Random Manhattan Indexing
An Incremental Word Space Model

Behrang Q. Zadeh*†*†

*University of Passau, Lower Bavaria, Germany
†National University of Ireland, Galway, Ireland

This paper introduces the Random Manhattan Indexing technique. Random Manhattan Indexing is an incremental—thus efficient and scalable—word space model that computes semantic similarities using the City-Block metric (Zadeh & Handschuh, 2014a,b).

Motivated by Harris's distributional hypothesis, distributional semantic models decipher meanings of linguistic entities, such as words and phrases, from their usages in large corpora. In distributional semantic models, vector spaces are a dominant tool for bridging the gap between distributional statistics and meanings. The collected distributional statistics are perceived in a high-dimensional vector space. In this vector space, a notion of distance—such as the cosine of the angles between the vectors, Euclidean distance or the City-Block metric—is employed to compute semantic similarities and explain meanings. This quantitative approach to model meanings is often known as the word space methodology.

A major barrier to the scalability of word space methods is the high dimensionality of vector spaces. Often, for example, due to the Zipfian distribution of words in documents, adding a new entity to a word space model results in a rapid increase in the dimensionality of the model. This phenomenon deteriorates the performance of word space models. To alleviate the problem, incremental methods, such as Random Indexing, have previously been proposed (Sahlgren, 2005). Although these incremental word space models have been successfully employed in many of applications, they can only be used when semantic similarities are computed using the cosine similarity and Euclidean distance. We tackle this limitation by introducing the Random Manhattan Indexing technique. In contrast to the Random Indexing technique and its variations, which employ very sparse Gaussian random projections, the proposed Random Manhattan Indexing method exploits Cauchy random projections. We delineate the introduced technique and show its ability in an experiment.

References

Sahlgren, Magnus. 2005. An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering (TKE)*, vol. 5, Copenhagen.

Zadeh, Behrang Q. & Siegfried Handschuh. 2014a. Random manhattan indexing. In *Proceedings of 25th international workshop on database and expert systems applications (DEXA)*, 203–208. Munich, Germany: IEEE Computer Society.

Zadeh, Behrang Q. & Siegfried Handschuh. 2014b. Random manhattan integer indexing: Incremental l1 normed vector space construction. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, Doha, Qatar: Association for Computational Linguistics.