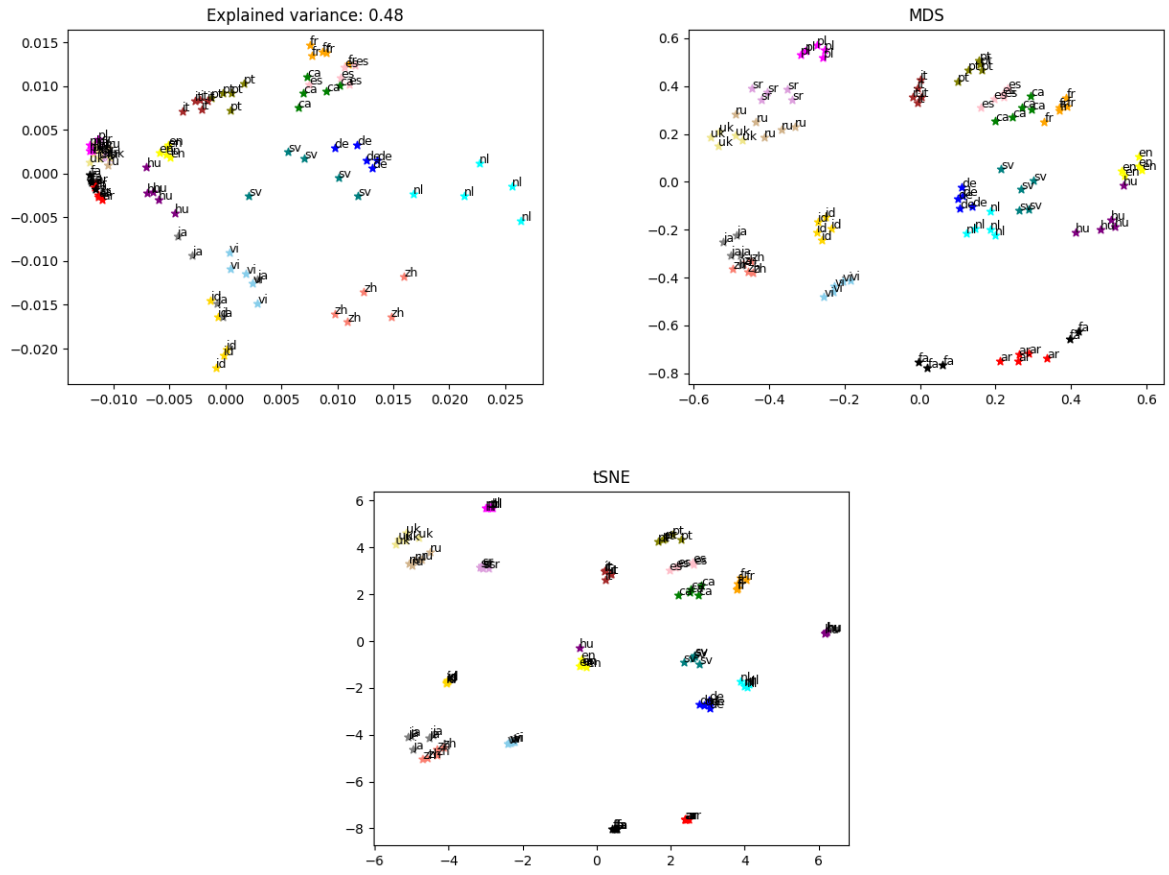
1 Rezultati projekcij



Slika 1: PCA, MDS, tSNE na 100 dokumentih

1.1 Komentar

Na prvem grafu(PCA) lahko razberemo tri osi. Prva predstavlja romanske jezike (francoski, španski, katalonski, portugalski, italijanski), druga os so germanski jeziki (švedski, nemški, nizozemski), zadnja tretja pa azijski jeziki (kitajski, vietnamski, indonezijski, japonski). Zanimivo mi je, da lahko prvi dve komponenti tako lepo razpenjanjo prostor na tri osi.

Pri MDS opazimo lepšo delitev. Na grafu razberemo kar 5 jezikovnih skupin:

- Slovanski (poljski, ukrajinski, srbski, ruski) - levo zgoraj
- Romanski - desno zgoraj
- Azijski - levo spodaj
- Germanski - sredina desno
- Arabski, perzijski jezik - spodaj

Pri tSNE vidimo še lepšo delitev jezikov. Smiselno mi je, da so zelo skupaj:

- ukrajinski, ruski
- japonski, kitajski
- portugalski, francoski, španski ter katalonski jezik

Vsi jeziki, ki so zelo skupaj pripadajo isti jezikovni skupini.

Mogoče kot zanimivost izpostavim članek v madžarskem jeziku(vijolična barva). Njegovo odstopanje opazimo na grafu MDS in tSNE, nahaja se blizu člankov v angleškem jeziku(rumena barva). Gre dejansko za članek, ki ima zelo veliko, skoraj večino citatov v angleščini, kar razloži odstopanje.