D:\Dropbox\Education\2016-2017\2017 Spring\COT 6930\Homework\Assignment - 02 - (Due 2017-02-26) - Due\libsvm-3.22\windows>svm-train.exe -s0 -t0 webkb-
train-stemmed-converted.txt webkb-test-stemmed-converted.txt
*
optimization finished, #iter = 789
nu = 0.703263
obj = -1098.117161, rho = 0.748900
nSV = 1311, nBSV = 1254
*
optimization finished, #iter = 426
nu = 0.440586
obj = -536.315821, rho = 1.418748
nSV = 656, nBSV = 599
*
optimization finished, #iter = 550
nu = 0.521675
obj = -708.547053, rho = 1.076774
nSV = 925, nBSV = 857
*
optimization finished, #iter = 457
nu = 0.606543
obj = -566.771772, rho = 0.528126
nSV = 692, nBSV = 620
*
optimization finished, #iter = 494
nu = 0.609791
obj = -636.724847, rho = 0.316891
nSV = 855, nBSV = 810
*
optimization finished, #iter = 401
nu = 0.611294
obj = -462.555050, rho = -0.025231
nSV = 609, nBSV = 562
Total nSV = 2377

D:\Dropbox\Education\2016-2017\2017 Spring\COT 6930\Homework\Assignment - 02 - (Due 2017-02-26) - Due\libsvm-3.22\windows>^V
```

After running svm-predict.exe the accuracy is 99.8551 (1378/1380)

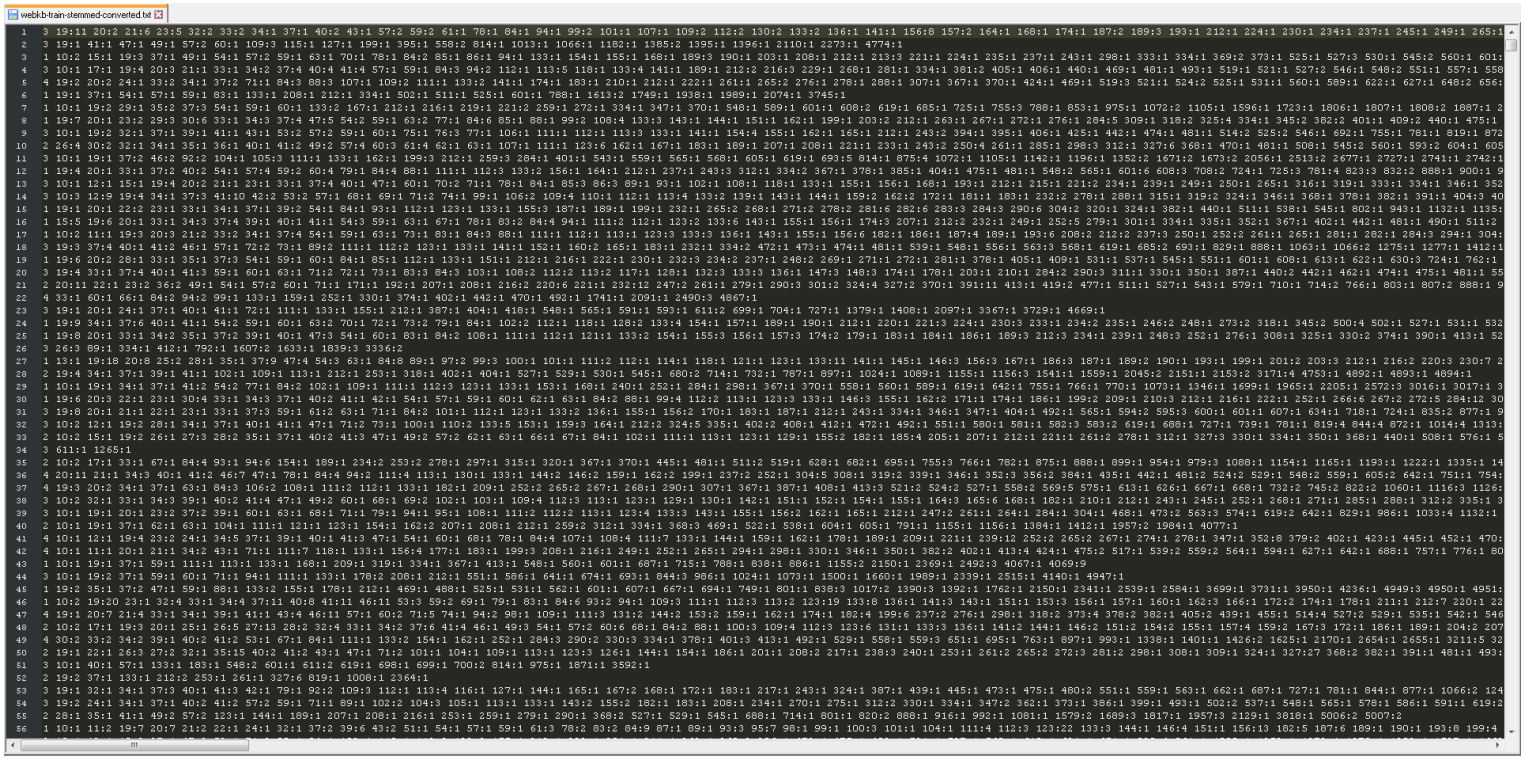
```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

D:\Dropbox\Education\2016-2017\2017 Spring\COT 6930\Homework\Assignment - 02 - (Due 2017-02-26) - Submitted\libsvm-3.22\windows>svm-predict.exe webkb-
test-stemmed-converted.txt webkb-test-stemmed-converted.txt.model model.out.txt
Accuracy = 99.8551% (1378/1380) (classification)

D:\Dropbox\Education\2016-2017\2017 Spring\COT 6930\Homework\Assignment - 02 - (Due 2017-02-26) - Submitted\libsvm-3.22\windows>
```

The supplied files required extensive pre-processing that involved python NTLK package and additional custom programming. In this assignment, I wrote a program that loaded each file for preprocessing and then tokenized it. Afterwards, I created dictionary based on all the words collected and assigned the word unique numerical value. Then I counted frequency of all words in each sentence and substituted words for dictionary derived values. In this way, I ended up with the desired input files that had a required format by LIBSVM.

Code available at link : https://www.dropbox.com/s/4jyy2k2gv9hewdv/Homework%20-%2002%20-%20Code.zip?dl=0



Below is a sample of converted “webkb-train-stemmed-converted.txt”



Below is a sample of converted “webkb-test-stemmed-converted.txt”

webkb-test-stemmed-converted.txt

```
3 11:2 12:1 13:1 14:1 15:1 16:1
3 19:4 20:1 21:1 22:1 23:1 24:1 25:2 26:1 27:1 28:1 29:3 30:2 31:3 32:1 33:1 34:1 35:1 36:1 37:3 38:1 39:2 40:1 41:1 42:4 43:1 44:1 45:1 46:1 47:1 48:1 49:3 50:1 51:1 52:1 53:1 54:1 55:2 56:1 57:2 58:1 59:2 60:2 61:3 62:1 63:1 64:1
3 10:1 12:3 15:1 19:1 20:2 30:2 40:1 41:1 50:1 69:1 70:2 71:4 72:2 73:1 74:1 75:1 76:1 77:1 78:2 79:1 80:1 81:1 82:1 83:1 84:1 85:1 86:1 87:2 88:1 89:1 90:4 91:5 92:1 93:3 94:1 95:1 96:1 97:3 98:1 99:1 100:1 101:1 102:1 103:1 104:2
3 10:1 19:4 32:1 34:3 37:3 39:1 40:2 41:2 60:1 68:1 79:1 83:1 109:3 110:1 113:2 128:5 129:1 130:1 131:4 132:1 133:2 134:1 135:2 136:1 137:1 138:1 139:1 140:1 141:1 142:1 143:1 144:1 145:1 146:1 147:3 148:5 149:1 150:1 151:3 152:1 1
2 19:1 34:1 94:1 146:2 156:2 160:1 185:2 186:1 187:1 188:1 189:1 190:1 191:1 192:1 193:1 194:1 195:1 196:1 197:2 198:1 199:1 200:1 201:1 202:1 203:1 204:1 205:1 206:1 207:1
1 19:10 20:1 23:1 33:1 37:1 84:1 94:1 99:2 112:1 156:6 189:6 194:1 209:1 210:1 211:1 212:1 213:6 214:1 215:2 216:5 217:5 218:1 219:1 220:2 221:1 222:1 223:2 224:1 225:1 226:2 227:2 228:1 229:4 230:2 231:3 232:2 233:2 234:4 235:1 23
2 10:2 19:2 20:1 23:1 26:1 34:1 37:2 39:11 49:1 54:1 57:1 59:1 60:1 93:4 111:5 114:1 123:5 155:2 174:1 187:1 188:1 190:1 199:5 208:1 231:1 243:1 252:2 253:2 254:2 255:1 256:1 257:1 258:1 259:1 260:1 261:2 262:1 263:1 264:1 265:10 2
3 10:5 19:9 20:2 24:5 34:1 37:3 39:1 40:7 41:8 46:1 48:1 53:1 54:1 59:1 60:1 66:1 67:1 68:1 69:1 70:4 71:2 72:1 73:1 78:2 79:1 84:1 85:3 86:3 94:1 102:1 104:1 109:2 110:1 111:1 112:3 113:3 123:1 130:3 133:2 136:1 141:6 143:1 153:3
1 19:2 20:3 23:1 29:2 33:2 37:2 41:1 54:1 78:1 84:4 112:2 133:1 146:3 151:1 154:5 159:1 162:3 209:1 216:2 232:1 234:1 252:1 266:1 272:1 276:1 302:1 334:1 335:1 354:1 378:1 382:4 409:1 440:3 491:1 509:1 510:5 511:1 512:1 513:1 514:1
4 19:2 33:1 34:1 37:2 39:4 67:3 71:1 77:2 84:3 111:1 133:3 155:3 163:2 216:1 230:1 240:1 275:1 281:1 330:1 334:1 347:2 400:1 442:3 460:1 548:1 552:3 553:1 554:2 555:2 556:1 557:1 558:1 559:2 560:1 561:1 562:1 563:1 564:1 565:1 566:
9 10:1 19:4 20:1 34:1 37:2 39:2 40:3 41:2 43:1 60:1 71:2 72:1 73:2 78:1 83:2 84:1 88:2 108:4 123:1 133:2 139:1 153:1 174:1 332:1 368:2 400:1 445:1 524:2 547:1 569:3 570:1 571:4 572:2 573:2 574:1 575:1 576:1 577:1 578:1 579:1 580:1
1 15:9 19:11 20:4 23:15 33:1 34:1 37:2 41:1 54:2 60:2 61:3 63:1 66:1 67:1 78:1 83:2 84:5 88:3 92:1 95:2 97:4 106:1 111:1 112:2 114:2 118:1 123:2 131:1 133:5 142:1 154:1 155:12 171:1 174:1 182:1 183:1 184:2 188:1 189:2 212:1 214:3 2
3 104:1 127:1 334:1 367:1 548:2 599:1 600:1 611:1 697:1 698:1 699:1 700:1 701:1 702:1 703:1
2 19:1 20:1 34:1 36:1 37:1 46:2 49:1 57:1 66:1 133:1 155:1 207:1 253:1 334:1 351:6 419:1 481:1 527:1 559:1 563:2 586:1 599:1 600:1 601:1 685:1 705:1 706:2 707:4 708:1 709:2 710:1 711:1 712:1 713:1 714:1 715:1 716:1 717:1 718:1 719:
1 15:2 19:9 20:7 21:1 29:2 32:1 34:2 37:5 50:1 54:2 60:2 71:3 73:3 84:6 88:1 94:1 97:1 99:2 106:1 109:1 111:3 112:2 113:1 118:2 123:3 133:7 141:1 155:3 157:1 167:1 183:1 186:1 187:3 199:1 201:1 208:1 209:1 210:2 212:3 216:1 230:2 2
2 19:2 26:10 27:5 34:1 37:2 46:3 52:1 54:1 100:3 101:1 113:1 133:1 155:1 208:1 212:1 243:1 261:1 278:1 368:1 391:1 417:1 508:1 527:6 528:1 793:2 794:1 795:1 796:1 797:1 798:2 799:2 800:1 801:2 802:4 803:1 804:2 805:5 806:1 807:1 80
2 19:1 26:1 29:15 34:1 36:2 46:3 49:2 57:4 59:1 60:1 62:1 63:1 111:11 123:2 170:1 207:1 253:2 284:1 309:3 315:1 327:15 424:1 548:1 714:7 802:10 803:1 813:1 814:2 815:1 816:1 817:1 818:26 819:41
3 10:2 19:7 20:2 21:1 33:2 34:2 37:3 39:3 41:2 43:1 60:1 71:1 83:3 84:2 88:2 95:1 100:1 101:1 102:1 106:1 109:1 112:3 113:1 118:1 122:1 143:3 155:3 157:3 174:2 210:1 230:1 231:1 246:1 265:1 276:5 278:1 311:1 332:1 335:3 345:1 350:1
3 30:1 49:1 57:1 71:1 84:1 88:1 112:2 142:1 143:1 146:1 207:1 216:1 500:1 519:1 531:1 546:1 642:1 670:1 708:1 710:1 873:1 874:1 875:2 876:1 877:1 878:1
4 19:1 20:3 21:2 32:2 34:1 37:1 39:4 43:1 71:3 83:1 84:5 93:2 107:2 110:1 111:6 114:1 123:5 124:1 131:1 133:2 144:1 156:2 181:1 189:1 193:1 199:3 202:1 212:1 224:1 232:1 238:3 243:2 250:1 252:5 259:1 264:1 265:1 267:4 272:8 275:1 2
3 19:5 20:1 22:1 33:1 34:5 37:6 39:12 41:1 57:2 59:1 60:1 71:1 72:2 78:5 84:8 85:1 88:1 89:2 95:1 100:1 111:1 112:2 122:2 133:1 155:6 157:1 162:1 182:1 184:1 187:2 209:1 210:2 212:1 230:1 234:1 246:1 248:1 249:2 268:1 271:4 272:4 2
2 23:1 35:3 40:1 41:2 60:3 68:1 109:2 110:1 183:2 208:1 271:1 327:2 334:1 368:2 369:3 481:3 642:1 714:1 721:1 792:2 796:1 802:1 809:1 832:3 838:2 888:1 985:1 986:2 987:1 988:1 989:2 990:1 991:2 992:1 993:1 994:2 995:1
3 15:1 19:4 23:1 29:1 33:1 37:1 40:1 41:3 42:1 59:1 60:1 70:3 71:1 78:1 85:4 86:4 87:1 88:1 103:1 111:1 113:2 133:1 155:1 156:2 160:1 212:1 221:1 232:1 319:1 333:1 334:1 368:1 369:1 370:1 402:1 468:1 472:2 473:1 524:2 525:1 558:2 6
2 12:1 22:1 26:1 30:1 34:1 35:1 59:1 111:1 121:1 123:1 167:1 207:1 208:1 211:2 232:1 312:1 368:3 387:1 390:2 520:2 521:2 799:1 805:1 806:1 808:1 814:1 1016:1 1017:2 1018:1 1019:1 1020:1 1021:1 1022:1 1023:1
1 19:6 20:1 33:1 35:1 37:2 51:2 63:1 84:1 112:1 123:1 133:2 136:2 155:5 157:2 207:1 212:1 221:1 234:2 248:3 261:1 265:2 276:1 282:2 284:1 345:1 391:1 569:3 607:3 613:1 719:1 739:1 782:1 791:1 833:1 928:1 963:2 996:1 1017:1 1025:1 1
1 19:3 32:2 33:1 34:5 37:5 39:6 42:1 49:1 54:1 57:2 59:1 63:1 83:2 84:2 99:1 101:1 109:1 110:1 111:1 112:2 123:3 143:1 155:5 182:1 203:2 208:2 210:2 212:2 234:1 243:1 246:3 252:1 261:1 267:1 268:2 318:1 334:1 382:1 386:1 402:2 413:
1 20:1 41:1 54:1 60:1 84:1 112:1 133:3 179:1 191:1 209:1 265:3 330:2 511:1 608:1 748:1 966:1 1045:1 1068:1 1069:1 1070:1 1071:1 1072:1 1073:2 1074:1 1075:1
2 10:5 15:2 19:26 20:4 22:2 23:3 26:35 27:3 30:3 32:2 34:5 35:21 37:5 39:1 40:4 41:2 42:1 45:1 46:9 49:5 51:1 54:1 57:5 60:2 66:1 67:4 71:11 81:1 82:1 83:2 93:2 94:1 100:1 101:1 102:1 104:6 106:3 109:1 110:5 111:1 121:2 123:27 126:
1 10:2 19:9 21:1 23:3 33:1 37:5 39:2 40:4 41:5 54:2 59:1 70:4 78:2 79:1 84:2 85:4 86:4 87:2 88:2 93:1 94:1 99:1 109:1 111:1 112:1 122:1 123:1 133:4 155:3 156:7 182:1 212:3 237:3 252:1 266:1 312:3 330:1 333:1 334:1 346:1 369:2 378:1
1 19:2 33:1 34:3 37:3 54:3 63:2 84:1 111:1 124:1 133:1 142:1 147:2 174:2 207:1 210:3 213:1 232:1 233:1 246:1 310:2 323:1 334:1 335:1 429:1 481:1 491:1 492:1 511:3 558:1 589:2 601:1 607:1 615:1 688:1 718:
2 10:1 19:2 20:1 26:1 34:2 36:1 37:2 40:2 41:2 49:2 57:1 109:2 133:1 155:2 157:3 160:1 162:1 167:2 183:1 188:1 199:3 204:1 207:1 212:2 232:1 253:1 261:1 265:1 312:1 330:1 369:3 378:1 387:1 442:2 446:1 469:1 481:1 502:2 522:1 531:1
4 10:2 15:1 19:1 20:4 26:2 33:1 34:1 37:1 40:2 41:3 60:1 67:1 68:1 71:3 84:2 114:1 133:2 170:1 183:1 186:1 189:2 199:2 212:1 231:5 232:17 247:4 252:2 281:1 282:1 287:1 298:1 312:3 325:1 330:1 367:3 402:1 413:8 475:13 517:1 559:2 56
1 10:1 19:2 26:1 34:4 35:2 37:1 40:3 41:2 42:1 53:1 54:1 60:1 68:1 71:1 76:1 92:1 109:2 113:1 133:1 149:1 162:1 185:2 208:4 212:1 284:1 327:1 352:1 368:2 370:2 382:1 387:1 404:2 417:1 469:1 529:1 553:1 569:1 604:1 792:1 796:1 801:1
3 10:1 19:3 29:1 37:3 40:1 41:1 71:1 73:2 84:2 103:1 112:1 118:1 133:3 212:1 370:1 378:1 440:1 475:1 520:1 521:1 601:1 619:1 715:1 788:1 882:1 941:1 1072:1 1206:1 1352:1 1442:3 1443:1 1444:1 1445:1 1446:1 1447:1 1448:1 1449:1 1450:
3 19:1 34:2 37:1 40:1 57:2 59:1 60:1 63:1 79:1 84:1 103:1 109:1 112:3 117:2 130:1 131:3 136:3 137:2 138:2 139:1 140:1 153:1 159:2 168:1 208:1 232:1 259:1 275:2 297:2 341:1 357:1 370:1 378:1 386:1 387:1 396:1 400:1 405:1 414:1 446:1
3 19:1 34:1 37:2 57:2 59:1 60:1 78:1 84:1 111:1 112:1 133:1 212:1 265:1 378:1 512:1 643:1 688:1 749:1 948:1 949:2 950:1 981:1 1043:1 1525:1 1526:1 1527:1
1 19:7 20:2 21:2 22:1 33:1 35:1 51:1 54:1 93:1 100:1 111:1 112:2 123:3 133:1 146:2 156:1 184:1 187:2 197:1 203:3 210:2 216:2 230:5 231:1 234:2 237:1 246:5 248:5 276:4 284:1 304:1 318:1 346:1 352:1 378:1 382:1 435:1 491:1 511:1 513:2 518:
4 10:1 41:1 57:1 59:1 60:1 100:1 209:1 214:1 552:1 558:1 727:2 1542:1 1543:1 1544:1
1 14:1 15:2 19:3 20:13 21:6 29:2 34:1 37:2 41:4 46:10 54:1 57:1 59:1 61:3 63:1 79:1 84:6 93:3 98:1 103:1 111:1 112:2 113:1 118:1 133:7 144:1 156:7 186:1 187:1 189:1 199:1 208:1 212:4 237:1 238:1 246:1 247:1 250:1 252:2 267:2 268:1
4 19:1 20:1 33:1 84:2 94:2 111:1 112:1 183:1 201:1 374:1 412:1 470:1 475:1 514:1 531:1 574:1 628:1 644:1 872:1 988:1 1304:1 1576:1
1 19:3 20:6 34:2 37:4 39:1 40:1 41:1 54:2 59:1 63:1 67:1 71:2 84:7 86:1 94:2 111:2 112:1 118:1 131:1 133:5 143:1 151:2 156:1 174:1 203:1 209:3 212:3 221:2 232:4 243:1 265:1 266:1 317:1 330:5 334:1 335:1 352:1 382:2 440:2 442:1 475:
3 10:1 19:2 33:1 37:2 41:1 57:1 59:1 63:1 69:1 71:2 76:2 84:2 103:1 131:1 133:3 212:1 252:1 290:1 334:1 384:1 438:1 440:1 475:2 524:2 558:1 579:1 607:1 1206:1 1606:1 1607:1 1608:1 1609:1 1610:1 1611:1 1612:1 1613:1 1614:1 1615:1 16
4 20:1 33:1 34:1 39:1 43:2 60:1 61:4 95:1 111:2 123:1 184:1 193:2 278:1 281:1 282:1 284:1 402:1 475:1 481:1 517:1 562:1 586:1 677:1 713:1 727:1 754:2 767:1 801:1 819:1 821:3 944:2 1027:1 1126:1 1256:1 1275:1 1314:3 1349:1 1621:1 16
1 10:1 19:5 37:2 54:1 84:2 93:1 94:2 112:1 123:1 133:2 210:1 212:1 237:2 265:1 284:2 310:1 334:1 548:1 558:1 574:1 600:1 601:1 608:2 656:1 811:3 948:1 1352:1 1500:1
1 19:1 33:1 37:1 53:2 54:1 57:1 59:1 60:1 76:1 84:1 112:2 133:2 208:1 212:2 335:1 413:1 473:1 511:1 560:1 601:1 696:1 755:1 894:1 1006:1 1170:1 1607:1 1608:1 1631:2 1632:1 1633:2
3 10:2 19:1 37:1 40:1 41:1 69:2 102:1 113:1 130:1 133:1 143:1 208:1 387:1 390:1 481:1 492:1 758:1 1076:2 1249:1 1277:1 1375:1 1635:1 1636:1 1637:1 1638:1
2 10:1 20:1 30:2 35:2 36:1 41:1 57:1 59:1 83:1 88:1 109:1 111:1 123:5 133:2 151:1 155:1 159:1 167:1 182:1 190:3 193:1 205:4 207:2 211:4 216:1 261:1 265:2 283:1 284:8 288:1 290:1 294:1 308:2 350:1 363:1 370:1 382:1 384:2 409:1 440:1
3 40:1 41:1 60:1 68:1 92:1 95:1 143:1 212:1 237:1 292:1 448:1 519:1 608:1 727:2 872:1 977:1 1060:4 1142:1 1182:1 1240:1 1258:1 1335:1 1401:1 1663:3 1664:1 1665:1 1666:1
3 10:2 19:1 37:1 40:1 41:1 59:1 113:1 127:1 133:2 139:1 159:1 167:2 168:1 212:1 406:1 464:1 473:1 492:1 500:1 548:1 619:1 727:1 781:1 872:1 878:3 1057:2 1072:2 1078:1 1379:1 1660:1 1668:2 1669:1
3 10:1 34:1 57:1 59:1 60:1 109:1 127:1 142:1 143:1 159:1 168:1 481:1 758:1 759:1 1072:1 1671:1 1672:1 1673:1 1674:1
2 20:1 21:1 28:1 29:2 34:1 35:2 36:3 42:1 44:1 49:7 57:7 59:5 60:5 61:5 100:3 101:3 123:10 167:1 183:2 186:1 199:2 207:2 208:7 232:1 247:2 253:2 308:1 310:1 312:1 327:5 368:13 391:21 431:1 470:1 508:3 528:2 529:1 530:1 559:1
3 10:1 12:2 19:4 37:3 40:3 41:5 57:3 59:1 60:1 71:2 74:1 84:1 89:1 99:2 102:1 112:3 113:1 115:1 117:1 123:2 130:1 133:3 151:1 162:2 165:1 172:1 212:1 237:1 268:2 271:1 284:1 336:1 339:1 347:2 368:1 387:1 417:1 424:1 472:1 473:1 491
4 19:5 20:1 32:1 37:1 39:2 40:1 41:2 47:1 54:1 84:3 95:1 100:1 111:1 113:1 133:1 155:1 174:2 183:1 210:1 212:1 281:1 282:1 330:2 332:1 402:1 469:1 483:1 522:1 524:1 558:2 574:1 614:1 718:1 731:1 744:1 897:1 941:1 994:1 1053:1 1335:
2 10:3 19:1 20:5 26:6 34:2 37:1 40:1 42:1 49:1 57:2 59:1 62:1 71:2 83:1 93:1 111:6 114:1 121:1 131:1 133:1 161:1 204:1 207:1 208:1 210:3 220:1 232:5 234:1 240:1 243:1 265:1 270:3 275:6 278:3 293:1 324:2 326:1 339:1 360:1 368:7 385:
3 21:2 29:1 33:1 41:1 109:1 112:1 127:1 154:1 170:1 182:1 199:1 237:1 280:1 284:1 404:1 468:1 586:1 688:1 859:1 1034:1 1764:1 1765:1 1766:1 1767:1
2 10:4 15:3 19:8 20:1 23:16 26:1 30:1 35:1 37:1 46:1 54:3 111:6 123:1 139:1 146:2 159:1 168:1 181:1 189:1 199:3 205:1 209:1 210:1 214:1 223:1 232:1 247:1 265:1 272:2 274:1 276:2 281:3 282:1 284:1 290:1 294:1 298:1 307:1 308:1 362:1
```

Below is a sample of model filed that includes the statistics of LIBSVM

webkb-test-stemmed-converted.txt.model

```
svm_type c_svc
kernel_type linear
nr_class 4
total_sv 649
rho -0.781458 -0.822187 -0.82931 -0.188367 -0.226841 -0.0233089
label 3 2 1 4
nr_sv 261 108 171 109
SV
0.0034540161441320145 0 0 10:1 19:4 32:1 34:3 37:3 39:1 40:2 41:2 60:1 68:1 79:1 83:1 109:3 110:1 113:2 128:5 129:1 130:1 131:4 132:1 133:2 134:1 135:2 136:1 137:1 138:1 139:1 140:1 141:1 142:1 143:1 144:1 145:1 146:1 147:3 148:5 14
0.0028072169189775554 0 10:1 19:4 20:1 34:1 37:2 39:2 40:3 41:2 43:1 60:1 71:2 72:1 73:2 78:1 83:2 84:1 88:2 108:4 123:1 133:2 139:1 153:1 174:1 332:1 368:2 400:1 445:1 524:2 547:1 569:3 570:1 571:4 572:2 573:2 574:1 575:1 576:1 5
0.1218395622671061 0 104:1 127:1 334:1 367:1 548:2 599:1 600:1 611:1 697:1 698:1 699:1 700:1 701:1 702:1 703:1
0.0729761650808651 0 30:1 49:1 57:1 71:1 84:1 88:1 112:2 142:1 143:1 146:1 207:1 216:1 500:1 519:1 531:1 546:1 642:1 670:1 708:1 710:1 873:1 874:1 875:2 876:1 877:1 878:1
0.0009380163516223674 0.00207695204924294 5 19:5 20:1 22:1 33:1 34:5 37:6 39:12 41:1 57:2 59:1 60:1 71:1 72:2 78:5 84:8 85:1 88:1 89:2 95:1 100:1 111:1 112:2 122:2 133:1 155:6 157:1 162:1 182:1 184:1 187:2 209:1 210:2 212:1 230:1
0.00432494065656165 25 15:1 19:4 23:1 29:1 33:1 37:1 40:1 41:3 42:1 59:1 60:1 70:3 71:1 78:1 85:4 86:4 87:1 88:1 103:1 111:1 113:2 133:1 155:1 156:2 160:1 212:1 221:1 232:1 319:1 333:1 334:1 368:1 369:1 370:1 402:1 468:1 472:2 47
0.0164812434997572 9 0 10:1 19:3 29:1 37:3 40:1 41:1 71:1 73:2 84:2 103:1 112:1 118:1 133:3 212:1 370:1 378:1 440:1 475:1 520:1 521:1 601:1 619:1 715:1 788:1 882:1 941:1 1072:1 1206:1 1352:1 1442:3 1443:1 1444:1 1445:1 1446:1 1447
0 0 0.00491401185538304 42 10:1 29:2 33:1 37:2 41:1 57:1 59:1 63:1 69:1 71:2 76:2 84:2 103:1 131:1 133:3 212:1 252:1 290:1 334:1 384:1 438:1 440:1 475:2 524:2 558:1 579:1 607:1 1206:1 1606:1 1607:1 1608:1 1609:1 1610:1 1611:1 1612:1
0 0.0929279647438244 8 0 10:1 19:5 37:2 54:1 84:2 93:1 94:2 112:1 123:1 133:2 210:1 212:1 237:2 265:1 284:2 310:1 334:1 548:1 558:1 574:1 600:1 601:1 608:2 656:1 811:3 948:1 1352:1 1500:1
0 0.00423558335801373 8 0.0208626500065594 6 40:1 41:1 60:1 68:1 92:1 95:1 143:1 212:1 237:1 292:1 448:1 519:1 688:1 727:2 872:1 977:1 1060:4 1142:1 1182:1 1240:1 1258:1 1335:1 1401:1 1663:3 1664:1 1665:1 1666:1
0 0 0.0879874604359561 1 41:1 60:1 127:1 727:1
0.00311850680790795 15 0.00187176840892230 9 0.00793258066737369 6 19:5 22:1 23:3 33:1 34:1 37:2 40:1 57:1 62:1 67:1 71:2 103:1 111:1 112:3 113:1 133:1 144:2 146:1 153:1 154:5 159:2 162:1 171:1 189:2 208:1 234:1 248:1 250:2 265:1 272:1
0.0568768786020159 8 0 19:1 20:2 24:1 40:1 41:2 60:1 127:1 232:1 237:2 367:1 368:1 405:2 475:1 548:2 569:1 582:1 600:1 611:1 688:1 698:1 699:1 700:1 701:1 703:1 704:1 727:1 821:1 1381:1 1869:1 1870:1 1871:1 1872:1 1873:1 1874:1
0 0 0.0110584881873128 9 40:3 41:3 61:1 111:1 113:1 123:1 284:1 391:2 481:1
0 0.00908861425226389 7 19:1 28:1 34:2 37:2 40:1 41:1 57:1 61:2 63:1 84:2 111:1 133:1 212:1 246:1 442:2 524:1 565:1 594:1 595:3 688:1 948:1 1126:1 1547:1 1548:1 1628:1 1767:1 2059:5 2060:1 2061:1
0 0.0148114116607708 7 19:1 20:1 32:1 33:1 37:2 40:1 57:1 59:2 84:1 104:2 109:1 123:1 133:1 141:1 153:1 156:1 166:1 212:1 221:1 224:1 232:1 234:1 237:1 245:1 252:1 323:1 334:1 431:1 475:5 514:1 522:1 524:1 567:1 569:1 575:1 585:2
0.00354750999050434 8 0.00542984667217 9 10:1 19:8 20:4 21:1 34:1 37:3 39:2 71:1 72:1 73:1 84:5 88:1 94:2 102:1 111:2 113:1 133:2 139:2 154:1 155:4 156:3 165:1 172:1 174:1 182:1 183:1 208:1 221:1 237:2 250:3 265:1 267:1 271:1 272:1
0.0570165105381140 9 0.0267447312433231 1 0 34:1 49:1 57:2 193:1 329:1 334:1 551:1 805:1 806:1 1127:3
0 0.00575756473657052 8 0.00683157516737751 9 15:1 19:5 20:1 30:1 34:2 37:5 41:1 59:1 68:1 71:2 83:1 84:2 107:1 111:1 112:1 123:8 125:2 126:1 133:1 143:1 183:3 217:1 221:1 224:1 261:1 274:1 304:1 314:1 334:1 346:1 362:1 368:1 399:2 4
0.0133414287427296 4 0 19:2 32:2 34:2 37:3 40:1 47:2 53:2 54:2 79:1 84:2 93:1 95:1 104:2 106:1 109:2 112:1 113:1 118:1 123:3 133:3 142:2 143:1 153:1 155:1 159:1 165:1 167:1 183:1 186:1 201:1 212:2 252:1 275:1 284:1 308:1 335:1 340
0.00052549298455579 8 0.0501121975756489 1 0.00449931823974539 10:1 19:1 20:1 28:1 34:1 37:3 60:1 62:1 84:2 94:2 112:2 133:1 143:1 156:2 197:1 211:2 212:1 239:1 237:1 304:1 312:1 445:1 469:2 524:2 569:1 575:1 627:1 661:1 740:1 799:
0.00132591533032408118 1 10:1 19:3 22:1 24:1 37:2 39:4 41:1 52:1 60:1 63:2 68:1 71:1 79:2 84:5 93:1 94:2 100:3 102:1 104:13 107:1 109:4 111:1 112:1 123:1 133:2 141:1 160:1 167:1 174:2 175:1 179:2 183:1 212:2 237:1 243:1 252:2 265:1
0 0.00720006568044062 19:3 21:1 24:1 34:1 37:1 39:2 40:1 41:1 67:1 68:2 71:2 77:3 79:1 84:2 94:4 102:1 103:1 111:2 112:3 124:1 133:4 135:1 139:1 141:1 142:1 146:1 155:2 156:1 167:1 183:1 187:1 346:2 378:2 382:1 387:1 404:3 418:1
0.0132802530268064 0 0 19:1 37:1 41:1 57:2 109:1 481:1 551:1 556:1 586:1 601:1 642:1 724:1 725:1 788:1 844:2 1058:1 1062:1 1260:1 1352:1 1489:1 1825:1 1938:1 2652:2
0.00426491510045721 0 0 10:1 19:1 20:1 37:1 57:1 59:1 84:1 111:1 112:2 123:1 133:1 212:1 237:1 288:1 334:1 469:1 574:1 601:1 611:2 688:1 814:1 821:1 994:1 1043:1 1227:1 1389:1 1405:1 1527:1 1686:1 2460:1 2629:1 2717:1 2718:1 2719:1
0.00099442735990273296 0 10:1 19:3 20:3 37:1 40:1 41:1 59:1 68:1 73:1 84:2 94:1 112:1 127:1 133:1 141:1 156:1 183:1 184:1 193:1 232:2 237:2 334:2 395:1 425:1 461:1 563:1 619:1 630:1 685:1 708:1 715:2 748:1 845:1 938:1 1182:1 1377:
0 3.78152592401154e-05 19:2 20:2 21:1 23:2 37:2 40:3 41:3 53:1 57:2 59:1 69:1 71:2 74:1 84:3 93:3 111:2 112:4 113:1 123:11 133:1 135:2 141:1 154:1 193:4 212:1 216:1 219:1 221:2 222:4 250:2 252:1 265:1 267:1 271:1 274:1 281:1 282
0.0125975869608709 7 0 19:3 20:1 23:1 37:2 40:1 41:1 47:3 54:1 59:1 60:1 71:1 103:1 112:1 113:1 133:1 155:2 212:2 267:1 330:1 359:1 467:1 569:2 615:1 644:1 669:1 772:1 838:2 894:1 1126:1 1145:1 1230:1 1280:1 1392:1 1408:1 1573:1 1
0.00364645465393949 4 0 10:1 12:4 19:9 20:1 24:1 32:1 37:3 41:5 46:1 57:3 59:1 60:3 71:1 72:1 73:1 75:1 76:2 77:3 84:4 91:1 94:2 104:1 109:1 110:1 112:3 113:5 123:2 133:1 135:1 143:3 144:1 153:2 156:1 159:1 160:1 167:1 168:2 170:1
0 0.0308807782282070 4 10:1 19:1 37:1 133:1 183:1 212:1 611:2 619:1 700:1 814:1 977:1
0.00067536819112685326 0.0307710600247611 6 0 19:2 33:1 37:1 40:1 41:1 57:2 59:1 60:1 62:1 77:1 84:1 103:1 107:1 111:1 112:1 113:1 133:1 152:1 157:1 184:2 208:1 216:1 221:1 232:1 233:1 245:2 250:1 290:1 347:2 367:1 481:1 545:1 548:1
0 0.00264048668263794 2 0 10:1 28:1 57:1 59:1 75:1 76:2 77:1 109:1 113:3 133:1 141:1 243:1 334:1 514:1 520:1 521:1 525:1 548:1 560:1 1517:1 1607:1 1611:1 1775:1 1858:1 1870:1
0 0 0.0242675087489217 4 92:1 563:1 1265:1 1352:1 1507:1 1881:1
0 0 0.00497662261755610 9 19:1 34:1 37:1 40:1 41:1 178:1 212:1 387:1 404:1 462:1 992:1 1113:1 1127:1 1277:1 1745:2 1898:3 1899:2
0.00781847958044494 9 0 19:2 20:1 33:1 37:4 60:1 68:1 71:3 72:1 104:1 133:2 580:1 607:1 642:1 749:1 788:1 994:1 1017:1 1072:2 1673:2 1711:1 2850:1
0 0 0.00204515708943348 1 10:1 19:3 20:2 21:2 27:1 37:2 39:1 41:1 62:2 63:2 83:4 84:3 88:1 94:1 104:1 106:1 111:4 112:3 118:1 123:1 133:2 155:3 156:4 168:1 182:2 186:3 193:1 199:3 207:2 208:2 212:1 232:1 237:3 238:1 243:1 252:2 265:
0 0.00462608134021100 2 40:2 41:3 104:1 109:1 111:1 152:1 154:1 155:1 159:1 160:1 162:1 184:1 356:1 402:1 435:1 469:1 582:1 585:1 727:1 781:1 872:1 1207:1 1421:1 1423:1 1541:2 1813:1 1857:2
0.0435102305598674 5 0 10:1 19:2 34:1 37:1 41:1 60:2 62:1 63:1 133:1 187:1 212:1 334:1 391:1 500:1 548:2 619:1 725:1 878:2 1072:1 1157:1 1668:1 2150:1 2327:1 2444:1 3037:1 3393:1
0.00660430584443038 0 0 34:1 37:1 40:3 41:4 47:1 57:1 59:1 84:1 94:1 111:1 133:1 265:1 311:2 334:1 351:2 367:1 467:1 518:2 537:1 558:1 563:1 586:1 601:1 615:1 717:1 751:1 761:3 841:1 848:1 1105:1 1215:1 1579:1 1599:1 1865:3
0.00088017699884604991 0 0.00642946407486475 9 19:3 20:3 21:1 32:1 37:1 39:4 40:1 41:1 42:1 43:1 47:4 60:1 71:1 72:1 78:1 106:1 111:3 124:1 127:1 133:1 155:2 156:1 158:1 189:2 197:1 199:3 208:1 212:1 243:1 247:5 252:2 259:1 276:2 309
0.0526727262958368 4 0 19:1 33:1 34:1 37:1 41:1 49:1 57:2 59:1 62:1 71:1 89:1 133:1 144:1 212:1 232:1 248:1 319:1 334:1 500:1 559:1 579:1 601:1 710:1 724:1 805:1 829:1 878:2 879:1 888:1 1706:1 1712:1 1850:1 2150:1 2944:1 3234:1 32
0.0367542125559465 5 0 0.0445789767267705 40:1 41:1 112:1 151:1 688:1 900:1 1014:1 1043:1 1379:1
0 0 0.00547503232416815 3 10:1 19:5 20:4 23:4 28:3 30:2 33:1 34:1 37:5 62:1 84:5 85:2 93:1 99:2 110:1 112:1 118:2 121:1 133:4 146:4 155:2 156:6 171:1 174:4 193:1 209:1 212:3 222:1 232:1 234:1 243:1 248:1 252:1 265:2 266:3 272:3 284:
0.0105128007866399 7 0.0132507011669726 0 10:1 19:1 34:1 37:1 41:1 54:1 67:1 102:1 123:1 133:2 135:1 158:1 167:1 183:1 212:1 216:1 221:1 237:1 242:1 366:1 387:1 497:1 562:1 571:1 607:2 619:1 648:1 668:1 680:1 716:1 792:1 905:2 1016
0.0102018559480142 2 0 0.0150245260049660 1 19:2 37:1 40:1 41:1 45:1 84:1 109:1 159:1 172:1 216:1 232:1 284:1 351:1 392:1 393:1 402:1 469:1 472:1 473:1 571:1 727:1 761:1 875:1 900:1 1015:1 1142:1 1745:1 2222:1 2330:1 2462:4 2500:1 37
0.0149600624848592 9 0 10:1 34:1 39:1 71:2 124:1 311:1 369:1 381:1 586:1 619:1 688:1 838:1 1528:1 1733:1 1887:1 1891:1 2332:1 2339:2 3477:1 3478:1 3697:1 3748:2 3749:1
0.0627077549673419 7 0 19:1 20:2 25:1 40:1 47:3 59:1 84:1 104:1 112:1 143:1 189:1 199:1 245:1 334:1 382:1 500:1 519:1 548:1 550:1 559:1 568:1 601:1 687:1 715:4 776:1 853:1 878:1 990:1 993:1 1266:1 1390:2 1544:1 1733:1 1746:1 1849:
0.0111633456133138 9 0 19:2 20:2 23:2 29:1 30:1 34:1 37:1 40:1 41:2 49:1 57:2 59:1 62:1 63:2 70:2 71:1 78:1 84:2 85:1 86:1 111:3 127:3 133:1 155:1 183:1 187:3 199:2 212:1 237:2 259:1 281:1 331:1 333:1 334:2 369:2 470:2 503:1 558:1
```

Both files available at link : https://www.dropbox.com/s/sox583lf0y3e0k2/Homework%20-%2002%20-%20Data.zip?dl=0

Below is a code used for program that did preprocessing.

```
import nltk, string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer

readfiles = ["webkb-test-stemmed.txt", "webkb-train-stemmed.txt"]
writefiles = ["webkb-test-stemmed-converted.txt", "webkb-train-stemmed-converted.txt"]
documents = []
documentsreorganized = []
dictionary = {}
```

```
def tokenize(text):
    stop_words = set(stopwords.words("english"))
    exclude = set(string.punctuation)
    stemmer = PorterStemmer()
    text = ''.join(ch for ch in text if ch not in exclude)
    document = nltk.word_tokenize(text)
    document_filtered = []
    for word in document:
        if word not in stop_words:
            document_filtered.append(word.lower())
    return [stemmer.stem(item) for item in document filtered]

def reconcile(documents):
    for i in range(0, len(documents)):
        document = []
        for j in range(0, len(documents[i])):
            string = ""
            strings = {}
            for k in range(0, len(documents[i][j])):
                if(k == 0):
                    keyword = documents[i][j][k]
                    if(keyword[0] == "f"):
                        string += "1 "
                    elif(keyword[0] == "c"):
                        string += "2 "
                    elif(keyword[0] == "s"):
                        string += "3 "
                    elif(keyword[0] == "p"):
                        string += "4 "
                else:
                    if dictionary[documents[i][j][k]] in strings.keys():
                        count = strings[dictionary[documents[i][j][k]]]
                        strings[dictionary[documents[i][j][k]]] = count + 1
                    else:
                        strings[dictionary[documents[i][j][k]]] = 1
            maximum = 0
            for key in strings:
                if key > maximum:
                    maximum = key;
            for k in range(0, maximum):
                if k in strings:
                    string += str(k) + ":" + str(strings[k]) + " "
            string = string[:-1]
            if len(string) > 3:
                document.append(string)
        documentsreorganized.append(document)

def createdictonary():
    tracker = 10
    for i in range(0, len(documents)):
        for j in range(0, len(documents[i])):
            for k in range(0, len(documents[i][j])):
                if documents[i][j][k] in dictionary.keys():
                    continue
                else:
                    dictionary[documents[i][j][k]] = tracker
                    tracker += 1

def loadfile(readfiles):
    for i in range(0, len(readfiles)):
        with open("files\\" + str(readfiles[i]), "r") as file:
        document = file.read().split("\n")
    context = []
    for i in range(0, len(document)):
        sentence = tokenize(document[i])
        context.append(sentence)
    documents.append(context)

def readfile(path):
    with open(path, "r") as file:
        document = file.read().split("\n")
    return document

def writefile():
    for i in range(0, len(writefiles)):
        with open("files\\" + str(writefiles[i]), "w") as file:
            for j in range(0, len(documentsreorganized[i])):
                file.write(documentsreorganized[i][j] + "\n")

print("Reading Files")
loadfile(readfiles)
```

```
print("Creating Dictionary")
createdictonary()
print("Reconciling Files With Dictionary")
reconcile(documents)
print("Writing Files")
writefile()
```