Notes on Natural Language Processing

Nils Kujath

v2026.01

From Sparse to Dense Word Representations

- Let $V = \{w_1, \ldots, w_{|V|}\}$ be a finite vocabulary. The canonical (standard) basis of $\mathbb{R}^{|V|}$ is the indexed set $\mathcal{B}_{\text{can}} = \{\mathbf{e}_1, \ldots, \mathbf{e}_{|V|}\}$, where $\mathbf{e}_i$ denotes the i-th basis vector, i.e. the unique (one-hot) vector with a 1 in coordinate i and zeros elsewhere. The BoW construction $\boxed{\iota : V \to \mathbb{R}^{|V|}, \quad w_i \mapsto \mathbf{e}_i}$ identifies each word $w_i \in V$ with the canonical basis vector $\mathbf{e}_i \in \mathbb{R}^{|V|}$.
- Under the standard inner product on $\mathbb{R}^{|V|}$, the BoW representation satisfies

$$\mathbf{e}_u{}^\top \mathbf{e}_v = \delta_{uv} = \begin{cases} 1 & \text{if } u = v, \\ 0 & \text{if } u \neq v \end{cases}, \quad \forall u, v \in \{1, \ldots, |V|\}.$$

  Consequently, all distinct word types are mutually orthogonal (for any $u \neq v$, $\mathbf{e}_u{}^\top \mathbf{e}_v = 0$) and equidistant in $\mathbb{R}^{|V|}$ (for any $u \neq v$, $\|\mathbf{e}_u - \mathbf{e}_v\| = \sqrt{(1)^2 + (-1)^2} = \sqrt{2}$). That is, the induced geometry encodes no graded notion of similarity between words in V.
- To obtain representations with a non-degenerate geometry, the BoW map ι can be replaced by a learned embedding $\boxed{\text{emb} : V \to \mathbb{R}^m, \quad w \mapsto \mathbf{v}_w, \quad m \ll |V|}$, where word vectors are dense and the embedding space offers the possibility of expressing similarity between words as geometric proximity.