

# Task 7. Word Embedding by Using Word2Vec and BERT

---

Github link: [https://github.com/Allen-ZKW/NLP\\_HZAU/tree/task7](https://github.com/Allen-ZKW/NLP_HZAU/tree/task7)

Author: Kewei Zhao

Date: 2021-5-22

## Abstract

---

In this task, we use two algorithms to complete Word Embedding. These two algorithms all try to transform words to vectors, these vectors will reflect the closeness and similarity between two different words.

## Principle

---

Both of these two algorithms' target is translating word to type of information which can be understood by computers and programs. We can use 'translate data' to calculate distance between two different words. BERT algorithm takes word in different context into concern, so the result of BERT will include more information. This model is trained by forwards information and afterwards information.

## Measure

---

### Data Preparation

```
unzip data/litcovid-trainingdata.zip
```

### Environmental Preparation

```
conda create -n NLP_task7 python=3.8
conda activate NLP_task7
conda install pytorch torchvision torchaudio cpuonly -c pytorch
conda install --yes --file requirements.txt
```

### Word2Vec

```
python3 Skip_Gram_basic.py
```

### BERT

```
python3 Bert_4_Litcovid_WordEmbedding.py
```

## Result

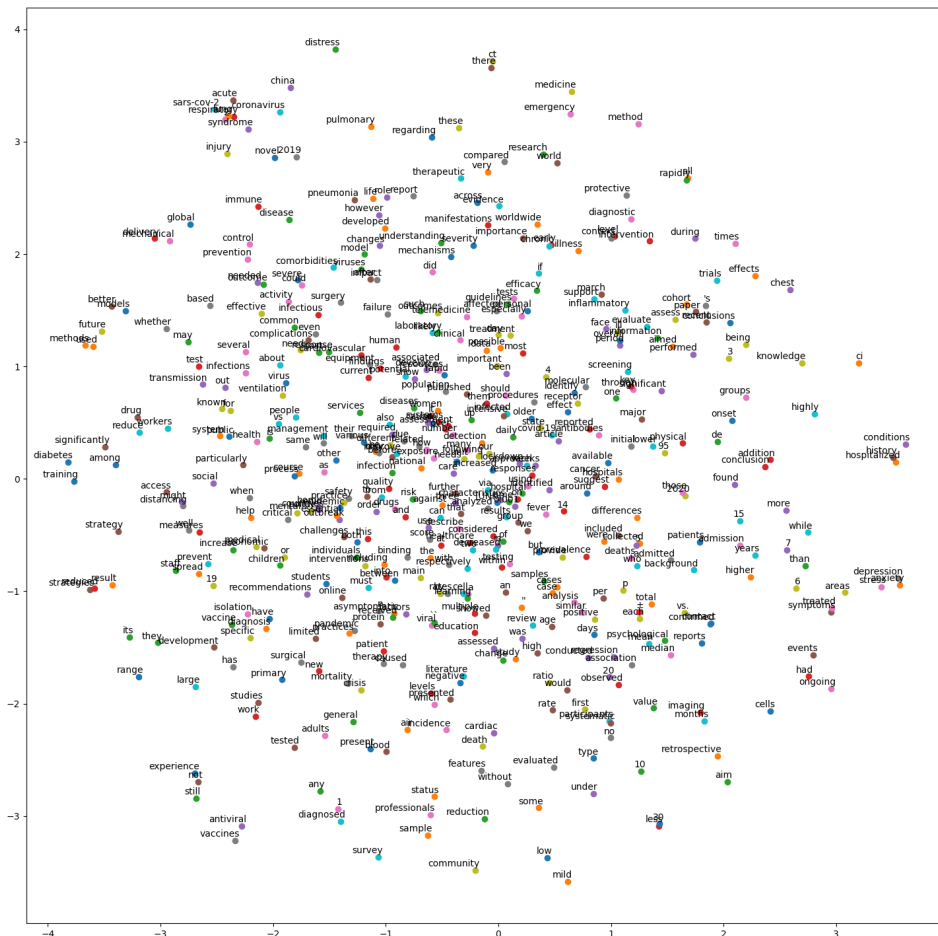
---

the 0.35628464818000793 -0.525038063155156 -0.2492087334410855 -0.805640151866537 -0.9308571815490723 1.630868166685104 0.13157081604003906 -0.0601432062685489  
0.08181028813127303 0.5495608448982239 0.316282063726105 0.04454610869288445 0.794414818286958 0.3056371808052063 0.3184380323287813 0.173862993717936 0.1565353;  
47905 -0.0421967947483063 0.14156823348999 -0.409829166195179504 0.3186410367488861 -1.165080544921875 -0.05547158420085907 0.7745139002799988 0.43387165665626526 -0.2;  
2594403326513813 0.5255263491791294 0.1762919505798035 -0.3134252727031708 -0.3362714502200314 0.204840442167663574 -0.1781500596046448 -0.958416235923767 -1.22800011;  
0.1184807771404785 0.05566768083572388 0.29235568165791104 -0.6948006947189331 -0.029690961837769 1.2941761445999146 0.605786991984558 0.596462824821472 -0.03026198;  
7 -1.5654470920562744 0.10546262799488068 -0.3829757869243622 -0.1052527189254767 0.1069300174713134 0.784586581488342 0.5608899993706824 0.2748860169649124 0.52680;  
864502906799 0.344282597303395 0.277339745888005 -0.715245004686279 -0.2468967556755726 -0.7228854958959164 -1.205928707122803 0.00330242700874805 1.121882319;  
32605 -1.6373316049575806 1.644091844587518 -0.43068304597936096 -0.47330057610022 -1.2751322984945345 -0.252463102304069824 1.0536020994186401 -0.9704072517967224 -0.3;  
2358283276654053 -0.4130273232950851 0.01643935342027773 0.216070756316105 0.184273370680902 0.4754025502570374 -0.566543872748764 -0.255428079607531 1.06030929088;  
62297058105 0.017322789877653122 0.1453235198888779 -0.728230714797936 -0.91285436602197876 -0.852407734675598 0.1137823760509491 -0.519420328140259 0.408936089096;  
97026824951 -0.03319588829576969 0.1041843812599182 0.5999304056167603 -0.0292342325015187 -0.387977193069458 -0.2799862361051949 -1.423446774482727 0.1086149772373881;  
-0.2590559422969818 0.5961945652961734 0.47084878975183105 0.4812585143455963 -0.473358631134003 0.00340399363015296 0.50695906444615 0.07217913866034091 0.49646727;  
28891 -0.5169460773468018 -1.1743741035461262 -0.711765451069641 -0.4870204291120605 -1.606303453454346 -0.1689727306359668 1.5204494978967285 -0.1661432683467865 -C  
80347061157 0.2954944656085968 -0.1853585512483215 0.735328104001917 0.3123325705528259 0.6593093872070312 0.3685827851295471 0.21299351571804352 0.9169759750366211  
04139697551727 -0.02953370990753174 0.5013274550437927 -0.21574245393276215 -0.5316971233467102 -0.9796431064605513 0.28284478187561035 -0.15776784718036652 0.99296814;  
of 0.2614080011844635 -0.057005424054274688 0.050627123564481395 -0.12816961109638214 -0.5257653594017029 0.0948389858007431 0.34602688003245276 0.343322604894638;  
928033351898 -0.1149474084372888 0.4615645110607147 0.34808754987716675 0.1813564515903473 0.06026953431198921 -0.9813010954856873 0.1087572202864487 0.0937142521144;  
095 005110456408461205 -0.29497521157504 0.485073036695488035 -0.7861011624356245 0.3377158033487809 -0.8752107024192810 0.5334144234657288 0.8603790998458662 0.183168;  
0.05487389964163303 -0.1913101353584595 -0.1402575075623673 0.22328968312740326 -0.113235095846417615 0.12927047908306122 -0.25044979810147 0.40160505122184755 -0.3;  
76 0.0342016385184284 -0.4061035931371749 -0.03724380955100595 0.33968245983123708 0.516175746917746 -0.4283783043368552 -0.5008599162107146 0.540671706199646 0.06227;  
-0.5704774260520995 -0.1018802668710358 -0.5763516426086426 -0.838429221748352 -0.28841865301312 0.05828624930143356 -0.0714394822710731 0.317959915961265564 -0.126;  
42 0.230325299441101074 0.29350471495682031 0.1147939413785934 0.68363495280265808 0.18847321847820282 0.1635154487375385 -0.38685430770392236 0.1314295083284378 0.5;  
-0.754610776601261 0.061656504896107 0.4786179184913635 -0.087644598340252 0.3273068770140076 0.983331741532898 -0.04582743719202164 -0.1225066874050974 -0.56;  
74774 0.058116699272637177 -0.08491975402832 -0.4840514063835144 -0.2196997255086898 -0.0087560626533124 -0.02952695101499557 0.2120148413658142 0.023833980783821;  
4354038239 0.11455037319539642 -0.2036601069847107 -0.186210058824157 -0.28095537424087524 -0.07885187864303589 0.1127619478702545 -0.198195107028305 -0.57789272;  
013820400977448 0.042733740992

## Word2Vec Vectors

[UNK] -0.04881330579519272 -0.0379069391010898 -0.05652253329753876 -0.09073057370729065 -0.11317484825849533 0.05839929729700885 0.020258576345443726 0.04351604357  
14718006134 -0.011797694489359856 0.10310874134302139 0.01187390339958660 0.11605213582515717 -0.00109352682209015 -0.04330721125006676 0.03871  
611858368  
the -0.143330588968680573 -0.1587565541267395 0.14328989847507477 -0.3877052068710327 -0.10853473864144012 -0.03758662149310110 0.04662489991052246 0.02643380529  
78235 -0.04327136278152466 -0.07235678285360336 -0.3090354800224304 0.12058231979608536 -0.02043553069238943 0.10384968668222427 0.28311672806739807 0.135265320539474  
966949463  
and 0.03092149645090103 0.06394444406032562 0.22052952647209167 0.12312621623277664 0.13519074022769928 -0.33134761452674866 -0.08193377405405045 0.17148441076278  
10996109992265701 0.08987542241811752 -0.06514536589384079 0.0856313705444336 -0.06625906378030771 0.11863439530134201 -0.1884736567735672 -0.07212437689304352 0.09053  
7  
-0.022927175748827267 -0.27186140418052673 -0.1382113109588623 0.2388809472322464 -0.0699752010450636 0.0157379353594894 -0.2610098421576369 -0.031904140  
710476875305 0.1495684082764587 -0.26795533299446106 -0.29544875025749207 0.1886664034524515 -0.15007612109184265 0.13896331232738495 -0.015536155551671982 0.1629645  
803672458976507  
in 0.07111769165969016 -0.09316769242286682 -0.024860944598913193 0.12261395901441574 -0.32726961374282837 0.15798386931419373 -0.0223913059949875 -0.03950957953  
1018983211368322232 -0.1676515047818298 -0.17081472277641296 0.033556707230388214 -0.23027724027633667 0.1060002226572037 0.17270736639312714 0.2785123884677887 -0.06  
to -0.0718165785074234 -0.06219632923603366 -0.11162675172209053 -0.0351937639859581 0.043366577529907 0.009930439293384552 -0.06467465311288834 -0.4905491769313  
26825 0.0883138179790527 0.025115570533466234 -0.434296139738678 -0.2750214220903966 -0.035340536386776 -0.0367852038227844 -0.095753255062651 0.13776637760737787  
-0.25891706347405155 -0.2228063438036730 -0.1143641273754127 -0.33120733499527 -0.0591867561807036 -0.2348829507827758 -0.2785216867937366 0.051779131644213195  
17957484722137 0.24900034070014954 0.11283621791978455 -0.1255256626343536 -0.07017210125923157 0.009417561814188957 0.10308688732981682 0.07272831350564957 0.141226  
with -0.0985988451363373 -0.2385247264729862 0.1175894141972064 0.06500596416130066 -0.0687106409549731 -0.2874845266342163 -0.0856295964479446 0.1696904450654  
164905246347018895 0.21965739130973816 -0.3124369978904742 -0.1029041260480807 -0.0461577698588373 0.04010459035639945 0.2183129227161407 0.041780966281991 0.22014  
for -0.0885828405618668 0.0496838003300949 0.0377464219972778 -0.06489270677478821 0.04539823531045 -0.11270886445056 -0.04672267660496819 0.1040973588824  
1584 -0.2120668433112946 -0.039591915905475616 -0.6262766969906681 -0.2139615416527944 0.05573865562677383 -0.08018310368061066 0.13496974753836972 0.12017136662715202  
covid-19 0.33330416673982324 0.01479049479856491 0.1470600680959691833 -0.023025231392703056 0.191682571331787 -0.11672676154646683 0.08899170397015734 0.146422351817  
55426 -0.04511909931898117 0.2939341962337494 0.24353934824466705 0.17898909747600555 -0.00947872167663956 -0.3241097927093506 -0.07146181166172028 -0.196622908115386  
patients 0.04939527046561844 -0.1442393667353493 -0.1095363081378937 -0.25172326910893564 -0.0855308729410172 -0.0432542273368835 0.195867627859156 0.3668800294939  
922166109085 -0.071305499131562184 0.0436086118221283 -0.02143552626831627 0.03789140389593545 0.0390790104866072 -0.0972571071789856 -0.0541148111224745 0.1774584  
16  
we 0.5449550747871999 0.12638379633426666 0.0373534739390373 0.08160431683063507 -0.11231604963541031 0.15202890336513552 -0.3664050400257106 0.1093474701046943  
883780777454376 -0.6140636205673218 -0.09264766424894333 0.023603010922670364 -0.13203932344913483 0.09088529646396637 0.2482601147651672 0.026199063286185265 0.11704  
2937  
is -0.201849744273529 -0.0146817662048364891 -0.27701535820861 -0.348742008292285 0.5759435891025103 0.1453063189983668 0.0859629511833109 0.4284841820659375 (3  
297199554443 0.04105306416749954 -0.1034683212679013 0.02682025730609887 -0.18854

## Word2Vec Visualization



## Discussion

In these two graphs, we can find that in BERT result word mainly separate in two parts while Word2Vec result stay in one part. Also, in BERT result, there exists some nodes which stay very close. Although, Word2Vec result's nodes all exist in one part, the distance between two nodes is longer than BERT model. BERT result is more obvious, which may be one reason of playing a better role in word embedding.