Input: Regional Spanish corpora

Each regional vocabulary V were computed from geolocated tweets to each region. We drop the top 100 most frequent terms and also drop terms with less than 10 occurrences.

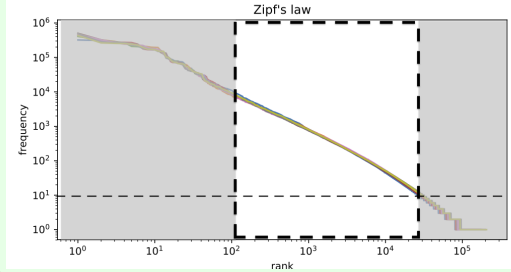
We define a global vocabulary L, of size ℓ, as the union of all vocabularies. We use L to convert each vocabulary into a high dimensional vector X. Each vector's coordinate corresponds with a term, and its value is the frequency of that term in the region corpus.

The distance between two region vocabularies X, Y is the cosine dissimilarity;

$$d_{\cos}(X,Y) = 1 - \frac{\sum_{i=1}^{\ell} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{\ell} X_i^2}\sqrt{\sum_{i=1}^{\ell} Y_i^2}}$$

Preprocessing and tokenization

- lower casing
- diacritic marks were removed
- group users, urls, and numbers
- normalize repetitions (2 max.)
- normalize blanks
- laughs were normalized to four letters
- words, punctuation, and emojis are tokens



Vocabularies

$$\begin{array}{c} \\ V_{\mathsf{AR}} = \\ V_{\mathsf{BO}} = \\ \vdots \end{array} \begin{array}{cccc} w_1 & w_2 & \cdots & w_\ell \\ f_1^{\mathsf{AR}} & f_2^{\mathsf{AR}} & \cdots & f_\ell^{\mathsf{AR}} \\ f_1^{\mathsf{BO}} & f_2^{\mathsf{BO}} & \cdots & f_\ell^{\mathsf{BO}} \\ \vdots & \vdots & & \end{array}$$

Output: Affinity matrix of regional vocabularies.