Notes on Natural Language Processing

Nils Kujath

v2026.03

Degenerate Geometry of One-Hot 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 $\mathbb{R}^{|V|}$, all distinct word types are mutually orthogonal:

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

  Fmr., the metric restricted to $\iota(V)$ is trivial; all word types are equidistant in $\mathbb{R}^{|V|}$: $\forall u, v \in \{1, \ldots, |V|\}$,

$$\boxed{d\big|_{\iota(V)\times\iota(V)}(\mathbf{e}_u, \mathbf{e}_v) = \|\mathbf{e}_u - \mathbf{e}_v\| = \begin{cases} 0 & \text{if } u = v, \\ \sqrt{\underbrace{(1-0)^2}_{\text{pos. } u} + \underbrace{(0-1)^2}_{\text{pos. } v} + \underbrace{(0-0)^2 + \cdots + (0-0)^2}_{|V|-2 \text{ pos.}}} = \sqrt{2} & \text{if } u \neq v \end{cases}}.$$

- Consequently, the geometry induced by ι is degenerate: all distinct words are equally dissimilar, and no notion of graded semantic proximity can be expressed under the BoW representation.