

Download a set of pretrained word embeddings (e.g. word2vec, GloVe in English) and optionally embeddings for another language that you are fluent in.

i) Use these embeddings with a cosine distance to solve the word analogy task for the

four word lists: [list1.txt](#) [\\_Download list1.txt](#), [list2.txt](#) [\\_Download](#)

[list2.txt](#), [list3.txt](#) [\\_Download list3.txt](#), [list4.txt](#) [\\_Download list4.txt](#). Specifically, for the "a b c d" analogy (e.g. boy girl brother sister), look up the vector embeddings for

words a, b, and c, compute an estimated embedding for word d, , and find the 3

words in the vocabulary with an embedding that is closest to according to the cosine distance. Assess the performance by computing the top 1 and top 3 accuracy rates (the relative frequency that the correct word d is the closest word and is among the top 3, respectively). Report results separately for each list, since the different lists correspond to different types of analogies.

ii) Create your own test set of 40 examples with some other type of analogy or a word similarity task (e.g. use a thesaurus to provide ground truth) and assess how well the vector space model works for your task.

Turn in a 1 page summary of your work, including a brief description of your method (what embeddings you used, equations for the distance, type of analogy or similarity task that you designed), the experiment results, and your observations about what the vector space model is or is not good for. Turn in your 40 sample text set in a separate txt file.