1.1 (see code)

1.2

	w = 3	w = 6
#(chicken, the)	52	103
#(chicken, wings)	6	7
#(chicago, chicago)	38	122
#(coffee, the)	95	201
#(coffee, cup)	10	14
#(coffee, coffee)	4	36

1.3

- simlex999: correlation=0.05876135331349779
- MEN: correlation=0.2251396048448754

Overall, we saw a much higher correlation score when the word vectors were evaluated on the MEN dataset compared to the simlex999 dataset. However, the absolute values of correlation were in the low range for both simlex999 and MEN datasets.

2.

IDF without logarithm transformations or other scaling techniques:

- simlex999: correlation=0.1643113945921928
- MEN: correlation=0.47281906258988254

We saw a large increase of correlation scores when the word vectors were evaluated on both simlex999 and MEN datasets. TF-IDF word vectors still performed much better on the MEN dataset than the simlex999 dataset.

3.1

highest PMI (from high to low):
('tea', 8.16600126243293)
('drinking', 7.58797865873193)
('shop', 7.411693771493207)
('costa', 7.350256393786161)
('shops', 7.260751873418467)
('sugar', 6.533949521544205)
('coffee', 6.501977131805925)
('mix', 6.131195903101976)
('seattle', 5.950816325067398)
('houses', 5.868161497268183)

lowest PMI (from low to high):
('he', -2.26033826495274)
('be', -2.1509730526875237)
('had', -1.9875291676196303)
('this', -1.979549817934235)
('not', -1.9115928402014317)

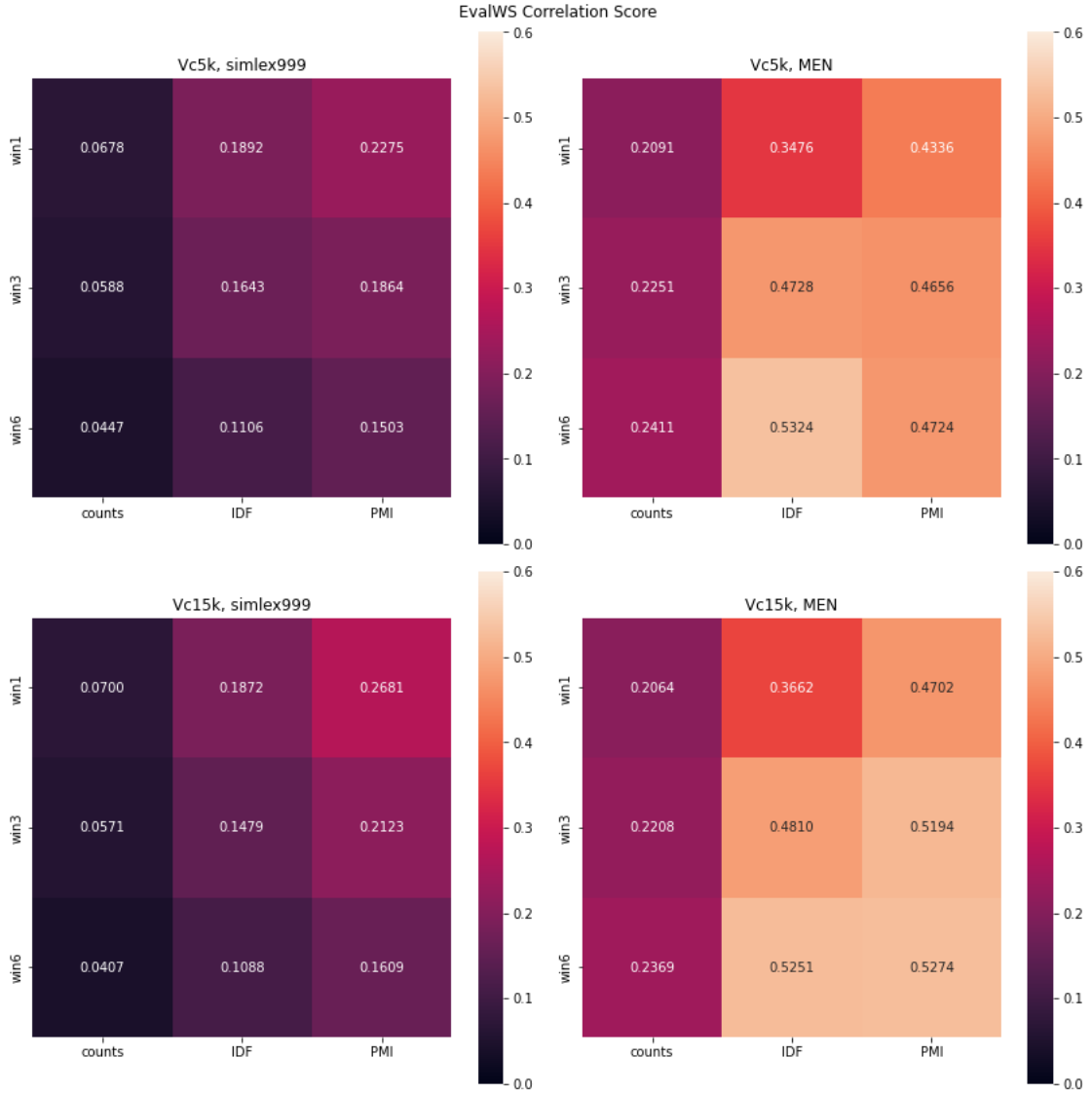
('its', -1.839457915441101)
('after', -1.598505205571959)
('more', -1.4785257922880328)
('when', -1.4043486976803334)
('page', -1.2805627423998573)

3.2

- simlex999: correlation=0.18643183126956037
- MEN: correlation=0.46563240836038006

Overall, PMI-based word vectors performed at a similar level as TF-IDF word vectors; although PMI-based word vectors achieve correlation scores slightly higher than TF-IDF word vectors on the simlex999 dataset, and slightly lower than TF-IDF word vectors on the MEN dataset. For both simlex999 and MEN datasets, PMI-based word vectors performed much better than the original distributional counting-based word vectors.

4.1



The figure above shows the EvalWS scores across difference window sizes and vector generating methods, faceted by context vocabulary choices and evaluation datasets with the same color scale across all four subplots. The highest EvalWS score was obtained using IDF method with a window size of 6 on vocab5k context vocabulary and MEN evaluation dataset. The lowest EvalWS score was obtained using raw counts method with a window size of 6 on vocab15k context vocabulary and simlex999 evaluation dataset.

Window size's effect on EvalWS score differs between evaluation datasets. For simlex999, as window size increases, EvalWS score *decreases* for all three methods (counts, IDF, PMI) and both context vocabulary sets (5k and 15k). But for MEN, as window size increases, EvalWS score *increases* for all three methods (counts, IDF, PMI) and both context vocabulary sets (5k and 15k). A potential reason for this trend difference is the difference in similarity definition between simlex999 and MEN upon examining the datasets: while simlex999 reflects the degree of semantic similarity, MEN's similarity scores are actually closer to the frequency of co-occurrence; the latter is close to how we compute word vectors with the three methods reported here. With a larger window size, our word vectors are able to capture more comprehensive and informative co-occurrence statistics; and since MEN's similarity scores are conceptually similar to co-occurrence statistics, we observed an increasing EvalWS score with increasing window size on MEN dataset. Semantically close words as captured by simlex999, however, do not tend to co-occur in the same sentence, which is probably why EvalWS score decreases on simlex999 dataset as we capture more co-occurrence statistics with a larger window size.

The effect of context vocabulary is more mixed. For raw distributional counting word vectors, large context vocabulary in general led to a tiny decrease of EvalWS score for all three window sizes and both simlex999 & MEN evaluation datasets (only exception being window size 1 on simlex999 dataset). For PMI word vectors, large context vocabulary led to a relatively large increase of EvalWS score for all three window sizes and both simlex999 & MEN evaluation datasets. The effect of context vocabulary on IDF word vectors differs by evaluation datasets: for simlex999, EvalWS score slightly decreases with larger context vocabulary; but for MEN, EvalWS score generally slightly increases with larger context vocabulary (only exception being window size 6). Among the three methods, the most prominent effect of context vocabulary size is seen on PMI; we observe a gradient of helpfulness of large context vocabulary from PMI (most helpful), to IDF (mixed effect), and finally raw distributional counts (not helpful, even very slightly harmful). In theory, increasing context vocabulary size helps capture more dimensions of a word's meaning, resulting in better EvalWS results. However, in raw distributional counting word vectors, highly frequent context words such as "a", "the", etc. dominate the word vectors. Thus, further increasing the dimensionality of raw distributional counting word vectors does not demonstrate a helping effect on capturing word meanings; it may actually appear as adding noises to the word vectors, resulting in slightly decreased EvalWS scores. IDF downweighs those highly frequent context words, thus we see slightly increased EvalWS scores with large context vocabulary for IDF word vectors in some cases; however, since IDF does not take each context word's paired center word into consideration, it still does not allow a large context vocabulary to take full effect. PMI word vector downweighs highly frequent context words while considering its relationship with the center word, therefore most informatively captures co-occurrence statistics among the three methods studied. PMI word vector allows the additional information captured by a large context vocabulary to be reflected efficiently; thus, we observe obviously improved EvalWS scores with larger context vocabulary across the board for PMI word vectors.

As discussed in section 4.1, Window size's effect on EvalWS score differs between evaluation datasets. For simlex999, as window size increases, EvalWS score *decreases* for all three methods (counts, IDF, PMI) and both context vocabulary sets (5k and 15k). But for MEN, as window size increases, EvalWS score *increases* for all three methods (counts, IDF, PMI) and both context vocabulary sets (5k and 15k).

Upon examining the two datasets, it appears that the two datasets do not encode the same type of similarity.

Simlex999 encodes semantic similarity, regardless of whether a pair of words tend to appear together. For example, "smart" and "intelligent" have similar meanings, and they get a high similarity score of 9.2/10; "large" and "huge" also have similar meanings, and they get a similarity score of 9.47/10. On the other hand, "beach" and "sea" frequently appears together in our lives, but their semantic meanings are not similar, so in simlex999 they only have a medium similarity score of 4.68/10.

MEN encodes contextual similarity, such that those words that appear together frequently are rated as more similar. For example, "sun" and "sunlight" are clearly different in meanings, but because they almost always appear together, they are given a similarity score of 50/50. The "beach" and "sea" pair are also assigned a high similarity score of 44/50 in MEN.

As discussed in section 4.1, the similarity encoded by MEN is closer to how we compute word vectors. With a larger window size, our word vectors are able to capture more comprehensive and informative contextual similarity; and since MEN's similarity scores are conceptually similar to contextual similarity, we observed an increasing EvalWS score with increasing window size on MEN dataset. Semantically close words as captured by simlex999, however, do not tend to co-occur in the same sentence, which is probably why EvalWS score decreases on simlex999 dataset as we capture more co-occurrence statistics with a larger window size.

5.1

Nearest neighbors for "judges", w = 1:
('judge', 0.16088226399322997),
('justices', 0.14678754041290754),
('arbitrators', 0.1372853856778373),
('players', 0.1324587858712465),
('trustees', 0.12963894816216848),
('contestants', 0.12422541827146137),
('officials', 0.12298001702204098),
('admins', 0.1204856574246801),
('appeals', 0.11843728431064837),
('officers', 0.11500945538099382)

Nearest neighbors for "judges", w = 6:
('judge', 0.20254689232438003),
('appeals', 0.17741896149362088),
('supreme', 0.1765936374973593),
('court', 0.1719953519361144),
('panel', 0.16925787572571646),
('courts', 0.1666058030872883),
('jury', 0.1652240379118564),
('contestants', 0.16440586358293963),
('justice', 0.1638721845764639),

('officials', 0.16358549055953772)

5.2

Nearest neighbors do not always have the same part-of-speech (POS) tags as the query word. In general, nearest neighbors generated by word vectors with a smaller window size are more likely to have the same POS tag as the query word than word vectors with a big window. For example, words with different POS tags from the query word are marked in bold in the tables below; we can see that for most of the query words I tried (e.g., freedom, transported, agreed, freed, great, angry, across, etc.), w=6 contains more nearest neighbors with different POS from the query word then w=1.

The POS of query words also impacts the POS of nearest neighbors. From the tables of examples below, we observe that noun query words are the least likely to have nearest neighbors with different POS compared to verb, adjective, or proposition query words – for both w=1 and w=6. This is especially the case for more concrete nouns like "transportation" and "classroom" – there are very few nearest neighbors with different POS for these concrete nouns for both w=1 and w=6. Abstract nouns like "freedom" also have few nearest neighbors with different POS when w=1, but start to have some nearest neighbors with different POS when w=6. Verbs and adjectives are the second least likely to have nearest neighbors with different POS compared to query words. In the examples shown below, most verbs and adjectives tested have very few nearest neighbors with different POS when w=1, but about half of the nearest neighbors have different POS when w=6. Propositions are the most likely to have nearest neighbors with different POS; from the examples below, the propositions have about half of the nearest neighbors with different POS for both w=1 and w=6.

In the examples I tried, there is no query word that have almost exactly the same nearest neighbors with the two window sizes; but query word "transportation" has quite a bit of overlapping nearest neighbor between w=1 and w=6. Overall, I find that the nearest neighbors when w=1 are more likely to capture formality similarities (e.g., same POS as discussed above, same stem such as "great" – "greatest" and "greater") between query words and neighbors, whereas the nearest neighbors when w=6 are more likely to capture contextual similarities between query words and neighbors (e.g., "freedom" usually co-occurrence with "religious"/"political"/"social"/" democracy"/"legal" movements).

Examples:

Nouns:

	w = 1	w = 6
transportation	('transport', 0.260727570275655), ('transit', 0.22027310941877104), ('rail', 0.19952301061484123), ('services', 0.18087111761167657), ('telecommunications', 0.17578274357408213), ('communications', 0.17267070384074792),	('transport', 0.28115971671885337), ('transit', 0.24402048491949901), ('services', 0.24038096148970992), ('management', 0.23616871474538897), ('rail', 0.23491574358411335), ('infrastructure', 0.2299490065930391),

	('security', 0.16057554628774476), ('management', 0.1592947101255434), ('aviation', 0.15920819039818648), ('health', 0.15651287795842775)	('facilities', 0.22664744356995903), ('systems', 0.22136871914615203), ('sector', 0.21880321102489644), ('equipment', 0.2144728437993003)
freedom	('independence', 0.16764953606669553), ('autonomy', 0.1519938883947579), ('unity', 0.1392942345850242), ('equality', 0.13106335537427613), ('fame', 0.11588302844521195), ('rights', 0.11534191122790885), ('alliance', 0.10851056251353822), ('liberty', 0.1076216194552954), ('liberties', 0.10434500596277292), ('peace', 0.10140643571806301)	('rights', 0.2287552013290907), (**'religious'**, 0.21256896071865847), (**'political'**, 0.20149221581983393), ('peace', 0.1952595673111546), (**'social'**, 0.18962858709107305), ('human', 0.18667648436081996), ('democracy', 0.18522742370266274), ('security', 0.18275246933692504), (**'legal'**, 0.1813474751180026), ('reform', 0.1802844376093805)
classroom	('schooling', 0.12179490819150403), ('journalism', 0.11155716277463625), (**'vocational'**, 0.10580622717455902), ('entrances', 0.09523912512888276), ('escalator', 0.09309693810002552), ('madness', 0.09225029864284869), ('physics', 0.09197235435956753), ('niche', 0.0896428084131407), ('education', 0.08912840663013388), ('köppen', 0.08842920935705238)	('classrooms', 0.19794106307410708), ('learning', 0.17680797623966277), ('curriculum', 0.17538834940909154), ('programs', 0.15444199601186487), ('teachers', 0.14879320180275646), ('rooms', 0.14656820180106678), ('educators', 0.14231899649249144), ('facilities', 0.14208349388499544), ('instruction', 0.14194894251926182), ('skills', 0.14168786363278388)

Verbs:

	w = 1	w = 6
transported	('marched', 0.21360483373321038),	('transport', 0.15627548919989045),

	('shipped', 0.187264472591486), ('deported', 0.18464850555312534), ('reassigned', 0.16471326782623108), ('detected', 0.15780940838624802), ('relegated', 0.1573823232125492), ('subjected', 0.15126412128931685), ('vanish', 0.14576885607822365), ('converted', 0.1435779108787988), ('dragged', 0.1411659595858849)	('**supplies**', 0.14467651312181792), ('transporting', 0.1434097429792131), ('**supply**', 0.14267513893608488), ('loading', 0.14038537363546771), ('**cargo**', 0.13890709446473615), ('carrying', 0.13621952321748967), ('**passengers**', 0.13427540811164368), ('**goods**', 0.13327231271114795), ('ships', 0.13286252763129813)
agreed	('decided', 0.1559450836605943), ('accepted', 0.15246284349029052), ('agree', 0.14893661460121138), ('returned', 0.1393087577863287), ('switched', 0.12638724455383724), ('agreeing', 0.12375469233761498), ('voted', 0.12162451944667299), ('approved', 0.11946380543433396), ('believed', 0.11719656575042452), ('condemned', 0.11438187491711355)	('decided', 0.19804003791907104), ('asked', 0.19732027971657448), ('**decision**', 0.193380830197811), ('refused', 0.19172915698202425), ('**agreement**', 0.19056207034233516), ('**contract**', 0.17587645937620155), ('**peace**', 0.1698037189265475), ('offered', 0.16941555972563463), ('wanted', 0.16854986998003035), ('allowed', 0.1677612902759754)
freed	('slain', 0.182503082931569), ('discharged', 0.1688551988352399), ('beaten', 0.1491041294390059), ('diagnosed', 0.13304642501949), ('superseded', 0.13096734316538058), ('liberated', 0.12993345988224286), ('escaped', 0.12761052752093263), ('condemned', 0.1255871995507472), ('convicted', 0.1254950010272328),	('**slaves**', 0.14067763533268154), ('imprisoned', 0.1395392856198543), ('**prisoners**', 0.12235898702211448), ('arrested', 0.1195745833669216), ('**soldiers**', 0.11681455739233375), ('escaped', 0.11662727973790814), ('attacked', 0.11642413298936793), ('killed', 0.11616173035715197), ('convicted', 0.11495728376773895),

	('repaired', 0.12513444885781433)	('**custody**', 0.11371974515669667)

Adjectives:

	w = 1	w = 6
great	('considerable', 0.2214628746900238), ('greatest', 0.1999112758713731), ('greater', 0.19678844306321594), ('significant', 0.19542548491316833), ('huge', 0.17646409550894857), ('little', 0.17601606311265375), ('major', 0.17315159260863996), ('good', 0.17056986133668614), ('large', 0.1646812207266528), ('high', 0.15244596718043418)	('**life**', 0.2533089021540761), ('**king**', 0.2341353732765489), ('**much**', 0.2287122627586876), ('**century**', 0.22406896915309402), ('important', 0.2204273573737884), ('**among**', 0.2198542957122009), ('large', 0.21944028212636377), ('**along**', 0.21749968034694744), ('**himself'**, 0.2146874666609595), ('western', 0.21261041487867344)
angry	('tired', 0.25134912033075635), ('jealous', 0.24263848349457098), ('frustrated', 0.21425333293839713), ('sick', 0.19315156949842907), ('worried', 0.186006119586205), ('confused', 0.18212798771706054), ('anxious', 0.17701862807399962), ('confident', 0.17438533727989555), ('drunk', 0.17333128562087383), ('pregnant', 0.16847741877675657)	('telling', 0.15161040072600415), ('tries', 0.14427305295328113), ('**believing**', 0.14162889207197188), ('jealous', 0.13956220356304685), ('worried', 0.139171448656322), ('furious', 0.13847447255177206), ('disappointed', 0.13493395104023678), ('upset', 0.1310173441047243), ('**tells**', 0.13090007571630388), ('**finds**', 0.12936476553530615)
exciting	('interesting', 0.14495508100384358), ('disturbing', 0.14049923265867495), ('informative', 0.13909893142638638), ('confusing', 0.1348483288610989), ('straightforward', 0.13316114680532734), ('annoying', 0.1320193075859962), ('attractive', 0.1292071421618473),	('interesting', 0.12287076411937113), ('fun', 0.10077041503700145), ('enthusiastic', 0.09771895095027622), ('useful', 0.09685521526724496), ('challenging', 0.0941294725548605), ('talented', 0.09397812942033491), ('innovative', 0.0927233193828527), ('attractive', 0.0915633617050935), ('**themes**', 0.0913789709530405 4),

	('useful', 0.12786685877365853), ('romanized', 0.11736902874939037), ('**progresses**', 0.11607836429383532)	('curious', 0.09010056200696873)

Propositions:

	w = 1	w = 6
above	('below', 0.33232768897992054), ('**here**', 0.22234009336238267), ('**same**', 0.1699280115865459), ('about', 0.16178440728759394), ('**following**', 0.1588562732053043), ('over', 0.1568150834338496), ('**talk**', 0.15422809352297664), ('**debate**', 0.15336785025067415), ('**should**', 0.15093301965320888), ('**page**', 0.1458036893500276)	('**discussion**', 0.36691671573061224), ('**page**', 0.36436143471029075), ('**should**', 0.35964114411147513), ('**talk**', 0.34746368501386615), ('below', 0.3455160530399381), ('here', 0.345139051831542), ('**link**', 0.3343384698968526), ('**do**', 0.32581726128725025), ('**debate**', 0.3184001727232551), ('**article**', 0.3164506797941051)
across	('throughout', 0.2766873325309047), ('around', 0.23536401463858897), ('through', 0.23139582983376572), ('along', 0.22177029295035205), ('within', 0.20622066144363427), ('between', 0.19476250005420045), ('into', 0.18424550161965633), ('**several**', 0.16949397087939208), ('**down**', 0.16927093232104196), ('**these**', 0.16424515986506924)	('along', 0.30412706484130203), ('around', 0.30204177385367226), ('**river**', 0.2877932456943227), ('near', 0.28181588000819785), ('throughout', 0.2793922818124401), ('**southern**', 0.2750467534900814), ('**east**', 0.2748219874171184), ('**areas**', 0.26977609968660327), ('**northern**', 0.2682784110397638), ('**large**', 0.2664842069171165)
between	('until', 0.2687721644508421), ('from', 0.266573987678875), ('around', 0.26178511131998483), ('through', 0.2528322864478568), ('**;**', 0.25217624372507924), ('**june**', 0.248154715568185), ('**october**', 0.24725771521431408), ('**november**', 0.24597797181565562), ('**july**', 0.24392863451818653), ('over', 0.24354820018939408)	('**north**', 0.28327558357534705), ('**south**', 0.2755193441542402), ('**west**', 0.2665415015635558), ('**east**', 0.2662795213372876), ('until', 0.25636909669077507), ('**line**', 0.25531959244511), ('around', 0.24586188554493751), ('through', 0.24481792741647054), ('along', 0.24418952389561116), ('near', 0.2389187972474669)

5.3

For words with multiple senses, the most commonly used sense tends to dominate the top nearest neighbors of the word when w=1. For example, when w=1, the “financial institution” sense dominates the top nearest neighbors of “bank”; the biological sense dominates the top nearest neighbors of “cell”, the fruit sense dominates the top nearest neighbors of “apple”, and the spatial sense of dominates the top nearest neighbors of “axes”.

As the window size gets larger, e.g., w=6 in our examples, the less frequent sense(s) may start to emerge in top nearest neighbors, or even dominate the top nearest neighbors. For example, the “riverside” sense of “bank” and the “tool” sense of “axes” appear in their top 10 nearest neighbors when w=6. For “apple”, the company name sense even fully dominates all top 10 nearest neighbors when w=6, completely reversing the pattern of w=1. But there are also words whose nearest neighbors’ dominate sense does not change between window sizes, such “cell”, for which the biological sense dominates both w=1 and w=6.

These differences might be due to the fact that larger window sizes can effectively encode more co-occurrence information, thus capturing more diverse meanings; but if one interpretation of a word largely dominates all of its usages (e.g., “cell” in our examples), this effect may not be obvious.

Another possible reason is that larger window sizes can capture more similarity in thematic contexts instead of merely formality (e.g., all same type of nouns since they tend to appear right next to verbs) as with small window sizes; this might be the case for “apple” – when w=1, its nearest neighbors are mostly words for other types of fruits, which tend to be used in a similar way as “apple” in a sentence; but when w=6, word vectors are no longer limited to identifying other words that, for example, appear after “eat” or before “juice”, so the “company name” meaning starts to dominate nearest neighbors.

Examples:

	w=1	w=6
bank	('banks', 0.18279205653327552), ('company', 0.14277384047080272), ('insurance', 0.13049722349957882), ('corporation', 0.1277524939152002), ('railway', 0.12268850302979893), ('government', 0.12231567199691551), ('banking', 0.1172845803417153), ('companies', 0.1120694659653104), ('institute', 0.11144760694487761), ('conference', 0.11055429400995352)	('corporation', 0.26188916237681464), ('banks', 0.24753450386935372), ('company', 0.24304049154326887), ('railway', 0.23959687310524896), ('river', 0.23952934368688034), ('capital', 0.23562226231918784), ('west', 0.234891029835324), ('central', 0.2294717459878127), ('east', 0.2284250990589614), ('northern', 0.2235443595241854)
cell	('cells', 0.27825498024044554),	('cells', 0.4206664569903039),

	('cellular', 0.1957797465237581), ('protein', 0.15501930521505625), ('tissue', 0.15453916660414266), ('brain', 0.12431467038168373), ('proteins', 0.12312275889052095), ('tissues', 0.12215081886245771), ('growth', 0.11580589956377381), ('human', 0.1108403422264825), ('enzyme', 0.11070403124912413)	('protein', 0.29795036670888936), ('membrane', 0.28173864921572395), ('proteins', 0.2790529621329563), ('cellular', 0.2689622324027659), ('dna', 0.2615435593904262), ('genes', 0.24887530746018272), ('function', 0.24692677216205808), ('tissue', 0.2448869921743174), ('brain', 0.24285347536380142)
apple	('cherry', 0.1441883773484989), ('chili', 0.13158226892869251), ('desktop', 0.11426697226625157), ('olive', 0.1047936529699098), ('tulip', 0.10406617695978132), ('orange', 0.10384858188444923), ('palm', 0.0950325057001701), ('pine', 0.09494292849310254), ('atari', 0.09327944092459393), ('wines', 0.09247491298338233)	('os', 0.22349178687178484), ('microsoft', 0.21797351180423946), ('macintosh', 0.20376424225413267), ('mac', 0.20067629605758588), ('ios', 0.19998904264792114), ('software', 0.1998374751022892), ('desktop', 0.19650194384710842), ('computers', 0.1854881332651733), ('linux', 0.18052016388955286), ('iphone', 0.1757837975076005)
axes	('phases', 0.1695062216346567), ('tributaries', 0.1352705806484195), ('qualities', 0.1215807093713247), ('paths', 0.11975062567100989), ('viewpoints', 0.11972740295337282), ('spells', 0.11164791559952274), ('sorts', 0.1086556311183971), ('branches', 0.10859294928483698), ('motifs', 0.10655919775107979), ('frames', 0.1038782505517733)	('angles', 0.12401253204025546), ('flint', 0.11839587778421004), ('neolithic', 0.11317106126229237), ('axe', 0.1083356697640823), ('symmetry', 0.10810610486154147), ('parallel', 0.10784579372981042), ('shapes', 0.10650966487929354), ('puzzle', 0.10222630940545176), ('knives', 0.10187960720045441), ('vectors', 0.09996523043837513)