Data Analysis

0: Annotations

Makes sense
Has some correlation
Does not make sense

1: Newsgroup Similarity Comparison: Jaccard vs Cosine vs Dice:

1.1: soc.religion.christian

Jaccard	Cosine	Dice
talk.politics.mideast(similarity=0.011324376199616123)	talk.religion.misc(similarity=0.9919183114052993)	talk.politics.mideast(similarity=0.02239514139305371)
rec.sport.hockey(similarity=0.0111303012857417)	alt.atheism(similarity=0.9891265888210407)	rec.sport.hockey(similarity=0.022015562725374836)
talk.politics.misc(similarity=0.0109847939751523)	talk.politics.misc(similarity=0.9862178149194812)	talk.politics.misc(similarity=0.021730878724615675)

1.2: rec.autos

Jaccard	Cosine	Dice
rec.sport.hockey(similarity=0.009435317783418747)	rec.motorcycles(similarity=0.9917449260765105)	rec.sport.hockey(similarity=0.018694249383184664)
talk.politics.mideast(similarity=0.009290297864189254)	rec.sport.baseball(similarity=0.9811252846768931)	talk.politics.mideast(similarity=0.018409565382425507)
soc.religion.christian(similarity=0.009097002776979795)	talk.politics.guns(similarity=0.9777878622971252)	soc.religion.christian(similarity=0.01802998671474663)

As shown in 1.1 and 1.2, Jaccard and Dice is unable to find closely related newsgroups. On the other hand, cosine similarity does a much better job. This is expected as Dice and Jaccard puts emphasis in the intersection of words. This is less effective in this case because there are many vocabulary in English that is unnecessary in determining the newsgroup similarity. In this case, it acts as noise and makes it worse.

2: Word document(cosine sim) vs Word Context(cosine sim) vs Word2Vec

2.1: money

Word Document	Word Context	Word2Vec
class(similarity=0.9767438403760214)	up(similarity=0.9603185130010726)	fun(similarity=0.6567017436027527)
increase(similarity=0.9634182042669379)	us(similarity=0.9601583637846574)	much(similarity=0.6381500959396362)
em(similarity=0.9597610728746588)	use(similarity=0.9601296809924658)	bike(similarity=0.6243270635604858)
care(similarity=0.9589233925927593)	then(similarity=0.9568131101947593)	weight(similarity=0.618684709072113)
pay(similarity=0.9572138758030033)	all(similarity=0.9558875026151603)	talent(similarity=0.6144708395004272)

2.2: Engineer

Word Document	Word Context	Word2Vec
makers(similarity=0.9665607275877093)	m(similarity=0.7614772816221814)	mellon(similarity=0.7146978974342346)
unit(similarity=0.9640027414815829)	sabre(similarity=0.7521588786978674)	corp(similarity=0.7134993076324463)
interior(similarity=0.9618744428455661)	d(similarity=0.7502770864330681)	austin(similarity=0.7006423473358154)
needing(similarity=0.9602778606944725)	e(similarity=0.74991351328366)	electronics(similarity=0.6859579682350159)
windows(similarity=0.956533701375456)	UNK(similarity=0.7368149595966172)	packard(similarity=0.6827089190483093)

It is obvious from 2.1 and 2.2 that word context performs the worse, followed by word document. It is no surprise that word2vec performs the best because it is more complicated in nature and does not look merely at the word count.

In 2.2, it is important to note that ‘mellon’ and ‘austin’ usually show up in the form of “carnegie **mellon** pittsburgh” and “**austin** ibm com”. When put together, these words are more engineer-like.

3: tf-idf comparison (newsgroup)

3.1: talk.politics.misc

tf-idf	non tf-idf
talk.politics.guns(similarity=0.07494059012684189)	talk.religion.misc(similarity=0.9892166807719729)
soc.religion.christian(similarity=0.029462289645630173)	talk.politics.guns(similarity=0.9865660369163336)
talk.religion.misc(similarity=0.025656448074236136)	soc.religion.christian(similarity=0.9862178149194812)

3.2: rec.sport.baseball

tf-idf	non tf-idf
rec.sport.hockey(similarity=0.05906752656305329)	rec.autos(similarity=0.9811252846768931)
rec.autos(similarity=0.022159636063377845)	talk.politics.guns(similarity=0.9759752875554392)
talk.politics.guns(similarity=0.011665303817526718)	rec.motorcycles(similarity=0.9750420488961419)

As shown in 3.1 and 3.2, I think tf-idf performs slightly better than non tf-idf. Looking at the tf-idf algorithm, we see that the ‘idf’ (inverse document frequency) part diminishes the weight of terms that occur very frequently and increases the weight of terms that occur rarely. As a result, it is no surprise to see tf-idf give better suggestions.

4: PPMI comparison (word context)

4.1: handgun

ppmi	non ppmi
handguns(similarity=0.429128475978102)	without(similarity=0.9176317357786625)
concealed(similarity=0.424120057931429)	gun(similarity=0.9117507309911148)
firearms(similarity=0.40896644913607116)	small(similarity=0.9116211591552732)
firearm(similarity=0.40735467266988645)	great(similarity=0.9106385956337493)
file(similarity=0.40045108175915245)	man(similarity=0.9097660974492888)

4.2: money

ppmi	non ppmi
care(similarity=0.4669921743051763)	up(similarity=0.9603185130010726)
want(similarity=0.4647178872041391)	us(similarity=0.9601583637846574)
lot(similarity=0.4646366424150167)	use(similarity=0.9601296809924658)
car(similarity=0.46341268251509643)	then(similarity=0.9568131101947593)
going(similarity=0.4624452732858013)	all(similarity=0.9558875026151603)

From 4.1 and 4.2, it is obvious using ppm yields better results.

5: Log

1. Data loaded.

2. Created term_newsgroup_mat

3. Created term_context_mat

4. Test Newsgroup similarities (cosine)

test_word_similarity | Top 3 similar words for soc.religion.christian is
['talk.religion.misc(similarity=0.9919183114052993)',
'alt.atheism(similarity=0.9891265888210407)',
'talk.politics.misc(similarity=0.9862178149194812)']
test_word_similarity | Last 3 similar words for soc.religion.christian is
['rec.sport.hockey(similarity=0.8915391730047594)',
'rec.motorcycles(similarity=0.9151165514765032)', 'rec.autos(similarity=0.9343163675563361)']
test_word_similarity | Top 3 similar words for rec.autos is
['rec.motorcycles(similarity=0.9917449260765105)',
'rec.sport.baseball(similarity=0.9811252846768931)',
'talk.politics.guns(similarity=0.9777878622971252)']
test_word_similarity | Last 3 similar words for rec.autos is
['soc.religion.christian(similarity=0.9343163675563361)',
'alt.atheism(similarity=0.9440258808670355)',
'talk.politics.mideast(similarity=0.951251048894889)']
test_word_similarity | Top 3 similar words for talk.politics.misc is
['talk.religion.misc(similarity=0.9892166807719729)',
'talk.politics.guns(similarity=0.9865660369163336)',
'soc.religion.christian(similarity=0.9862178149194812)']
test_word_similarity | Last 3 similar words for talk.politics.misc is
['rec.sport.hockey(similarity=0.9103587788043602)',
'rec.motorcycles(similarity=0.9357207850074614)', 'rec.autos(similarity=0.9524401741067724)']
test_word_similarity | Top 3 similar words for rec.sport.hockey is
['rec.sport.baseball(similarity=0.9740641525100004)',
'rec.autos(similarity=0.952038820849965)', 'rec.motorcycles(similarity=0.9498687411455566)']
test_word_similarity | Last 3 similar words for rec.sport.hockey is
['alt.atheism(similarity=0.8911479472362545)',
'soc.religion.christian(similarity=0.8915391730047594)',
'talk.politics.misc(similarity=0.9103587788043602)']
test_word_similarity | Top 3 similar words for alt.atheism is
['talk.religion.misc(similarity=0.9902247786108441)',
'soc.religion.christian(similarity=0.9891265888210407)',
'talk.politics.misc(similarity=0.9839637429053031)']
test_word_similarity | Last 3 similar words for alt.atheism is
['rec.sport.hockey(similarity=0.8911479472362545)',
'rec.motorcycles(similarity=0.9261325125888774)', 'rec.autos(similarity=0.9440258808670355)']
test_word_similarity | Top 3 similar words for rec.sport.baseball is
['rec.autos(similarity=0.9811252846768931)',
'talk.politics.guns(similarity=0.9759752875554392)',
'rec.motorcycles(similarity=0.9750420488961419)']

test_word_similarity | Last 3 similar words for rec.sport.baseball is
['soc.religion.christian(similarity=0.9414533520073202)',
'alt.atheism(similarity=0.9465506374413303)',
'talk.politics.mideast(similarity=0.9547631765203182)']
test_word_similarity | Top 3 similar words for talk.politics.mideast is
['talk.religion.misc(similarity=0.9839241688803108)',
'talk.politics.misc(similarity=0.9839129050563167)',
'talk.politics.guns(similarity=0.9835764301830736)']
test_word_similarity | Last 3 similar words for talk.politics.mideast is
['rec.sport.hockey(similarity=0.9190893129740321)',
'rec.motorcycles(similarity=0.9366922026910317)', 'rec.autos(similarity=0.951251048894889)']
test_word_similarity | Top 3 similar words for rec.motorcycles is
['rec.autos(similarity=0.9917449260765105)',
'rec.sport.baseball(similarity=0.9750420488961419)',
'talk.politics.guns(similarity=0.9672540187483335)']
test_word_similarity | Last 3 similar words for rec.motorcycles is
['soc.religion.christian(similarity=0.9151165514765032)',
'alt.atheism(similarity=0.9261325125888774)',
'talk.politics.misc(similarity=0.9357207850074614)']
test_word_similarity | Top 3 similar words for talk.politics.guns is
['talk.religion.misc(similarity=0.9869326016798315)',
'talk.politics.misc(similarity=0.9865660369163336)',
'talk.politics.mideast(similarity=0.9835764301830736)']
test_word_similarity | Last 3 similar words for talk.politics.guns is
['rec.sport.hockey(similarity=0.9391679426479208)',
'rec.motorcycles(similarity=0.9672540187483335)',
'soc.religion.christian(similarity=0.9719267074671791)']
test_word_similarity | Top 3 similar words for talk.religion.misc is
['soc.religion.christian(similarity=0.9919183114052993)',
'alt.atheism(similarity=0.9902247786108441)',
'talk.politics.misc(similarity=0.9892166807719729)']
test_word_similarity | Last 3 similar words for talk.religion.misc is
['rec.sport.hockey(similarity=0.9179238170558549)',
'rec.motorcycles(similarity=0.9460793902775345)', 'rec.autos(similarity=0.960075252861954)']

5. Test Word similarities (cosine)
test_word_similarity | Top 10 similar words for lethal is ['waco(similarity=0.9965612343379819)',
'kerosene(similarity=0.9959075764325377)', 'idbsu(similarity=0.9956824505818589)',
'uxh(similarity=0.9947897254651769)', 'batf(similarity=0.9939985333599834)',
'compound(similarity=0.9911789699188875)', 'encryption(similarity=0.9910860560585223)',
'tennessee(similarity=0.9908298963032667)', 'federalist(similarity=0.9903123096385865)',
'bills(similarity=0.9901482076345278)']
test_word_similarity | Last 10 similar words for lethal is ['nmm(similarity=0.0)',
'tharp(similarity=0.0)', 'cds(similarity=0.0)', 'davet(similarity=0.0)', 'virago(similarity=0.0)',
'sturges(similarity=0.0)', '3com(similarity=0.0)', 'motorcyclists(similarity=0.0)',
'denizens(similarity=0.0)', 'sears(similarity=0.0)']
test_word_similarity | Top 10 similar words for handgun is
['firearm(similarity=0.9999624232692909)', 'cipriani(similarity=0.9999226257946291)',
'dividian(similarity=0.999919765307815)', 'crphilli(similarity=0.999854872649987)',

'hound(similarity=0.999854872649987)', 'dazixca(similarity=0.999854872649987)',
'jrm(similarity=0.999854872649987)', 'u28037(similarity=0.999854872649987)',
'rkba(similarity=0.999854872649987)', 'bms(similarity=0.999854872649987)']
test_word_similarity | Last 10 similar words for handgun is ['nmm(similarity=0.0)',
'tharp(similarity=0.0)', 'cds(similarity=0.0)', 'davet(similarity=0.0)', 'virago(similarity=0.0)',
'sturges(similarity=0.0)', '3com(similarity=0.0)', 'motorcyclists(similarity=0.0)',
'denizens(similarity=0.0)', 'sears(similarity=0.0)']
test_word_similarity | Top 10 similar words for money is
['class(similarity=0.9767438403760214)', 'increase(similarity=0.9634182042669379)',
'em(similarity=0.9597610728746588)', 'care(similarity=0.9589233925927593)',
'pay(similarity=0.9572138758030033)', 'board(similarity=0.9550166338088882)',
'forward(similarity=0.9549707090282492)', 'virtually(similarity=0.9513969848573453)',
'mine(similarity=0.9512385045020354)', 'considering(similarity=0.950378913903427)']
test_word_similarity | Last 10 similar words for money is
['schismatic(similarity=0.11585657190134323)', 'chalcedon(similarity=0.11585657190134323)',
'sarto(similarity=0.11585657190134323)', 'jhpb(similarity=0.11585657190134323)',
'scroggs(similarity=0.11585657190134323)', 'whitsell(similarity=0.11585657190134323)',
'sirach(similarity=0.11585657190134323)', 'caralv(similarity=0.11585657190134323)',
'crossroads(similarity=0.11585657190134324)', 'nasb(similarity=0.11585657190134324)']
test_word_similarity | Top 10 similar words for engineer is
['makers(similarity=0.9665607275877093)', 'unit(similarity=0.9640027414815829)',
'interior(similarity=0.9618744428455661)', 'needing(similarity=0.9602778606944725)',
'windows(similarity=0.956533701375456)', 'equipped(similarity=0.9547802207405517)',
'electronics(similarity=0.9540012488175563)', 'worn(similarity=0.9535492477839298)',
'assembly(similarity=0.9489765796783615)', 'spelling(similarity=0.9445004164979767)']
test_word_similarity | Last 10 similar words for engineer is ['nmm(similarity=0.0)',
'tharp(similarity=0.0)', 'cds(similarity=0.0)', 'davet(similarity=0.0)', 'virago(similarity=0.0)',
'sturges(similarity=0.0)', '3com(similarity=0.0)', 'motorcyclists(similarity=0.0)',
'denizens(similarity=0.0)', 'sears(similarity=0.0)']
test_word_similarity | Top 10 similar words for rich is ['vastly(similarity=0.9751798419308304)',
'facilities(similarity=0.9700429851443605)', 'primary(similarity=0.9521586978939848)',
'clue(similarity=0.9496313894101055)', 'variety(similarity=0.9448880416076804)',
'advocate(similarity=0.9397600578188596)', 'routine(similarity=0.9390552131003703)',
'indicate(similarity=0.938431508021737)', 'stands(similarity=0.938350782598531)',
'closely(similarity=0.937277502003693)']
test_word_similarity | Last 10 similar words for rich is
['shamim(similarity=0.05770041704858152)', 'nye(similarity=0.05770041704858152)',
'nyeda(similarity=0.05770041704858152)', 'buphy(similarity=0.05770041704858152)',
'jaeger(similarity=0.05770041704858152)', 'karner(similarity=0.057700417048581525)',
'fanatism(similarity=0.057700417048581525)', 'beauchaine(similarity=0.057700417048581525)',
'bobbe(similarity=0.057700417048581525)', 'ingles(similarity=0.057700417048581525)']
test_word_similarity | Top 10 similar words for republican is
['feasible(similarity=0.9952608536713679)', 'resign(similarity=0.9946943735478533)',
'reagan(similarity=0.9942214424884378)', 'libertarian(similarity=0.9939863401803372)',
'govt(similarity=0.9936207312752264)', 'senior(similarity=0.9900270171791653)',
'drugs(similarity=0.9897188177865796)', 'adjective(similarity=0.989096333033071)',
'desy(similarity=0.9879753580966648)', 'funds(similarity=0.987613041514261)']
test_word_similarity | Last 10 similar words for republican is ['balltown(similarity=0.0)',
'cabot(similarity=0.0)', 'welty(similarity=0.0)', 'heiser(similarity=0.0)', 'wb3ffv(similarity=0.0)',

'goucher(similarity=0.0)', 'delco(similarity=0.0)', 'qazi(similarity=0.0)', 'delcoelect(similarity=0.0)',
'kocrsv01(similarity=0.0)']

6. Test Context similarities (cosine)

test_word_similarity | Top 10 similar words for lethal is
['practice(similarity=0.7760815184415686)', 'similar(similarity=0.7678458112828889)',
'non(similarity=0.766693911463068)', 'certain(similarity=0.7634279655156974)',
'common(similarity=0.7605236126470136)', 'religion(similarity=0.7601873271891662)',
'particular(similarity=0.759560407488776)', 'different(similarity=0.7593902209027847)',
'legal(similarity=0.7553267399385606)', 'religious(similarity=0.7540666559524279)']
test_word_similarity | Last 10 similar words for lethal is ['cunixc(similarity=0.0)',
'8563446(similarity=0.0012744094606823616)', 'z1dan(similarity=0.0026756601486263125)',
'syl(similarity=0.004095281052907502)', 'terresterial(similarity=0.0042163429564793554)',
'fxwg(similarity=0.0043017218937217 74)', 'uu4(similarity=0.00527396295344497)',
'rscharfy(similarity=0.005817451418636954)', '92717(similarity=0.007428324807994075)',
'budd(similarity=0.011260599584471797)']
test_word_similarity | Top 10 similar words for handgun is
['without(similarity=0.9176317357786625)', 'gun(similarity=0.9117507309911148)',
'small(similarity=0.9116211591552732)', 'great(similarity=0.9106385956337493)',
'man(similarity=0.9097660974492888)', 'complete(similarity=0.908899254710954)',
'short(similarity=0.9077055210898648)', 'having(similarity=0.9070794064107418)',
'run(similarity=0.9062774316984719)', 'religious(similarity=0.9059989660666309)']
test_word_similarity | Last 10 similar words for handgun is ['cunixc(similarity=0.0)',
'z1dan(similarity=0.003175407279514281)', '8563446(similarity=0.004761377706023825)',
'terresterial(similarity=0.0049406715821844055)', '92717(similarity=0.00740087833621886)',
'syl(similarity=0.008772321735855359)', 'rscharfy(similarity=0.010007081777016655)',
'uu4(similarity=0.010031273325943224)', 'fxwg(similarity=0.014464668553373073)',
'budd(similarity=0.02039539845586068)']
test_word_similarity | Top 10 similar words for money is ['up(similarity=0.9603185130010726)',
'us(similarity=0.9601583637846574)', 'use(similarity=0.9601296809924658)',
'then(similarity=0.9568131101947593)', 'all(similarity=0.9558875026151603)',
'also(similarity=0.9537642808755673)', 'one(similarity=0.95348772750585)',
'them(similarity=0.9529053349522391)', 'out(similarity=0.9523257128525326)',
'right(similarity=0.9501889814811495)']
test_word_similarity | Last 10 similar words for money is
['8563446(similarity=0.0027701844398604876)',
'terresterial(similarity=0.0032812141159155165)', 'z1dan(similarity=0.006842451429008159)',
'cunixc(similarity=0.009558886192622719)', '92717(similarity=0.011460755218058769)',
'uu4(similarity=0.017104698266207242)', 'syl(similarity=0.019039695819677408)',
'1777(similarity=0.021003289106963724)', 'budd(similarity=0.02361043078390783)',
'rscharfy(similarity=0.023845915910570125)']
test_word_similarity | Top 10 similar words for engineer is ['m(similarity=0.7614772816221814)',
'sabre(similarity=0.7521588786978674)', 'd(similarity=0.7502770864330681)',
'e(similarity=0.74991351328366)', 'UNK(similarity=0.736814959596 6172)',
'wrote(similarity=0.730849306916171)', 'dod(similarity=0.7301109376393984)',
'k(similarity=0.7300238384715432)', '750(similarity=0.7299068208026112)',
'from(similarity=0.7288771814134118)']
test_word_similarity | Last 10 similar words for engineer is ['cunixc(similarity=0.0)',
'budd(similarity=0.004770469597970687)', 'wb3ffv(similarity=0.006796220309799095)',

'fxwg(similarity=0.007798145139346011)', 'chatham(similarity=0.011697217709019015)',
'1708(similarity=0.012457711206722162)', '1778(similarity=0.013212502205942825)',
'z1dan(similarity=0.01355266248311483)', 'terresterial(similarity=0.01476560215176502)',
'8302(similarity=0.015596290278692021)']
test_word_similarity | Top 10 similar words for rich is ['UNK(similarity=0.9509490411842516)',
'in(similarity=0.9448963115754381)', 'air(similarity=0.942888583550633)',
'from(similarity=0.940655021833566 1)', 'etc(similarity=0.9397053035083023)',
'hot(similarity=0.9377332099692838)', 'by(similarity=0.9367258289250366)',
'black(similarity=0.9364368661160101)', 'path(similarity=0.934604138361302)',
'and(similarity=0.9313585171295478)']
test_word_similarity | Last 10 similar words for rich is
['cunixc(similarity=0.004480648311401667)', '8563446(similarity=0.011903371049267574)',
'terresterial(similarity=0.012394523206605607)', 'z1dan(similarity=0.017400327058557118)',
'unforgiven(similarity=0.019156390749255264)', '92717(similarity=0.019279149810062336)',
'wb3ffv(similarity=0.025813099564786907)', 'fxwg(similarity=0.027256192848312957)',
'1777(similarity=0.028455255311720398)', 'munny(similarity=0.028867838224488126)']
test_word_similarity | Top 10 similar words for republican is
['man(similarity=0.9150809337027437)', 'with(similarity=0.9150108650273447)',
'dog(similarity=0.9133518585955777)', 'while(similarity=0.9131535790201056)',
'for(similarity=0.9129540476896036)', 'big(similarity=0.9125071446834768)',
'without(similarity=0.9112853152961086)', 'great(similarity=0.909495600774423)',
'real(similarity=0.907350744069411 3)', 'making(similarity=0.905573880246982)']
test_word_similarity | Last 10 similar words for republican is ['cunixc(similarity=0.0)',
'terresterial(similarity=0.0)', '8563446(similarity=0.0017302024229862193)',
'syl(similarity=0.0018870166142313663)', 'z1dan(similarity=0.002797304634740165)',
'fxwg(similarity=0.004247443012572129)', 'uu4(similarity=0.0059175186251546694)',
'rscharfy(similarity=0.0061030500957245494)', '92717(similarity=0.006890073743199138)',
'budd(similarity=0.009909979853607996)']

7. Test Word2Vec

test_word2vec_similarity | Top 10 similar words for lethal is
['toxic(similarity=0.8075752258300781)', 'weaponry(similarity=0.772251009941101 1)',
'competitive(similarity=0.76609867811203)', 'heavily(similarity=0.7441263794898987)',
'warfare(similarity=0.7415521144866943)', 'deadly(similarity=0.7395095825195312)',
'excessive(similarity=0.7388725876808167)', 'violent(similarity=0.7287948727607727)',
'armed(similarity=0.7261874079704285)', 'whites(similarity=0.7179344892501831)']
test_word2vec_similarity | Last 10 similar words for lethal is
['ms(similarity=-0.4632258713245392)', 'joe(similarity=-0.4449475109577179)',
'morning(similarity=-0.4119894206523895)', 'mark(similarity=-0.40045252442359924)',
'ps(similarity=-0.3948546350002289)', 'matthew(similarity=-0.38946157693862915)',
'george(similarity=-0.3883707523345947)', 'wrote(similarity=-0.3873569965362549)',
'hey(similarity=-0.37755218148231506)', 'myers(similarity=-0.3660753071308136)']
test_word2vec_similarity | Top 10 similar words for handgun is
['rates(similarity=0.7769370079040527)', 'firearm(similarity=0.7702646255493164)',
'income(similarity=0.7640714049339294)', 'homicide(similarity=0.7539728879928589)',
'criminal(similarity=0.7475253343582153)', 'governmental(similarity=0.7387802004814148)',
'concealed(similarity=0.7302584648132324)', 'nuclear(similarity=0.7292414307594299)',
'regulations(similarity=0.7130141854286194)', 'tax(similarity=0.7105146646499634)']

test_word2vec_similarity | Last 10 similar words for handgun is
['ra(similarity=-0.5353955030441284)', 'wrote(similarity=-0.5193497538566589)',
'hey(similarity=-0.42737656831741333)', 'uh(similarity=-0.419799268245697)',
'writes(similarity=-0.4178798794746399)', 'mr(similarity=-0.4060622453689575)',
'ya(similarity=-0.3958871364593506)', 'sorry(similarity=-0.38822048902511597)',
'john(similarity=-0.38796043395996094)', 'nick(similarity=-0.3837873935699463)']
test_word2vec_similarity | Top 10 similar words for money is
['fun(similarity=0.6567017436027527)', 'much(similarity=0.6381500959396362)',
'bike(similarity=0.6243270635604858)', 'weight(similarity=0.618684709072113)',
'talent(similarity=0.6144708395004272)', 'attention(similarity=0.6121313571929932)',
'credit(similarity=0.6057361364364624)', 'car(similarity=0.5950636267662048)',
'gloves(similarity=0.5895965695381165)', 'land(similarity=0.5868417620658875)']
test_word2vec_similarity | Last 10 similar words for money is
['ap(similarity=-0.46097975969314575)', 'originator(similarity=-0.3963715434074402)',
'federalist(similarity=-0.37642720341682434)', 'de(similarity=-0.3592833876609802)',
'remus(similarity=-0.3589582505226135)', 'kent(similarity=-0.3583595156696167)',
'uoregon(similarity=-0.3555840849876404)', 'rutgers(similarity=-0.3544737696647644)',
'csuohio(similarity=-0.3489276170730591)', 'mchp(similarity=-0.3466157615184784)']
test_word2vec_similarity | Top 10 similar words for engineer is
['mellon(similarity=0.7146978974342346)', 'corp(similarity=0.7134993076324463)',
'austin(similarity=0.7006423473358154)', 'electronics(similarity=0.6859579682350159)',
'packard(similarity=0.6827089190483093)', 'services(similarity=0.6732348203659058)',
'communications(similarity=0.666509747505188)', 'software(similarity=0.6657020449638367)',
'cellular(similarity=0.6615795493125916)', 'carnegie(similarity=0.6595965623855591)']
test_word2vec_similarity | Last 10 similar words for engineer is
['jesus(similarity=-0.47231894731521606)', 'everything(similarity=-0.4722181260585785)',
'already(similarity=-0.46030524373054504)', 'him(similarity=-0.4192686378955841)',
'koresh(similarity=-0.4094349145889282)', 'ever(similarity=-0.4090379476547241)',
'meaning(similarity=-0.40610525012016296)', 'everybody(similarity=-0.4056099057197571)',
'woman(similarity=-0.40444162487983704)', 'satan(similarity=-0.40183064341545105)']
test_word2vec_similarity | Top 10 similar words for rich is
['behanna(similarity=0.6416579484939575)', 'armstrong(similarity=0.6265410780906677)',
'syl(similarity=0.6200705170631409)', 'rain(similarity=0.6176596879959106)',
'rm(similarity=0.6124000549316406)', 'cy(similarity=0.6122149229049683)',
'milk(similarity=0.6100950241088867)', 'medraut(similarity=0.6078882217407227)',
'sequent(similarity=0.607221245765686)', 'blaine(similarity=0.6041014790534973)']
test_word2vec_similarity | Last 10 similar words for rich is
['missed(similarity=-0.4057275056838989)', 'particular(similarity=-0.3322140872478485)',
'verse(similarity=-0.33209195733070374)', 'since(similarity=-0.2928088903427124)',
'section(similarity=-0.2846943736076355)', 'original(similarity=-0.2741057276725769)',
'earlier(similarity=-0.27385446429252625)', 'which(similarity=-0.2724696397781372)',
'although(similarity=-0.26663047075271606)', 'however(similarity=-0.2616586983203888)']
test_word2vec_similarity | Top 10 similar words for republican is
['cancer(similarity=0.7324565649032593)', 'socialized(similarity=0.7232665419578552)',
'arkansas(similarity=0.7222924828529358)', 'pricing(similarity=0.7157039046287537)',
'irresponsible(similarity=0.7020460963249207)', 'critique(similarity=0.6989775896072388)',
'images(similarity=0.695737898349762)', 'damages(similarity=0.694169819355011)',
'zoroastrians(similarity=0.6937937140464783)', 'scholar(similarity=0.6931765079498291)']

test_word2vec_similarity | Last 10 similar words for republican is
['let(similarity=-0.2049948275089264)', 'until(similarity=-0.1917157769203186)',
'check(similarity=-0.18432560563087463)', 'show(similarity=-0.17653144896030426)',
'give(similarity=-0.17203079164028168)', 'turn(similarity=-0.17136111855506897)',
'leave(similarity=-0.1661013662815094)', 'spend(similarity=-0.16378051042556763)',
'back(similarity=-0.15800540149211884)', 'shoot(similarity=-0.1572718471288681)']

8. Compare sim functions

8a. Test Newsgroup similarities.

Jaccard
test_word_similarity | Top 3 similar words for soc.religion.christian is
['talk.politics.mideast(similarity=0.011324376199616123)',
'rec.sport.hockey(similarity=0.0111303012857417)',
'talk.politics.misc(similarity=0.0109847939751523)']
test_word_similarity | Last 3 similar words for soc.religion.christian is
['rec.autos(similarity=0.009097002776979795)',
'rec.motorcycles(similarity=0.00919364106493009)',
'talk.religion.misc(similarity=0.009435317783418747)']
test_word_similarity | Top 3 similar words for rec.autos is
['rec.sport.hockey(similarity=0.009435317783418747)',
'talk.politics.mideast(similarity=0.009290297864189254)',
'soc.religion.christian(similarity=0.009097002776979795)']
test_word_similarity | Last 3 similar words for rec.autos is
['rec.motorcycles(similarity=0.007746007459118294)',
'talk.religion.misc(similarity=0.008131636850664882)',
'rec.sport.baseball(similarity=0.008276323972635507)']
test_word_similarity | Top 3 similar words for talk.politics.misc is
['soc.religion.christian(similarity=0.0109847939751523)',
'talk.politics.mideast(similarity=0.010887812365101444)',
'rec.sport.hockey(similarity=0.010790849359742938)']
test_word_similarity | Last 3 similar words for talk.politics.misc is
['rec.autos(similarity=0.009000382995021065)',
'rec.motorcycles(similarity=0.009338633207221876)',
'talk.religion.misc(similarity=0.009386973180076629)']
test_word_similarity | Top 3 similar words for rec.sport.hockey is
['talk.politics.mideast(similarity=0.011227329430956723)',
'soc.religion.christian(similarity=0.0111303012857417)',
'talk.politics.guns(similarity=0.01108179419525066)']
test_word_similarity | Last 3 similar words for rec.sport.hockey is
['rec.motorcycles(similarity=0.009145319607373713)',
'rec.autos(similarity=0.009435317783418747)',
'talk.religion.misc(similarity=0.009773859716366424)']
test_word_similarity | Top 3 similar words for alt.atheism is
['talk.politics.mideast(similarity=0.010403183278201255)',
'rec.sport.hockey(similarity=0.010306313216049088)',
'soc.religion.christian(similarity=0.010209461726501463)']

test_word_similarity | Last 3 similar words for alt.atheism is
['rec.autos(similarity=0.008324562242847574)',
'rec.motorcycles(similarity=0.00866235941612826)',
'talk.religion.misc(similarity=0.00919364106493009)']
test_word_similarity | Top 3 similar words for rec.sport.baseball is
['talk.politics.mideast(similarity=0.010403183278201255)',
'soc.religion.christian(similarity=0.010161042944785276)',
'talk.politics.misc(similarity=0.010015814443858724)']
test_word_similarity | Last 3 similar words for rec.sport.baseball is
['rec.autos(similarity=0.008276323972635507)',
'talk.religion.misc(similarity=0.008565822845384504)',
'rec.motorcycles(similarity=0.008807198927819261)']
test_word_similarity | Top 3 similar words for talk.politics.mideast is
['soc.religion.christian(similarity=0.011324376199616123)',
'rec.sport.hockey(similarity=0.011227329430956723)',
'talk.politics.misc(similarity=0.010887812365101444)']
test_word_similarity | Last 3 similar words for talk.politics.mideast is
['rec.autos(similarity=0.009290297864189254)',
'talk.religion.misc(similarity=0.009677110280732011)',
'rec.motorcycles(similarity=0.00972548268097542 2)']
test_word_similarity | Top 3 similar words for rec.motorcycles is
['talk.politics.mideast(similarity=0.00972548268097542 2)',
'talk.politics.misc(similarity=0.009338633207221876)',
'soc.religion.christian(similarity=0.00919364106493009)']
test_word_similarity | Last 3 similar words for rec.motorcycles is
['rec.autos(similarity=0.007746007459118294)',
'talk.religion.misc(similarity=0.008469304751423512)',
'alt.atheism(similarity=0.00866235941612826)']
test_word_similarity | Top 3 similar words for talk.politics.guns is
['rec.sport.hockey(similarity=0.01108179419525066)',
'talk.politics.mideast(similarity=0.010548523206751054)',
'soc.religion.christian(similarity=0.010451625275673603)']
test_word_similarity | Last 3 similar words for talk.politics.guns is
['rec.autos(similarity=0.00866235941612826)',
'talk.religion.misc(similarity=0.009048690573083737)',
'rec.motorcycles(similarity=0.009048690573083737)']
test_word_similarity | Top 3 similar words for talk.religion.misc is
['rec.sport.hockey(similarity=0.009773859716366424)',
'talk.politics.mideast(similarity=0.009677110280732011)',
'soc.religion.christian(similarity=0.009435317783418747)']
test_word_similarity | Last 3 similar words for talk.religion.misc is
['rec.autos(similarity=0.008131636850664882)',
'rec.motorcycles(similarity=0.008469304751423512)',
'rec.sport.baseball(similarity=0.008565822845384504)']

Cosine
test_word_similarity | Top 3 similar words for soc.religion.christian is
['talk.religion.misc(similarity=0.9919183114052993)',

'alt.atheism(similarity=0.9891265888210407)',
'talk.politics.misc(similarity=0.9862178149194812)']
test_word_similarity | Last 3 similar words for soc.religion.christian is
['rec.sport.hockey(similarity=0.8915391730047594)',
'rec.motorcycles(similarity=0.9151165514765032)', 'rec.autos(similarity=0.9343163675563361)']
test_word_similarity | Top 3 similar words for rec.autos is
['rec.motorcycles(similarity=0.9917449260765105)',
'rec.sport.baseball(similarity=0.9811252846768931)',
'talk.politics.guns(similarity=0.9777878622971252)']
test_word_similarity | Last 3 similar words for rec.autos is
['soc.religion.christian(similarity=0.9343163675563361)',
'alt.atheism(similarity=0.9440258808670355)',
'talk.politics.mideast(similarity=0.951251048894889)']
test_word_similarity | Top 3 similar words for talk.politics.misc is
['talk.religion.misc(similarity=0.9892166807719729)',
'talk.politics.guns(similarity=0.9865660369163336)',
'soc.religion.christian(similarity=0.9862178149194812)']
test_word_similarity | Last 3 similar words for talk.politics.misc is
['rec.sport.hockey(similarity=0.9103587788043602)',
'rec.motorcycles(similarity=0.9357207850074614)', 'rec.autos(similarity=0.9524401741067724)']
test_word_similarity | Top 3 similar words for rec.sport.hockey is
['rec.sport.baseball(similarity=0.9740641525100004)',
'rec.autos(similarity=0.952038820849965)', 'rec.motorcycles(similarity=0.9498687411455566)']
test_word_similarity | Last 3 similar words for rec.sport.hockey is
['alt.atheism(similarity=0.8911479472362545)',
'soc.religion.christian(similarity=0.8915391730047594)',
'talk.politics.misc(similarity=0.9103587788043602)']
test_word_similarity | Top 3 similar words for alt.atheism is
['talk.religion.misc(similarity=0.9902247786108441)',
'soc.religion.christian(similarity=0.9891265888210407)',
'talk.politics.misc(similarity=0.9839637429053031)']
test_word_similarity | Last 3 similar words for alt.atheism is
['rec.sport.hockey(similarity=0.8911479472362545)',
'rec.motorcycles(similarity=0.9261325125888774)', 'rec.autos(similarity=0.9440258808670355)']
test_word_similarity | Top 3 similar words for rec.sport.baseball is
['rec.autos(similarity=0.9811252846768931)',
'talk.politics.guns(similarity=0.9759752875554392)',
'rec.motorcycles(similarity=0.9750420488961419)']
test_word_similarity | Last 3 similar words for rec.sport.baseball is
['soc.religion.christian(similarity=0.9414533520073202)',
'alt.atheism(similarity=0.9465506374413303)',
'talk.politics.mideast(similarity=0.9547631765203182)']
test_word_similarity | Top 3 similar words for talk.politics.mideast is
['talk.religion.misc(similarity=0.9839241688803108)',
'talk.politics.misc(similarity=0.9839129050563167)',
'talk.politics.guns(similarity=0.9835764301830736)']
test_word_similarity | Last 3 similar words for talk.politics.mideast is
['rec.sport.hockey(similarity=0.9190893129740321)',
'rec.motorcycles(similarity=0.9366922026910317)', 'rec.autos(similarity=0.951251048894889)']

test_word_similarity | Top 3 similar words for rec.motorcycles is
['rec.autos(similarity=0.9917449260765105)',
'rec.sport.baseball(similarity=0.9750420488961419)',
'talk.politics.guns(similarity=0.9672540187483335)']
test_word_similarity | Last 3 similar words for rec.motorcycles is
['soc.religion.christian(similarity=0.9151165514765032)',
'alt.atheism(similarity=0.9261325125888774)',
'talk.politics.misc(similarity=0.9357207850074614)']
test_word_similarity | Top 3 similar words for talk.politics.guns is
['talk.religion.misc(similarity=0.9869326016798315)',
'talk.politics.misc(similarity=0.9865660369163336)',
'talk.politics.mideast(similarity=0.9835764301830736)']
test_word_similarity | Last 3 similar words for talk.politics.guns is
['rec.sport.hockey(similarity=0.9391679426479208)',
'rec.motorcycles(similarity=0.9672540187483335)',
'soc.religion.christian(similarity=0.9719267074671791)']
test_word_similarity | Top 3 similar words for talk.religion.misc is
['soc.religion.christian(similarity=0.9919183114052993)',
'alt.atheism(similarity=0.9902247786108441)',
'talk.politics.misc(similarity=0.9892166807719729)']
test_word_similarity | Last 3 similar words for talk.religion.misc is
['rec.sport.hockey(similarity=0.9179238170558549)',
'rec.motorcycles(similarity=0.9460793902775345)', 'rec.autos(similarity=0.960075252861954)']

Dice
test_word_similarity | Top 3 similar words for soc.religion.christian is
['talk.politics.mideast(similarity=0.02239514139305371)',
'rec.sport.hockey(similarity=0.022015562725374836)',
'talk.politics.misc(similarity=0.021730878724615675)']
test_word_similarity | Last 3 similar words for soc.religion.christian is
['rec.autos(similarity=0.01802998671474663)',
'rec.motorcycles(similarity=0.01821977604858607)',
'talk.religion.misc(similarity=0.018694249383184664)']
test_word_similarity | Top 3 similar words for rec.autos is
['rec.sport.hockey(similarity=0.018694249383184664)',
'talk.politics.mideast(similarity=0.018409565382425507)',
'soc.religion.christian(similarity=0.01802998671474663)']
test_word_similarity | Last 3 similar words for rec.autos is
['rec.motorcycles(similarity=0.015372936040994496)',
'talk.religion.misc(similarity=0.016132093376352248)',
'rec.sport.baseball(similarity=0.016416777377111405)']
test_word_similarity | Top 3 similar words for talk.politics.misc is
['soc.religion.christian(similarity=0.021730878724615675)',
'talk.politics.mideast(similarity=0.021541089390776237)',
'rec.sport.hockey(similarity=0.0213513000569368)']
test_word_similarity | Last 3 similar words for talk.politics.misc is
['rec.autos(similarity=0.017840197380907193)',
'rec.motorcycles(similarity=0.01850446004934526)',
'talk.religion.misc(similarity=0.018599354716264945)']

test_word_similarity | Top 3 similar words for rec.sport.hockey is
['talk.politics.mideast(similarity=0.022205352059214273)',
'soc.religion.christian(similarity=0.022015562725374836)',
'talk.politics.guns(similarity=0.021920668058455117)']
test_word_similarity | Last 3 similar words for rec.sport.hockey is
['rec.motorcycles(similarity=0.01812488138166635)',
'rec.autos(similarity=0.018694249383184664)',
'talk.religion.misc(similarity=0.019358512051622 7)']
test_word_similarity | Top 3 similar words for alt.atheism is
['talk.politics.mideast(similarity=0.020592142721579047)',
'rec.sport.hockey(similarity=0.02040235338773961)',
'soc.religion.christian(similarity=0.02021256405390017)']
test_word_similarity | Last 3 similar words for alt.atheism is
['rec.autos(similarity=0.016511672044031124)',
'rec.motorcycles(similarity=0.01717593471246916)',
'talk.religion.misc(similarity=0.01821977604858607)']
test_word_similarity | Top 3 similar words for rec.sport.baseball is
['talk.politics.mideast(similarity=0.020592142721579047)',
'soc.religion.christian(similarity=0.020117669386980452)',
'talk.politics.misc(similarity=0.019832985386221295)']
test_word_similarity | Last 3 similar words for rec.sport.baseball is
['rec.autos(similarity=0.016416777377111405)',
'talk.religion.misc(similarity=0.016986145378629722)',
'rec.motorcycles(similarity=0.017460618713228317)']
test_word_similarity | Top 3 similar words for talk.politics.mideast is
['soc.religion.christian(similarity=0.02239514139305371)',
'rec.sport.hockey(similarity=0.022205352059214273)',
'talk.politics.misc(similarity=0.021541089390776237)']
test_word_similarity | Last 3 similar words for talk.politics.mideast is
['rec.autos(similarity=0.018409565382425507)',
'talk.religion.misc(similarity=0.01916872271778326)',
'rec.motorcycles(similarity=0.019263617384702978)']
test_word_similarity | Top 3 similar words for rec.motorcycles is
['talk.politics.mideast(similarity=0.019263617384702978)',
'talk.politics.misc(similarity=0.018504460049345226)',
'soc.religion.christian(similarity=0.01821977604858607)']
test_word_similarity | Last 3 similar words for rec.motorcycles is
['rec.autos(similarity=0.015372936040994496)',
'talk.religion.misc(similarity=0.016796356044790284)',
'alt.atheism(similarity=0.01717593471246916)']
test_word_similarity | Top 3 similar words for talk.politics.guns is
['rec.sport.hockey(similarity=0.021920668058455117)',
'talk.politics.mideast(similarity=0.020876826722338204)',
'soc.religion.christian(similarity=0.020687037388498766)']
test_word_similarity | Last 3 similar words for talk.politics.guns is
['rec.autos(similarity=0.01717593471246916)',
'talk.religion.misc(similarity=0.017935092047826912)',
'rec.motorcycles(similarity=0.017935092047826912)']

test_word_similarity | Top 3 similar words for talk.religion.misc is
['rec.sport.hockey(similarity=0.0193585120516227)',
'talk.politics.mideast(similarity=0.01916872271778326)',
'soc.religion.christian(similarity=0.018694249383184664)']
test_word_similarity | Last 3 similar words for talk.religion.misc is
['rec.autos(similarity=0.016132093376352248)',
'rec.motorcycles(similarity=0.016796356044790284)',
'rec.sport.baseball(similarity=0.016986145378629722)']

8a. Test Word similarities.

Jaccard
test_word_similarity | Top 10 similar words for lethal is
['nancy(similarity=0.3333333333333333)', 'cdt(similarity=0.3333333333333333)',
'facilities(similarity=0.3333333333333333)', '1968(similarity=0.3333333333333333)',
'journal(similarity=0.3333333333333333)', 'bearing(similarity=0.3333333333333333)',
'gray(similarity=0.3333333333333333)', 'strict(similarity=0.3333333333333333)',
'ethical(similarity=0.3333333333333333)', 'square(similarity=0.3333333333333333)']
test_word_similarity | Last 10 similar words for lethal is ['UNK(similarity=0.0)',
'dream(similarity=0.0)', 'fun(similarity=0.0)', 'favor(similarity=0.0)', 'fairly(similarity=0.0)',
'59(similarity=0.0)', 'green(similarity=0.0)', 'regularly(similarity=0.0)', 'advice(similarity=0.0)',
'request(similarity=0.0)']
test_word_similarity | Top 10 similar words for handgun is
['lethal(similarity=0.1111111111111111)', 'handguns(similarity=0.1111111111111111)',
'cse(similarity=0.1111111111111111)', 'craig(similarity=0.1111111111111111)',
'holland(similarity=0.1111111111111111)', 'bullets(similarity=0.1111111111111111)',
'rubber(similarity=0.1111111111111111)', 'plastic(similarity=0.1111111111111111)',
'stopping(similarity=0.1111111111111111)', 'democrat(similarity=0.1111111111111111)']
test_word_similarity | Last 10 similar words for handgun is ['UNK(similarity=0.0)',
'accurately(similarity=0.0)', 'cheers(similarity=0.0)', 'burden(similarity=0.0)', 'utah(similarity=0.0)',
'bnr(similarity=0.0)', 'fixed(similarity=0.0)', 'appearance(similarity=0.0)', 'likes(similarity=0.0)',
'angels(similarity=0.0)']
test_word_similarity | Top 10 similar words for money is ['face(similarity=0.3333333333333333)',
'couple(similarity=0.25)', 'fall(similarity=0.17647058823529413)',
'our(similarity=0.17647058823529413)', 'colorado(similarity=0.17647058823529413)',
'harvard(similarity=0.17647058823529413)', 'means(similarity=0.17647058823529413)',
'proof(similarity=0.17647058823529413)', 'sometimes(similarity=0.17647058823529413)',
'thinking(similarity=0.17647058823529413)']
test_word_similarity | Last 10 similar words for money is ['UNK(similarity=0.0)',
'locutus(similarity=0.0)', 'bolsheviks(similarity=0.0)', 'argentine(similarity=0.0)',
'argentina(similarity=0.0)', 'reisman(similarity=0.0)', 'cartel(similarity=0.0)',
'malpractice(similarity=0.0)', 'wiretapping(similarity=0.0)', 'borden(similarity=0.0)']
test_word_similarity | Top 10 similar words for engineer is
['image(similarity=0.5384615384615384)', 'rely(similarity=0.42857142857142855)',
'assumptions(similarity=0.42857142857142855)', 'implies(similarity=0.42857142857142855)',
'prejudice(similarity=0.42857142857142855)', 'guns(similarity=0.3333333333333333)',
'avenue(similarity=0.3333333333333333)', 'pretend(similarity=0.3333333333333333)',
'lawyer(similarity=0.3333333333333333)', 'sudden(similarity=0.3333333333333333)']

test_word_similarity | Last 10 similar words for engineer is ['UNK(similarity=0.0)', 'tongue(similarity=0.0)', 'bringing(similarity=0.0)', 'stick(similarity=0.0)', 'signal(similarity=0.0)', 'topic(similarity=0.0)', 'rochester(similarity=0.0)', 'kills(similarity=0.0)', 'promise(similarity=0.0)', 'exception(similarity=0.0)']
test_word_similarity | Top 10 similar words for rich is ['minor(similarity=0.42857142857142855)', 'currently(similarity=0.3333333333333333)', 'keeping(similarity=0.3333333333333333)', 'speaking(similarity=0.3333333333333333)', 'count(similarity=0.3333333333333333)', 'cold(similarity=0.3333333333333333)', 'bother(similarity=0.3333333333333333)', 'exact(similarity=0.3333333333333333)', 'eat(similarity=0.3333333333333333)', 'pa(similarity=0.3333333333333333)']
test_word_similarity | Last 10 similar words for rich is ['UNK(similarity=0.0)', 'locutus(similarity=0.0)', 'reisman(similarity=0.0)', 'cartel(similarity=0.0)', 'malpractice(similarity=0.0)', 'wiretapping(similarity=0.0)', 'borden(similarity=0.0)', 'pyotr(similarity=0.0)', 'stephanopoulos(similarity=0.0)', 'orchid(similarity=0.0)']
test_word_similarity | Top 10 similar words for republican is ['lo(similarity=0.42857142857142855)', 'handling(similarity=0.42857142857142855)', 'exciting(similarity=0.42857142857142855)', 'participants(similarity=0.42857142857142855)', 'steel(similarity=0.42857142857142855)', 'exceptions(similarity=0.42857142857142855)', 'assist(similarity=0.42857142857142855)', 'racial(similarity=0.42857142857142855)', '137(similarity=0.42857142857142855)', 'med(similarity=0.42857142857142855)']
test_word_similarity | Last 10 similar words for republican is ['UNK(similarity=0.0)', 'fun(similarity=0.0)', 'fairly(similarity=0.0)', '59(similarity=0.0)', 'green(similarity=0.0)', 'request(similarity=0.0)', 'surprise(similarity=0.0)', 'closed(similarity=0.0)', 'basically(similarity=0.0)', 'seemed(similarity=0.0)']

Cosine
test_word_similarity | Top 10 similar words for lethal is ['waco(similarity=0.9965612343379819)', 'kerosene(similarity=0.9959075764325377)', 'idbsu(similarity=0.9956824505818589)', 'uxh(similarity=0.9947897254651769)', 'batf(similarity=0.9939985333599834)', 'compound(similarity=0.9911789699188875)', 'encryption(similarity=0.9910860560585223)', 'tennessee(similarity=0.9908298963032667)', 'federalist(similarity=0.9903123096385865)', 'bills(similarity=0.9901482076345278)']
test_word_similarity | Last 10 similar words for lethal is ['nmm(similarity=0.0)', 'tharp(similarity=0.0)', 'cds(similarity=0.0)', 'davet(similarity=0.0)', 'virago(similarity=0.0)', 'sturges(similarity=0.0)', '3com(similarity=0.0)', 'motorcyclists(similarity=0.0)', 'denizens(similarity=0.0)', 'sears(similarity=0.0)']
test_word_similarity | Top 10 similar words for handgun is ['firearm(similarity=0.9999624232692909)', 'cipriani(similarity=0.9999226257946291)', 'dividian(similarity=0.9999197653078815)', 'crphilli(similarity=0.999854872649987)', 'hound(similarity=0.999854872649987)', 'dazixca(similarity=0.999854872649987)', 'jrm(similarity=0.999854872649987)', 'u28037(similarity=0.999854872649987)', 'rkba(similarity=0.999854872649987)', 'bms(similarity=0.999854872649987)']
test_word_similarity | Last 10 similar words for handgun is ['nmm(similarity=0.0)', 'tharp(similarity=0.0)', 'cds(similarity=0.0)', 'davet(similarity=0.0)', 'virago(similarity=0.0)', 'sturges(similarity=0.0)', '3com(similarity=0.0)', 'motorcyclists(similarity=0.0)', 'denizens(similarity=0.0)', 'sears(similarity=0.0)']
test_word_similarity | Top 10 similar words for money is ['class(similarity=0.9767438403760214)', 'increase(similarity=0.9634182042669379)', 'em(similarity=0.9597610728746588)', 'care(similarity=0.9589233925927593)',

'pay(similarity=0.9572138758030033)', 'board(similarity=0.9550166338088882)',
'forward(similarity=0.9549707090282492)', 'virtually(similarity=0.9513969848573453)',
'mine(similarity=0.9512385045020354)', 'considering(similarity=0.950378913903427)']
test_word_similarity | Last 10 similar words for money is
['schismatic(similarity=0.11585657190134323)', 'chalcedon(similarity=0.11585657190134323)',
'sarto(similarity=0.11585657190134323)', 'jhpb(similarity=0.11585657190134323)',
'scroggs(similarity=0.11585657190134323)', 'whitsell(similarity=0.11585657190134323)',
'sirach(similarity=0.11585657190134323)', 'caralv(similarity=0.11585657190134323)',
'crossroads(similarity=0.11585657190134324)', 'nasb(similarity=0.11585657190134324)']
test_word_similarity | Top 10 similar words for engineer is
['makers(similarity=0.9665607275877093)', 'unit(similarity=0.9640027414815829)',
'interior(similarity=0.9618744428455661)', 'needing(similarity=0.9602778606944725)',
'windows(similarity=0.956533701375456)', 'equipped(similarity=0.9547802207405517)',
'electronics(similarity=0.9540012488175563)', 'worn(similarity=0.9535492477839298)',
'assembly(similarity=0.9489765796783615)', 'spelling(similarity=0.9445004164979767)']
test_word_similarity | Last 10 similar words for engineer is ['nmm(similarity=0.0)',
'tharp(similarity=0.0)', 'cds(similarity=0.0)', 'davet(similarity=0.0)', 'virago(similarity=0.0)',
'sturges(similarity=0.0)', '3com(similarity=0.0)', 'motorcyclists(similarity=0.0)',
'denizens(similarity=0.0)', 'sears(similarity=0.0)']
test_word_similarity | Top 10 similar words for rich is ['vastly(similarity=0.9751798419308304)',
'facilities(similarity=0.9700429851443605)', 'primary(similarity=0.9521586978939848)',
'clue(similarity=0.9496313894101055)', 'variety(similarity=0.9448880416076804)',
'advocate(similarity=0.9397600578188596)', 'routine(similarity=0.9390552131003703)',
'indicate(similarity=0.938431508021737)', 'stands(similarity=0.938350782598531)',
'closely(similarity=0.937277502003693)']
test_word_similarity | Last 10 similar words for rich is
['shamim(similarity=0.05770041704858152)', 'nye(similarity=0.05770041704858152)',
'nyeda(similarity=0.05770041704858152)', 'buphy(similarity=0.05770041704858152)',
'jaeger(similarity=0.05770041704858152)', 'karner(similarity=0.057700417048581525)',
'fanatism(similarity=0.057700417048581525)', 'beauchaine(similarity=0.057700417048581525)',
'bobbe(similarity=0.057700417048581525)', 'ingles(similarity=0.057700417048581525)']
test_word_similarity | Top 10 similar words for republican is
['feasible(similarity=0.9952608536713679)', 'resign(similarity=0.9946943735478533)',
'reagan(similarity=0.9942214424884378)', 'libertarian(similarity=0.9939863401803372)',
'govt(similarity=0.9936207312752264)', 'senior(similarity=0.9900270171791653)',
'drugs(similarity=0.9897188177865796)', 'adjective(similarity=0.989096333033071)',
'desy(similarity=0.9879753580966648)', 'funds(similarity=0.987613041514261)']
test_word_similarity | Last 10 similar words for republican is ['balltown(similarity=0.0)',
'cabot(similarity=0.0)', 'welty(similarity=0.0)', 'heiser(similarity=0.0)', 'wb3ffv(similarity=0.0)',
'goucher(similarity=0.0)', 'delco(similarity=0.0)', 'qazi(similarity=0.0)', 'delcoelect(similarity=0.0)',
'kocrsv01(similarity=0.0)']

Dice
test_word_similarity | Top 10 similar words for lethal is ['nancy(similarity=0.5)',
'cdt(similarity=0.5)', 'facilities(similarity=0.5)', '1968(similarity=0.5)', 'journal(similarity=0.5)',
'bearing(similarity=0.5)', 'gray(similarity=0.5)', 'strict(similarity=0.5)', 'ethical(similarity=0.5)',
'square(similarity=0.5)']
test_word_similarity | Last 10 similar words for lethal is ['UNK(similarity=0.0)',
'dream(similarity=0.0)', 'fun(similarity=0.0)', 'favor(similarity=0.0)', 'fairly(similarity=0.0)',

'59(similarity=0.0)', 'green(similarity=0.0)', 'regularly(similarity=0.0)', 'advice(similarity=0.0)', 'request(similarity=0.0)']
test_word_similarity | Top 10 similar words for handgun is ['lethal(similarity=0.2)', 'handguns(similarity=0.2)', 'cse(similarity=0.2)', 'craig(similarity=0.2)', 'holland(similarity=0.2)', 'bullets(similarity=0.2)', 'rubber(similarity=0.2)', 'plastic(similarity=0.2)', 'stopping(similarity=0.2)', 'democrat(similarity=0.2)']
test_word_similarity | Last 10 similar words for handgun is ['UNK(similarity=0.0)', 'accurately(similarity=0.0)', 'cheers(similarity=0.0)', 'burden(similarity=0.0)', 'utah(similarity=0.0)', 'bnr(similarity=0.0)', 'fixed(similarity=0.0)', 'appearance(similarity=0.0)', 'likes(similarity=0.0)', 'angels(similarity=0.0)']
test_word_similarity | Top 10 similar words for money is ['face(similarity=0.5)', 'couple(similarity=0.4)', 'fall(similarity=0.3)', 'our(similarity=0.3)', 'colorado(similarity=0.3)', 'harvard(similarity=0.3)', 'means(similarity=0.3)', 'proof(similarity=0.3)', 'sometimes(similarity=0.3)', 'thinking(similarity=0.3)']
test_word_similarity | Last 10 similar words for money is ['UNK(similarity=0.0)', 'locutus(similarity=0.0)', 'bolsheviks(similarity=0.0)', 'argentine(similarity=0.0)', 'argentina(similarity=0.0)', 'reisman(similarity=0.0)', 'cartel(similarity=0.0)', 'malpractice(similarity=0.0)', 'wiretapping(similarity=0.0)', 'borden(similarity=0.0)']
test_word_similarity | Top 10 similar words for engineer is ['image(similarity=0.7)', 'rely(similarity=0.6)', 'assumptions(similarity=0.6)', 'implies(similarity=0.6)', 'prejudice(similarity=0.6)', 'guns(similarity=0.5)', 'avenue(similarity=0.5)', 'pretend(similarity=0.5)', 'lawyer(similarity=0.5)', 'sudden(similarity=0.5)']
test_word_similarity | Last 10 similar words for engineer is ['UNK(similarity=0.0)', 'tongue(similarity=0.0)', 'bringing(similarity=0.0)', 'stick(similarity=0.0)', 'signal(similarity=0.0)', 'topic(similarity=0.0)', 'rochester(similarity=0.0)', 'kills(similarity=0.0)', 'promise(similarity=0.0)', 'exception(similarity=0.0)']
test_word_similarity | Top 10 similar words for rich is ['minor(similarity=0.6)', 'currently(similarity=0.5)', 'keeping(similarity=0.5)', 'speaking(similarity=0.5)', 'count(similarity=0.5)', 'cold(similarity=0.5)', 'bother(similarity=0.5)', 'exact(similarity=0.5)', 'eat(similarity=0.5)', 'pa(similarity=0.5)']
test_word_similarity | Last 10 similar words for rich is ['UNK(similarity=0.0)', 'locutus(similarity=0.0)', 'reisman(similarity=0.0)', 'cartel(similarity=0.0)', 'malpractice(similarity=0.0)', 'wiretapping(similarity=0.0)', 'borden(similarity=0.0)', 'pyotr(similarity=0.0)', 'stephanopoulos(similarity=0.0)', 'orchid(similarity=0.0)']
test_word_similarity | Top 10 similar words for republican is ['lo(similarity=0.6)', 'handling(similarity=0.6)', 'exciting(similarity=0.6)', 'participants(similarity=0.6)', 'steel(similarity=0.6)', 'exceptions(similarity=0.6)', 'assist(similarity=0.6)', 'racial(similarity=0.6)', '137(similarity=0.6)', 'med(similarity=0.6)']
test_word_similarity | Last 10 similar words for republican is ['UNK(similarity=0.0)', 'fun(similarity=0.0)', 'fairly(similarity=0.0)', '59(similarity=0.0)', 'green(similarity=0.0)', 'request(similarity=0.0)', 'surprise(similarity=0.0)', 'closed(similarity=0.0)', 'basically(similarity=0.0)', 'seemed(similarity=0.0)']

8c. Test Context similarities.

Jaccard
test_word_similarity | Top 10 similar words for lethal is ['re(similarity=0.0008547820305822015)', 'non(similarity=0.0008547820305822015)', 'lethal(similarity=0.0008547820305822015)', 'to(similarity=0.0008547820305822015)', 'from(similarity=0.0008547820305822015)',

'edu(similarity=0.0008547820305822015)', 'think(similarity=0.0008547820305822015)',
'nntp(similarity=0.0008547820305822015)', 'posting(similarity=0.0008547820305822015)',
'host(similarity=0.0008547820305822015)']
test_word_similarity | Last 10 similar words for lethal is
['8302(similarity=4.744958481613286e-05)', 'fxwg(similarity=4.744958481613286e-05)',
'odwyer(similarity=4.744958481613286e-05)', 'hatching(similarity=4.744958481613286e-05)',
'ksand(similarity=4.744958481613286e-05)', 'cunixc(similarity=4.744958481613286e-05)',
'alink(similarity=4.744958481613286e-05)', 'chatham(similarity=4.744958481613286e-05)',
'englishman(similarity=4.744958481613286e-05)',
'rethought(similarity=4.744958481613286e-05)']
test_word_similarity | Top 10 similar words for handgun is
['to(similarity=0.0012351543942992875)', 'from(similarity=0.0012351543942992875)',
'edu(similarity=0.0012351543942992875)', 'nntp(similarity=0.0012351543942992875)',
'1(similarity=0.0012351543942992875)', 'cs(similarity=0.0012351543942992875)',
'what(similarity=0.0012351543942992875)', 'about(similarity=0.0012351543942992875)',
'with(similarity=0.0012351543942992875)', 'like(similarity=0.0012351543942992875)']
test_word_similarity | Last 10 similar words for handgun is
['8302(similarity=4.744958481613286e-05)', 'fxwg(similarity=4.744958481613286e-05)',
'odwyer(similarity=4.744958481613286e-05)', 'hatching(similarity=4.744958481613286e-05)',
'ksand(similarity=4.744958481613286e-05)', 'cunixc(similarity=4.744958481613286e-05)',
'chatham(similarity=4.744958481613286e-05)',
'englishman(similarity=4.744958481613286e-05)',
'rethought(similarity=4.744958481613286e-05)',
'rethinking(similarity=4.744958481613286e-05)']
test_word_similarity | Top 10 similar words for money is
['money(similarity=0.00275953944238271 95)', 'a(similarity=0.0027118321518626006)',
'the(similarity=0.0027118321518626006)', 'is(similarity=0.0027118321518626006)',
'UNK(similarity=0.0027118321518626006)', 'to(similarity=0.002664129400570885)',
'for(similarity=0.002664129400570885)', 'in(similarity=0.0026164311878597592)',
'and(similarity=0.0026164311878597592)', 'that(similarity=0.0026164311878597592)']
test_word_similarity | Last 10 similar words for money is
['ksand(similarity=4.744958481613286e-05)', '8302(similarity=9.490367277213628e-05)',
'fxwg(similarity=9.490367277213628e-05)', 'odwyer(similarity=9.490367277213628e-05)',
'hatching(similarity=9.490367277213628e-05)', 'cunixc(similarity=9.490367277213628e-05)',
'chatham(similarity=9.490367277213628e-05)',
'englishman(similarity=9.490367277213628e-05)',
'rethought(similarity=9.490367277213628e-05)',
'rethinking(similarity=9.490367277213628e-05)']
test_word_similarity | Top 10 similar words for engineer is
['to(similarity=0.0009498480243161094)', 'from(similarity=0.0009498480243161094)',
'edu(similarity=0.0009498480243161094)', 'nntp(similarity=0.0009498480243161094)',
'1(similarity=0.0009498480243161094)', 'lines(similarity=0.0009498480243161094)',
'cs(similarity=0.0009498480243161094)', 'what(similarity=0.0009498480243161094)',
'with(similarity=0.0009498480243161094)', 'or(similarity=0.0009498480243161094)']
test_word_similarity | Last 10 similar words for engineer is
['8302(similarity=4.744958481613286e-05)', 'fxwg(similarity=4.744958481613286e-05)',
'02238(similarity=4.744958481613286e-05)', '382761(similarity=4.744958481613286e-05)',
'jaskew(similarity=4.744958481613286e-05)', 'ksand(similarity=4.744958481613286e-05)',
'cunixc(similarity=4.744958481613286e-05)', '0002(similarity=9.490367277213628e-05)',

'maidenhead(similarity=9.490367277213628e-05)',
'urbanachampaign(similarity=9.490367277213628e-05)']
test_word_similarity | Top 10 similar words for rich is ['edu(similarity=0.0014254490164401787)',
'in(similarity=0.0014254490164401787)', 'of(similarity=0.0014254490164401787)',
'it(similarity=0.0014254490164401787)', 'the(similarity=0.0014254490164401787)',
'this(similarity=0.0014254490164401787)', 'is(similarity=0.0014254490164401787)',
'but(similarity=0.0014254490164401787)', 'are(similarity=0.0014254490164401787)',
'rich(similarity=0.0014254490164401787)']
test_word_similarity | Last 10 similar words for rich is
['8302(similarity=4.744958481613286e-05)', 'fxwg(similarity=4.744958481613286e-05)',
'odwyer(similarity=4.744958481613286e-05)', 'hatching(similarity=4.744958481613286e-05)',
'ksand(similarity=4.744958481613286e-05)', 'cunixc(similarity=4.744958481613286e-05)',
'chatham(similarity=4.744958481613286e-05)',
'englishman(similarity=4.744958481613286e-05)',
'rethought(similarity=4.744958481613286e-05)',
'rethinking(similarity=4.744958481613286e-05)']
test_word_similarity | Top 10 similar words for republican is
['re(similarity=0.0006647042066280505)', 'to(similarity=0.0006647042066280505)',
'from(similarity=0.0006647042066280505)', 'edu(similarity=0.0006647042066280505)',
'think(similarity=0.0006647042066280505)', 'nntp(similarity=0.0006647042066280505)',
'posting(similarity=0.0006647042066280505)', 'host(similarity=0.0006647042066280505)',
'version(similarity=0.0006647042066280505)', '1(similarity=0.0006647042066280505)']
test_word_similarity | Last 10 similar words for republican is
['8302(similarity=4.744958481613286e-05)', 'fxwg(similarity=4.744958481613286e-05)',
'odwyer(similarity=4.744958481613286e-05)', 'hatching(similarity=4.744958481613286e-05)',
'ksand(similarity=4.744958481613286e-05)', 'cunixc(similarity=4.744958481613286e-05)',
'alink(similarity=4.744958481613286e-05)', 'chatham(similarity=4.744958481613286e-05)',
'englishman(similarity=4.744958481613286e-05)',
'rethought(similarity=4.744958481613286e-05)']

Cosine
test_word_similarity | Top 10 similar words for lethal is
['practice(similarity=0.7760815184415686)', 'similar(similarity=0.7678458112828889)',
'non(similarity=0.766693911463068)', 'certain(similarity=0.7634279655156974)',
'common(similarity=0.7605236126470136)', 'religion(similarity=0.7601873271891662)',
'particular(similarity=0.759560407488776)', 'different(similarity=0.7593902209027847)',
'legal(similarity=0.7553267399385606)', 'religious(similarity=0.7540666559524279)']
test_word_similarity | Last 10 similar words for lethal is ['cunixc(similarity=0.0)',
'8563446(similarity=0.0012744094606823616)', 'z1dan(similarity=0.0026756601486263125)',
'syl(similarity=0.004095281052907502)', 'terresterial(similarity=0.0042163429564793554)',
'fxwg(similarity=0.0043017218937217774)', 'uu4(similarity=0.00527396295344497)',
'rscharfy(similarity=0.005817451418636954)', '92717(similarity=0.007428324807994075)',
'budd(similarity=0.011260599584471797)']
test_word_similarity | Top 10 similar words for handgun is
['without(similarity=0.9176317357786625)', 'gun(similarity=0.9117507309911148)',
'small(similarity=0.9116211591552732)', 'great(similarity=0.9106385956337493)',
'man(similarity=0.9097660974492888)', 'complete(similarity=0.908899254710954)',
'short(similarity=0.9077055210898648)', 'having(similarity=0.9070794064107418)',
'run(similarity=0.9062774316984719)', 'religious(similarity=0.9059989660666309)']

test_word_similarity | Last 10 similar words for handgun is ['cunixc(similarity=0.0)', 'z1dan(similarity=0.003175407279514281)', '8563446(similarity=0.004761377706023825)', 'terresterial(similarity=0.0049406715821844055)', '92717(similarity=0.00740087833621886)', 'syl(similarity=0.008772321735855359)', 'rscharfy(similarity=0.010007081777016655)', 'uu4(similarity=0.010031273325943224)', 'fxwg(similarity=0.014464668553373073)', 'budd(similarity=0.02039539845586068)']
test_word_similarity | Top 10 similar words for money is ['up(similarity=0.9603185130010726)', 'us(similarity=0.9601583637846574)', 'use(similarity=0.9601296809924658)', 'then(similarity=0.9568131101947593)', 'all(similarity=0.9558875026151603)', 'also(similarity=0.9537642808755673)', 'one(similarity=0.95348772750585)', 'them(similarity=0.9529053349522391)', 'out(similarity=0.9523257128525326)', 'right(similarity=0.9501889814811495)']
test_word_similarity | Last 10 similar words for money is ['8563446(similarity=0.0027701844398604876)', 'terresterial(similarity=0.0032812141159155165)', 'z1dan(similarity=0.006842451429008159)', 'cunixc(similarity=0.009558886192622719)', '92717(similarity=0.011460755218058769)', 'uu4(similarity=0.017104698266207242)', 'syl(similarity=0.019039695819677408)', '1777(similarity=0.021003289106963724)', 'budd(similarity=0.02361043078390783)', 'rscharfy(similarity=0.023845915910570125)']
test_word_similarity | Top 10 similar words for engineer is ['m(similarity=0.7614772816221814)', 'sabre(similarity=0.7521588786978674)', 'd(similarity=0.7502770864330681)', 'e(similarity=0.74991351328366)', 'UNK(similarity=0.7368149595966172)', 'wrote(similarity=0.730849306916171)', 'dod(similarity=0.7301109376393984)', 'k(similarity=0.7300238384715432)', '750(similarity=0.7299068208026112)', 'from(similarity=0.7288771814134118)']
test_word_similarity | Last 10 similar words for engineer is ['cunixc(similarity=0.0)', 'budd(similarity=0.004770469597970687)', 'wb3ffv(similarity=0.006796220309799095)', 'fxwg(similarity=0.007798145139346011)', 'chatham(similarity=0.011697217709019015)', '1708(similarity=0.012457711206722162)', '1778(similarity=0.013212502205942825)', 'z1dan(similarity=0.01355266248311483)', 'terresterial(similarity=0.01476560215176502)', '8302(similarity=0.015596290278692021)']
test_word_similarity | Top 10 similar words for rich is ['UNK(similarity=0.9509490411842516)', 'in(similarity=0.9448963115754381)', 'air(similarity=0.942888583550633)', 'from(similarity=0.9406550218335661)', 'etc(similarity=0.9397053035083023)', 'hot(similarity=0.9377332099692838)', 'by(similarity=0.9367258289250366)', 'black(similarity=0.9364368661160101)', 'path(similarity=0.934604138361302)', 'and(similarity=0.9313585171295478)']
test_word_similarity | Last 10 similar words for rich is ['cunixc(similarity=0.004480648311401667)', '8563446(similarity=0.011903371049267574)', 'terresterial(similarity=0.012394523206605607)', 'z1dan(similarity=0.017400327058557118)', 'unforgiven(similarity=0.019156390749255264)', '92717(similarity=0.019279149810062336)', 'wb3ffv(similarity=0.025813099564786907)', 'fxwg(similarity=0.027256192848312957)', '1777(similarity=0.028455255311720398)', 'munny(similarity=0.028867838224488126)']
test_word_similarity | Top 10 similar words for republican is ['man(similarity=0.9150809337027437)', 'with(similarity=0.9150108650273447)', 'dog(similarity=0.9133518585955777)', 'while(similarity=0.9131535790201056)', 'for(similarity=0.9129540476896036)', 'big(similarity=0.9125071446834768)', 'without(similarity=0.9112853152961086)', 'great(similarity=0.909495600774423)', 'real(similarity=0.9073507440694113)', 'making(similarity=0.905573880246982)']

test_word_similarity | Last 10 similar words for republican is ['cunixc(similarity=0.0)', 'terresterial(similarity=0.0)', '8563446(similarity=0.0017302024229862193)', 'syl(similarity=0.0018870166142313663)', 'z1dan(similarity=0.002797304634740165)', 'fxwg(similarity=0.00424744301257212 9)', 'uu4(similarity=0.0059175186251546694)', 'rscharfy(similarity=0.0061030500957245494)', '92717(similarity=0.006890073743199138)', 'budd(similarity=0.009909979853607996)']

Dice
test_word_similarity | Top 10 similar words for lethal is ['re(similarity=0.001708104004554944)', 'non(similarity=0.001708104004554944)', 'lethal(similarity=0.001708104004554944)', 'to(similarity=0.001708104004554944)', 'from(similarity=0.001708104004554944)', 'edu(similarity=0.001708104004554944)', 'think(similarity=0.001708104004554944)', 'nntp(similarity=0.001708104004554944)', 'posting(similarity=0.001708104004554944)', 'host(similarity=0.001708104004554944)']
test_word_similarity | Last 10 similar words for lethal is ['8302(similarity=9.48946669197191e-05)', 'fxwg(similarity=9.48946669197191e-05)', 'odwyer(similarity=9.48946669197191e-05)', 'hatching(similarity=9.48946669197191e-05)', 'ksand(similarity=9.48946669197191e-05)', 'cunixc(similarity=9.48946669197191e-05)', 'alink(similarity=9.48946669197191e-05)', 'chatham(similarity=9.48946669197191e-05)', 'englishman(similarity=9.48946669197191e-05)', 'rethought(similarity=9.48946669197191e-05)']
test_word_similarity | Top 10 similar words for handgun is ['to(similarity=0.002467261339912697)', 'from(similarity=0.002467261339912697)', 'edu(similarity=0.002467261339912697)', 'nntp(similarity=0.002467261339912697)', '1(similarity=0.002467261339912697)', 'cs(similarity=0.002467261339912697)', 'what(similarity=0.002467261339912697)', 'about(similarity=0.002467261339912697)', 'with(similarity=0.002467261339912697)', 'like(similarity=0.002467261339912697)']
test_word_similarity | Last 10 similar words for handgun is ['8302(similarity=9.48946669197191e-05)', 'fxwg(similarity=9.48946669197191e-05)', 'odwyer(similarity=9.48946669197191e-05)', 'hatching(similarity=9.48946669197191e-05)', 'ksand(similarity=9.48946669197191e-05)', 'cunixc(similarity=9.48946669197191e-05)', 'chatham(similarity=9.48946669197191e-05)', 'englishman(similarity=9.48946669197191e-05)', 'rethought(similarity=9.48946669197191e-05)', 'rethinking(similarity=9.48946669197191e-05)']
test_word_similarity | Top 10 similar words for money is ['money(similarity=0.005503890681343709)', 'a(similarity=0.005408996014423989)', 'the(similarity=0.005408996014423989)', 'is(similarity=0.005408996014423989)', 'UNK(similarity=0.005408996014423989)', 'to(similarity=0.00531410134750427)', 'for(similarity=0.00531410134750427)', 'in(similarity=0.005219206680584551)', 'and(similarity=0.005219206680584551)', 'that(similarity=0.005219206680584551)']
test_word_similarity | Last 10 similar words for money is ['ksand(similarity=9.48946669197191e-05)', '8302(similarity=0.0001897893338394382)', 'fxwg(similarity=0.0001897893338394382)', 'odwyer(similarity=0.0001897893338394382)', 'hatching(similarity=0.0001897893338394382)', 'cunixc(similarity=0.0001897893338394382)', 'chatham(similarity=0.0001897893338394382)', 'englishman(similarity=0.0001897893338394382)', 'rethought(similarity=0.0001897893338394382)', 'rethinking(similarity=0.0001897893338394382)']
test_word_similarity | Top 10 similar words for engineer is ['to(similarity=0.0018978933383943823)', 'from(similarity=0.0018978933383943823)', 'edu(similarity=0.0018978933383943823)', 'nntp(similarity=0.0018978933383943823)',

'1(similarity=0.0018978933383943823)', 'lines(similarity=0.0018978933383943823)',
'cs(similarity=0.0018978933383943823)', 'what(similarity=0.0018978933383943823)',
'with(similarity=0.0018978933383943823)', 'or(similarity=0.0018978933383943823)']
test_word_similarity | Last 10 similar words for engineer is
['8302(similarity=9.48946669197191e-05)', 'fxwg(similarity=9.48946669197191e-05)',
'02238(similarity=9.48946669197191e-05)', '382761(similarity=9.48946669197191e-05)',
'jaskew(similarity=9.48946669197191e-05)', 'ksand(similarity=9.48946669197191e-05)',
'cunixc(similarity=9.48946669197191e-05)', '0002(similarity=0.0001897893338394382)',
'maidenhead(similarity=0.0001897893338394382)',
'urbanachampaign(similarity=0.0001897893338394382)']
test_word_similarity | Top 10 similar words for rich is ['edu(similarity=0.0028468400075915734)',
'in(similarity=0.0028468400075915734)', 'of(similarity=0.0028468400075915734)',
'it(similarity=0.0028468400075915734)', 'the(similarity=0.0028468400075915734)',
'this(similarity=0.0028468400075915734)', 'is(similarity=0.0028468400075915734)',
'but(similarity=0.0028468400075915734)', 'are(similarity=0.0028468400075915734)',
'rich(similarity=0.0028468400075915734)']
test_word_similarity | Last 10 similar words for rich is ['8302(similarity=9.48946669197191e-05)',
'fxwg(similarity=9.48946669197191e-05)', 'odwyer(similarity=9.48946669197191e-05)',
'hatching(similarity=9.48946669197191e-05)', 'ksand(similarity=9.48946669197191e-05)',
'cunixc(similarity=9.48946669197191e-05)', 'chatham(similarity=9.48946669197191e-05)',
'englishman(similarity=9.48946669197191e-05)', 'rethought(similarity=9.48946669197191e-05)',
'rethinking(similarity=9.48946669197191e-05)']
test_word_similarity | Top 10 similar words for republican is
['re(similarity=0.0013285253368760675)', 'to(similarity=0.0013285253368760675)',
'from(similarity=0.0013285253368760675)', 'edu(similarity=0.0013285253368760675)',
'think(similarity=0.0013285253368760675)', 'nntp(similarity=0.0013285253368760675)',
'posting(similarity=0.0013285253368760675)', 'host(similarity=0.0013285253368760675)',
'version(similarity=0.0013285253368760675)', '1(similarity=0.0013285253368760675)']
test_word_similarity | Last 10 similar words for republican is
['8302(similarity=9.48946669197191e-05)', 'fxwg(similarity=9.48946669197191e-05)',
'odwyer(similarity=9.48946669197191e-05)', 'hatching(similarity=9.48946669197191e-05)',
'ksand(similarity=9.48946669197191e-05)', 'cunixc(similarity=9.48946669197191e-05)',
'alink(similarity=9.48946669197191e-05)', 'chatham(similarity=9.48946669197191e-05)',
'englishman(similarity=9.48946669197191e-05)', 'rethought(similarity=9.48946669197191e-05)']

9. Interesting Findings

9a. non-tf-idf vs tf-idf (newsgroup)

non-tf-idf
test_word_similarity | Top 3 similar words for soc.religion.christian is
['talk.religion.misc(similarity=0.9919183114052993)',
'alt.atheism(similarity=0.9891265888210407)',
'talk.politics.misc(similarity=0.9862178149194812)']
test_word_similarity | Last 3 similar words for soc.religion.christian is
['rec.sport.hockey(similarity=0.8915391730047594)',
'rec.motorcycles(similarity=0.9151165514765032)', 'rec.autos(similarity=0.9343163675563361)']
test_word_similarity | Top 3 similar words for rec.autos is
['rec.motorcycles(similarity=0.9917449260765105)',

'rec.sport.baseball(similarity=0.9811252846768931)',
'talk.politics.guns(similarity=0.9777878622971252)']
test_word_similarity | Last 3 similar words for rec.autos is
['soc.religion.christian(similarity=0.9343163675563361)',
'alt.atheism(similarity=0.9440258808670355)',
'talk.politics.mideast(similarity=0.951251048894889)']
test_word_similarity | Top 3 similar words for talk.politics.misc is
['talk.religion.misc(similarity=0.9892166807719729)',
'talk.politics.guns(similarity=0.9865660369163336)',
'soc.religion.christian(similarity=0.9862178149194812)']
test_word_similarity | Last 3 similar words for talk.politics.misc is
['rec.sport.hockey(similarity=0.9103587788043602)',
'rec.motorcycles(similarity=0.9357207850074614)', 'rec.autos(similarity=0.9524401741067724)']
test_word_similarity | Top 3 similar words for rec.sport.hockey is
['rec.sport.baseball(similarity=0.9740641525100004)',
'rec.autos(similarity=0.952038820849965)', 'rec.motorcycles(similarity=0.9498687411455566)']
test_word_similarity | Last 3 similar words for rec.sport.hockey is
['alt.atheism(similarity=0.8911479472362545)',
'soc.religion.christian(similarity=0.8915391730047594)',
'talk.politics.misc(similarity=0.9103587788043602)']
test_word_similarity | Top 3 similar words for alt.atheism is
['talk.religion.misc(similarity=0.9902247786108441)',
'soc.religion.christian(similarity=0.9891265888210407)',
'talk.politics.misc(similarity=0.9839637429053031)']
test_word_similarity | Last 3 similar words for alt.atheism is
['rec.sport.hockey(similarity=0.8911479472362545)',
'rec.motorcycles(similarity=0.9261325125888774)', 'rec.autos(similarity=0.9440258808670355)']
test_word_similarity | Top 3 similar words for rec.sport.baseball is
['rec.autos(similarity=0.9811252846768931)',
'talk.politics.guns(similarity=0.9759752875554392)',
'rec.motorcycles(similarity=0.9750420488961419)']
test_word_similarity | Last 3 similar words for rec.sport.baseball is
['soc.religion.christian(similarity=0.9414533520073202)',
'alt.atheism(similarity=0.9465506374413303)',
'talk.politics.mideast(similarity=0.9547631765203182)']
test_word_similarity | Top 3 similar words for talk.politics.mideast is
['talk.religion.misc(similarity=0.9839241688803108)',
'talk.politics.misc(similarity=0.9839129050563167)',
'talk.politics.guns(similarity=0.9835764301830736)']
test_word_similarity | Last 3 similar words for talk.politics.mideast is
['rec.sport.hockey(similarity=0.9190893129740321)',
'rec.motorcycles(similarity=0.9366922026910317)', 'rec.autos(similarity=0.951251048894889)']
test_word_similarity | Top 3 similar words for rec.motorcycles is
['rec.autos(similarity=0.9917449260765105)',
'rec.sport.baseball(similarity=0.9750420488961419)',
'talk.politics.guns(similarity=0.9672540187483335)']
test_word_similarity | Last 3 similar words for rec.motorcycles is
['soc.religion.christian(similarity=0.9151165514765032)',

'alt.atheism(similarity=0.9261325125888774)',
'talk.politics.misc(similarity=0.9357207850074614)']
test_word_similarity | Top 3 similar words for talk.politics.guns is
['talk.religion.misc(similarity=0.9869326016798315)',
'talk.politics.misc(similarity=0.9865660369163336)',
'talk.politics.mideast(similarity=0.9835764301830736)']
test_word_similarity | Last 3 similar words for talk.politics.guns is
['rec.sport.hockey(similarity=0.9391679426479208)',
'rec.motorcycles(similarity=0.9672540187483335)',
'soc.religion.christian(similarity=0.9719267074671791)']
test_word_similarity | Top 3 similar words for talk.religion.misc is
['soc.religion.christian(similarity=0.9919183114052993)',
'alt.atheism(similarity=0.9902247786108441)',
'talk.politics.misc(similarity=0.9892166807719729)']
test_word_similarity | Last 3 similar words for talk.religion.misc is
['rec.sport.hockey(similarity=0.9179238170558549)',
'rec.motorcycles(similarity=0.9460793902775345)', 'rec.autos(similarity=0.960075252861954)']

tf-idf
test_word_similarity | Top 3 similar words for soc.religion.christian is
['talk.religion.misc(similarity=0.21606046026534503)',
'alt.atheism(similarity=0.13920518748945393)',
'talk.politics.misc(similarity=0.02946228964563017 3)']
test_word_similarity | Last 3 similar words for soc.religion.christian is
['rec.motorcycles(similarity=0.007013487601447647)',
'rec.sport.hockey(similarity=0.008558749332142488)',
'rec.sport.baseball(similarity=0.011024056607403642)']
test_word_similarity | Top 3 similar words for rec.autos is
['rec.motorcycles(similarity=0.1302567294035502)',
'talk.politics.guns(similarity=0.027689276179750748)',
'rec.sport.baseball(similarity=0.022159636063377845)']
test_word_similarity | Last 3 similar words for rec.autos is
['talk.politics.mideast(similarity=0.005470790340600049)',
'rec.sport.hockey(similarity=0.012443619794183701)',
'talk.religion.misc(similarity=0.012538008742865196)']
test_word_similarity | Top 3 similar words for talk.politics.misc is
['talk.politics.guns(similarity=0.07494059012684189)',
'soc.religion.christian(similarity=0.02946228964563017 3)',
'talk.religion.misc(similarity=0.025656448074236136)']
test_word_similarity | Last 3 similar words for talk.politics.misc is
['rec.motorcycles(similarity=0.008330532152695472)',
'rec.sport.baseball(similarity=0.01073609584237602)',
'rec.autos(similarity=0.012697061161404496)']
test_word_similarity | Top 3 similar words for rec.sport.hockey is
['rec.sport.baseball(similarity=0.05906752656305329)',
'talk.politics.misc(similarity=0.01581871661518779)',
'talk.politics.guns(similarity=0.012676884063855384)']
test_word_similarity | Last 3 similar words for rec.sport.hockey is
['talk.politics.mideast(similarity=0.004647362654504037)',

'talk.religion.misc(similarity=0.006703763118731284)',
'alt.atheism(similarity=0.008108232967097196)']
test_word_similarity | Top 3 similar words for alt.atheism is
['talk.religion.misc(similarity=0.23457992516524093)',
'soc.religion.christian(similarity=0.13920518748945393)',
'talk.politics.mideast(similarity=0.02935877073176335)']
test_word_similarity | Last 3 similar words for alt.atheism is
['rec.motorcycles(similarity=0.006060863661748232)',
'rec.sport.hockey(similarity=0.008108232967097196)',
'rec.sport.baseball(similarity=0.008490784004000553)']
test_word_similarity | Top 3 similar words for rec.sport.baseball is
['rec.sport.hockey(similarity=0.05906752656305329)',
'rec.autos(similarity=0.022159636063377845)',
'talk.politics.guns(similarity=0.011665303817526718)']
test_word_similarity | Last 3 similar words for rec.sport.baseball is
['talk.politics.mideast(similarity=0.0047645569405877)',
'alt.atheism(similarity=0.008490784004000553)',
'talk.religion.misc(similarity=0.009923571128611612)']
test_word_similarity | Top 3 similar words for talk.politics.mideast is
['alt.atheism(similarity=0.02935877073176335)',
'soc.religion.christian(similarity=0.02805505083560433)',
'talk.religion.misc(similarity=0.027917074199105742)']
test_word_similarity | Last 3 similar words for talk.politics.mideast is
['rec.motorcycles(similarity=0.003365653025163428)',
'rec.sport.hockey(similarity=0.004647362654504037)',
'rec.sport.baseball(similarity=0.0047645569405877)']
test_word_similarity | Top 3 similar words for rec.motorcycles is
['rec.autos(similarity=0.1302567294035502)',
'talk.politics.guns(similarity=0.019602257839956436)',
'rec.sport.baseball(similarity=0.011663817098955788)']
test_word_similarity | Last 3 similar words for rec.motorcycles is
['talk.politics.mideast(similarity=0.003365653025163428)',
'alt.atheism(similarity=0.006060863661748232)',
'soc.religion.christian(similarity=0.0070134876014476 47)']
test_word_similarity | Top 3 similar words for talk.politics.guns is
['talk.politics.misc(similarity=0.07494059012684189)',
'talk.religion.misc(similarity=0.04690001781885539)',
'rec.autos(similarity=0.027689276179750748)']
test_word_similarity | Last 3 similar words for talk.politics.guns is
['rec.sport.baseball(similarity=0.011665303817526718)',
'rec.sport.hockey(similarity=0.012676884063855384)',
'alt.atheism(similarity=0.01500885128497656)']
test_word_similarity | Top 3 similar words for talk.religion.misc is
['alt.atheism(similarity=0.23457992516524093)',
'soc.religion.christian(similarity=0.2160604602653 4503)',
'talk.politics.guns(similarity=0.04690001781885539)']
test_word_similarity | Last 3 similar words for talk.religion.misc is
['rec.sport.hockey(similarity=0.006703763118731284)',

'rec.sport.baseball(similarity=0.009923571128611612)',
'rec.motorcycles(similarity=0.01094504288938469)']

9b. non-ppmi vs ppmi (word context)

ppmi
test_word_similarity | Top 10 similar words for lethal is
['injection(similarity=0.4559971235513492)', 'objection(similarity=0.40786885219547964)',
'safer(similarity=0.40385293347061935)', 'premises(similarity=0.4007968968844087)',
'credible(similarity=0.39627859442575386)', 'conscience(similarity=0.3955973996459534)',
'insult(similarity=0.39233248085726674)', 'nasty(similarity=0.3921326337048823)',
'unacceptable(similarity=0.3911722216521524)',
'justification(similarity=0.3911146045986 1647)']
test_word_similarity | Last 10 similar words for lethal is ['cunixc(similarity=0.0)',
'reichel(similarity=0.01145006455818481)', 'wsh(similarity=0.0119705856335962 7)',
'terresterial(similarity=0.01624947894309831)', 'ans(similarity=0.016392361406872325)',
'hfd(similarity=0.018038806028689562)', 'semak(similarity=0.018915177710577782)',
'unassisted(similarity=0.01965395009381966)', 'askeri(similarity=0.02010493194699972)',
'429(similarity=0.02097848459043039)']
test_word_similarity | Top 10 similar words for handgun is
['handguns(similarity=0.429128475978102)', 'concealed(similarity=0.424120057931429)',
'firearms(similarity=0.4089644913607116)', 'firearm(similarity=0.40735467266988645)',
'file(similarity=0.40045108175915245)', 'crime(similarity=0.3969786540169863)',
'homicides(similarity=0.3925517622632677)', 'weapons(similarity=0.384522700210 9911)',
'weapon(similarity=0.38432474943451583)', 'rates(similarity=0.380954711742 3649)']
test_word_similarity | Last 10 similar words for handgun is ['cunixc(similarity=0.0)',
'02238(similarity=0.019416086496666575)', 'fait(similarity=0.019850497663186077)',
'comme(similarity=0.0211238789438036)', 'askeri(similarity=0.029306420953084018)',
'belgeleri(similarity=0.03399481338892189)', 'umumiye(similarity=0.0385708125318 4156)',
'jyusenkyou(similarity=0.038827832272916696)', 'aucun(similarity=0.03918015540467066)',
'1778(similarity=0.03930517751869393)']
test_word_similarity | Top 10 similar words for money is ['care(similarity=0.4669921743051763)',
'want(similarity=0.4647178872041391)', 'lot(similarity=0.4646366424150167)',
'car(similarity=0.46341268251509643)', 'going(similarity=0.4624452732858013)',
'better(similarity=0.46190217769192565)', 'getting(similarity=0.4617538559485 9364)',
'someone(similarity=0.46146744693123215)', 'government(similarity=0.461140373053381 35)',
'give(similarity=0.4606811214339154)']
test_word_similarity | Last 10 similar words for money is
['fait(similarity=0.011610181962456183)', 'cunixc(similarity=0.012221132301620946)',
'comme(similarity=0.01235495867417425)', 'aucun(similarity=0.022915734470937497)',
'1778(similarity=0.02356840857979871)', '1708(similarity=0.02408998545583741)',
'ksand(similarity=0.0315468692712967)', 'belgeleri(similarity=0.03358294480289 5316)',
'nezareti(similarity=0.03496636678974724)', 'askeri(similarity=0.035151209360799913)']
test_word_similarity | Top 10 similar words for engineer is
['mechanical(similarity=0.32026116770765783)', 'senior(similarity=0.3145919733033882)',
'curiosity(similarity=0.31277192995307024)', 'chin(similarity=0.3116799022560252)',
'aa(similarity=0.31110810038707226)', 'jaeger(similarity=0.31054334035251385)',
'um(similarity=0.3092191502315358)', 'ear(similarity=0.30895631203466456)',
'hurts(similarity=0.30829972456928445)', 'competent(similarity=0.307144155897637 1)']

test_word_similarity | Last 10 similar words for engineer is ['cunixc(similarity=0.0)', '02238(similarity=0.027542167433522873)', 'b64635(similarity=0.028250215601854173)', 'jyusenkyou(similarity=0.033930513597831786)', 'fxwg(similarity=0.034168557259359515)', 'terresterial(similarity=0.03693204321856655)', 'bhm116e(similarity=0.03697456136160614)', '05pm(similarity=0.03702476422307454)', 'reichel(similarity=0.03927162889185381)', 'chatham(similarity=0.039955161379241134)']
test_word_similarity | Top 10 similar words for rich is ['stopped(similarity=0.39344616917092146)', 'job(similarity=0.3896140959239139)', 'knowing(similarity=0.3874341620229573)', 'matthew(similarity=0.3871588397066845)', 'moment(similarity=0.38704434603007776)', 'hey(similarity=0.3869136423235455)', 'thank(similarity=0.3837708513739156)', 'sell(similarity=0.3823211862030554)', 'tough(similarity=0.38230677195295854)', 'learned(similarity=0.38191416693260694)']
test_word_similarity | Last 10 similar words for rich is ['terresterial(similarity=0.01890793330385688)', '1778(similarity=0.020222138531622396)', '1708(similarity=0.020669661316473725)', 'fait(similarity=0.030897334355288495)', 'cunixc(similarity=0.03264617324228877)', 'comme(similarity=0.03287935454725431)', 'belgeleri(similarity=0.03721993692601051)', 'chatham(similarity=0.03844806569765402)', '9937(similarity=0.039425307120147156)', 'mecmuasi(similarity=0.04140171797475569)']
test_word_similarity | Top 10 similar words for republican is ['republicans(similarity=0.4300529645550248)', 'voted(similarity=0.41015449903939044)', 'wasted(similarity=0.40626241122333473)', 'stimulus(similarity=0.4058059068406606)', 'package(similarity=0.4021166024982317)', 'badly(similarity=0.4020888986462517)', 'knife(similarity=0.39908407534339063)', 'debt(similarity=0.3983422267402182)', 'aid(similarity=0.39707187137127353)', 'senate(similarity=0.396106465959893)']
test_word_similarity | Last 10 similar words for republican is ['cunixc(similarity=0.0)', 'terresterial(similarity=0.0)', 'ans(similarity=0.014865492444557916)', 'bhm116e(similarity=0.017231281763602307)', 'carderock(similarity=0.023792535388831047)', 'b64635(similarity=0.024868107269376466)', '32bis(similarity=0.025809110159192926)', 'fxwg(similarity=0.02602226297029974)', 'fait(similarity=0.026199784893901564)', 'comme(similarity=0.027880463948209224)']