Results

Use 10-fold cross-validation

Use ‘random.shuffle’ for shuffle the sequence of files in each group so that each time the files in 10 folds are randomly.

Use ‘random.seed’ to control the random situation

Here is the random seed number and their average accuracy and f1 score

|  |  |  |
| --- | --- | --- |
| Seed | Average Accuracy | Average F1 Score |
| 10 | 0.884850977 | 0.876856115 |
| 15 | 0.886350804 | 0.878470376 |
| 20 | 0.884643586 | 0.876842605 |
| 25 | 0.883474266 | 0.875754527 |
| 30 | 0.88670445 | 0.879004404 |

More detail about counts of words of each fold for each seed show blow pages

Use group number 1 – 20 to represent 20 news groups

|  |  |
| --- | --- |
| alt.atheism | 1 |
| comp.graphics | 2 |
| comp.os.ms-windows.misc | 3 |
| comp.sys.ibm.pc.hardware | 4 |
| comp.sys.mac.hardware | 5 |
| comp.windows.x | 6 |
| misc.forsale | 7 |
| rec.autos | 8 |
| rec.motorcycles | 9 |
| rec.sport.baseball | 10 |
| rec.sport.hockey | 11 |
| sci.crypt | 12 |
| sci.electronics | 13 |
| sci.med | 14 |
| sci.space | 15 |
| soc.religion.christian | 16 |
| talk.politics.guns | 17 |
| talk.politics.mideast | 18 |
| talk.politics.misc | 19 |
| talk.religion.misc | 20 |

**Random seed = 10**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Counts of words | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 124702 | 127670 | 127438 | 124474 | 127977 | 122706 | 126877 | 126903 | 123019 | 127217 |
| 2 | 129571 | 138917 | 133885 | 139549 | 131215 | 130562 | 141562 | 141296 | 130029 | 135673 |
| 3 | 293477 | 307727 | 295479 | 297897 | 293623 | 308632 | 302479 | 265854 | 294591 | 320438 |
| 4 | 96293 | 96194 | 95462 | 96778 | 95547 | 96271 | 99009 | 96125 | 97801 | 97768 |
| 5 | 82283 | 87632 | 86668 | 85569 | 81574 | 85996 | 85114 | 84999 | 86140 | 85371 |
| 6 | 161764 | 156407 | 167011 | 154743 | 157387 | 164816 | 161621 | 155315 | 164350 | 162843 |
| 7 | 74481 | 74201 | 73129 | 73773 | 73870 | 74328 | 71565 | 71587 | 73450 | 74079 |
| 8 | 104186 | 104101 | 101104 | 103809 | 105559 | 105688 | 105715 | 104362 | 106051 | 104685 |
| 9 | 98569 | 98826 | 98836 | 97465 | 94836 | 98550 | 97038 | 97408 | 97968 | 97319 |
| 10 | 108243 | 109338 | 108312 | 109216 | 108417 | 107876 | 111109 | 109507 | 108789 | 107554 |
| 11 | 130285 | 134065 | 131951 | 135303 | 134112 | 135578 | 132737 | 133544 | 131028 | 126035 |
| 12 | 162438 | 158878 | 165191 | 165057 | 168259 | 163492 | 164196 | 165346 | 161336 | 164275 |
| 13 | 99892 | 99923 | 99660 | 94554 | 99091 | 99295 | 100968 | 100158 | 100430 | 99665 |
| 14 | 144250 | 141973 | 142005 | 139514 | 146151 | 142290 | 147573 | 142316 | 145487 | 145003 |
| 15 | 147287 | 144000 | 146802 | 143197 | 146510 | 147192 | 144541 | 148499 | 145003 | 149064 |
| 16 | 170645 | 172606 | 169092 | 170051 | 170875 | 169894 | 165397 | 169952 | 169424 | 172938 |
| 17 | 152978 | 152448 | 148122 | 155080 | 151506 | 150434 | 156531 | 151923 | 154275 | 147757 |
| 18 | 214832 | 218325 | 219997 | 220282 | 219975 | 222720 | 218335 | 221914 | 218378 | 220234 |
| 19 | 156608 | 162174 | 156935 | 159178 | 159632 | 162169 | 150952 | 150384 | 159136 | 156986 |
| 20 | 103104 | 100319 | 103469 | 101713 | 105291 | 104371 | 101852 | 105010 | 105489 | 104049 |
| Total | 2755888 | 2785724 | 2770548 | 2767202 | 2771407 | 2792860 | 2785171 | 2742402 | 2772174 | 2798953 |

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1 Score |
| 1 | 0.882072977 | 0.872842422 |
| 2 | 0.881885593 | 0.873086907 |
| 3 | 0.889183457 | 0.882195539 |
| 4 | 0.893899204 | 0.888554186 |
| 5 | 0.887413702 | 0.880699557 |
| 6 | 0.879851143 | 0.871845173 |
| 7 | 0.888888889 | 0.879951021 |
| 8 | 0.883448643 | 0.87612557 |
| 9 | 0.869542066 | 0.861961544 |
| 10 | 0.892324094 | 0.881299229 |
| Average | 0.884850977 | 0.876856115 |

**Random Seed = 15**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Counts of words | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 127975 | 125093 | 123263 | 127181 | 126141 | 121625 | 127592 | 126323 | 127507 | 126283 |
| 2 | 135459 | 135371 | 140762 | 119598 | 134654 | 140773 | 131778 | 139574 | 134723 | 139567 |
| 3 | 290237 | 319475 | 311786 | 307121 | 280033 | 297354 | 302688 | 294983 | 295550 | 280970 |
| 4 | 97819 | 97954 | 94589 | 97610 | 97635 | 98833 | 97607 | 94152 | 95652 | 95397 |
| 5 | 86208 | 85721 | 85216 | 85711 | 81864 | 86540 | 86666 | 86510 | 82218 | 84692 |
| 6 | 162887 | 156271 | 156703 | 166126 | 161348 | 163652 | 158959 | 160281 | 154708 | 165322 |
| 7 | 74000 | 71838 | 74255 | 73330 | 72897 | 74288 | 71648 | 73700 | 74298 | 74209 |
| 8 | 105275 | 104572 | 104631 | 104791 | 104735 | 102978 | 104275 | 105779 | 106382 | 101842 |
| 9 | 97183 | 97830 | 97086 | 99369 | 97807 | 98203 | 95381 | 98008 | 99056 | 96892 |
| 10 | 107809 | 109396 | 109083 | 110382 | 107620 | 107032 | 109212 | 110262 | 108043 | 109522 |
| 11 | 125960 | 132656 | 134434 | 132225 | 133478 | 134132 | 132902 | 132786 | 132708 | 133357 |
| 12 | 162263 | 163958 | 166510 | 163530 | 158656 | 160795 | 166370 | 166297 | 166269 | 163820 |
| 13 | 92541 | 100431 | 100142 | 98747 | 100929 | 100080 | 100168 | 100582 | 99873 | 100143 |
| 14 | 146372 | 143129 | 142361 | 147843 | 140275 | 145985 | 140323 | 147258 | 140412 | 142604 |
| 15 | 143528 | 147770 | 146452 | 141070 | 149368 | 147900 | 147076 | 144582 | 147911 | 146438 |
| 16 | 167882 | 173207 | 170199 | 167827 | 167265 | 173636 | 168964 | 170872 | 170132 | 170890 |
| 17 | 153225 | 146534 | 147908 | 153336 | 154026 | 151469 | 152307 | 152398 | 154645 | 155206 |
| 18 | 223043 | 218722 | 221361 | 221574 | 219169 | 223127 | 219245 | 219454 | 217267 | 212030 |
| 19 | 158104 | 159724 | 160427 | 157822 | 156415 | 162869 | 156374 | 154762 | 151067 | 156590 |
| 20 | 105051 | 102149 | 102560 | 100650 | 102037 | 103669 | 104053 | 105707 | 104814 | 103977 |
| Total | 2762821 | 2791801 | 2789728 | 2775843 | 2746352 | 2794940 | 2773588 | 2784270 | 2763235 | 2759751 |

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1 Score |
| 1 | 0.875198308 | 0.869458838 |
| 2 | 0.882415254 | 0.874120719 |
| 3 | 0.880699894 | 0.871643836 |
| 4 | 0.884880637 | 0.875387762 |
| 5 | 0.895379713 | 0.888459603 |
| 6 | 0.881446039 | 0.872572482 |
| 7 | 0.890483785 | 0.882144675 |
| 8 | 0.889835019 | 0.88196841 |
| 9 | 0.887113951 | 0.878434802 |
| 10 | 0.896055437 | 0.890512638 |
| Average | 0.886350804 | 0.878470376 |

**Random seed = 20**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Counts of words | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 128291 | 124582 | 126146 | 120334 | 124508 | 128001 | 125438 | 126885 | 126867 | 127931 |
| 2 | 135547 | 135364 | 140420 | 139776 | 121501 | 140312 | 135645 | 136376 | 138618 | 128700 |
| 3 | 308089 | 296900 | 287590 | 310213 | 311225 | 289013 | 251783 | 297328 | 319918 | 308138 |
| 4 | 95686 | 93185 | 99062 | 96692 | 97632 | 97628 | 98316 | 96478 | 97113 | 95456 |
| 5 | 81514 | 86914 | 84925 | 84857 | 87691 | 85756 | 80177 | 86490 | 86423 | 86599 |
| 6 | 168915 | 152130 | 167909 | 158604 | 163554 | 165162 | 158358 | 155619 | 151830 | 164176 |
| 7 | 72906 | 73270 | 74320 | 74871 | 72959 | 72696 | 72986 | 74190 | 72722 | 73543 |
| 8 | 103892 | 105024 | 103199 | 104996 | 104759 | 103479 | 104221 | 105024 | 105969 | 104697 |
| 9 | 98468 | 96886 | 97477 | 97254 | 99091 | 98396 | 97740 | 95357 | 98353 | 97793 |
| 10 | 109300 | 107904 | 108209 | 107824 | 107575 | 108819 | 110793 | 109508 | 109737 | 108692 |
| 11 | 133035 | 134972 | 134142 | 127409 | 133018 | 131214 | 133859 | 132618 | 134319 | 130052 |
| 12 | 166899 | 157182 | 164680 | 165134 | 160516 | 163774 | 163355 | 164801 | 163095 | 169032 |
| 13 | 94423 | 100339 | 99711 | 100066 | 99025 | 99588 | 99926 | 99211 | 100297 | 101050 |
| 14 | 146965 | 145573 | 147740 | 137907 | 141469 | 144644 | 145054 | 138155 | 143960 | 145095 |
| 15 | 143913 | 147175 | 148850 | 146987 | 148892 | 146284 | 141763 | 142623 | 149799 | 145809 |
| 16 | 168038 | 171690 | 171649 | 169427 | 172887 | 170273 | 172952 | 165595 | 169526 | 168837 |
| 17 | 152952 | 146169 | 151151 | 153879 | 151432 | 153330 | 153373 | 149499 | 155069 | 154200 |
| 18 | 216903 | 223264 | 216497 | 220005 | 219918 | 220199 | 223213 | 216340 | 217225 | 221428 |
| 19 | 162634 | 156039 | 159615 | 156576 | 156498 | 153721 | 153273 | 159686 | 154984 | 161128 |
| 20 | 103409 | 103895 | 104319 | 102573 | 99737 | 101995 | 105828 | 106058 | 103828 | 103025 |
| Total | 2791779 | 2758457 | 2787611 | 2775384 | 2773887 | 2774284 | 2728053 | 2757841 | 2799652 | 2795381 |

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1 Score |
| 1 | 0.890005288 | 0.88137299 |
| 2 | 0.878177966 | 0.87122347 |
| 3 | 0.880699894 | 0.871022659 |
| 4 | 0.876392573 | 0.869054939 |
| 5 | 0.890069039 | 0.880968006 |
| 6 | 0.883040936 | 0.873405875 |
| 7 | 0.878256247 | 0.874830936 |
| 8 | 0.898882384 | 0.89331505 |
| 9 | 0.884984026 | 0.876249329 |
| 10 | 0.885927505 | 0.876982791 |
| Average | 0.884643586 | 0.876842605 |

**Random seed = 25**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Counts of words | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 124116 | 127517 | 124160 | 122015 | 126283 | 125426 | 125464 | 126462 | 129088 | 128452 |
| 2 | 140346 | 141444 | 136324 | 132771 | 127398 | 122750 | 141192 | 131728 | 141323 | 136983 |
| 3 | 286915 | 271740 | 301068 | 285797 | 296787 | 280794 | 321401 | 320209 | 308067 | 307419 |
| 4 | 97068 | 96995 | 94341 | 97778 | 97338 | 98147 | 93440 | 94999 | 97960 | 99182 |
| 5 | 86135 | 85832 | 85560 | 85692 | 86251 | 82077 | 85447 | 82305 | 86452 | 85595 |
| 6 | 142667 | 162426 | 159344 | 150514 | 168447 | 159926 | 163357 | 163752 | 167346 | 168478 |
| 7 | 73775 | 73239 | 73690 | 73192 | 74114 | 74503 | 73857 | 71246 | 72850 | 73997 |
| 8 | 105144 | 104370 | 103966 | 105123 | 105891 | 105029 | 101934 | 104275 | 103521 | 106007 |
| 9 | 97172 | 98536 | 98268 | 98480 | 98148 | 97863 | 95782 | 98786 | 95393 | 98387 |
| 10 | 107440 | 111515 | 106124 | 108810 | 108790 | 108904 | 107843 | 108505 | 110199 | 110231 |
| 11 | 132739 | 131518 | 133515 | 134521 | 132358 | 126027 | 132053 | 133837 | 133752 | 134318 |
| 12 | 164838 | 167940 | 165017 | 163945 | 155727 | 160941 | 166470 | 162936 | 167344 | 163310 |
| 13 | 100963 | 100338 | 100635 | 99056 | 101223 | 100231 | 100841 | 97965 | 94129 | 98255 |
| 14 | 147993 | 145129 | 143595 | 146304 | 147942 | 143197 | 135768 | 134273 | 145640 | 146721 |
| 15 | 148994 | 144304 | 144037 | 145652 | 148282 | 149061 | 143570 | 146716 | 143710 | 147769 |
| 16 | 169572 | 171493 | 167557 | 173011 | 171729 | 171551 | 169401 | 171401 | 165638 | 169521 |
| 17 | 150499 | 154871 | 152950 | 152778 | 151724 | 145221 | 154473 | 156391 | 147615 | 154532 |
| 18 | 221753 | 213928 | 218145 | 222214 | 221142 | 223264 | 213265 | 222616 | 219086 | 219579 |
| 19 | 160381 | 162527 | 161371 | 156022 | 153597 | 153068 | 159045 | 152953 | 157704 | 157486 |
| 20 | 95373 | 105230 | 101952 | 106251 | 102444 | 104032 | 103873 | 105336 | 104909 | 105267 |
| Total | 2753883 | 2770892 | 2771619 | 2759926 | 2775615 | 2732012 | 2788476 | 2786691 | 2791726 | 2811489 |

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1 Score |
| 1 | 0.880486515 | 0.873431701 |
| 2 | 0.887182203 | 0.878542631 |
| 3 | 0.87804878 | 0.871973654 |
| 4 | 0.885411141 | 0.879768501 |
| 5 | 0.884758364 | 0.876677416 |
| 6 | 0.880382775 | 0.872032982 |
| 7 | 0.882509304 | 0.871760537 |
| 8 | 0.881852049 | 0.875505852 |
| 9 | 0.883386581 | 0.874044291 |
| 10 | 0.890724947 | 0.883807703 |
| Average | 0.883474266 | 0.875754527 |

**Random seed = 30**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Counts of words | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 123778 | 126654 | 122132 | 127118 | 126970 | 126193 | 126170 | 128285 | 125799 | 125884 |
| 2 | 131350 | 125656 | 140861 | 138367 | 140148 | 134827 | 133986 | 141299 | 135057 | 130708 |
| 3 | 315103 | 299114 | 295706 | 271945 | 287968 | 286841 | 296700 | 322135 | 282195 | 322490 |
| 4 | 95691 | 98340 | 96316 | 95251 | 97007 | 96507 | 96584 | 97997 | 95506 | 98049 |
| 5 | 85948 | 81151 | 86367 | 85348 | 81421 | 86658 | 87115 | 86502 | 85633 | 85203 |
| 6 | 159831 | 158328 | 138900 | 165811 | 167733 | 168541 | 162738 | 159527 | 165716 | 159132 |
| 7 | 72498 | 73575 | 73419 | 74288 | 74618 | 72244 | 73368 | 74124 | 73791 | 72538 |
| 8 | 104611 | 104334 | 105369 | 105056 | 104466 | 101105 | 104295 | 104904 | 104594 | 106526 |
| 9 | 97769 | 97035 | 98529 | 97748 | 94821 | 98074 | 98541 | 97574 | 99006 | 97718 |
| 10 | 109301 | 109119 | 108777 | 107285 | 108647 | 110167 | 108292 | 109002 | 108744 | 109027 |
| 11 | 133639 | 133574 | 132694 | 133727 | 131767 | 131290 | 133160 | 124763 | 133836 | 136188 |
| 12 | 164346 | 163348 | 169369 | 159244 | 164033 | 169274 | 166479 | 158333 | 158126 | 165916 |
| 13 | 100012 | 100351 | 99972 | 100355 | 100025 | 94497 | 98754 | 99560 | 100876 | 99234 |
| 14 | 141912 | 142452 | 147195 | 146897 | 143706 | 139620 | 145542 | 141832 | 141810 | 145596 |
| 15 | 147514 | 147253 | 143603 | 148641 | 143277 | 149776 | 144574 | 148286 | 141465 | 147706 |
| 16 | 168807 | 168124 | 172583 | 168381 | 170570 | 170751 | 171677 | 170238 | 170838 | 168905 |
| 17 | 155837 | 152655 | 143734 | 153173 | 152547 | 153734 | 152592 | 153402 | 150192 | 153188 |
| 18 | 215600 | 221441 | 216101 | 220540 | 215996 | 214856 | 221577 | 222127 | 221947 | 224807 |
| 19 | 158157 | 157112 | 159313 | 153740 | 153630 | 156002 | 161583 | 158546 | 154998 | 161073 |
| 20 | 106048 | 103751 | 100016 | 101687 | 104469 | 106300 | 101719 | 104443 | 100614 | 105620 |
| Total | 2787752 | 2763367 | 2750956 | 2754602 | 2763819 | 2767257 | 2785446 | 2802879 | 2750743 | 2815508 |

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1 Score |
| 1 | 0.888418826 | 0.880137431 |
| 2 | 0.896186441 | 0.890141745 |
| 3 | 0.880699894 | 0.874806015 |
| 4 | 0.890185676 | 0.885327 |
| 5 | 0.892724376 | 0.884832859 |
| 6 | 0.887293993 | 0.880129068 |
| 7 | 0.879851143 | 0.871924196 |
| 8 | 0.896221394 | 0.885921697 |
| 9 | 0.875931842 | 0.867290142 |
| 10 | 0.879530917 | 0.869533883 |
| Average | 0.88670445 | 0.879004404 |