Universidad del Desarrollo

Documento Resumen (Cuaderno Jupyter)
Proyecto Final
Asignatura: Procesamiento de Lenguaje Natural

1. ¿Qué sesgo se han identificado en la literatura?

- El **sesgo de género** es el más reportado en embeddings (asociación de masculino con "carrera/profesión" y femenino con "familia/hogar").
- También se han descrito sesgos raciales, religiosos, políticos y de estatus socioeconómico, pero en este proyecto nos centramos en **género** porque:
    - Está bien documentado en inglés y otros idiomas (Caliskan et al., 2017; Garg et al., 2018).
    - Existen recursos previos en español que permiten hacer la prueba (SBW, fastText).

2. ¿Con qué corpus se trabajó?

Se emplearon **dos modelos pre-entrenados en español**, cada uno con su corpus específico:

- **fastText cc.es.300.vec**
    - Entrenado por Facebook AI sobre **Common Crawl + Wikipedia** en español
    - Vocabulario enorme (~2M palabras).
    - Captura un lenguaje más general y web-scale.

- **SBW (Spanish Billion Words)**
    - Entrenado por Cardellino (2016) sobre el **Spanish Billion Words Corpus** (~1.5B tokens, 1M vocabulario).
    - Más específico, corpus recopilado de noticias y textos web en español.

3. ¿Con qué métricas?

- Se utilizó el **Word Embedding Association Test (WEAT)**, propuesto por Caliskan et al. (2017).
- **Fórmulas incluidas en el informe**:
    - $s(w,A,B)$ (diferencia de similitudes coseno con atributos A y B).
    - **Effect size (d de Cohen)** → magnitud de la diferencia de asociaciones.
      Prueba de permutaciones (p-valor) → significancia estadística.

4. ¿Qué sesgo se escogió para testear?

- **Sesgo de género**:
    - Listas X (masculino) vs Y (femenino).
    - Listas A (carrera/profesión) vs B (familia/hogar).

5. ¿Qué modelo pre-entrenado en español se usó?

- **Principal:** fastText-es (cc.es.300.vec).
- **Comparación:** SBW (SBW-vectors-300-min5.txt).

Resumen

En este trabajo se evaluó el **sesgo de género** en *word embeddings* en español, específicamente la asociación entre términos masculinos con el ámbito laboral/profesional y términos femeninos con el ámbito familiar/doméstico. Para ello, se utilizarón dos modelos pre-entrenados en español: **fastText cc.es.300.vec** (entrenado sobre *Common Crawl* y *Wikipedia*) y **SBW-vectors-300-min5.txt** (entrenado sobre el *Spanish Billion Words Corpus*). La métrica seleccionada para detectar sesgos es el **Word Embedding Association Test (WEAT)**, que mide diferencias de similitud coseno entre conjuntos de palabras objetivos y atributos, y calcula tanto el **tamaño de efecto (d de Cohen)** como la **significancia mediante prueba de permutaciones**.