Mineração de Texto

Word embeddings

Prof. Walmes Marques Zeviani

JUSTIÇA 4.0: INOVAÇÃO E EFETIVIDADE NA REALIZAÇÃO DA JUSTIÇA PARA TODOS
PROJETO DE EXECUÇÃO NACIONAL BRA/20/015

Justificativa e objetivos

- Até aqui, o espaço vetorial foi criado a partir de frequências.
- Não foi considerado o contexto.
- Word embeddings são úteis:
  - Um vetor denso é criado para cada token.
  - O contexto é explorado para criação da representação.
  - Essa representação é mais interessante por contemplar contexto.

Abordagens anteriores

A abordagem bag of words

- Separação do documento em tokens.
- Etapas de processamento visando reduzir variantes de escrita e tokens de pouca contribuição.
- Tipos de ponderação:
  - **Frequência** do termo.
  - **Ocorrência** do termo.
  - **TF-IDF**.
- Não leva o contexto de um termo em consideração.
- Como saber que "quarto" e "dormitório" são sinônimos?
- A matriz de documentos e termos é bastante esparsa.

A abordagem bag of words

- Se "quarto" e "dormitório" são sinônimos, trocar um token apenas.
- Mas e "bom" e "proeficiente", são realmente sinônimos?
- E antônimos? Usar -1 na matriz de documentos e termos?
- Porém, vetores ainda são interessantes.
  - Similaridade entre termos: distância do coseno.
  - Detecção de sinônimos/antônimos.

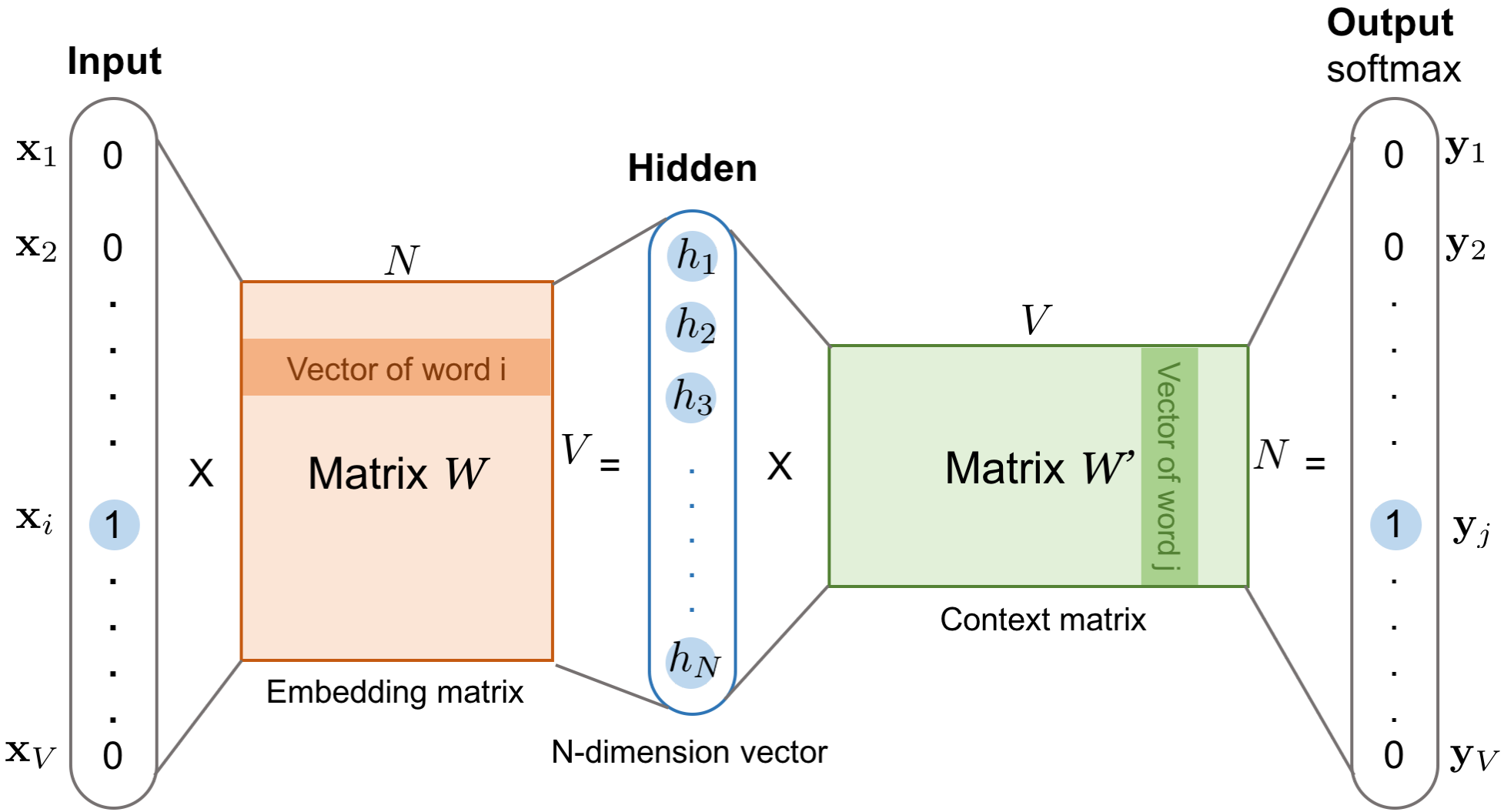
Modelos de linguagem

- Sequência de palavras.
- O significado de uma palavra pode ser inferido pelas palavras vizinhas.
- "You shall know a word by the company it keeps" (J. R. Firth)
- Dessa forma, podemos usar os vários contextos em que uma palavra ocorre para construir uma representação dela.
- Diferente da representação BOW, esses vetores serão densos.

word2vec

word2vec

- A ideia base de métodos word2vec é treinar uma rede neural rasa (uma camada oculta) para uma tarefa de classificação.
- O truque é não usar o classificador obtido, mas simplesmente pegar os pesos das camadas de entrada e oculta que serão os vetores word embedding.
- O objetivo é ter um vetor para cada termo do vocabulário.
- Uma janela móvel se desloca e a cada posição há um termo central c a os termos vizinhos o.
- A similaridade entre os vetores para c e o podem ser usados para calcular probabilidades de sentenças.
- Para aprender, o modelo otimiza os componentes dos vetores para maximizar a probabilidade.



Modelo de um skip-grama. Fonte: https://lilianweng.github.io/posts/2017-10-15-word-embedding/.

A lógica de porque funciona

- Palavras de contexto o coerente com o termo central c terão alta probabilidade.
- Já palavras de contextos exógenos terão baixa probabilidade.
- Exemplos:
  - Os astronautas viajam para Marte.
  - Os astronautas venceram as olimpíadas.
- Palavras de contexto similares serão vetores próximos.
  - Márcia levou o gato ao veterinário.
  - Márcia levou o cachorro ao veterinário.
- O mesmo vale para sinônimos.
  - Apartamento com três quartos e dois banheiros.
  - Apartamento com três dormitórios e dois BWC.

Hiperparâmetros

- Dimensionalidade: qual o tamanho do vetor?
- Tamanho de janela: como delimitar o contexto/vizinhança?
- Frequência mínima: quantas vezes o termo deve aparecer para ter um vetor?
- Número de interações: epochs do treinamento da rede neural.

Propriedade de linearidade

- Uma das propriedades mais interessantes de word2vec é a da linearidade.
- Para um corpus grande, os vetores podem resolver analogias usando soma e subtração.

$$\mathbf{vec}(\mathbf{king}) - \mathbf{vec}(\mathbf{man}) + \mathbf{vec}(\mathbf{woman}) \approx \mathbf{vec}(\mathbf{queen}).$$

Outras abordagens

- doc2vec (Le and Mikolov (2014))
- GloVe (Pennington, Socher, and Manning (2014))
- fastText (Bojanowski, Grave, Joulin, and Mikolov (2016))
- Latent semantic analysis (Deerwester, Dumais, Furnas, Harshman, Landauer, Lochbaum, and Streeter (1988))
- BERT (Devlin, Chang, Lee, and Toutanova (2018))