CS 512

Assignment 2 Extra Credit

Bing Liang NetID: bingl3

Result

	F1 (micro)	F1 (macro)
word2vec-feature (128-d)	0.23475766878038942	0.15646372745195353

Analysis

I utilized word2vec model to generate a 128-d embedding for each word, and for each word in the title (after cleaning), I sum the embedding of each word up and divide by the word count of the title. Theoretically, using word2vec to extra the title feature should be better that the text-based feature because it takes similarity of words in consider. However, the experimental result shows that the performance of using word2vec feature is worse than text-based feature. After some consideration, I believe it may because, when training the the word2vec model, I used the cleaned training data, which is about 50M. This size of training set can not make word2vec model to generate a good embedding. As a result, the embedding representation of feature is somehow not very accurate .

Further Improvement

We can firstly train the word2vec model by a large corpus, let's say, in gigabyte level. Turns out, the word2vec model could generate a more accurate embedding for words, so we can get a more accurate representation of title (feature). In addition, we could also utilize the HIN, embedding not only the title, but also the cited paper venue. But there should be a

weight, which determine which one (title and cited paper's venue) has more impact. There is the hyperparameters we need to tune.

Precision and Recall Per Venue

Text-based features (could also be found in output/result_w2c_clf.txt):

Venue Precision Recall

aamas 0.28444444444444444 0.1833810888252149

acc 0.018518518518518517 0.009174311926605505

acm_multimedia 0.12704174228675136 0.18469656992084432

acm_trans._graph. 0.0 0.0

amcis 0.21379310344827587 0.09323308270676692

amia 0.25675675675675674 0.15702479338842976

asp-dac 0.11475409836065574 0.0219435736677116

bioinformatics 0.0 0.0

cdc 0.47860962566844920.4246737841043891

chi 0.14424410540915394 0.24880382775119617

chi_extended_abstracts 0.21367521367521367 0.176056338028169

cikm 0.12605042016806722 0.03978779840848806

cogsci 0.39368421052631580.4308755760368664

coling 0.32291666666666670.16893732970027248

commun._acm0.0 0.0

compsac 0.18367346938775510.030405405405405407

comput._graph._forum 0.0 0.0

comput._j. 0.0 0.0

computer_communications 0.0 0.0

computer_networks 0.0 0.0

corr 0.0 0.0

cvpr 0.26732673267326734 0.13659359190556492

dac 0.21942446043165467 0.1367713004484305

date 0.208955223880597 0.05714285714285714

ecai 0.06306306306306306 0.0374331550802139

ecis 0.18674698795180722 0.08051948051948052

embc 0.36666666666666664 0.1286549707602339

encyclopedia_of_database_systems 0.40677966101694920.15946843853820597

etfa 0.25641025641025640.0784313725490196

eurospeech 0.10810810810810811 0.012578616352201259

eusipco 0.09090909090909091 0.08215962441314555

expert_syst._appl. 0.0 0.0

focs 0.23255813953488372 0.03861003861003861

fskd 0.02912621359223301 0.010380622837370242

fundam._inform. 0.0 0.0

fusion 0.35616438356164380.2727272727272727

fuzz-ieee 0.46946564885496184 0.5020408163265306

gecco 0.37037037037037035 0.2857142857142857

globecom 0.10230489284270117 0.1912320483749055

hicss 0.23370638578011850.36187563710499493

icalt 0.40293040293040294 0.3254437869822485

icarcv 0.03896103896103896 0.022304832713754646

icassp 0.08668076109936575 0.22202166064981949

icc 0.10081053698074975 0.14420289855072463

iccad 0.18471337579617833 0.09602649006622517

iccs 0.06 0.014218009478672985

iccv 0.15463917525773196 0.05514705882352941

icdar 0.45634920634920634 0.40492957746478875

icde 0.19070904645476772 0.22740524781341107

icecs 0.05263157894736842 0.0034129692832764505

icip 0.12637362637362637 0.024287222808870117

icis 0.225 0.022113022113022112

icmc 0.5553055350553351 0.6919540229885057

icme 0.09216589861751152 0.18518518518518517

icml 0.20105820105820105 0.14393939393939395

icnc 0.28 0.03333333333333333

icpr 0.14 0.015350877192982455

icra 0.39983129481231155 0.5287228109313998

icse 0.38686131386861317 0.16358024691358025

icslp 0.22159090909090911 0.13780918727915195

ieee_computer 0.0 0.0

ieee_congress_on_evolutionary_computation 0.3547297297297297 0.26515151515151514

ieee_journal_on_selected_areas_in_communications 0.0 0.0

ieee_software 0.0 0.0

ieee_trans._computers 0.0 0.0

ieee_trans._information_theory 0.0 0.0

ieee_trans._knowl._data_eng. 0.0 0.0

ieee_trans._parallel_distrib._syst. 0.0 0.0

ieee_trans._pattern_anal._mach._intell. 0.0 0.0

ieee_trans._software_eng. 0.0 0.0

igarss 0.6696061140505585 0.859622641509434

ijcai 0.11193058568329718 0.34308510638297873

ijcnn 0.15625 0.04604051565377532

inf._process._lett. 0.0 0.0

inf._sci. 0.0 0.0

infocom 0.21290322580645160.09565217391304348

int._cmg_conference 0.61410788381742740.4134078212290503

interspeech 0.4527027027027027 0.481629392971246

ipdps 0.24045801526717558 0.1403118040089087

iros 0.18155053974484790.1322373123659757

isbi 0.48177083333333330.44364508393285373

iscas 0.26014760147601473 0.47581552305961755

iscc 0.12820512820512820.015479876160990712

isit 0.43734015345268540.3717391304347826

itc 0.604591836734693 9 0.5243362831858407

j._acm 0.0 0.0

j._parallel_distrib._comput. 0.0 0.0

j._symb._log. 0.0 0.0

journal_of_systems_and_software 0.0 0.0

kdd 0.19230769230769232 0.04184100418410042

lcn 0.07009345794392523 0.05244755244755245

lrec 0.50366748166259170.44685466377440347

multimedia_tools_appl. 0.0 0.0

neuroimage 0.0 0.0

nips 0.14385474860335196 0.17517006802721088

pacis 0.16860465116279070.09931506849315068

pattern_recognition 0.0 0.0

pdpta 0.10967741935483871 0.05647840531561462

pimrc 0.11624203821656051 0.16258351893095768

robio 0.12727272727272726 0.017326732673267328

sac 0.03225806451612903 0.0831889081455806

siam_j._comput. 0.0 0.0

sigcse 0.705882352941176 50.4897959183673469

sigir 0.29498525073746310.2849002849002849

sigmod_conference 0.12589928057553956 0.11705685618729098

smc 0.04005252790544977 0.07003444316877153

soda 0.28877005347593580.18493150684931506

softw._pract._exper. 0.0 0.0

stoc 0.27513227513227510.16720257234726688

theor._comput._sci. 0.0 0.0

vlsi_design 0.178082191780821900.04797047970479705

vtc_fall 0.12116788321167883 0.15930902111324377

vtc_spring 0.19230769230769232 0.033112582781456956

wcnc 0.128686327077748 0.12291933418693982

winter_simulation_conference 0.4466374269005848 0.699885452462772