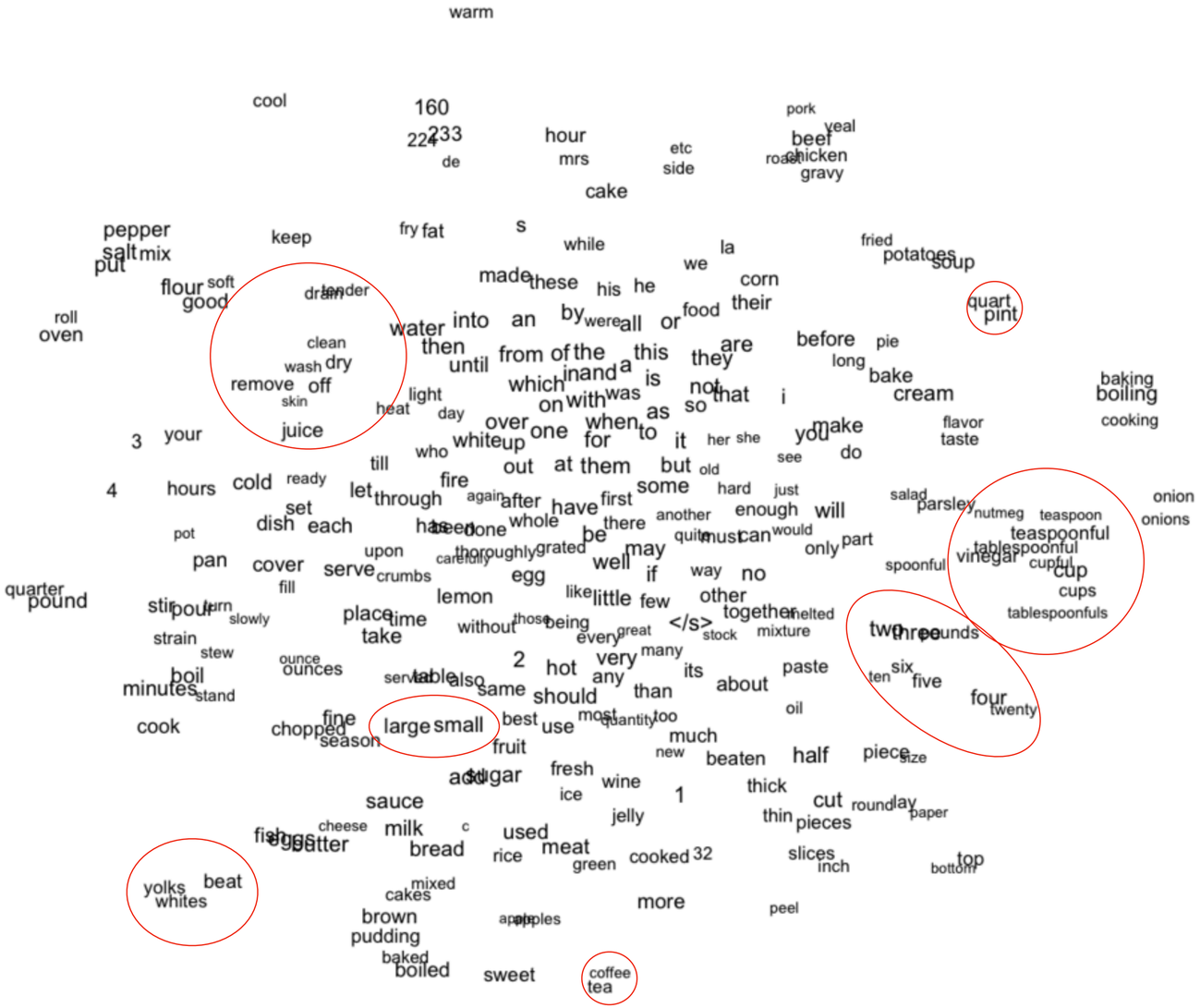
Word2Vec: learn a feature representation that preserves semantic relationships between words based on distance



Produced using Word2Vec word vector representations, compressed to 2D using tSNE

Algorithm for the skip gram word2vec model

let $\mathbf{U} \in \mathbb{R}^{K \times N}$ be our center word embeddings corresponding to $\mathbf{x}_w$

let $\mathbf{V} \in \mathbb{R}^{K \times N}$ be our center word embeddings corresponding to $\mathbf{x}_c$

$\forall (\mathbf{x}_w, \mathbf{x}_c) \in D$ do:

set $\mathbf{u}_w = \mathbf{U}_w$ via the non-zero indice of our one-hot center word $\mathbf{x}_w$

compute inner product between $\mathbf{u}_w$ and all context vectors $\mathbf{V}$: $\mathbf{u}_w \cdot \mathbf{V} \in \mathbb{R}^N$

encodes distributional hypothesis

compute probability over all context words given center word : $\frac{\exp(\mathbf{u}_w \cdot \mathbf{V})}{\sum_{j=1}^{N} \exp(\mathbf{u}_w \cdot \mathbf{V})_j}$

cross entropy loss: $\mathbf{L}(U, V \,|\, \mathbf{x}_w, \mathbf{x}_c) = -\,\mathbf{x}_c \cdot \log p(\mathbf{x}_c \,|\, \mathbf{x}_w \,; \mathbf{U}, \mathbf{V})$

gradients: $\nabla_{U_w} NLL = \mathbf{V} \cdot \left(P_{\mathbf{x}_c | \mathbf{x}_w} - \mathbf{x}_c\right)^T \in \mathbb{R}^K$

$\nabla_V NLL = \mathbf{u}_w \cdot \left(P_{\mathbf{x}_c | \mathbf{x}_w} - \mathbf{x}_c\right) \in \mathbb{R}^{K \times N}$

gradient descent: $\mathbf{U}_w \leftarrow \mathbf{U}_w - \eta \nabla_{\mathbf{U}_w} NLL$ only w^{th} row of $\mathbf{U}$ gets updated

$\mathbf{V} \leftarrow \mathbf{V} - \eta \nabla_{\mathbf{V}} NLL$ entire $\mathbf{V}$ gets updated