Supplementary Materials to: "A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection"

Gabriella Lapesa and Stefan Evert

1 Distribution of Performance

1.1 TOEFL

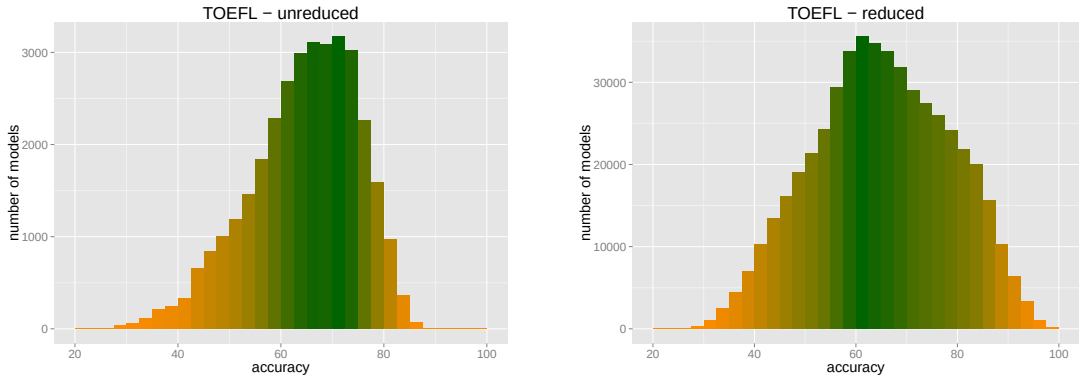

Figure 1.1: TOEFL: distribution of % accuracy

1.2 Ratings

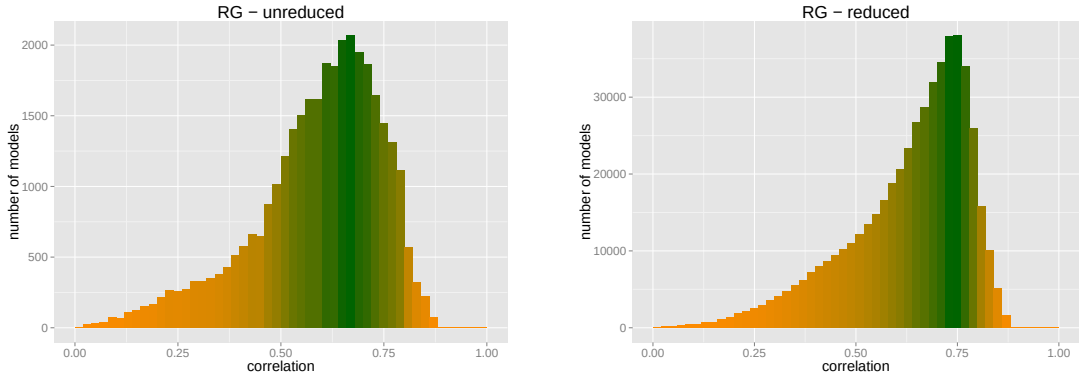

Figure 1.2: Rubenstein and Goodenough: distribution of Pearson's r

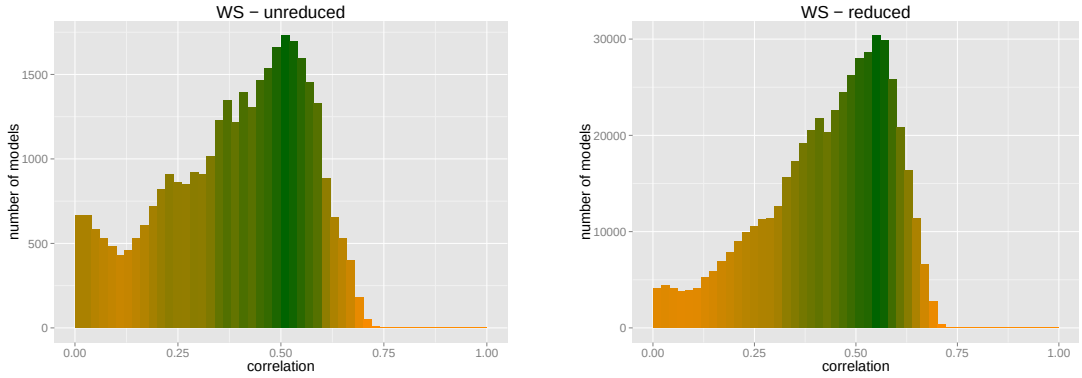

Figure 1.3: WordSim353: distribution of Pearson's r

1.3 Clustering

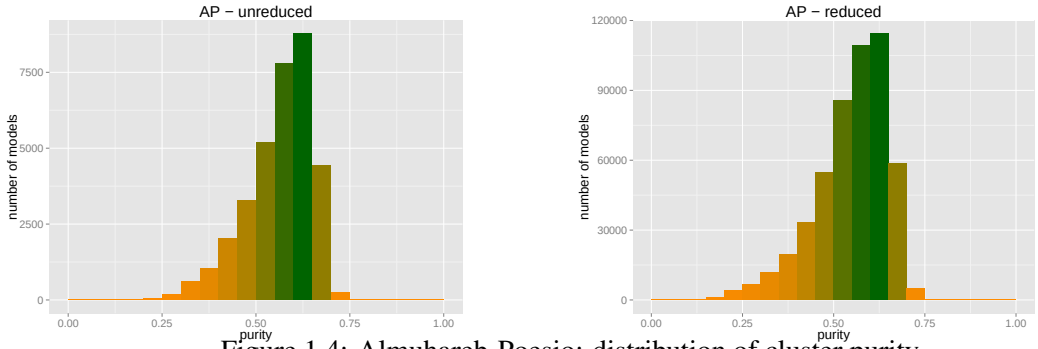

Figure 1.4: Almuhareb-Poesio: distribution of cluster purity

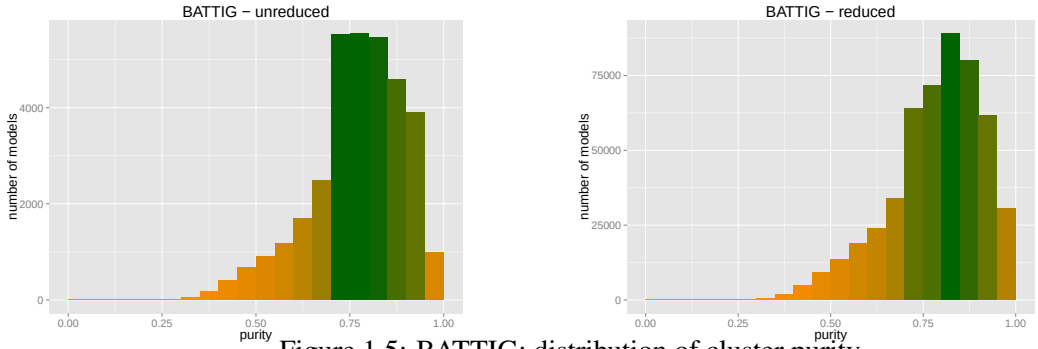

Figure 1.5: BATTIG: distribution of cluster purity

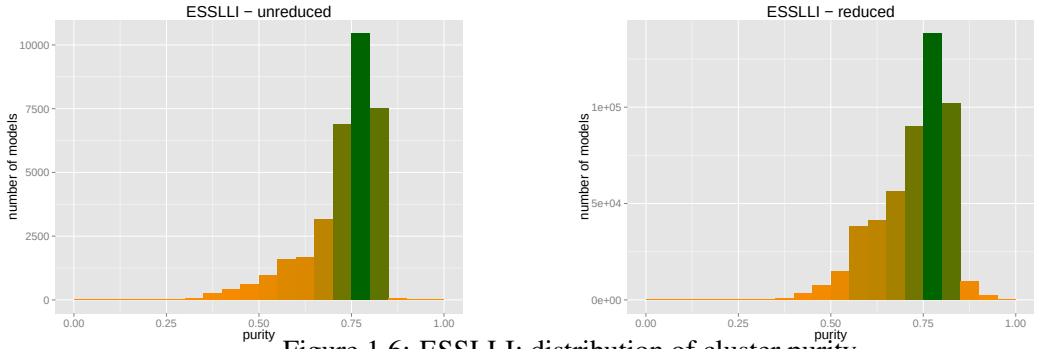

Figure 1.6: ESSLLI: distribution of cluster purity

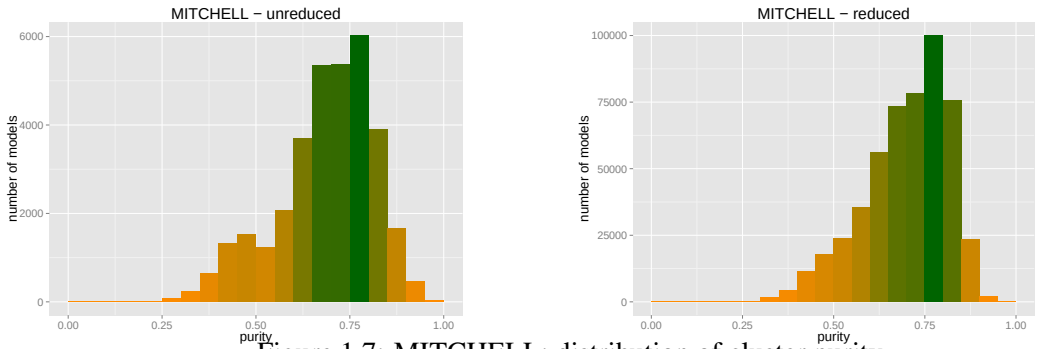

Figure 1.7: MITCHELL: distribution of cluster purity

2 Explanatory Power of DSM Parameters: Unreduced vs Reduced Runs

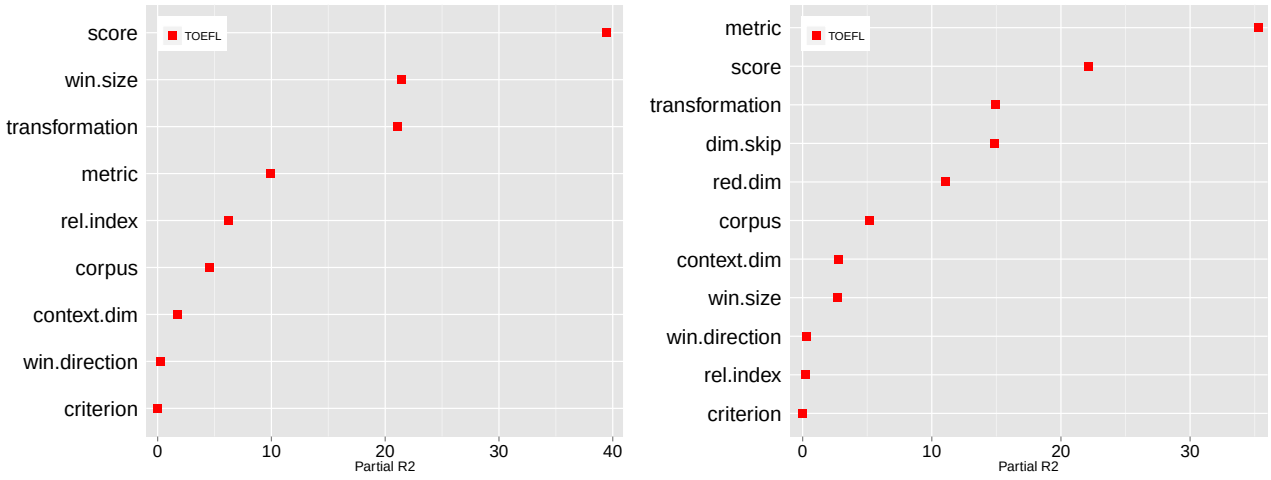

Figure 2.8: TOEFL: DSM parameters and their explanatory power. Left: unreduced; Right: reduced.

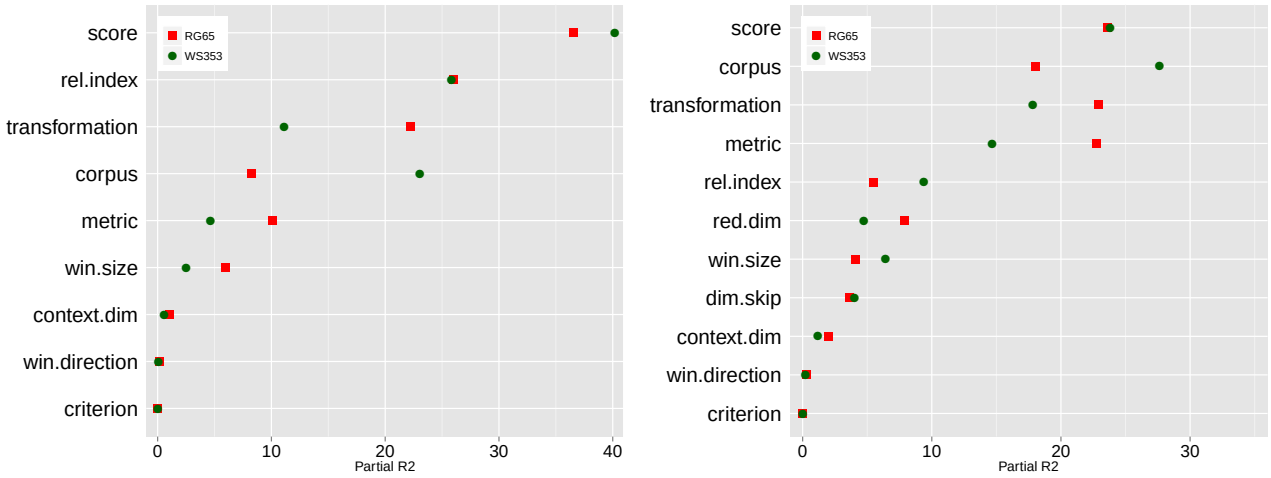

Figure 2.9: Ratings Datasets: DSM parameters and their explanatory power. Left: unreduced; Right: reduced.

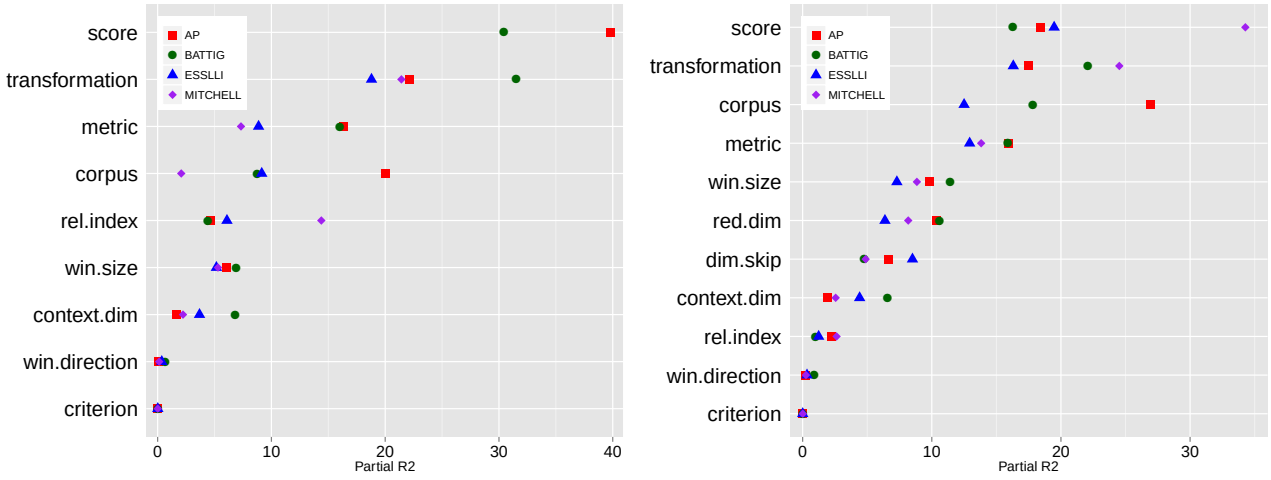

Figure 2.10: Clustering Datasets: DSM parameters and their explanatory power. Left: unreduced; Right: reduced.

3 Effect plots

3.1 TOEFL

Main Effects

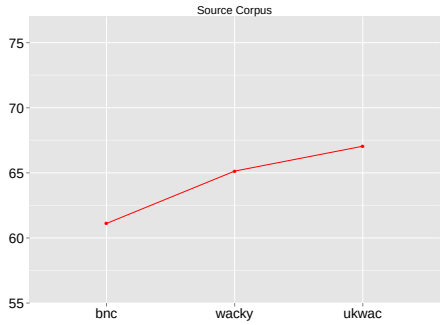

Figure 3.1.1

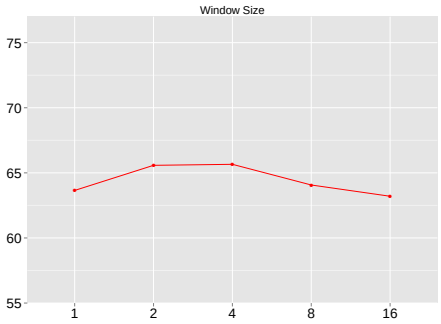

Figure 3.1.2

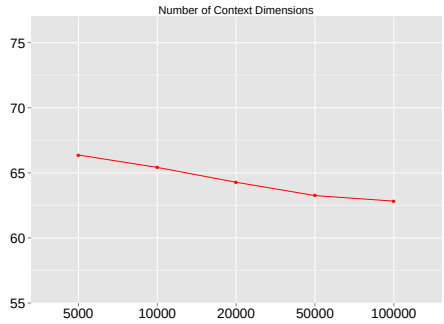

Figure 3.1.3

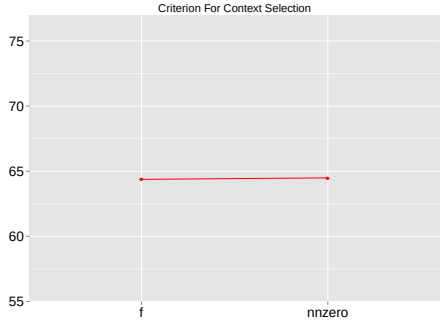

Figure 3.1.4

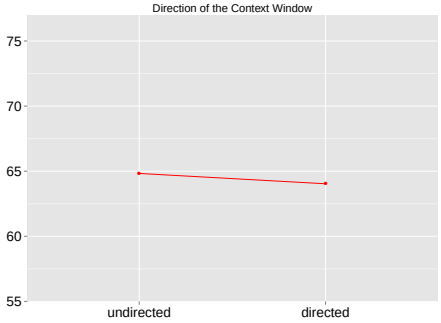

Figure 3.1.5

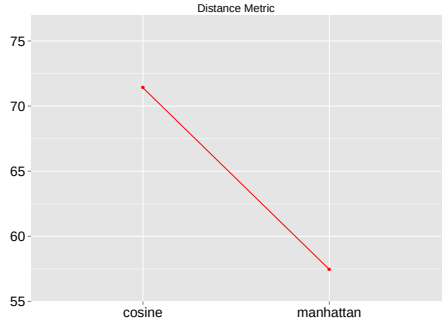

Figure 3.1.6

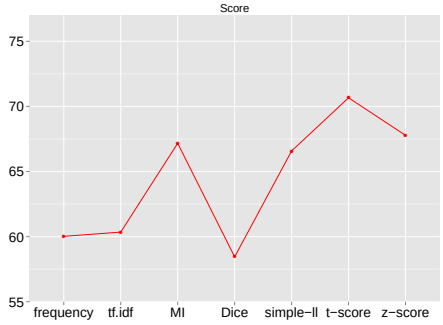

Figure 3.1.7

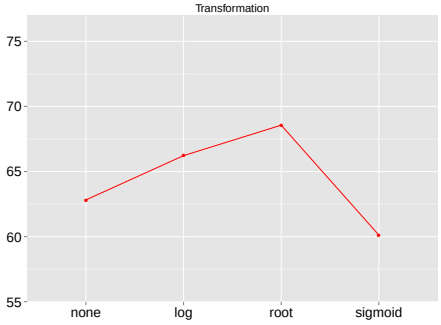

Figure 3.1.8

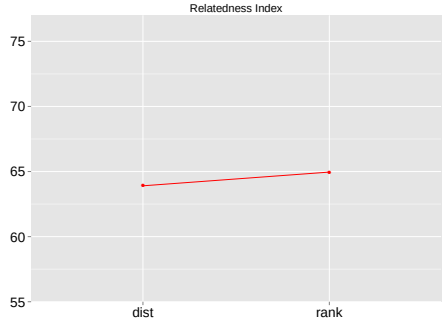

Figure 3.1.9

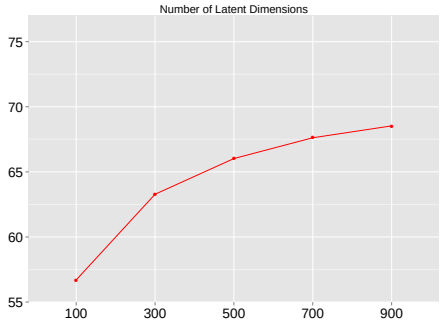

Figure 3.1.10

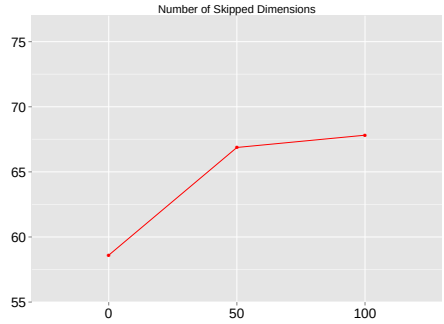

Figure 3.1.11

Interactions

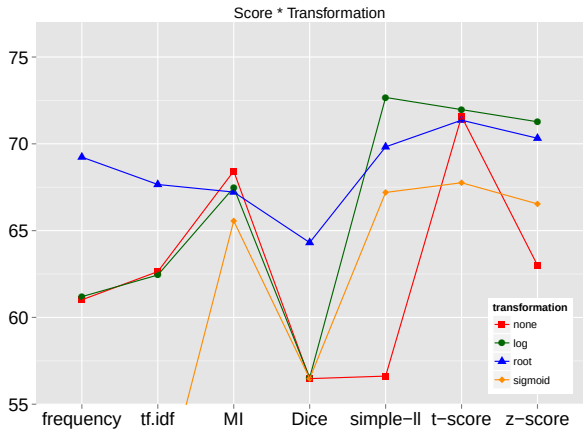

Figure 3.1.12

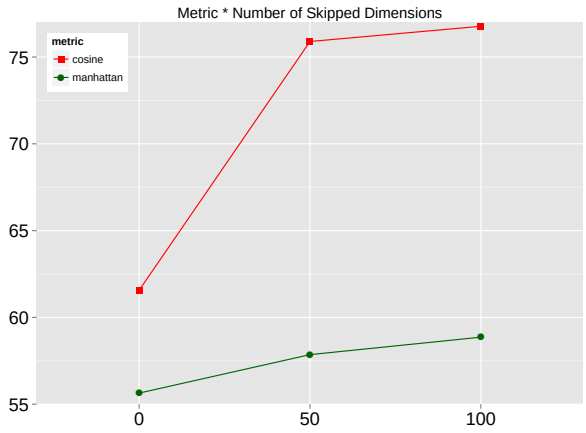

Figure 3.1.13

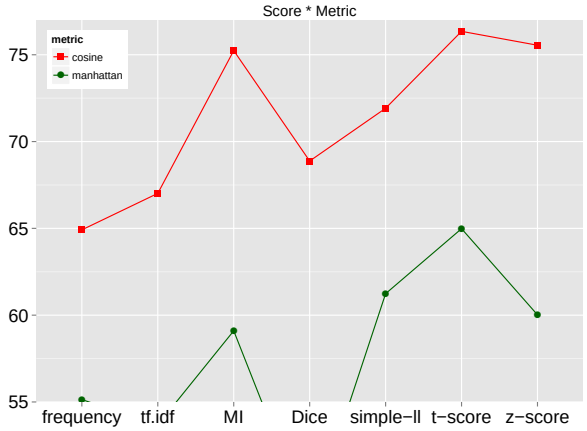

Figure 3.1.14

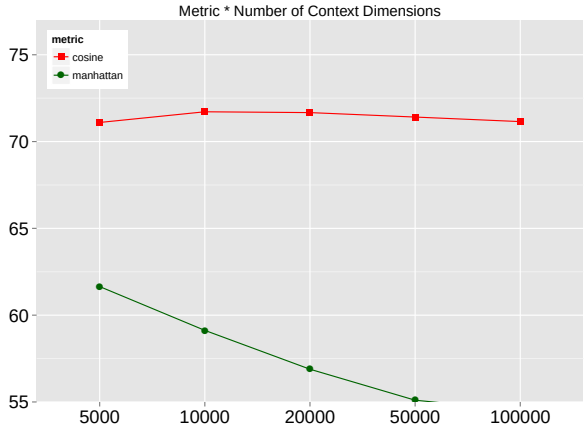

Figure 3.1.15

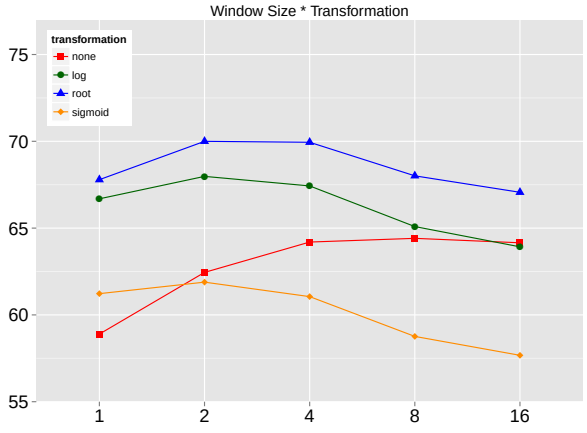

Figure 3.1.16

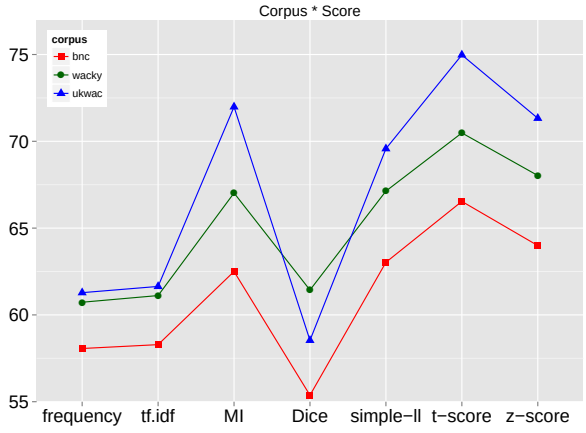

Figure 3.1.17

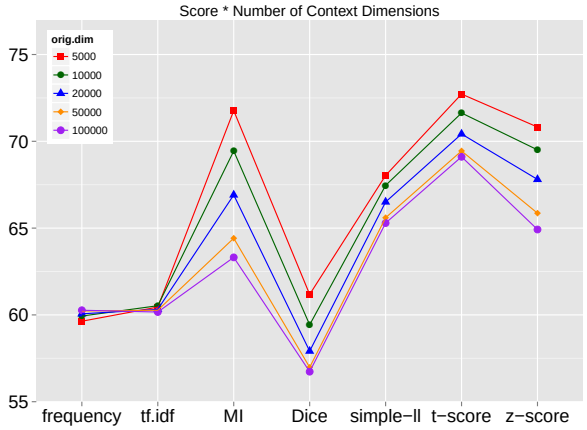

Figure 3.1.18

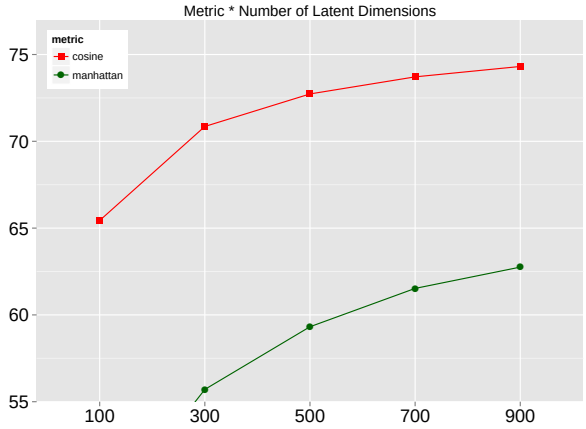

Figure 3.1.19

3.2 Ratings: Rubenstein-Goodenough dataset

Main Effects

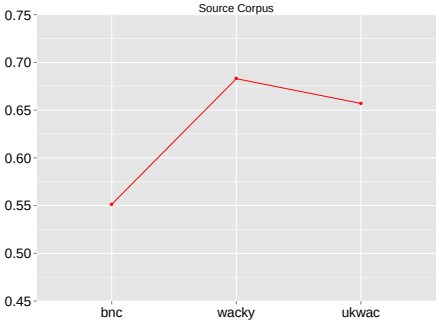

Figure 3.2.1

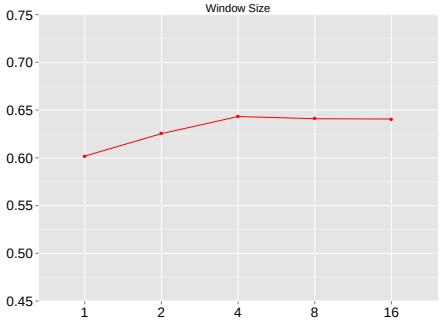

Figure 3.2.2

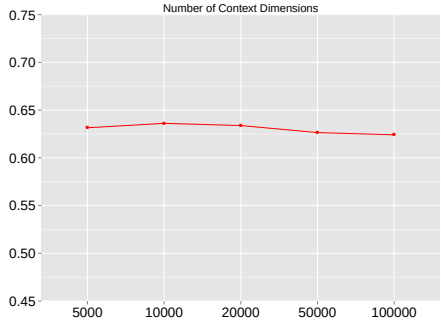

Figure 3.2.3

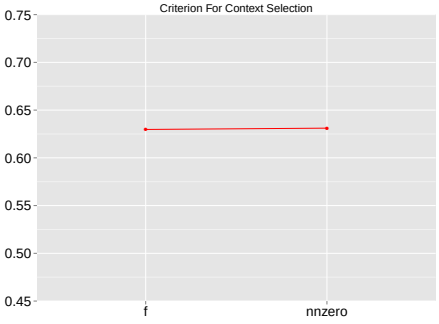

Figure 3.2.4

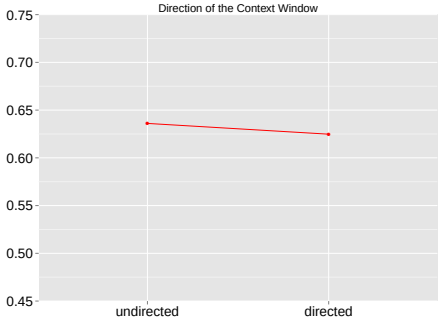

Figure 3.2.5

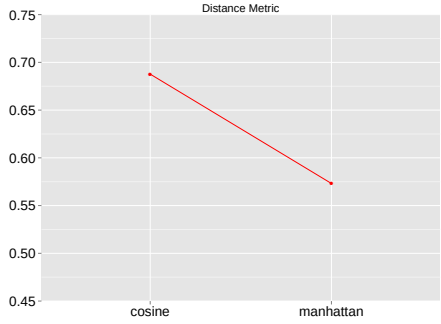

Figure 3.2.6

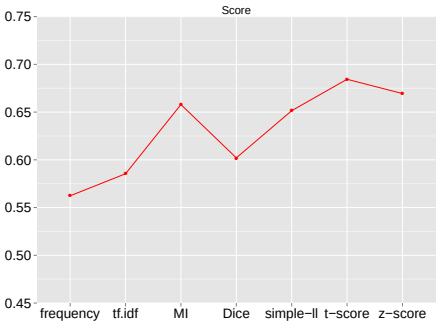

Figure 3.2.7

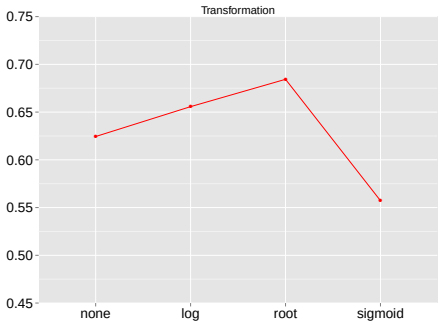

Figure 3.2.8

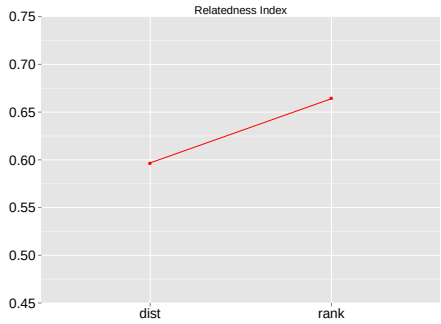

Figure 3.2.9

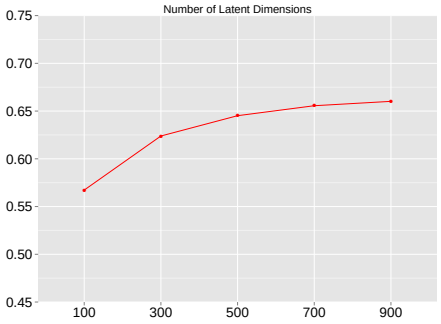

Figure 3.2.10

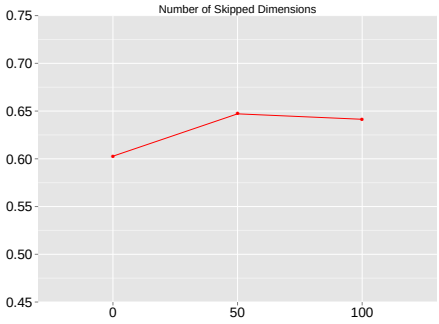

Figure 3.2.11

Interactions

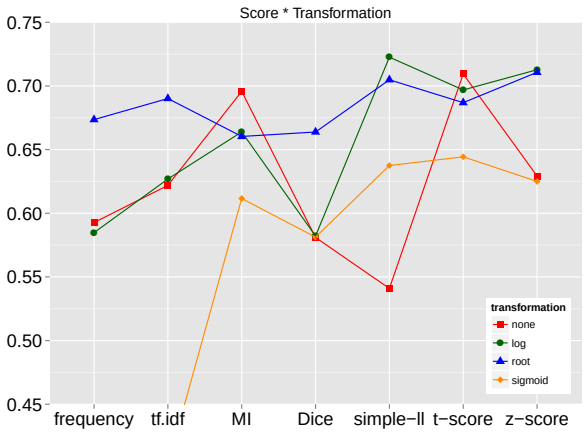

Figure 3.2.12

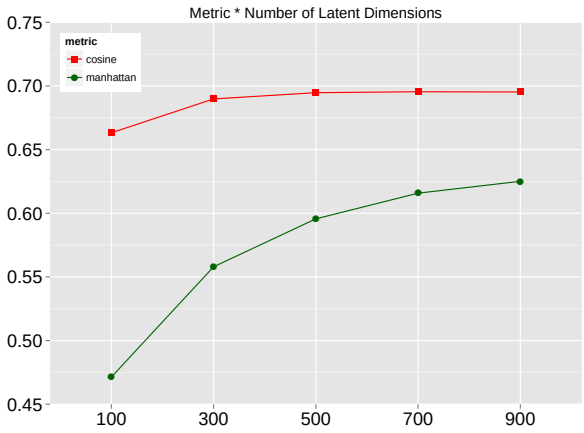

Figure 3.2.13

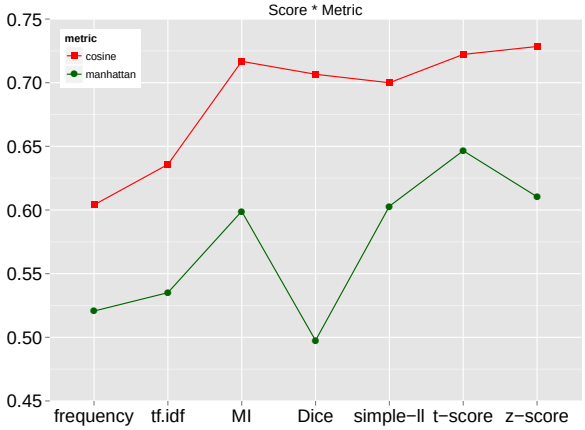

Figure 3.2.14

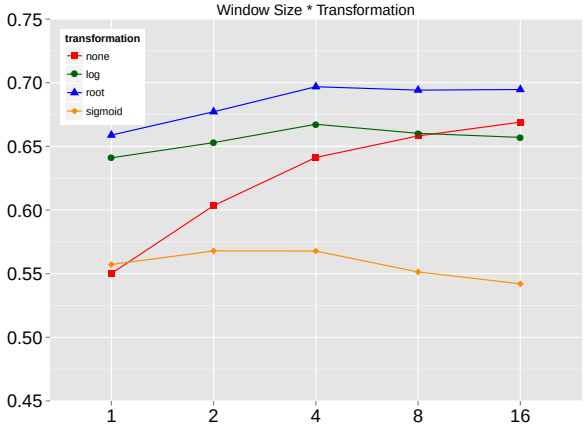

Figure 3.2.15

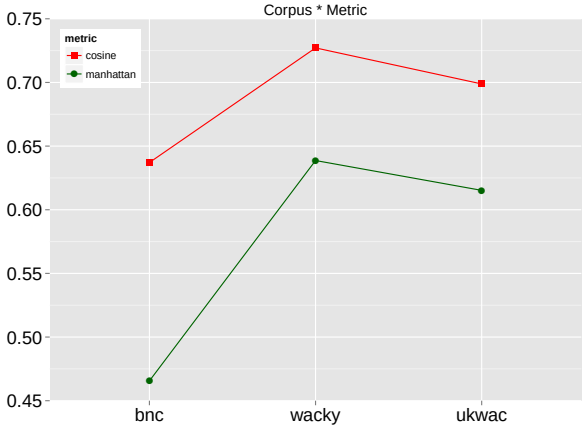

Figure 3.2.16

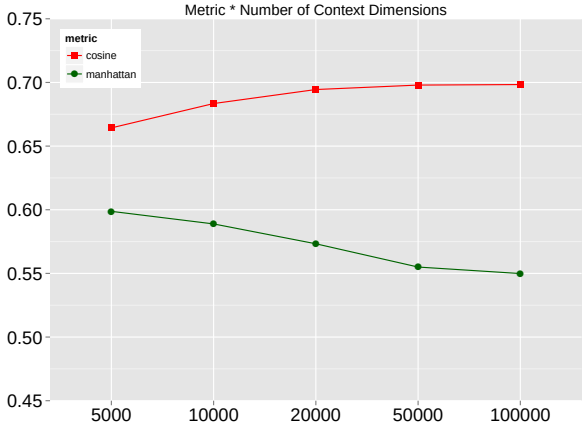

Figure 3.2.17

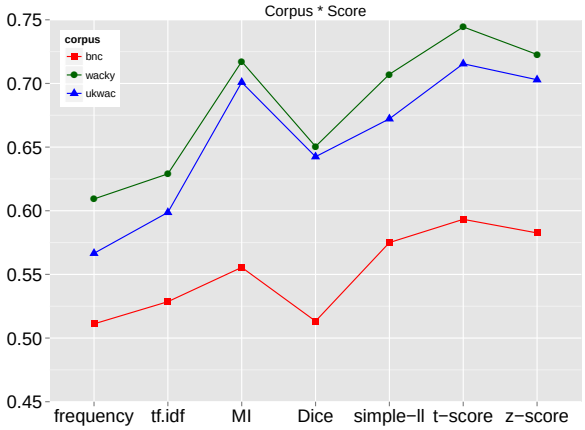

Figure 3.2.18

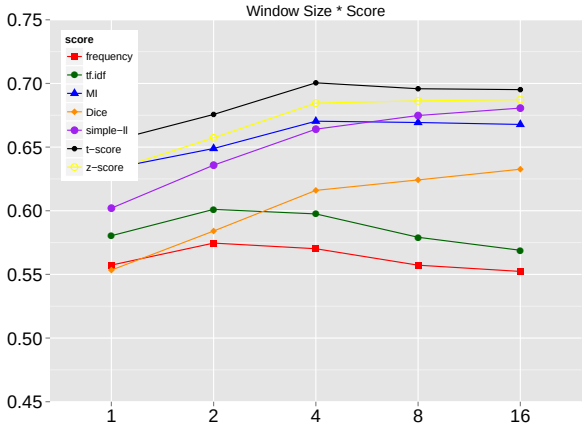

Figure 3.2.19

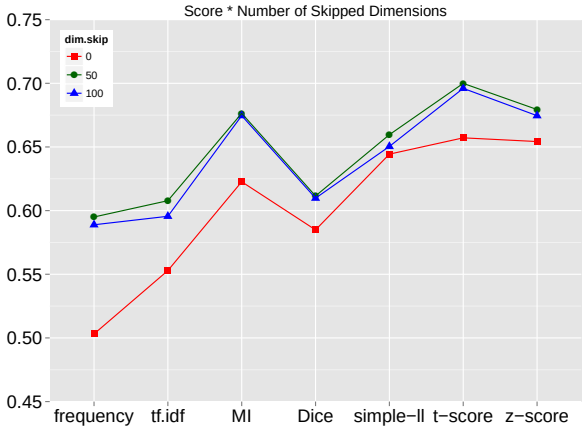

Figure 3.2.20

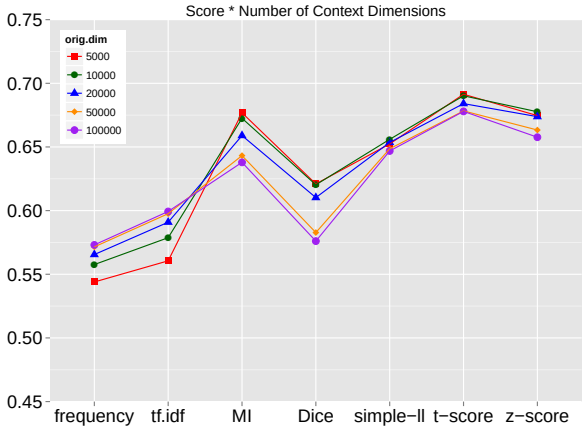

Figure 3.2.21

3.3 Ratings: WordSim353 dataset

Main Effects

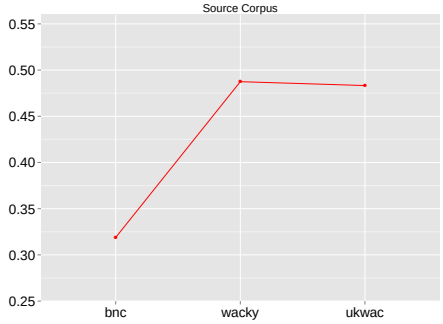

Figure 3.3.1

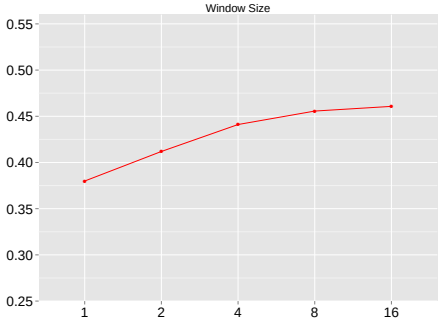

Figure 3.3.2

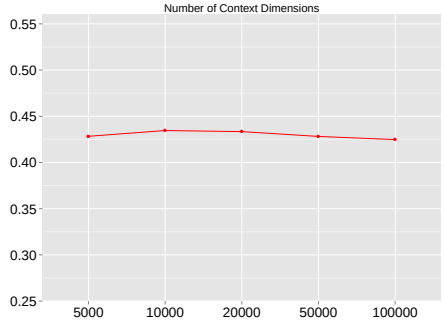

Figure 3.3.3

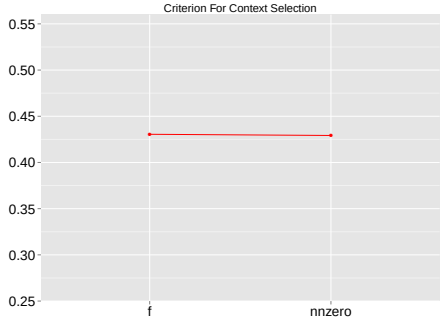

Figure 3.3.4

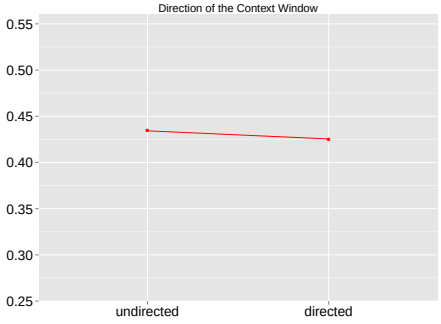

Figure 3.3.5

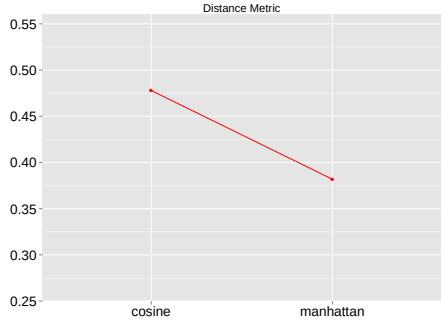

Figure 3.3.6

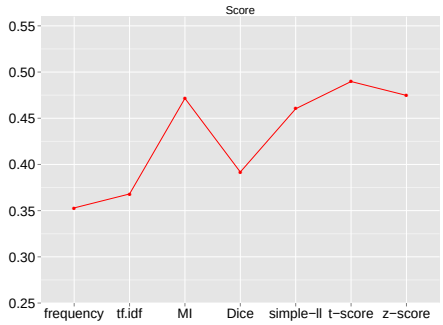

Figure 3.3.7

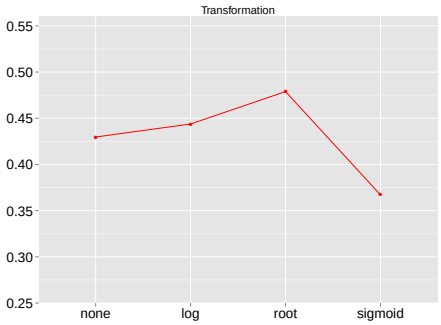

Figure 3.3.8

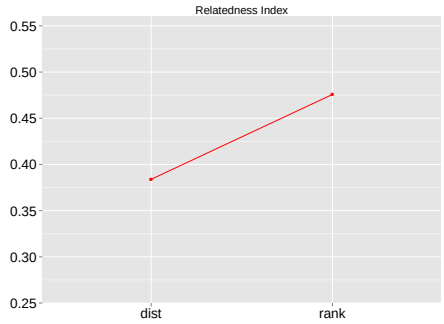

Figure 3.3.9

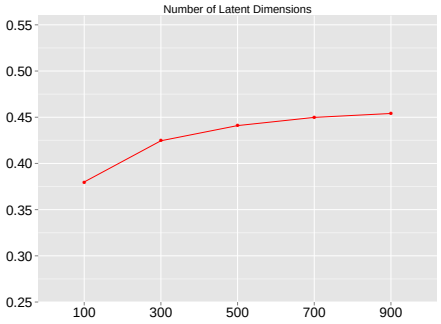

Figure 3.3.10

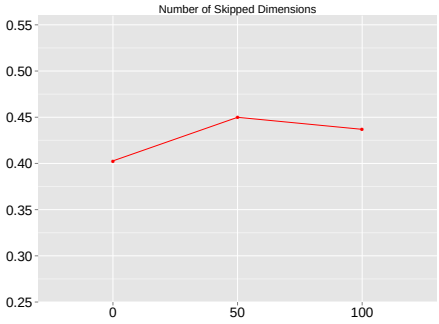

Figure 3.3.11

Interactions

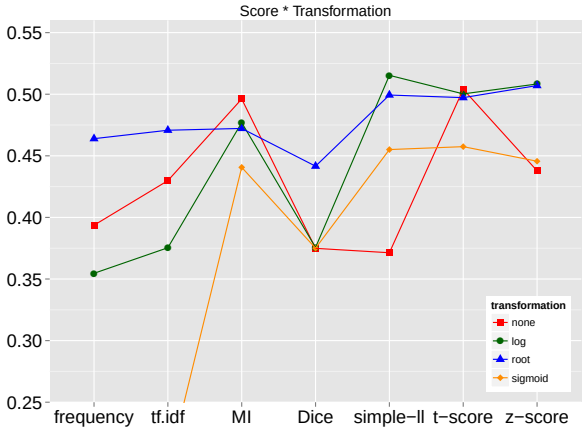

Figure 3.3.12

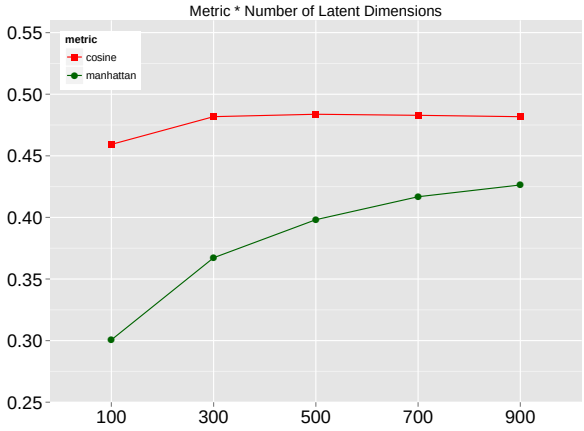

Figure 3.3.13

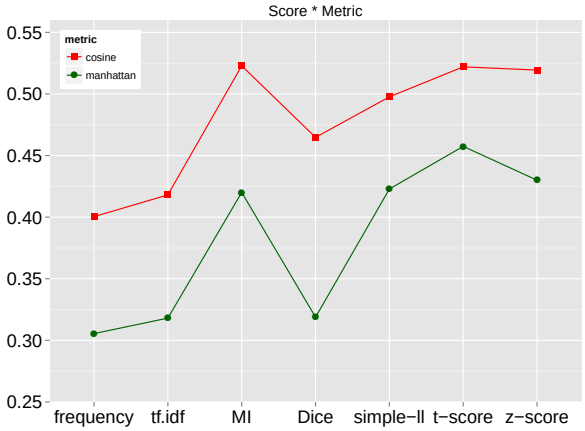

Figure 3.3.14

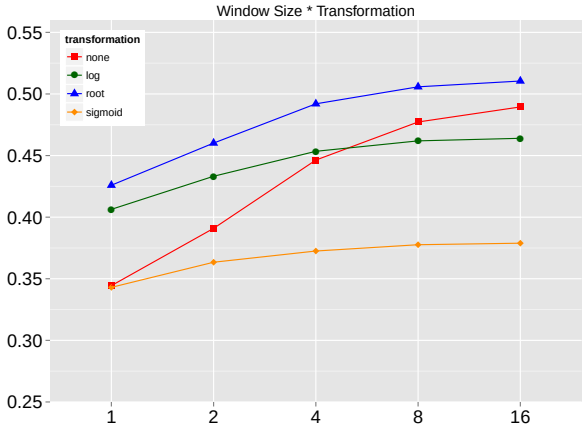

Figure 3.3.15

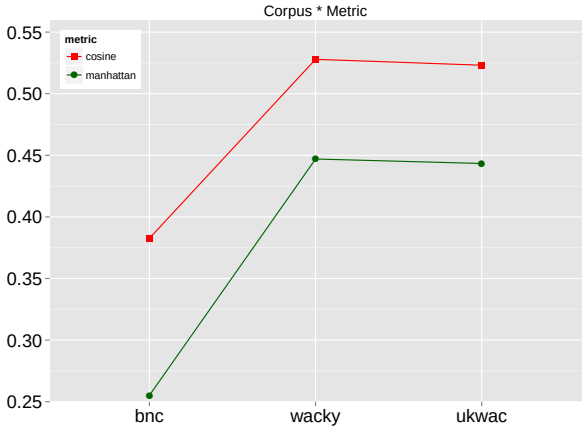

Figure 3.3.16

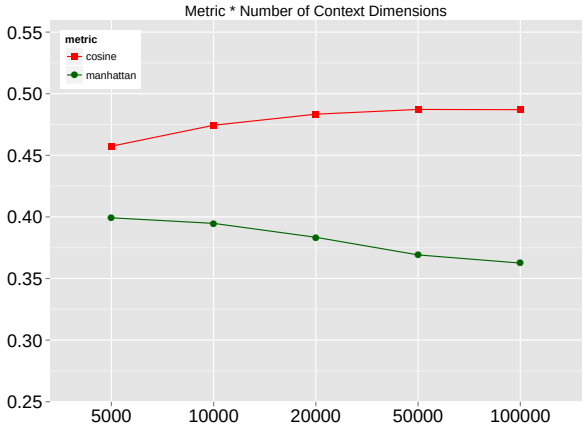

Figure 3.3.17

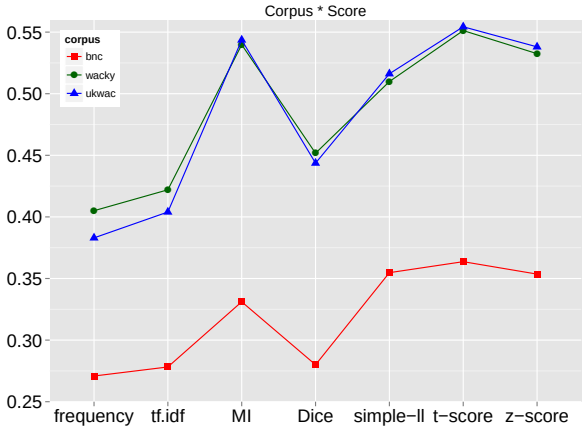

Figure 3.3.18

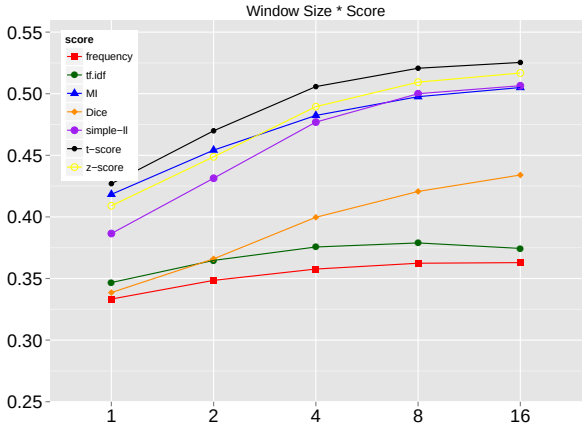

Figure 3.3.19

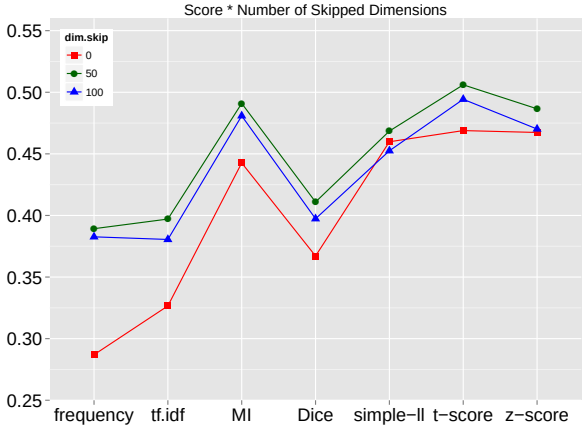

Figure 3.3.20

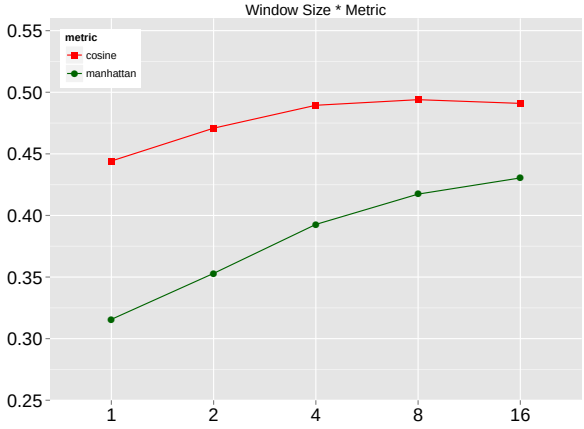

Figure 3.3.21

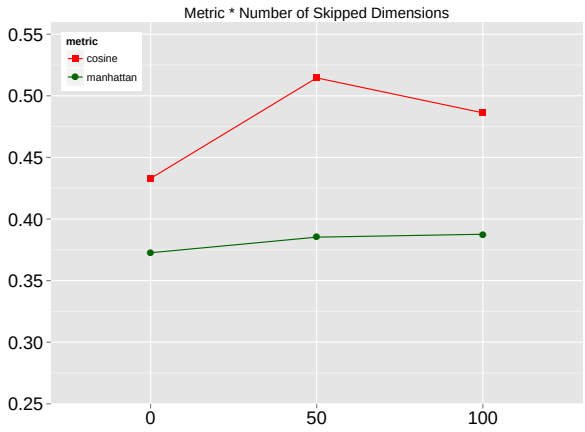

Figure 3.3.22

3.4 Clustering: Almuhareb-Poesio dataset

Main Effects

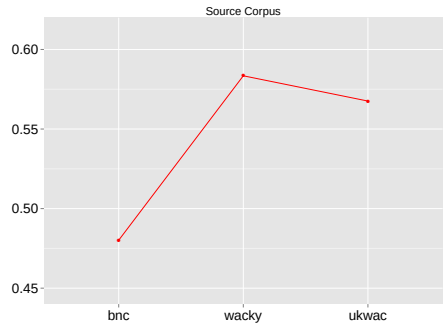

Figure 3.4.1

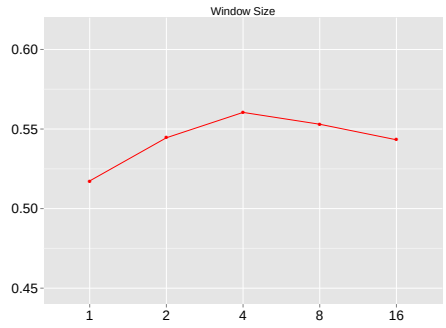

Figure 3.4.2

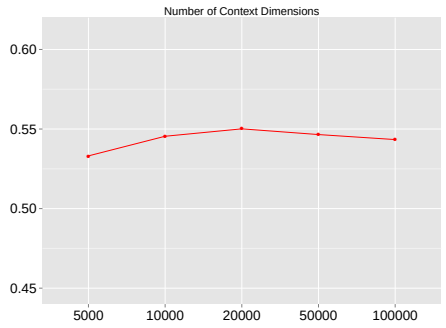

Figure 3.4.3

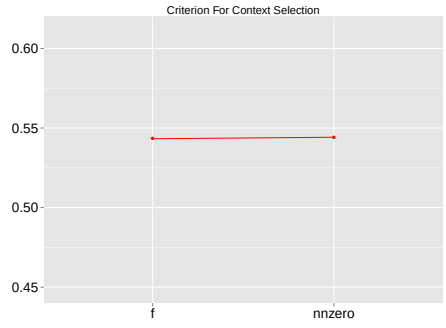

Figure 3.4.4

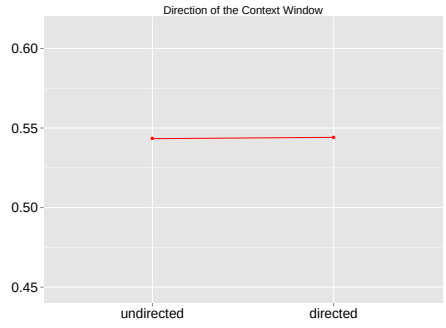

Figure 3.4.5

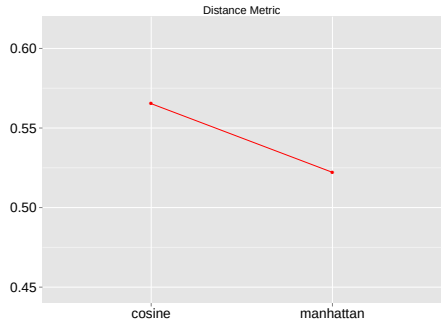

Figure 3.4.6

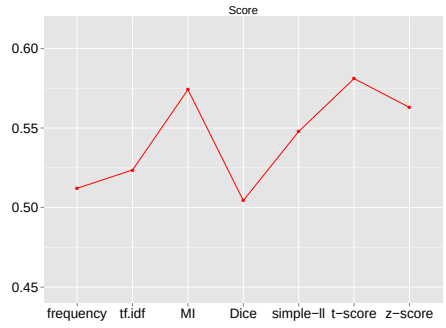

Figure 3.4.7

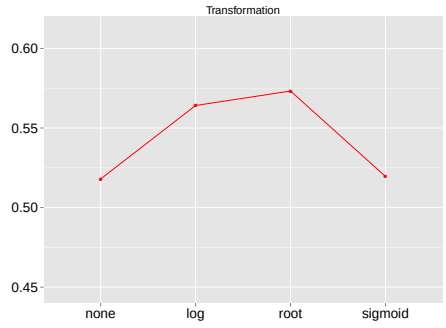

Figure 3.4.8

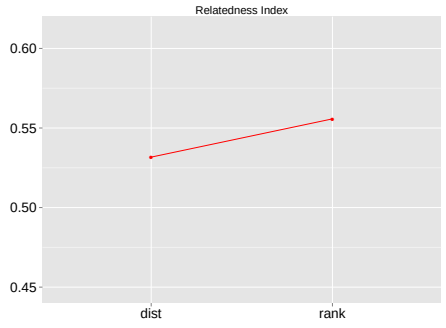

Figure 3.4.9

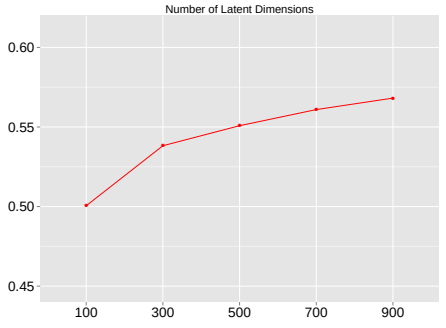

Figure 3.4.10

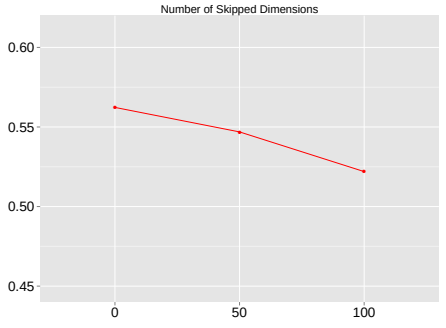

Figure 3.4.11

Interactions

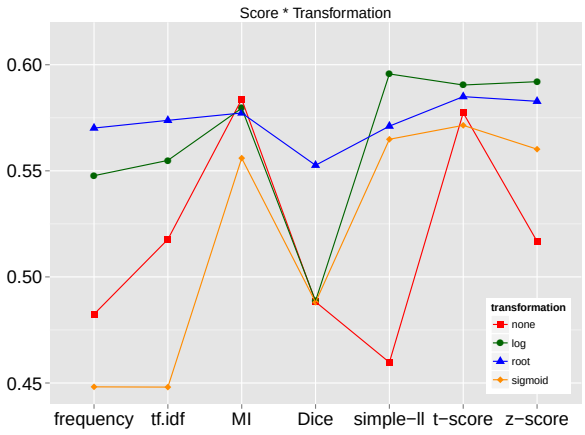

Figure 3.4.12

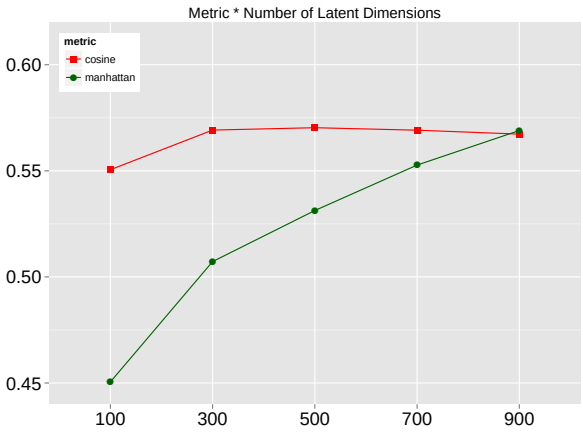

Figure 3.4.13

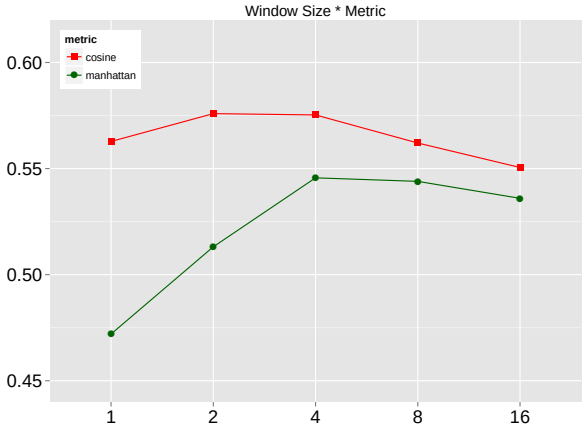

Figure 3.4.14

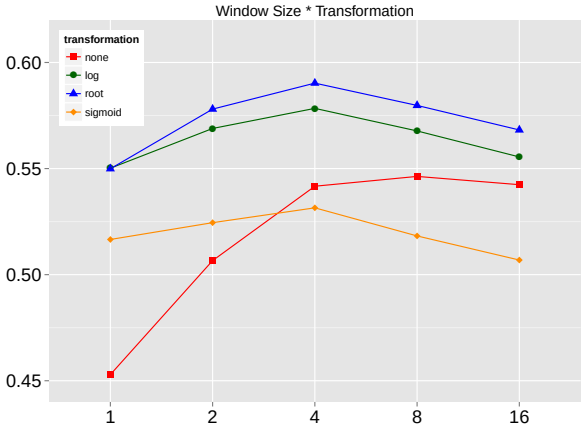

Figure 3.4.15

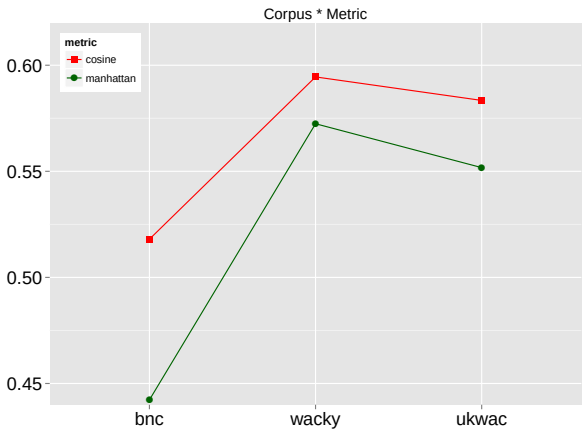

Figure 3.4.16

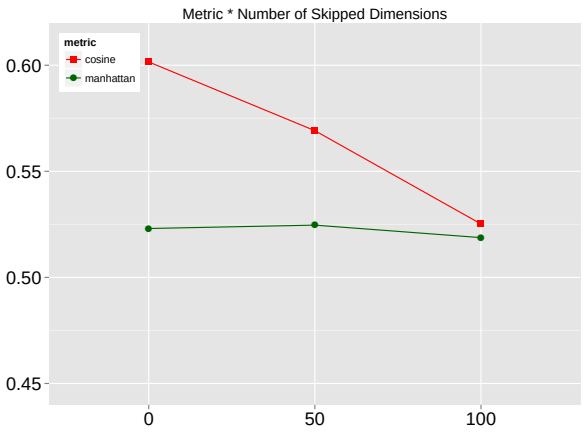

Figure 3.4.17

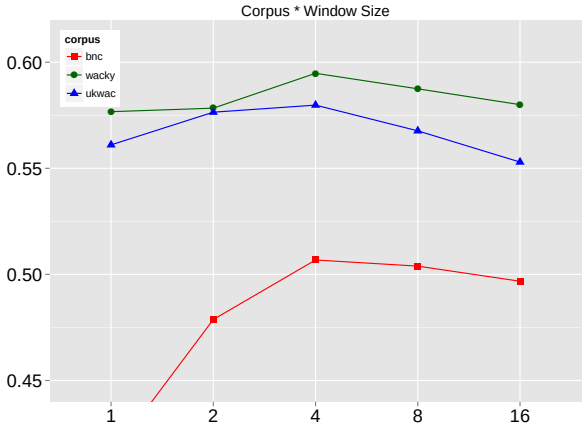

Figure 3.4.18

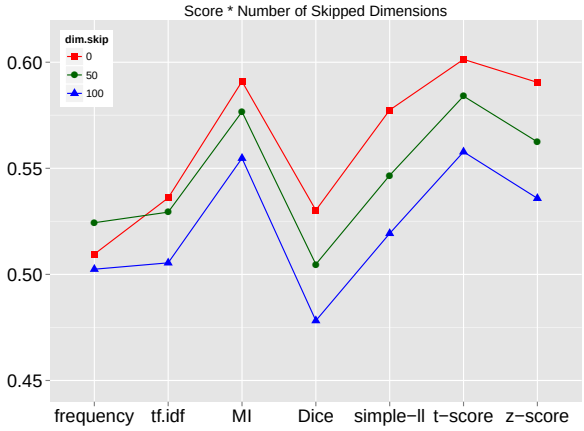

Figure 3.4.19

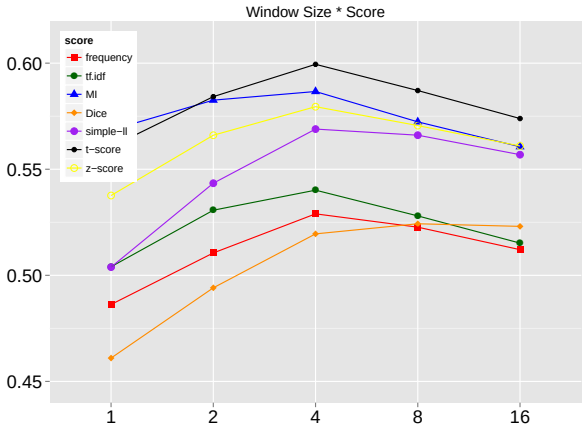

Figure 3.4.20

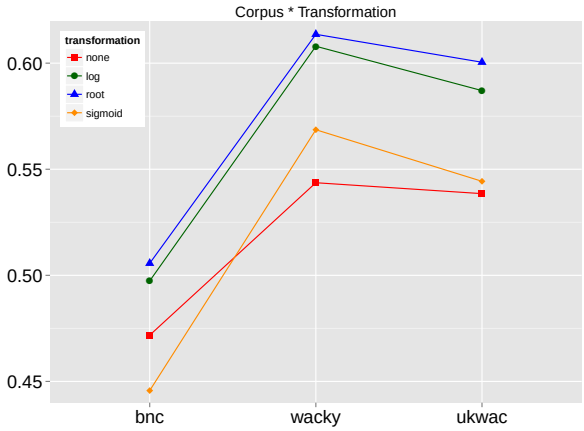

Figure 3.4.21

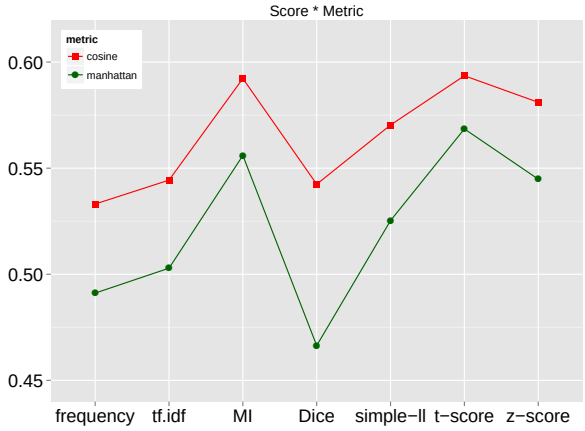

Figure 3.4.22

3.5 Clustering: BATTIG dataset

Main Effects

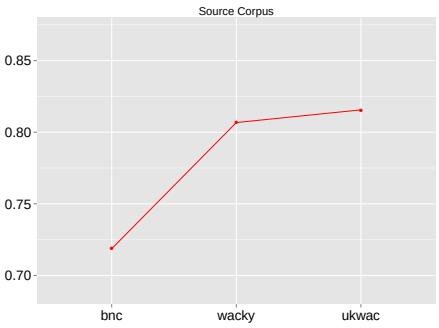

Figure 3.5.1

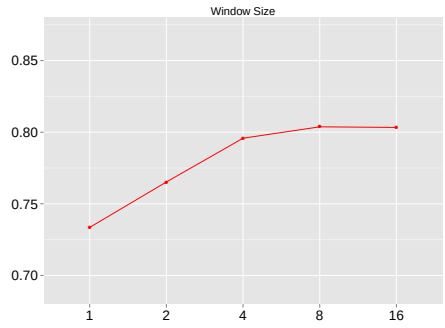

Figure 3.5.2

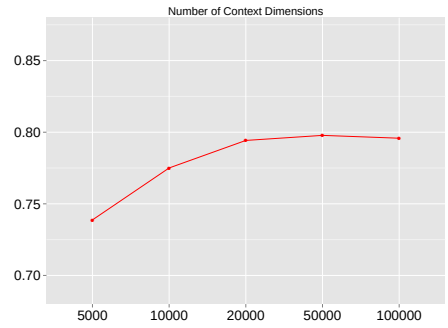

Figure 3.5.3

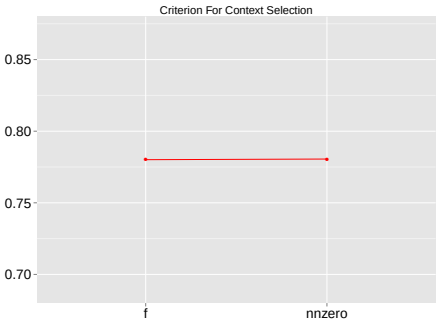

Figure 3.5.4

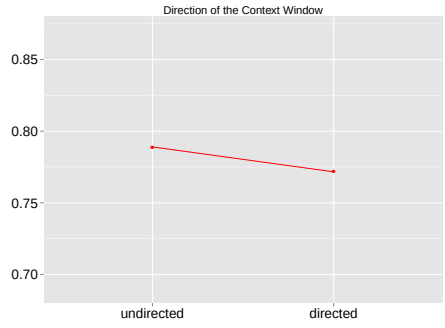

Figure 3.5.5

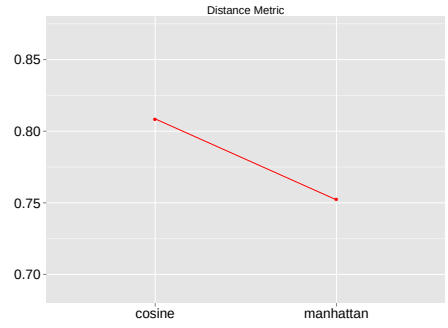

Figure 3.5.6

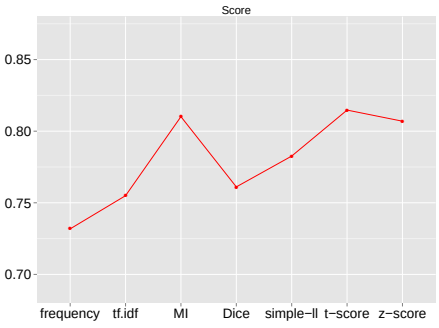

Figure 3.5.7

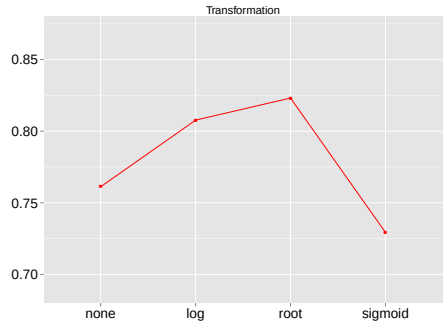

Figure 3.5.8

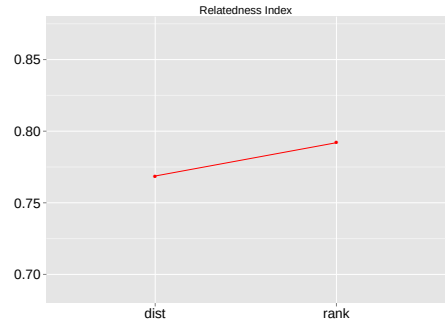

Figure 3.5.9

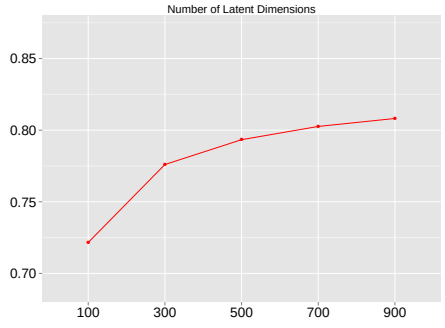

Figure 3.5.10

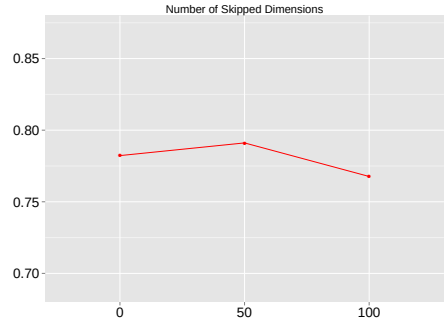

Figure 3.5.11

Interactions

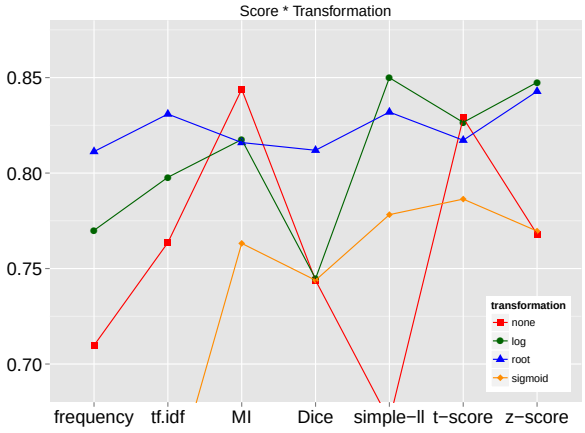

Figure 3.5.12

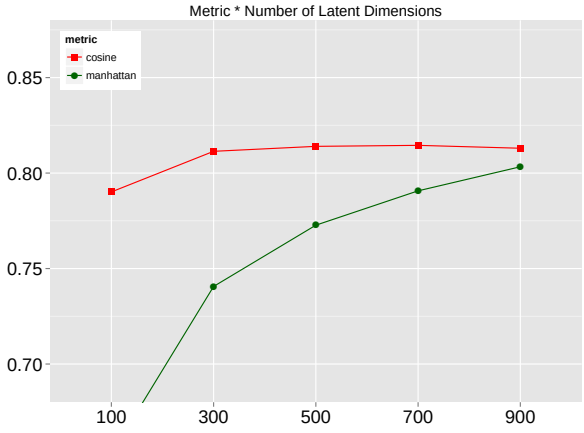

Figure 3.5.13

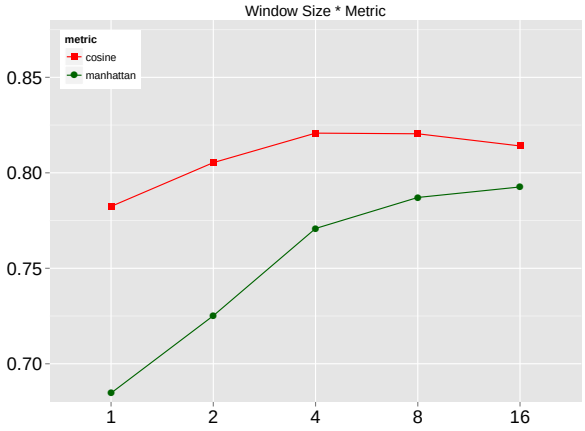

Figure 3.5.14

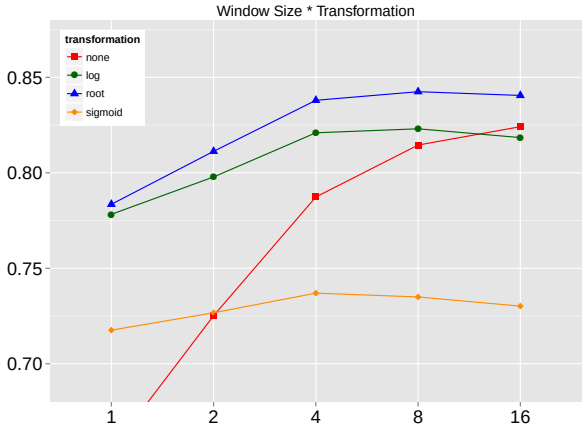

Figure 3.5.15

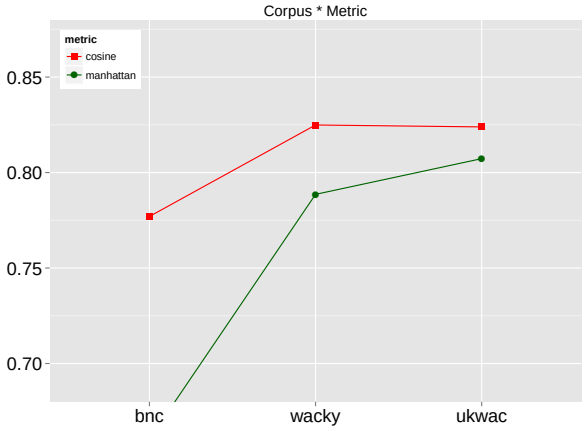

Figure 3.5.16

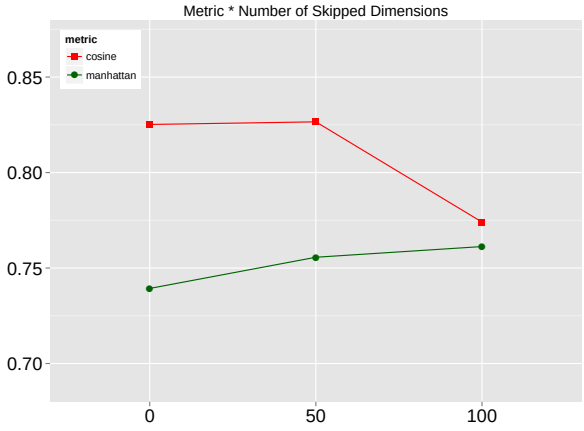

Figure 3.5.17

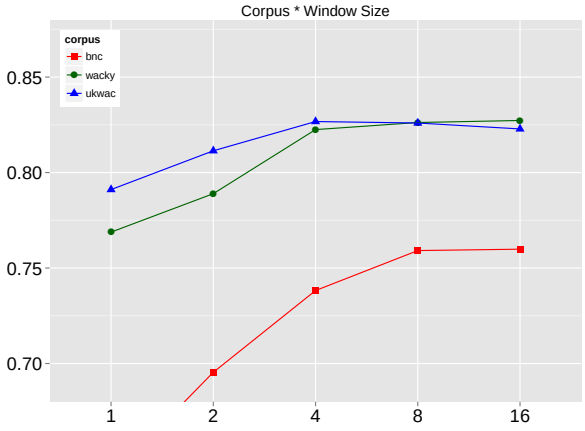

Figure 3.5.18

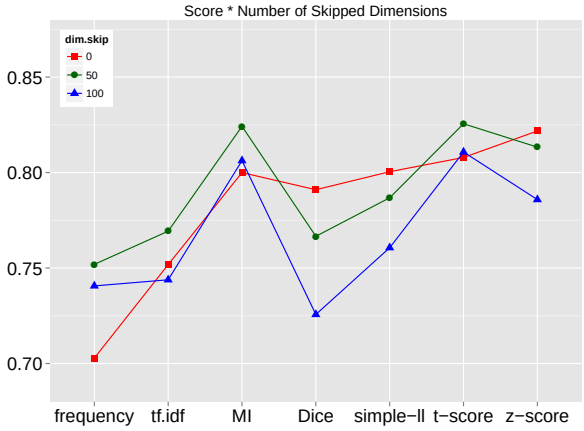

Figure 3.5.19

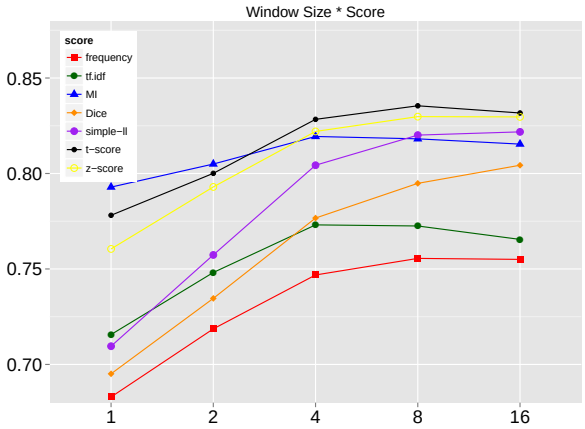

Figure 3.5.20

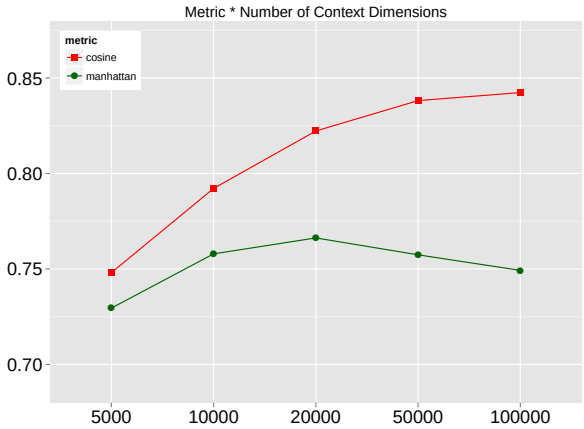

Figure 3.5.21

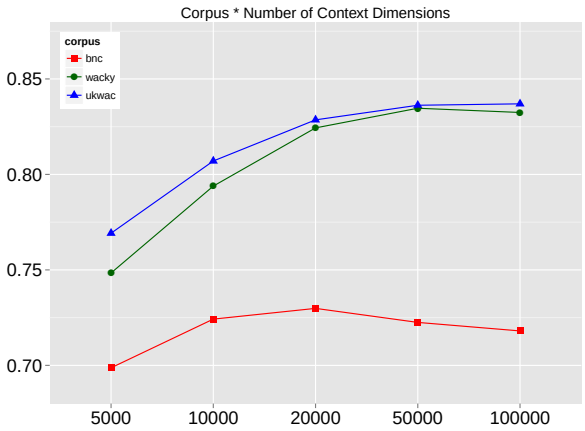

Figure 3.5.22

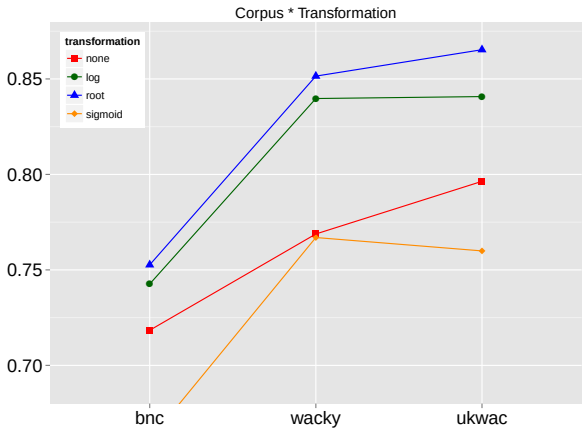

Figure 3.5.23

3.6 Clustering: ESSLLI dataset

Main Effects

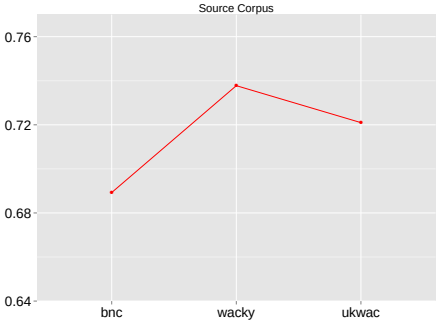

Figure 3.6.1

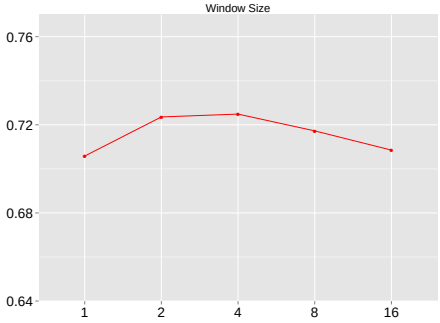

Figure 3.6.2

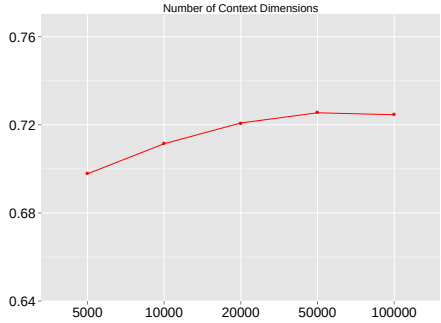

Figure 3.6.3

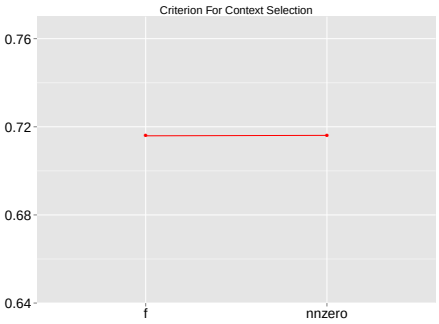

Figure 3.6.4

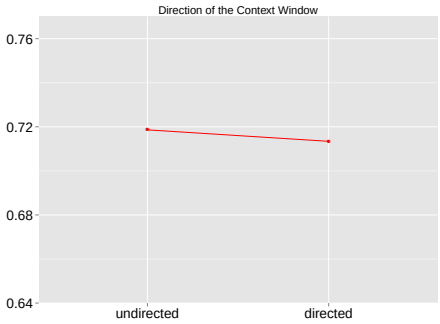

Figure 3.6.5

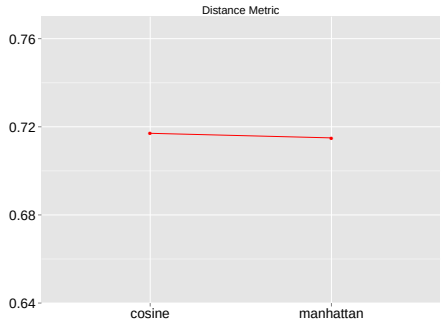

Figure 3.6.6

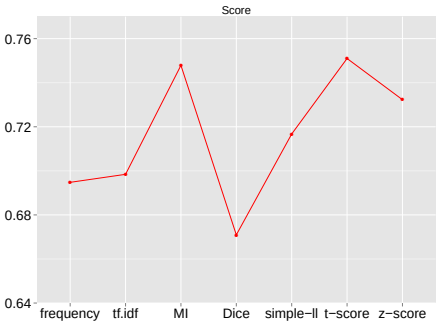

Figure 3.6.7

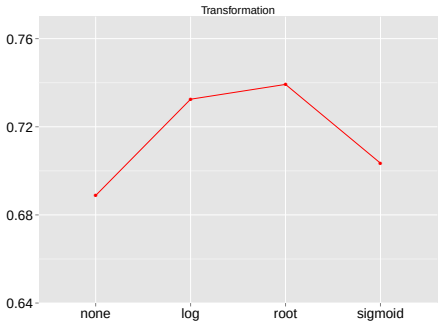

Figure 3.6.8

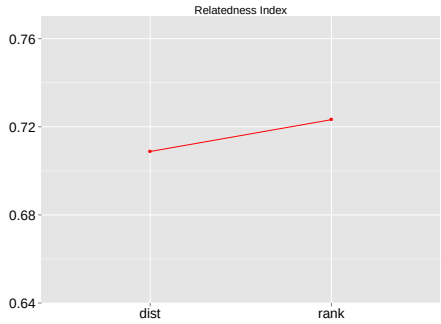

Figure 3.6.9

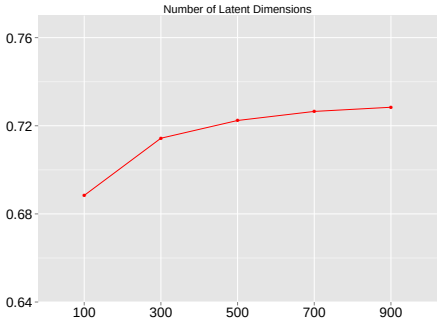

Figure 3.6.10

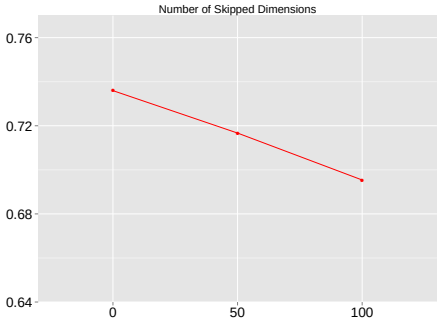

Figure 3.6.11

Interactions

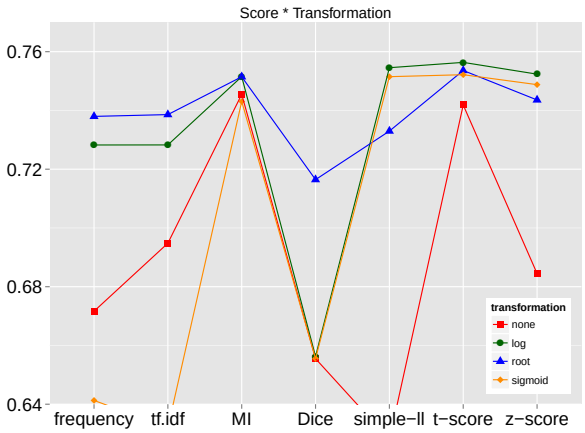

Figure 3.6.12

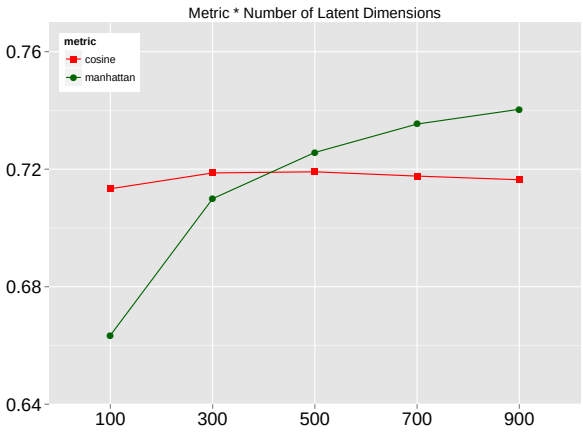

Figure 3.6.13

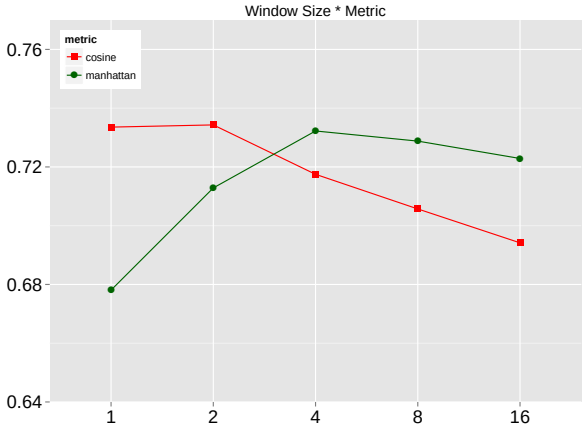

Figure 3.6.14

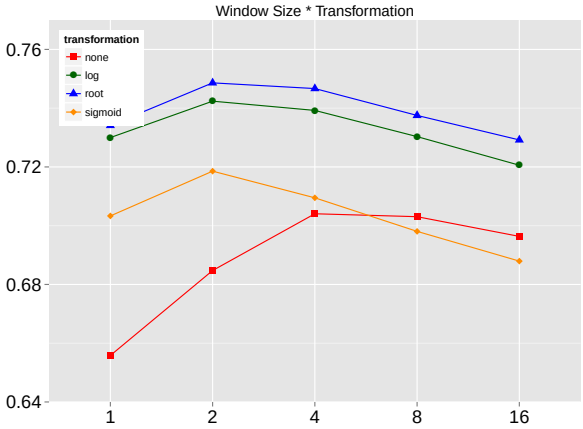

Figure 3.6.15

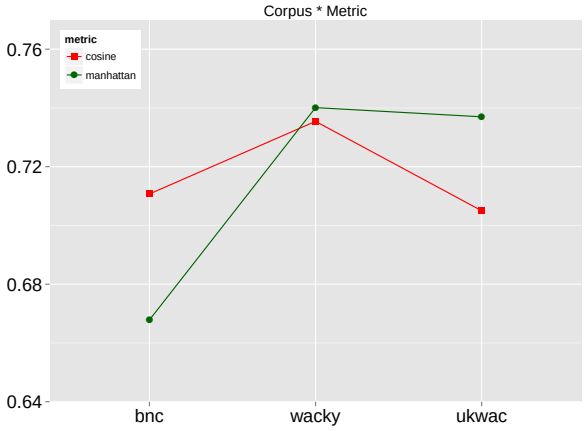

Figure 3.6.16

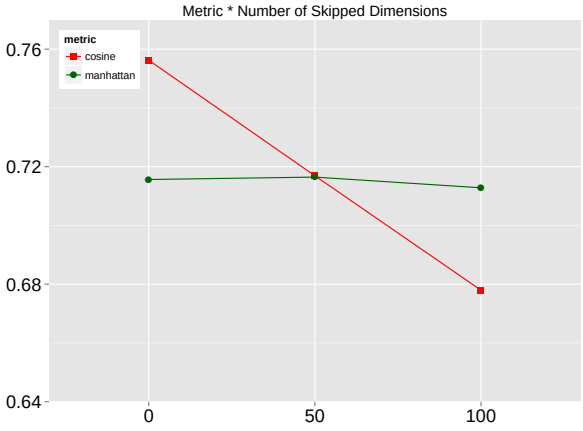

Figure 3.6.17

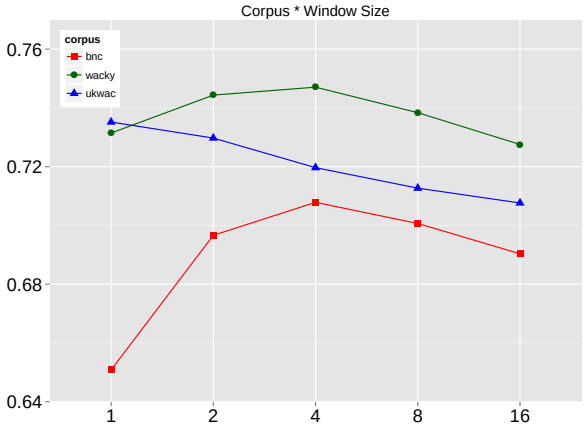

Figure 3.6.18

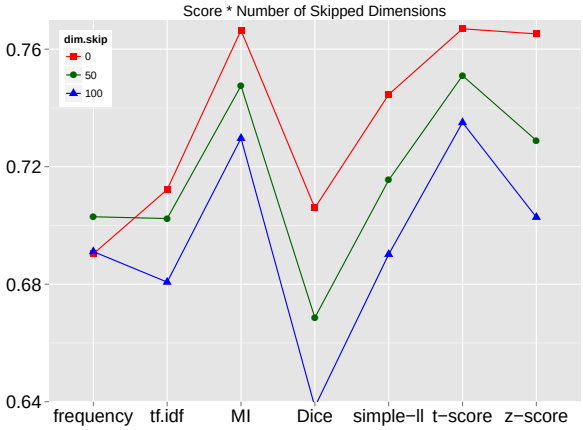

Figure 3.6.19

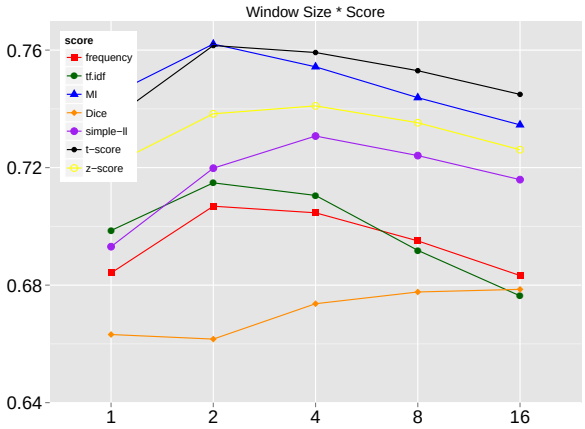

Figure 3.6.20

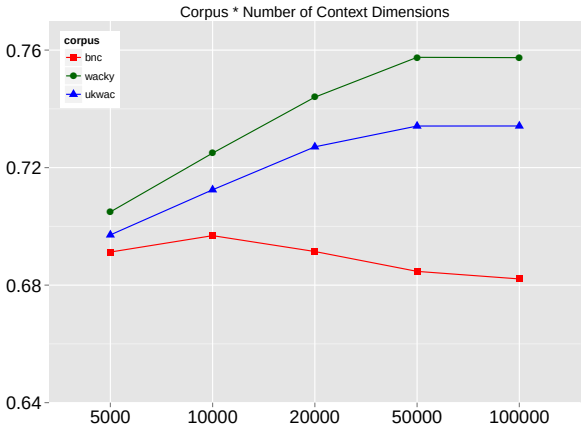

Figure 3.6.21

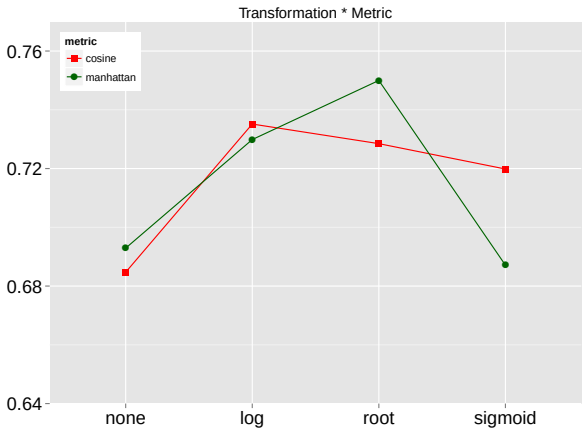

Figure 3.6.22

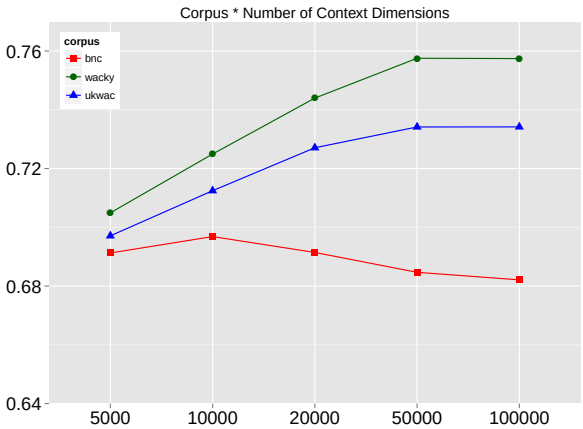

Figure 3.6.23

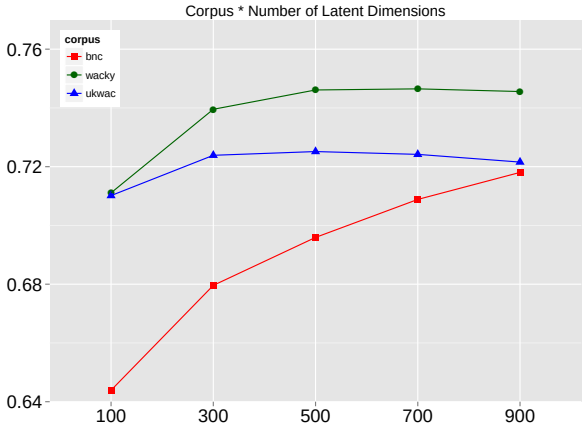

Figure 3.6.24

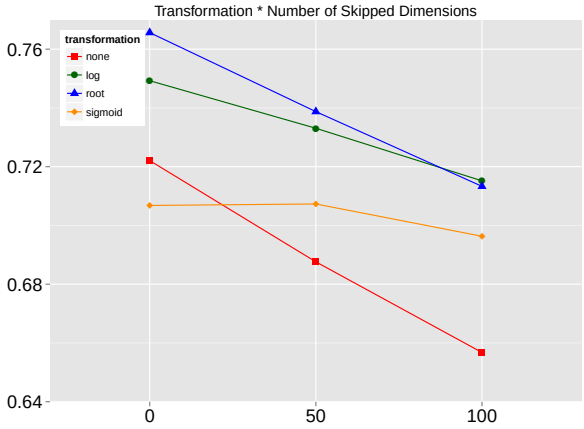

Figure 3.6.25

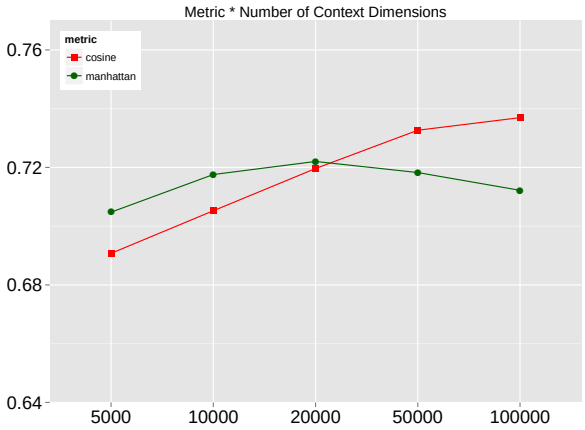

Figure 3.6.26

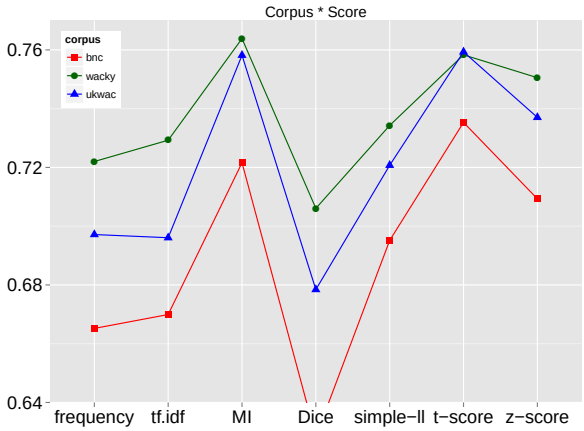

Figure 3.6.27

3.7 Clustering: MITCHELL dataset

Main Effects

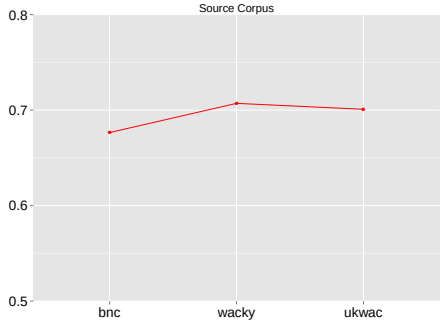

Figure 3.7.1

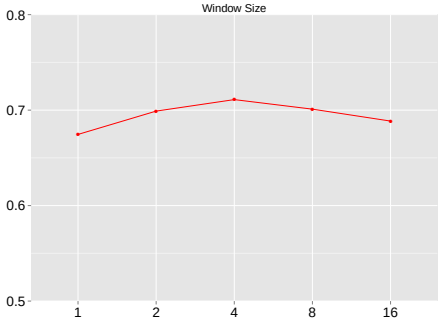

Figure 3.7.2

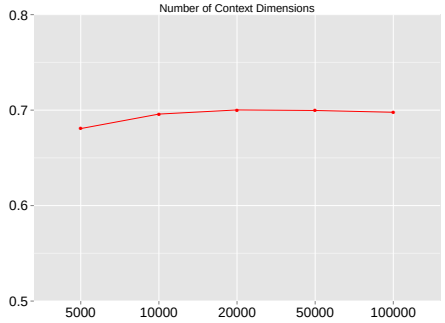

Figure 3.7.3

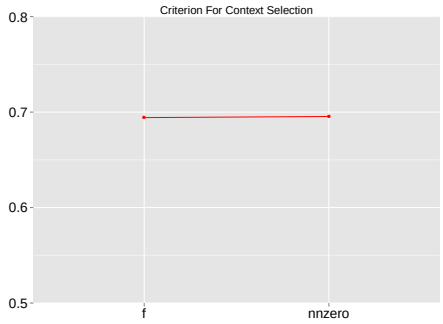

Figure 3.7.4

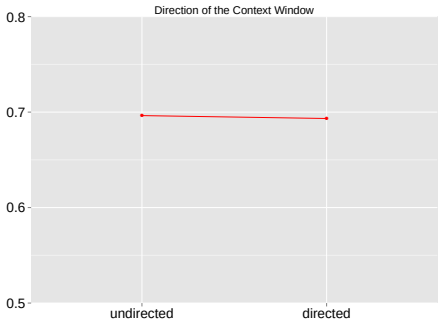

Figure 3.7.5

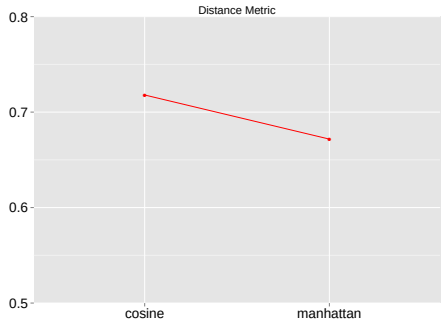

Figure 3.7.6

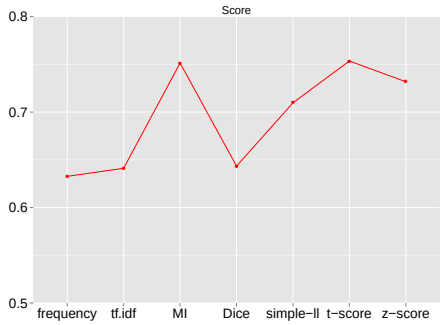

Figure 3.7.7

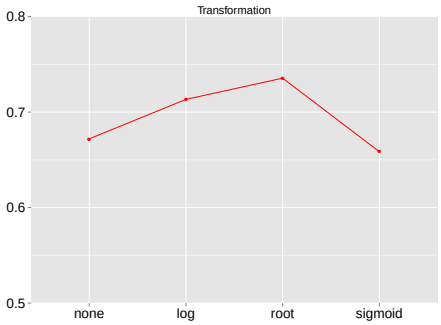

Figure 3.7.8

Figure 3.7.9

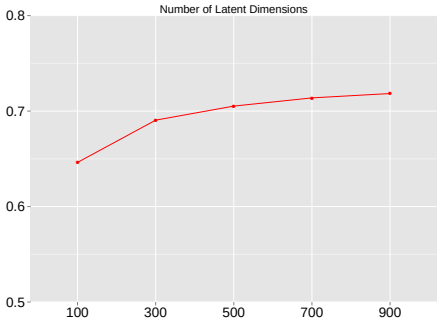

Figure 3.7.10

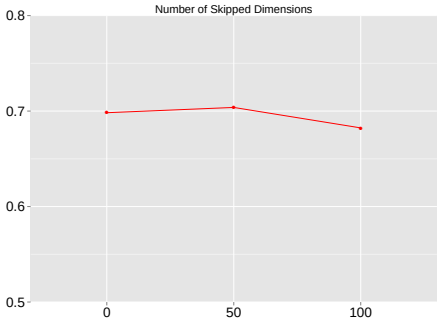

Figure 3.7.11

Interactions

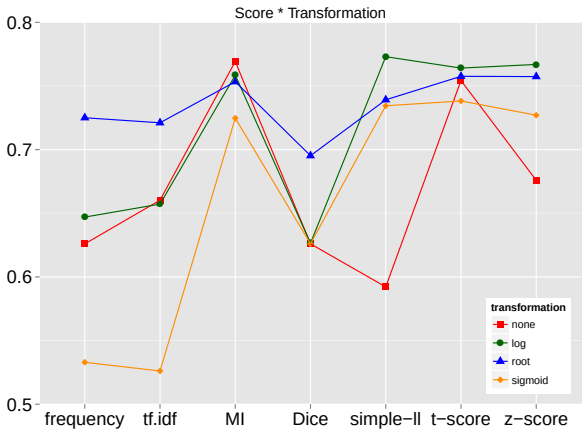

Figure 3.7.12

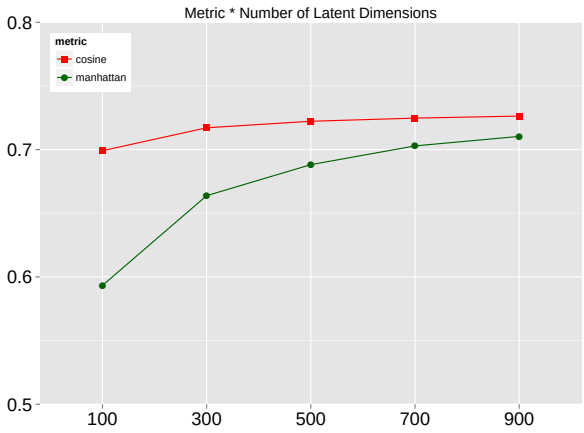

Figure 3.7.13

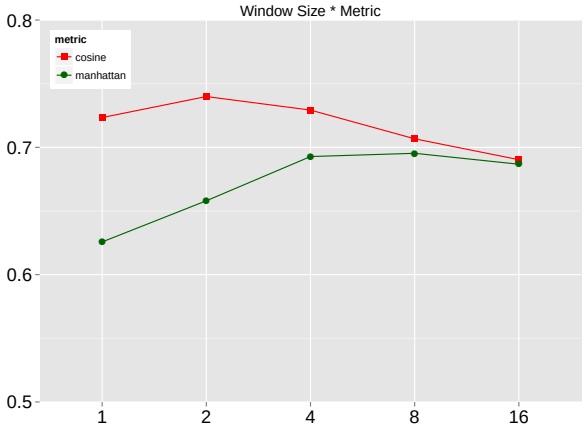

Figure 3.7.14

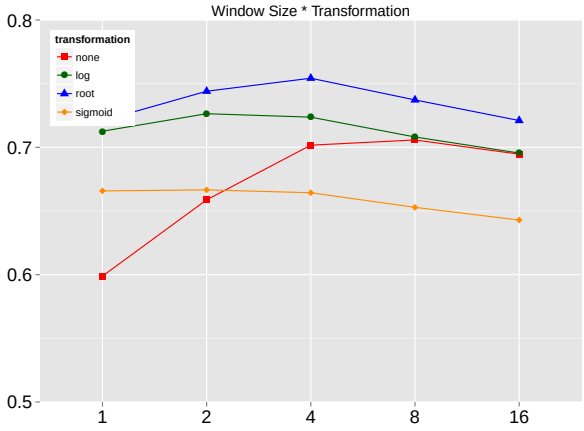

Figure 3.7.15

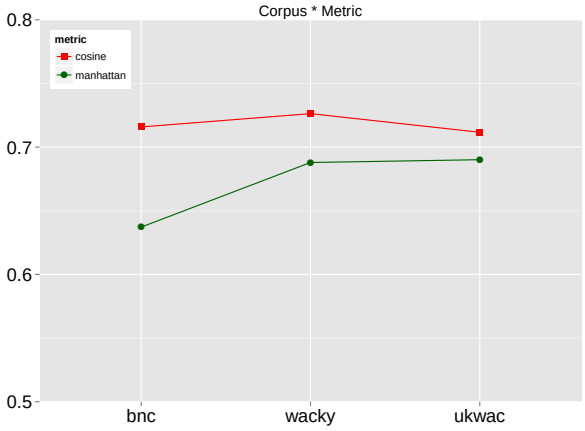

Figure 3.7.16

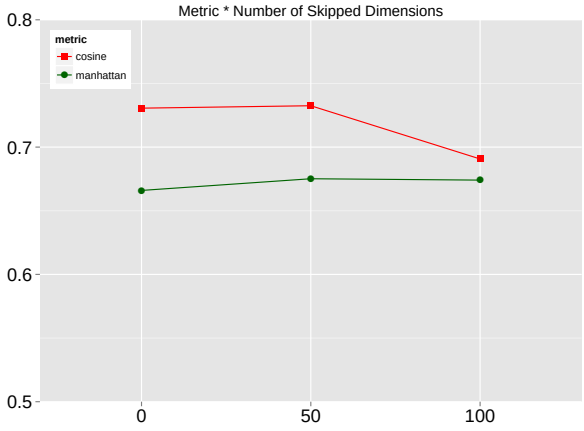

Figure 3.7.17

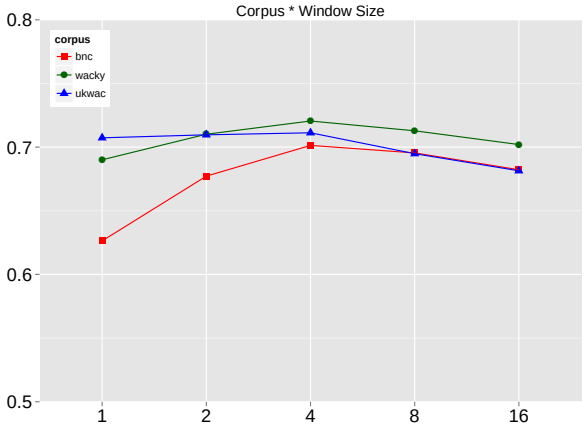

Figure 3.7.18

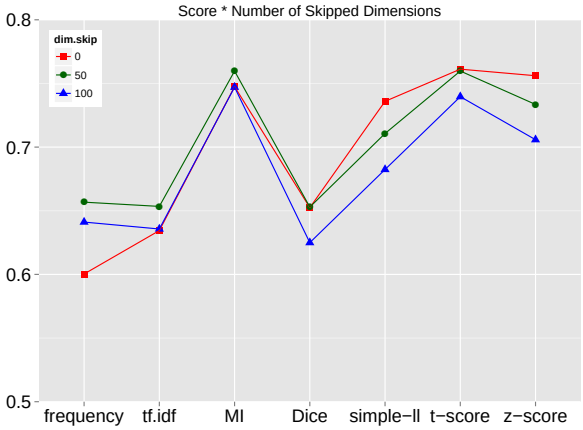

Figure 3.7.19

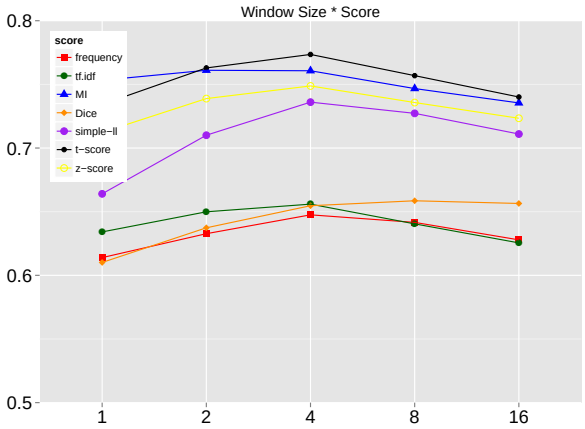

Figure 3.7.20

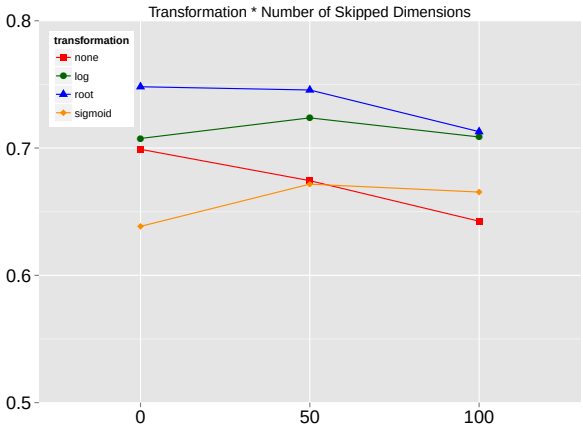

Figure 3.7.21

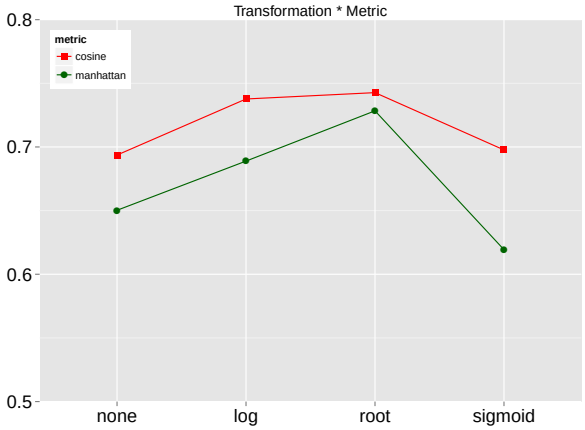

Figure 3.7.22

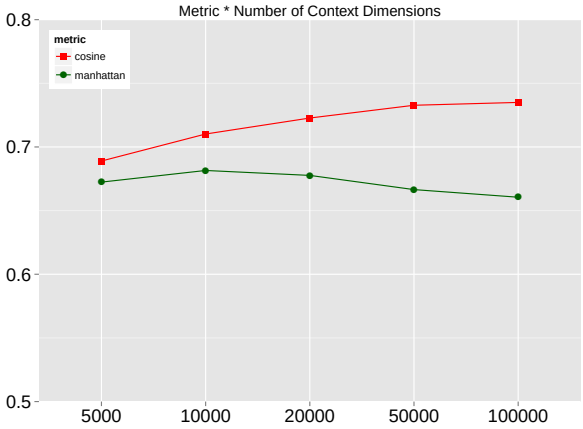

Figure 3.7.23

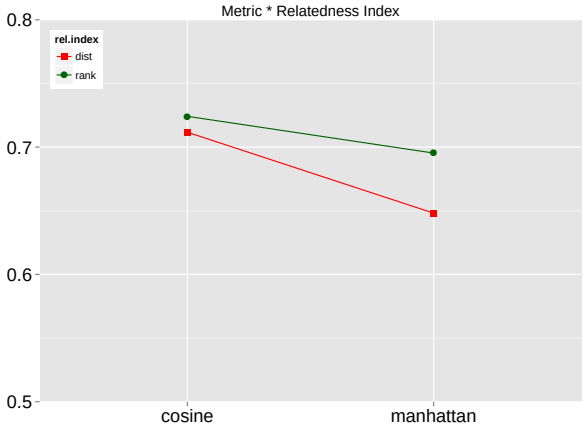

Figure 3.7.24

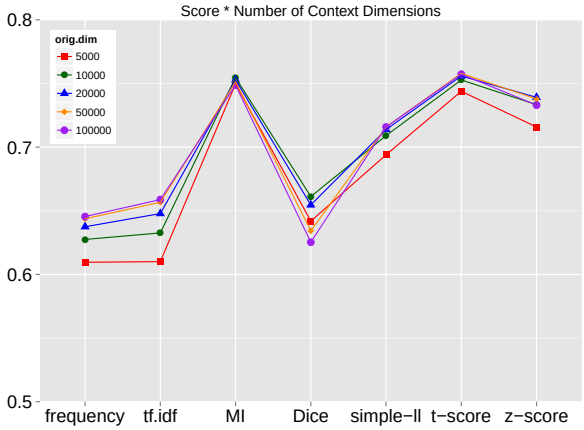

Figure 3.7.25

4 Interactions: Overview

4.1 Score * Transformation

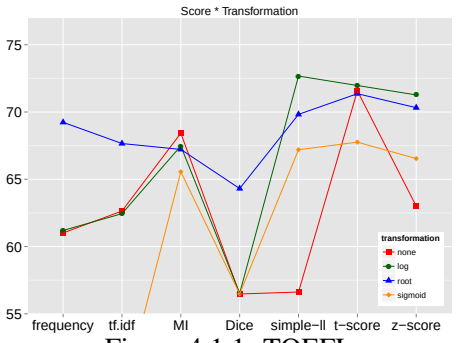

Figure 4.1.1: TOEFL

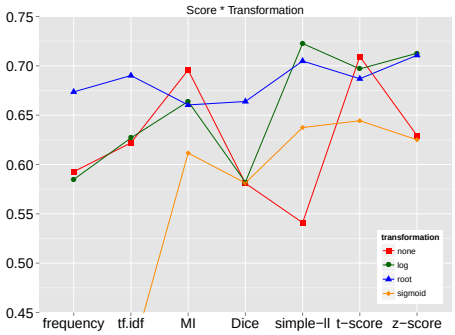

Figure 4.1.2: RG65

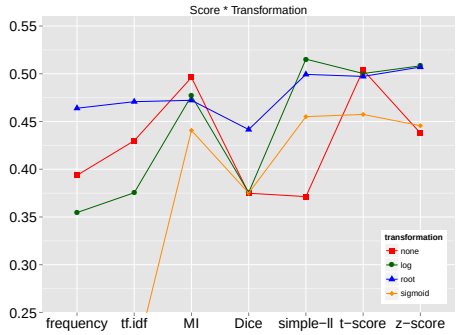

Figure 4.1.3: WS353

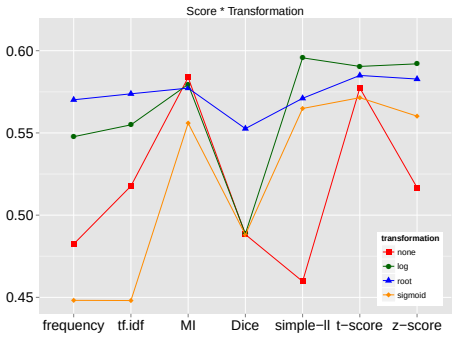

Figure 4.1.4: AP

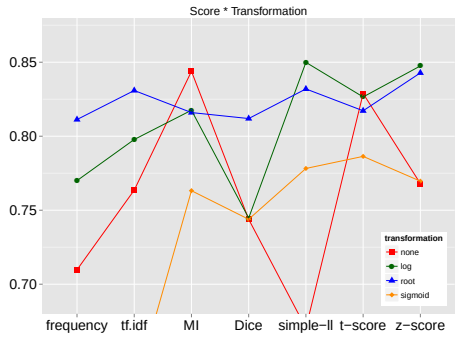

Figure 4.1.5: BATTIG

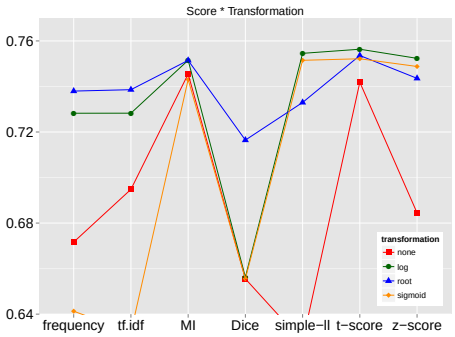

Figure 4.1.6: ESSLLI

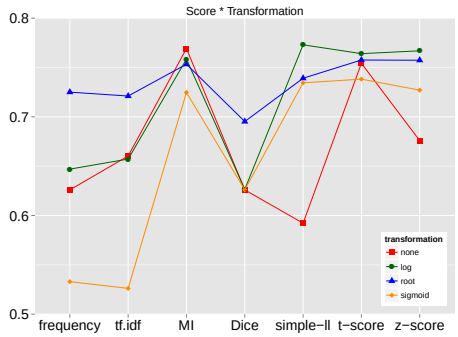

Figure 4.1.7: MITCHELL

4.2 Window * Transformation

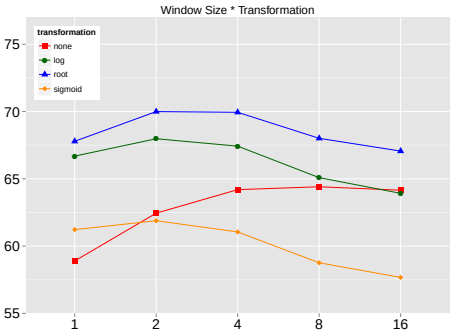

Figure 4.2.1: TOEFL

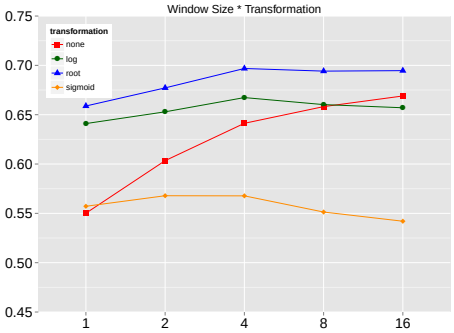

Figure 4.2.2: RG65

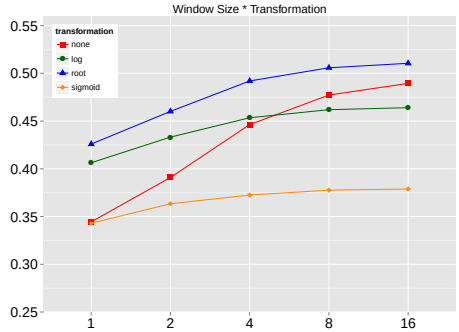

Figure 4.2.3: WS353

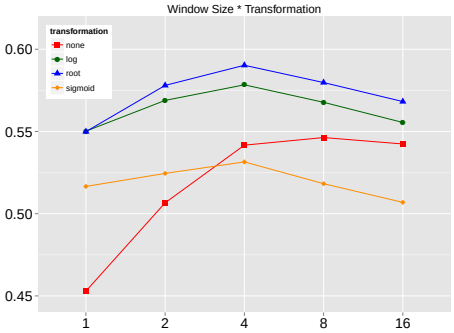

Figure 4.2.4: AP

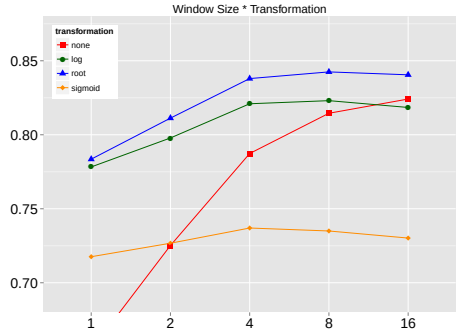

Figure 4.2.5: BATTIG

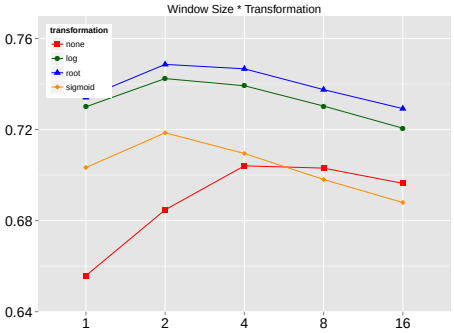

Figure 4.2.6: ESSLLI

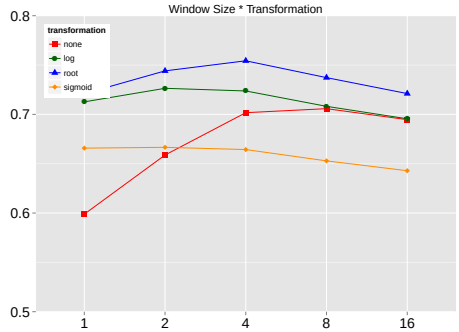

Figure 4.2.7: MITCHELL

4.3 Metric * Number of Latent Dimensions

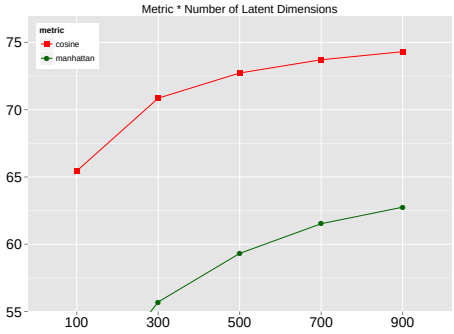

Figure 4.3.1: TOEFL

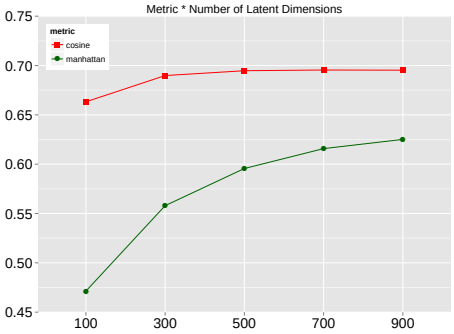

Figure 4.3.2: RG65

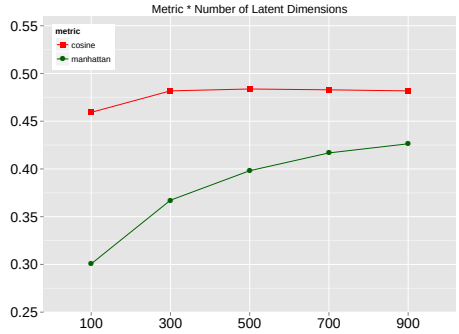

Figure 4.3.3: WS353

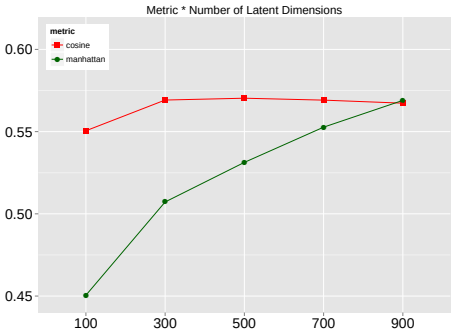

Figure 4.3.4: AP

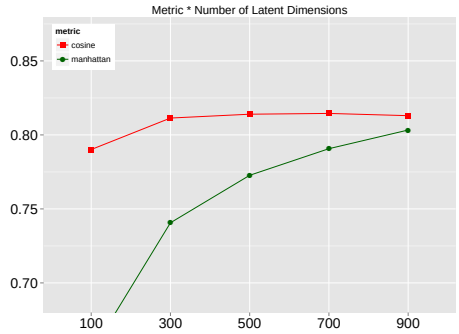

Figure 4.3.5: BATTIG

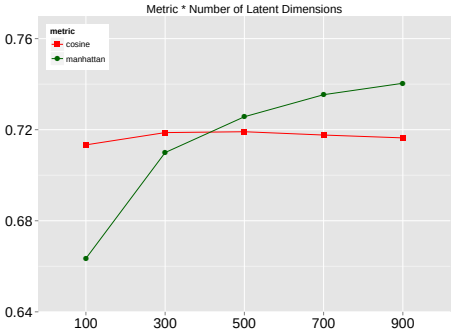

Figure 4.3.6: ESSLLI

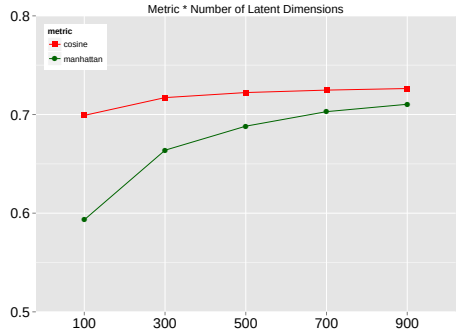

Figure 4.3.7: MITCHELL

4.4 Metric * Number of Skipped Dimensions

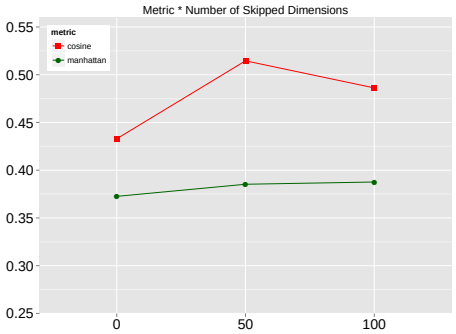

Figure 4.4.1: WS353

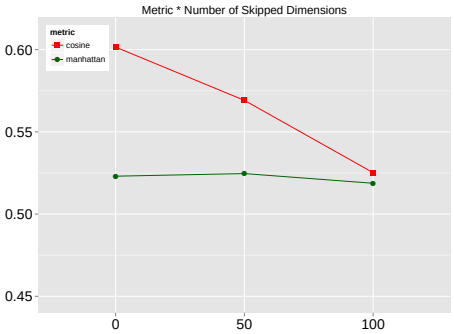

Figure 4.4.2: AP

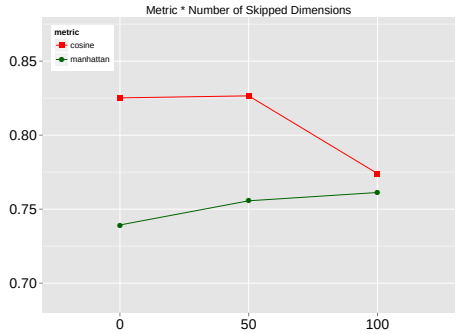

Figure 4.4.3: BATTIG

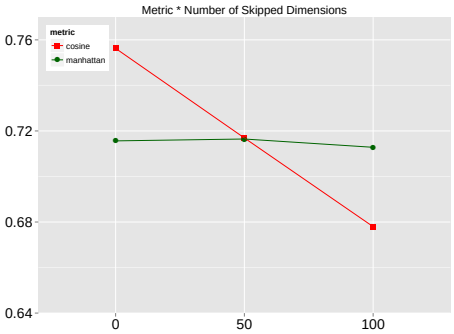

Figure 4.4.4: ESSLLI

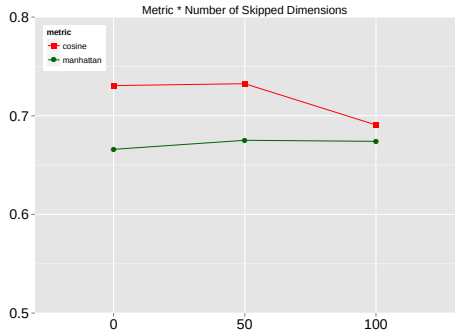

Figure 4.4.5: MITCHELL

4.5 Metric * Number of Context Dimensions

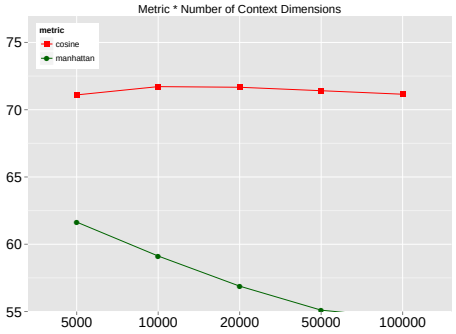

Figure 4.5.1: TOEFL

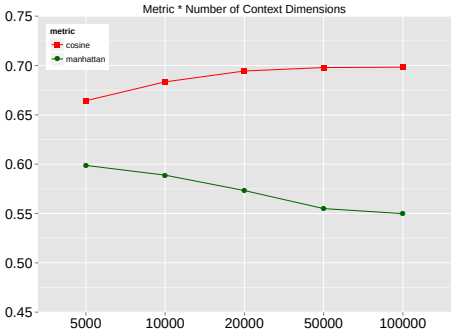

Figure 4.5.2: RG65

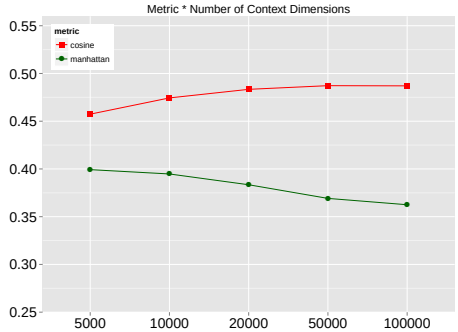

Figure 4.5.3: WS353

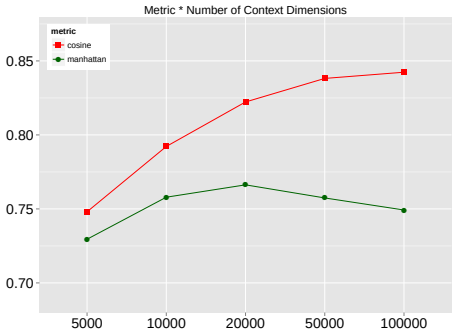

Figure 4.5.4: BATTIG

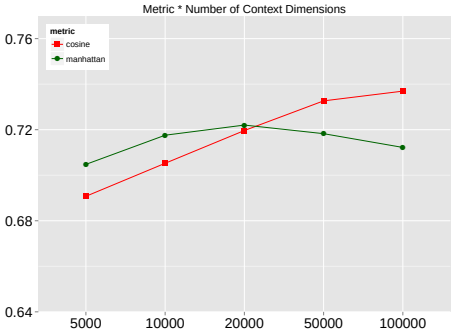

Figure 4.5.5: ESSLLI

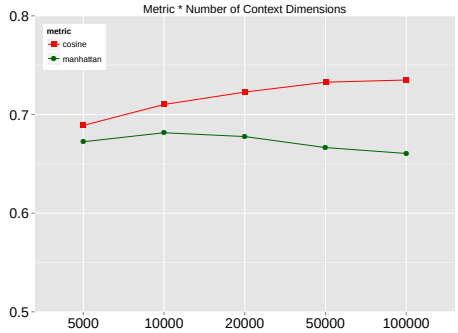

Figure 4.5.6: MITCHELL

4.6 Corpus * Metric

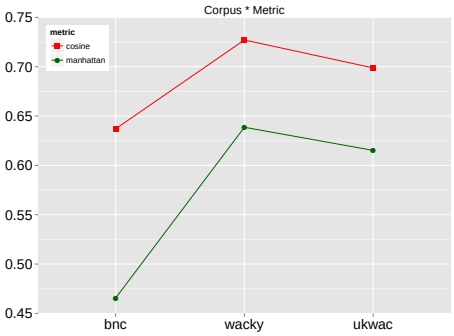

Figure 4.6.1: RG65

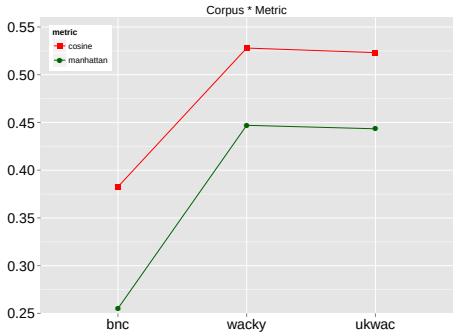

Figure 4.6.2: WS353

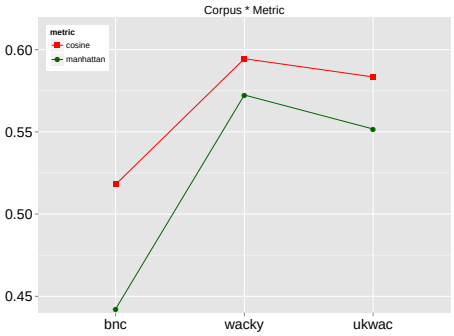

Figure 4.6.3: AP

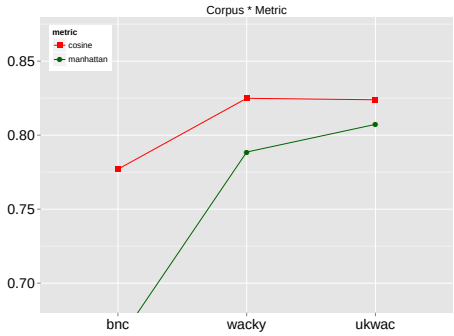

Figure 4.6.4: BATTIG

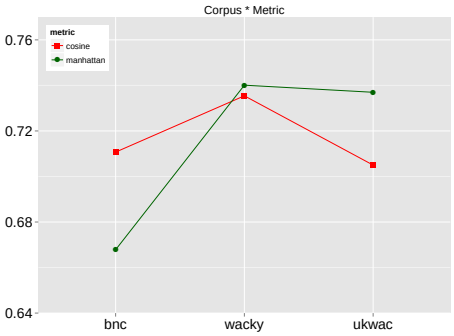

Figure 4.6.5: ESSLLI

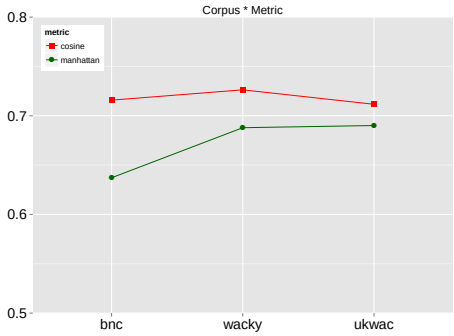

Figure 4.6.6: MITCHELL

5 Clustering algorithms and packages: PAM vs CLUTO

This appendix summarizes the results of the comparison between the clustering performance achieved by partitioning around medoids using the `pam` function from the R package `cluster` (with standard settings), and the performance achieved for the same vectors by the CLUTO toolkit with standard settings (Karypis, 2003). Table 5.1 and 5.2 report the results achieved with the *cosine distance metric* for the unreduced and reduced experimental runs. We conduct a paired t-test to check for significant differences between `pam` and CLUTO in our clustering experiments: table 5.1 and 5.2 report, for every comparison, the difference of means (`pam` minus CLUTO) and the significance value. The cases in which `pam` turned out to be better are highlighted in bold.

Dataset	Distance		Rank	
	diff.means	p	diff.means	p
AP	**0.0105**	***	**0.0459**	***
Battig	-0.0001		**0.0558**	***
Esslli	**0.0105**	***	**0.0459**	***
Mitchell	-0.0168	***	**0.0728**	***

Table 5.1: `pam` vs CLUTO: unreduced runs

Dataset	Distance		Rank	
	diff.means	p	diff.means	p
AP	**0.02814**	***	**0.03918**	***
Battig	**0.02310**	***	**0.03979**	***
Esslli	-0.00740	***	**0.00066**	***
Mitchell	**0.03999**	***	**0.05249**	***

Table 5.2: `pam` vs CLUTO: reduced runs

Even if CLUTO is the standard tool used in DSM evaluation on clustering tasks, we believe that the reassuring performance achieved by `pam` and its superiority in combination with neighbor *rank* are compelling arguments in favour of `pam`. Unlike CLUTO's standard algorithm, this choice allows a systematic evaluation of all DSM parameters, including different *metrics* and *relatedness indexes*.

The following plots display the distribution of the difference in clustering purity between `pam` and CLUTO, all clustering datasets. We report results for the *cosine* metric, comparing *distance* to *neighbor rank*, for both reduced and unreduced runs.

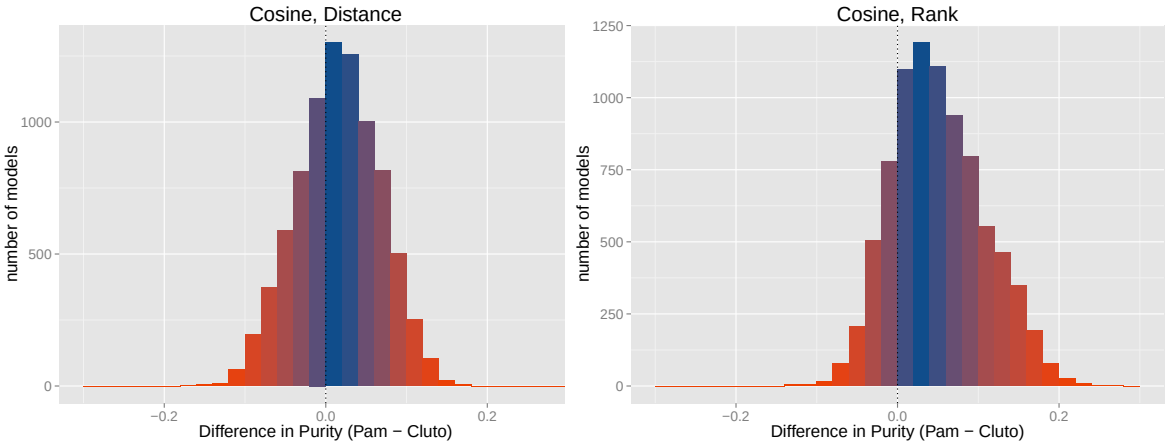

Figure 5.1: PAM vs CLUTO: Almuhareb-Poesio dataset, unreduced runs

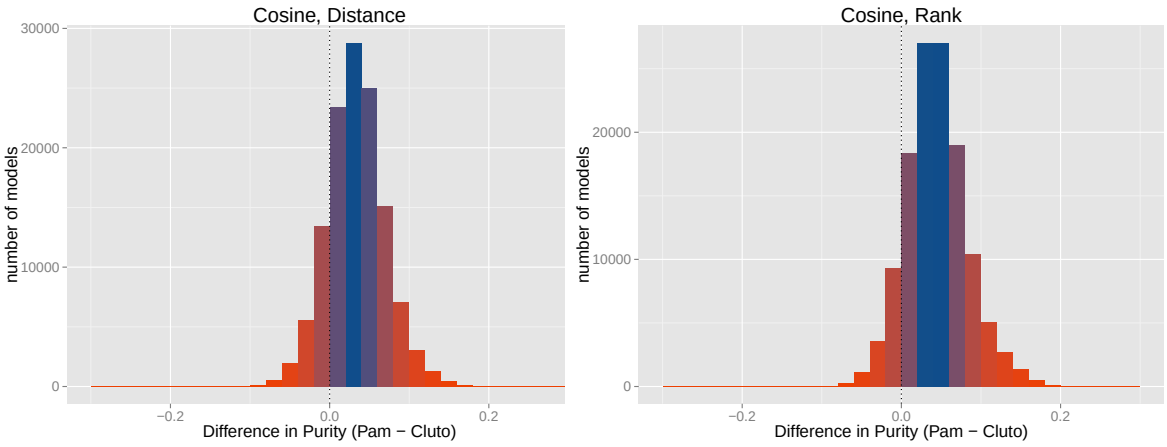

Figure 5.2: PAM vs CLUTO: Almuhareb-Poesio dataset, reduced runs

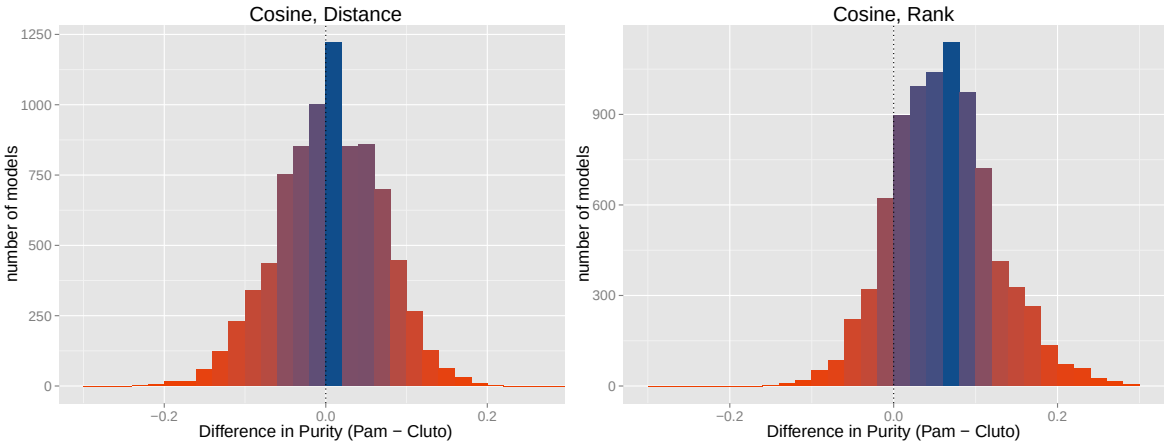

Figure 5.3: PAM vs CLUTO: Battig dataset, unreduced runs

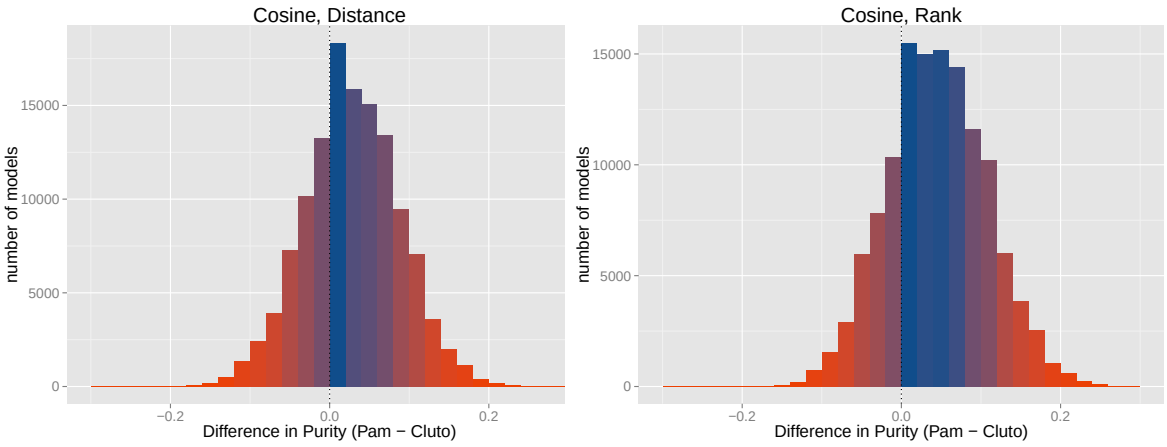

Figure 5.4: PAM vs CLUTO: Battig dataset, reduced runs

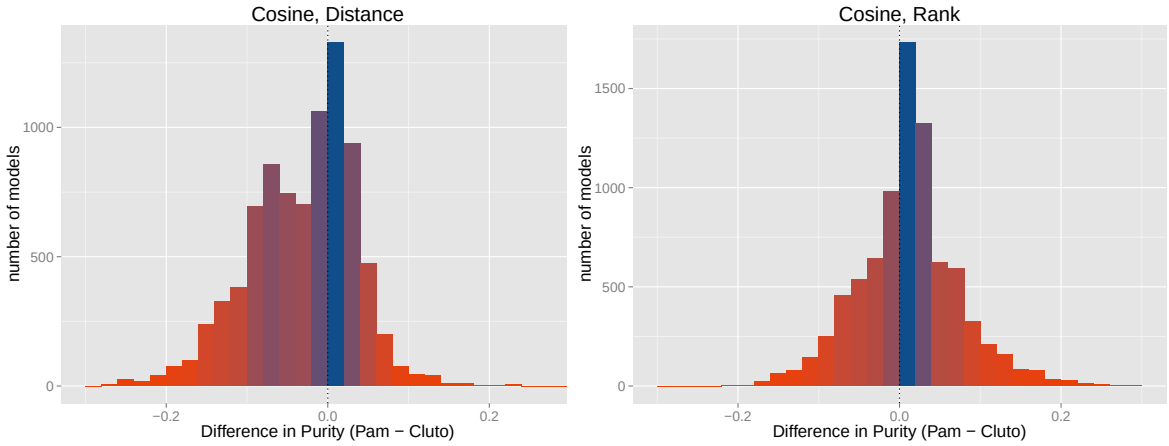

Figure 5.5: PAM vs CLUTO: ESSLLI dataset, unreduced runs

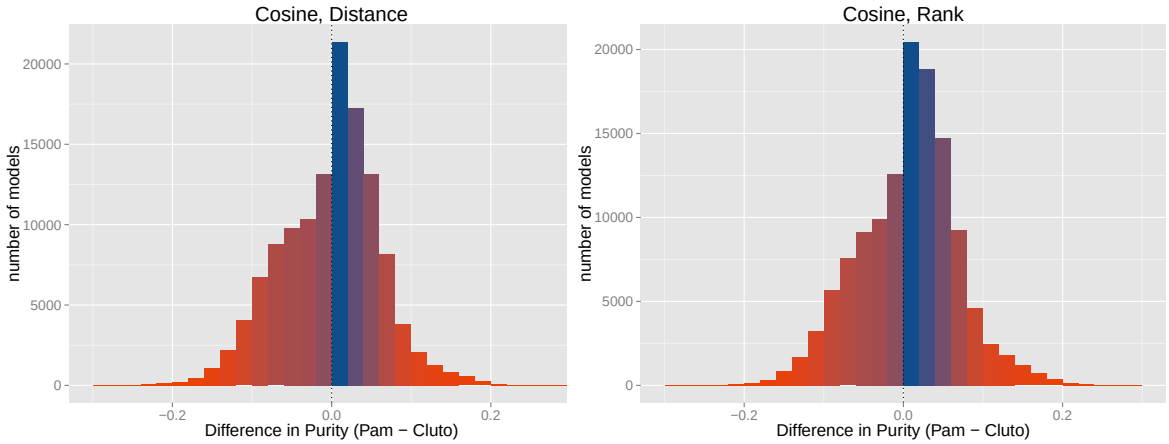

Figure 5.6: PAM vs CLUTO: ESSLLI dataset, reduced runs

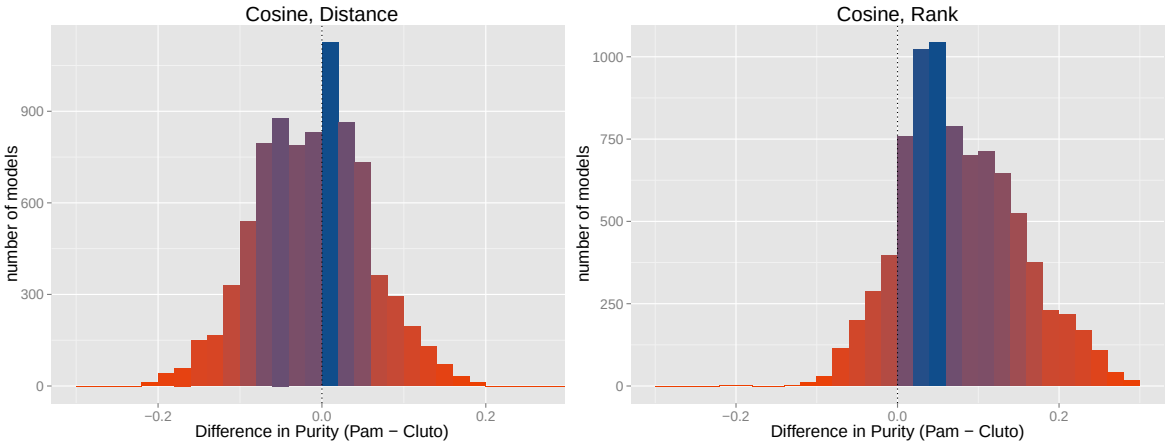

Figure 5.7: PAM vs CLUTO: Mitchell dataset, unreduced runs

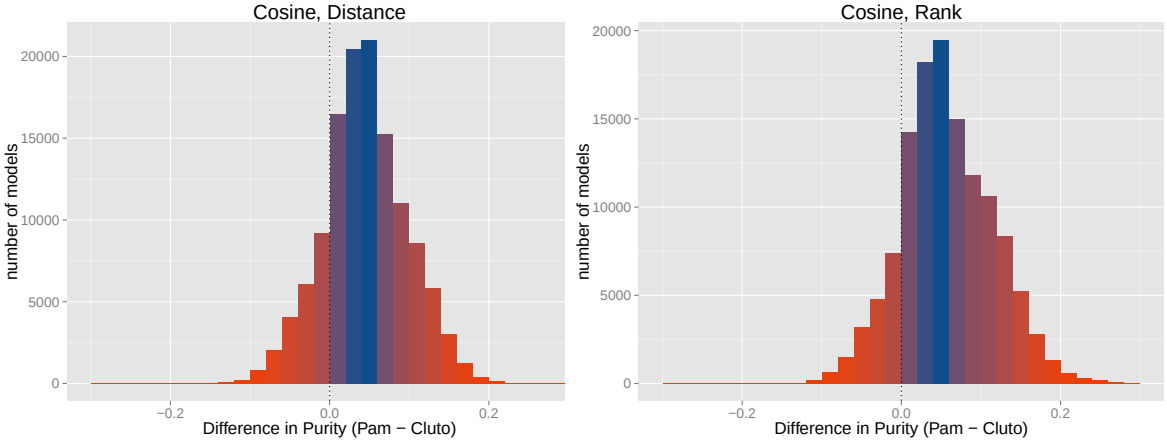

Figure 5.8: PAM vs CLUTO: Mitchell dataset, reduced runs