Cosine similarity of e_{cat} and e_{dog} is closest to 1 (0.92)

but cos similarity of e_{cat} and $e_{antarctica}$ is 0.2

↳ embeddings put similar words near each other.

↳ in word embedding space, direction is meaningful.

---

analogy: cos similarity of

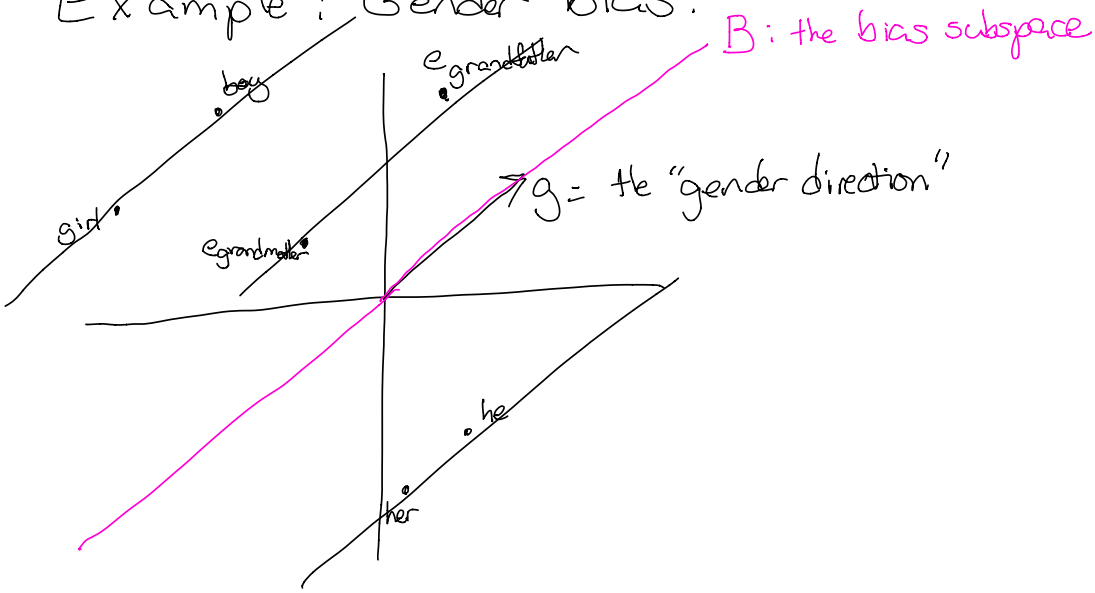
$(e_{paris} - e_{france})$ and $(e_{rome} - e_{italy})$ is 0.675

less cute analogy: cos similarity of

$(e_{man} - e_{doctor})$ and $(e_{woman} - e_{nurse})$ is 0.683

A model trained on sexist data will be sexist.

Example: Gender bias:



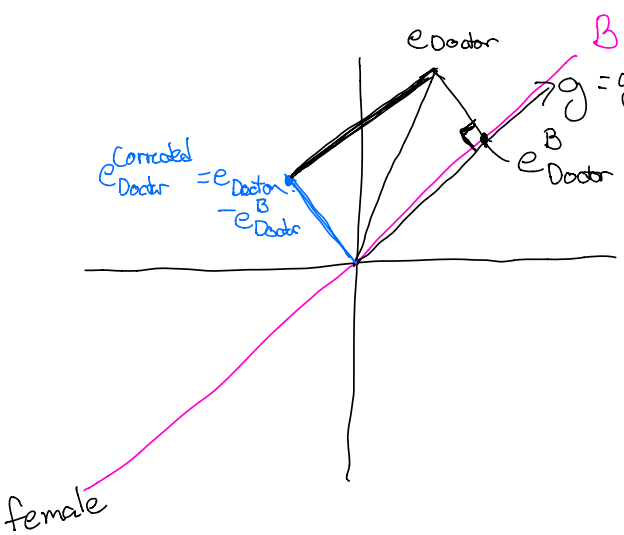
Develop 2 classes of words

1) "Neutral words" words you think should not be gendered. Examples: doctor, nurse, babysit.

2) "Equal words" word pairs that are gendered, but should have same association with neutral words. Examples: grandmother & grandfather.

Procedure (2 steps):

Step 1: "Neutralize": Remove gender association of all neutral words.

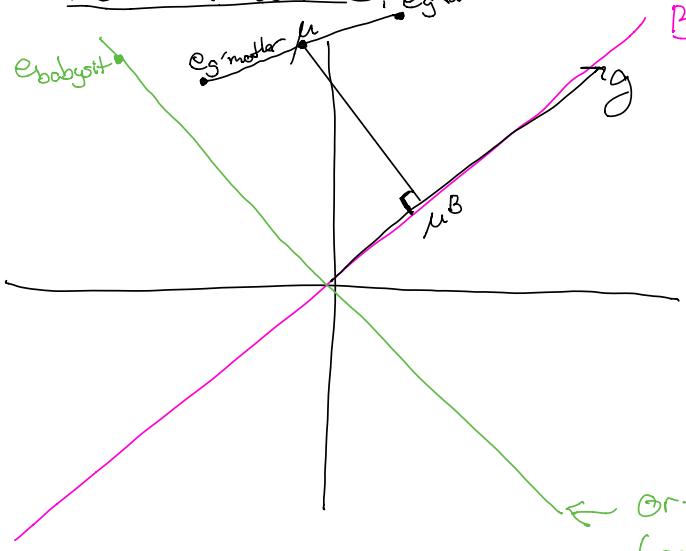


$$e^{corrected}_{Doctor} = e_{Doctor} - e^{B}_{Doctor}$$

$$= e_{Doctor} - \frac{gg^T}{g^Tg} \cdot e_{Doctor}$$

$$= \left(I_d - \frac{gg^T}{g^Tg}\right) \cdot e_{Doctor}$$

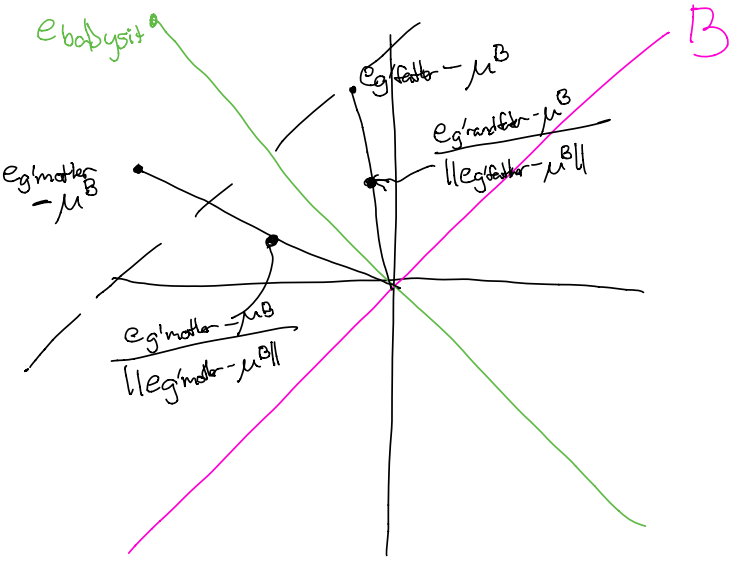
Step 2: "Equalize": Ensure that a pair of gendered words (like grandfather and grandmother) are <u>equidistant</u> <u>from all neutral words</u>.



B = bias subspace (dimension 1) a line

← orthogonal complement to B (perpendicular to B) (dimension 49)

First part of equalize step:
- shift $e_{g'mother}$ and $e_{g'father}$ down



Second step scale to unit vectors.