Tensorboard visualisaties

Tensorboard heeft een functie waar je embeddings en corresponderende woorden kan uploaden om ze op een interactieve manier te kunnen laten zien in de vectorruimte.

**Gender Specific Words before debiasing:**

<https://projector.tensorflow.org/?config=https://gist.githubusercontent.com/lysa-n/645e4b3b46530d69a6d6ffd7538c8d63/raw/5b84d80b354851689b7bc62dc99dabab8ad00779/genderSpecificConfig.json>

* Hier zien we goed dat de gender specifieke woorden ook daadwerkelijk een gender bias hebben
* Je kan linksboven door ‘color by’ = gender kiezen om de woorden een kleur te geven. Blauw is dan mannelijke biased, rood vrouwelijk
* Klikken op een punt in de space laat alle woorden zien
* Rechtsboven bij search GENDER DIRECTION intypen geeft je de gender direction vectoren.

**Personality Traits before debiasing:**

<https://projector.tensorflow.org/?config=https://gist.githubusercontent.com/lysa-n/0a4f104b5eac7f55b7f369817955dd8b/raw/11da95243ea172f276ccc45862dde6d176cd264a/config.json>

* Hier maakt de kleur van de woorden niet uit. De kleur geeft namelijk de gender aan van de woorden die volgens het paper in die kleur waren. In het nederlands zijn de woorden anders gender biased dus kleur zal niet zo veel zeggen.
* Wel zou je rechtsboven bij search weer GENDER DIRECTION in kunnen typen om de MALE en FEMALE direction te lokaliseren. Als je vervolgens op een bolletje in de ruimte klikt komen de woorden tevoorschijn. Dan kan je duidelijk zien welke woorden in het nederlands meer female of male biased zijn.

**Personality Traits after debiasing:**

<https://projector.tensorflow.org/?config=https://gist.githubusercontent.com/lysa-n/897b8c4733f26b355487efb602a8e4fe/raw/6f6fd7017cbed422079ef7f5fe6b711c5ed337db/personality_traits_After_debiasing.json>

* In deze visualisatie hoeven we niet te filteren op gender. Personality traits in het nederlands hadden sowieso al niet dezelfde bias als de engels dus de gender labels hebben geen betekenis.
* We kunnen deze embedding vergelijken met de personality traits before debiasing om te zien dat de embedding anders is geworden
* In custom kan je laten zien dat de personality traits gemapt zijn naar 0 op de gender as door voor **left=GENDER DIRECTION MAN** en **right=GENDER DIRECTION** WOMAN en voor **up=GENDER DIRECTION MAN** en **down=GENDER DIRECTION MAN** in te vullen. Maar hier hebben we ook een afbeelding van voor op de poster