Вывод даётся в формате

подвал : [('угол', 660), ('половина', 649), ('смерть', 646), ('сын', 644), ('лагерь', 618)]

Где подвал =yi

Угол, половина = yj, yk

А числа это количество общих x-ов у каждое пары то есть чем больше тем вроде бы должны быть «ближе» понятия но так происходит не всегда, разное бывает

Получается слишком расплывчато, наверное всё же из за небольшого датасета.

Ну вот например

processed word: андрей : [('иван', 1593), ('сын', 1548), ('полковник', 1428), ('лейтенант', 1323), ('сергей', 1318)]

Казалось бы корректные ассоциации одного имени с другим. Но потом почему то сюда присоединяется другой смысл в виде полковник и лейтенант - в датасете всё это де факто означает человек, поэтому сравнивая наборы y=(x1,x2…) когда y примерно означает человек «не даёт» корректно распознать сущности дальше, например фамильное родство, личное имя или звание. Чтобы это различить наверное нужен намного больший датасет, и то не факт что там будет.

Подумаю о более интеллектуальных методах анализа

Впрочим иногда выпадают прикольные ассоциации

Вроде

processed word: алкаш : [('матрос', 46), ('сын', 46), ('андрей', 44), ('дракон', 44), ('иван', 44)]

Правда представляю почему тут дракон