

Notas TFM

Estructura del GitHub

- **notas_tfm**: este archivo.
- **py**: archivos de python
 - **filter_unimorph** para extraer categorías específicas de archivos de UniMorph.
 - **main_triplets/biplets.py**: código principal. Triplets es para analizar archivos de tripletes, biplelets para archivos de parejas (base+inflection/verbo+flexión)
 - **datasets**: # *derinet debería estar en esta carpeta*
 - **spa**:
 - **spa.txt**: archivo de unimorph
 - **filtered**: solo los tiempos presente, pasado imperfecto y futuro.
 - **_small**: dataset pequeño para testear el código.
 - **50_triplets.csv**: archivo de tripletes que hice para el trabajo de clase de NLP.
 - **pol**:
 - **pol.txt**: archivo de unimorph
 - **filtered**
 - **derinet**: archivos de derinet (*TODAVÍA NO HE EMPEZADO A MIRARLOS*)
 - **embeddings**: modelos de vectores (W2V y FT) separados por idioma.
 - **results**: archivos csv con los resultados del código por fila.

Research question:

How do different word embedding models (Word2Vec, FastText, and BERT) capture the distinction between inflection and derivation in vector space representations?

Methodology

- Using word embedding models (static vs. contextual) to identify the distinction between inflection and derivation in Polish and Spanish.
 - Implemented:
 - Word2Vec (**Spanish**: Spanish Billion Words, **Polish**: WIP)
 - FastText (Spanish, Polish)
 - Multilingual BERT (WIP).
- Dataset from **UniMorph** of **Base/Inflection (2) V:V**.
 - Filtered **Spanish** UniMorph to create a Base/Inflection dataset of V:V in present, past imperfect and future.
 - **PRESENT** V;IND;PRS
 - **PAST IMPERFECT** V;IND;PST;IPFV
 - **FUTURE** V;IND;FUT
 - Result: 148051 rows (Base/Inflection/Category)
 - Filtered **Polish** UniMorph data:
 - **PRESENT** V;PRS
 - **PAST** V;PST
 - **FUTURE** V;FUT
 - Result: 23615 rows (Base/Inflection/Category)

Results

MEAN SIMILARITY biplets dataset (infinitive+inflection).

	WORD2VEC	FASTTEXT	MULTILINGUAL BERT
SPA	0.539277	0.555474	
POL		0.486620	

- Dataset from UniMorph/DeriNet/VeLeSpa Pivot/Inflection/Derivation (3). For derivations I need a derivations lexicon:
 -