Zhihong Li Project 5 - Readme

PS: I used vector_model_5_10.txt

Which similarity metric seemed to work better? Why do you think that is?

Output00\eval.txt: Total Accuracy: 15.3574% (5595/36432)

Output01\eval.txt: Total Accuracy: 15.3574% (5595/36432)

Output02\eval.txt: Total Accuracy: 15.0637% (5488/36432)

Output10\eval.txt: Total Accuracy: 15.3574% (5595/36432)

Output11\eval.txt: Total Accuracy: 15.3574% (5595/36432)

Output12\eval.txt: Total Accuracy: 15.0637% (5488/36432)

Overall, Euclidean distance and Manhattan Distance works a bit better than Cosine distance. I think because cosine distance is a bit more harder to distinguish than the other two. Cosine distance is based on the angle between two vectors.

What input files did better than others?

Gram6-nationality-adjective.txt has a accuracy of 64.22764227642277% (1027/1599). It did better than others

Did normalization help? In what cases?

Normalization did not really help, there is no big difference.

If you used a different vector model (more below) what did you notice about your Results?

I did not use a different vector model.