The project has the purpose of introducing and testing word embeddings, examining their potential, and how biases may affect the result.

After running many cases, I will talk about 2 that I think are the most useful – first, names\_male and names\_female with career/family attributes, and second, names\_europe and names\_africa with pleasant/unpleasant.

**Case 1:** Running names\_male and names\_female with career/family attributes with Twitter showed that Male names demonstrate a bias towards career words and female names are neutral. Difference scores confirm that – ranging (0.2; -0.2) for male names mean is 0.1 = career, and female – mean is 0 – neutral. Similarities values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| target\attributes | Family | | Career | |
| Male names | range: 0.1 – 0.2 | mean: 0.15 | range: 0.15 – 0.3 | mean: 0.24 |
| Female names | range: 0.1 – 0.2 | mean: 0.15 | range: 0.05 – 0.24 | mean: 0.1 |

When running with Wikipedia there are changes related to female names results. Similarities between female names and careers are negative so it means there is none and in difference plots, it shows a strong bias towards family-related words.

**Case 2:** Running names\_europe and names\_africa with pleasant/unpleasant with Twitter showed that European names demonstrate a bias towards pleasant words and African names, little bias towards pleasant or we can say that they are neutral. Difference scores confirm that – ranging (0.2; -0.2) for European names mean is 0.04 = pleasant, ranging from 0.15 to 0, and African - mean is 0.01 – neutral/slightly towards pleasant because of the range that is (0.1; -0.01) . Similarities values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| target\attributes | Pleasant | | Unpleasant | |
| European names | range: 0.05 – 0.35 | mean: 0.15 | range: 0.0 – 0.2 | mean: 0.07 |
| African names | range: 0.05 – 0.15 | mean: 0.1 | range: -0.03 – 0.26 | mean: 0.07 |

The data confirms that Europen names have string association with pleasant while African has a very weak one. Whe it comes to similarity to unpleasant words, based on the range of African names, we can says that there is a little similarity but nothing major. After rerunning the test with Wikipedia, there are changes that confuse our previous result and I can say that as Wikipedia has less information – 6B compared to 27B in Twitter, this scores are not as good. They are the following: the similarity plot shows no similarity between any of our targets or attributes – all the mean values are 0 or negative. When it comes to the results in difference scores plot – both types of names shows negative scores, meaning better association with unpleasant words, African names are more towards the neutral position but still negative.

The individual experiment that I conducted had me create a new WEAT paring, generate its data, and test it with the existing attribute pair. My target pair is Traditional Values vs Modern Values and the attribute pair that will be used is Family/ Career. The pairs’ data is the following:

* **traditional\_values:** Heritage, Ancestry, Family, Ceremony, Community, Stability, Conservatism, Modesty, Religion, Duty, Honor, Patriotism, Marriage, Craftsmanship, Tradition
* **modern\_values:** Innovation, Individualism, Technology, Progress, Independence, Equality, Entrepreneurship, Flexibility, Secularism, Democracy, Urbanization, Digital, Expression, Minimalism
* **family:** Home, Parenting, Child, Nurture, Relationship, Love, Marriage, Domestic, Caregiving, Spouse, Sibling, Generations, Kin, Household, Support, Togetherness, Dependents, Community
* **career:** Job, Office, Promotion, Work, Ambition, Business, Employment, Success, Corporate, Profession, Achievement, Salary, CEO, Goals, Opportunity, Workplace, Manager

In the world nowadays, the culture and values have changed. The traditional values used to be family first and the monetary world currently live in it changed and has created modern values, that put work and career first place. My hypothesis is that if something is related to traditional values, it will be related and similar to family, and if something is related to modern values, it will be related to work and career. I wrote down a list of words that are related to both types of values and used AI – Chat GPT to expand this list so I could have more data. The table provides similarity scores for the target/attribute pair.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| target\attributes | Family | | Career | |
| traditional | range: 0.15 – 0.4 | mean: 0.24 | range: 0.05 – 0.28 | mean: 0.21 |
| modern | range: 0.05 – 0.43 | mean: 0.26 | range: 0.05 – 0.44 | mean: 0.3 |

Based on the generated plot again, the table below shows the difference scores for each score. The data means the following: bigger than 0 – similar to family-related words, lower than 0 – similar to career-career-related words, 0 – neutral.

|  |  |  |  |
| --- | --- | --- | --- |
| Traditional values | | Modern values | |
| range: (0.1; -0.1) | mean: 0.00 | range: (0.5; -1.7) | mean: -0.5 |
| Neutral – as similar to family as to career | | Similar to career – almost no similarities to family | |

The data shows traditional values show neutral bias, by demonstrating similar association with both family and career concepts. But on the other hand, modern values demonstrate a significant bias towards career and little or none towards family.