Programming for Social Scientists: Text Mining V

黃瀚萱

Agenda

- Word embeddings
- Train word embeddings
- Visualization of word embeddings

Word Representations

- One hot
  - Treat each word as an individual symbol
- Distributional word representation (word embeddings)
  - Represent each word in a dense vector space

Word Embeddings

- Word embeddings (or distributed word representations) are trained to predict well words that appear in its context.

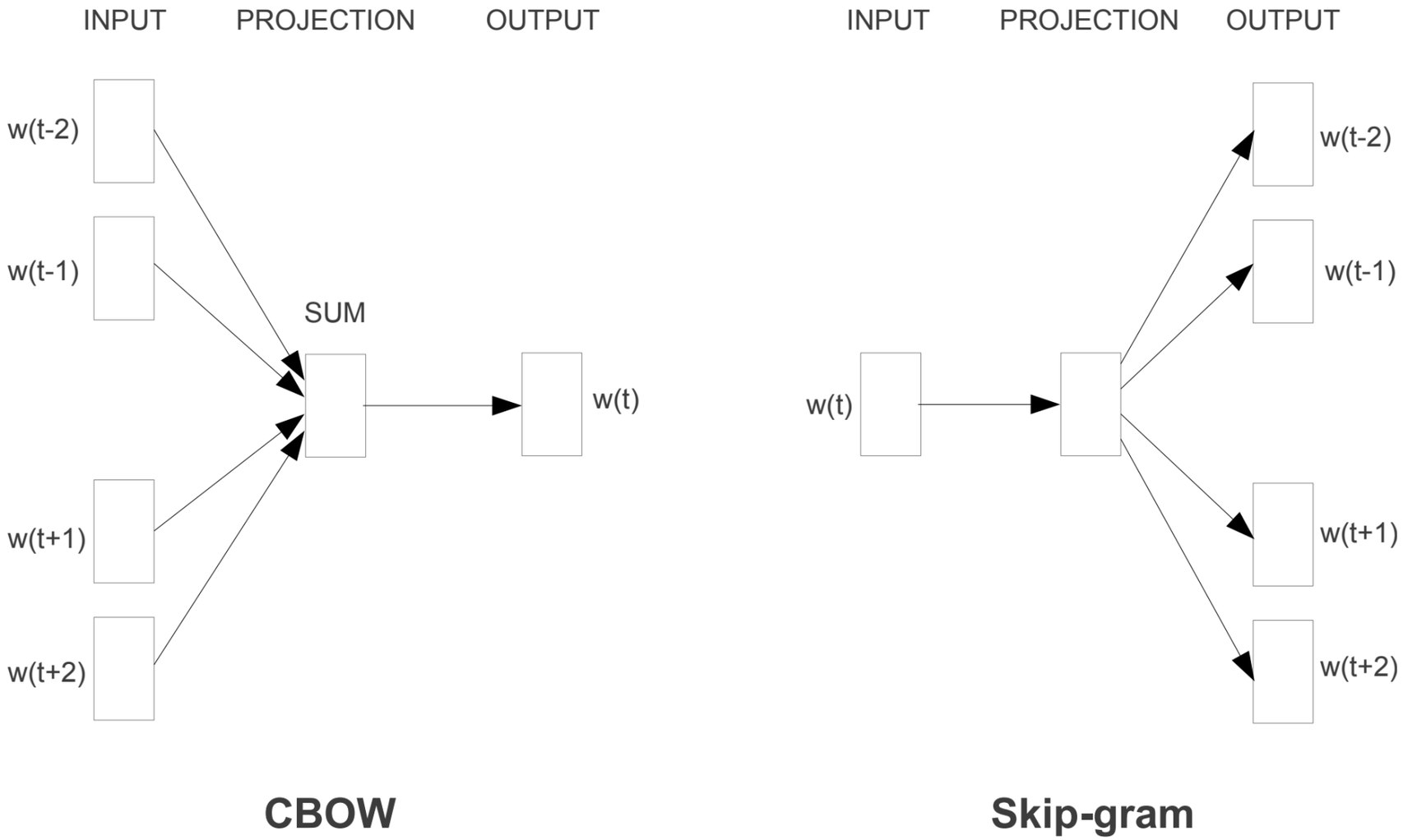
- Given a set of sentences w_1, ..., w_T, the objective of the skip-gram model is to maximize the log-likelihood:

$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c \mid w_t)$$

- With a scoring function s maps pairs of a target word and a contextual word to a real number.

$$p(w_c \mid w_t) = \frac{e^{s(w_t,\, w_c)}}{\sum_{j=1}^{W} e^{s(w_t,\, j)}}$$

CBOW vs Skip-gram



- CBOW predicts the current word based on the context.
- Skip-gram predicts surrounding words given the current word.

Properties of Word Embeddings

- Each word is represented in a vector with a dimension in between 50 and 1000 in a dense space.

- Similarity
    - Similar or related words are close in the vector space.

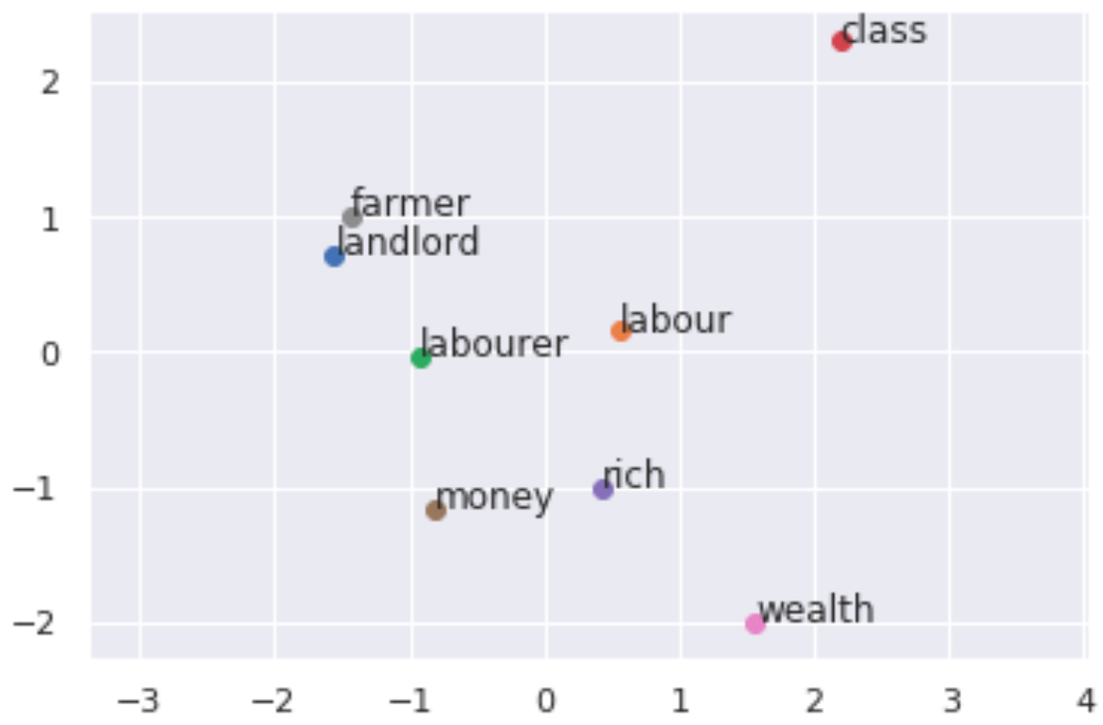
- Regularity
    - Rome : Italy = Paris : ?
    - vec(Rome) - vec(Italy) + vec(France) ~ vec(Paris)

Facebook FastText

- Consider the subword information.
- Taking the n-grams character sequences of a word as additional contextual information.
- The word “algorithm” with n=3
  - al, alg, lgo, gor, ori, rit, ith, thm, hm
- Especially useful for Chinese
  - Each *hanzi* is a subword

Visualization

- Project each word in the word embedding space to a low (2 or 3, typically) dimensional space.
    - Principal component analysis
- Scatter plot shows the relations of the words.



Assignment

- Train a word embedding with the full text of *communist manifesto*
    - /text_mining_2/corpus.txt
- Generating a 2D scatter plot for visualization
    - Show the top 20 *content words* in the text