* 1. (see code)

|  |  |  |
| --- | --- | --- |
|  | 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

A screenshot of a computer screen

Description automatically generated

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.

4.2

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

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.

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

5.3