Disciplina:

Processamento de linguagem natural

Professor: Gabriel Assunção

PUC Minas

Apresentação do curso

Módulos

1. Introdutório
2. Expressão regular
3. Processamento de texto
4. **Representação textual**
5. Classificação de texto
6. Modelos de NLP

PUC Minas

Representação Textual

Representação textual

- Precisamos transformar textos em valores numéricos para fácil compreensão pela máquina
- Modelos e técnicas de inteligência artificial usualmente é construída através de representações numéricas dos textos.
- A forma como é representado impacta nas análises e resultados.
- Através de representação textual podemos também verificar a similaridade entre as palavras e textos.

Representação textual

- Formas de representação:
  - TF-IDF
  - Bag of words
  - One hot encoding
  - Embedding

PUC Minas

One hot encoding

One hot encoding

- Representação mais clássica dos textos
- Representação mais simples, de fácil compreensão e implementação.
- Método depende de um vocabulário de palavras e a frequência das palavras.

One hot encoding

- Construção do modelo:
  - Possui um vocabulário com palavras únicas.
  - É criada uma matriz onde cada coluna representa uma palavra do vocabulário e cada linha é uma frase (sentença, texto)
  - Se a palavra **x** está na sentença **y** o campo da coluna **x** e linha **y** é preenchido com o valor 1 os demais serão preenchidos com zero.

One hot encoding

- Exemplos de frases:
  Este filme é muito assustador e longo
  Este filme não é assustador e é lento
  Este filme é assustador e bom
  Este filme é show
  Gosto de pizza

Vocabulário: {este, filme, é, muito, assustador, e, longo, não, lento, bom}

One hot encoding

	este	filme	é	muito	assustador	e	longo	não	lento	bom
Este filme é muito assustador e longo	1	1	1	1	1	1	1	0	0	0
Este filme não é assustador e é lento	1	1	1	0	1	1	0	1	1	0
Este filme é assustador e bom	1	1	1	0	1	1	0	0	0	1
Este filme é show	1	1	1	0	0	0	0	0	0	0
Gosto de pizza	0	0	0	0	0	0	0	0	0	0

One hot encoding

Prós

- Simples de entender
- Textos com as mesmas palavras terão representações parecidas.

Contra

- O tamanho da representação depende do vocabulário
- Não capta sinônimos
- Não é capaz de associar palavras novas
- Não traz informação semântica

Bag of words

Bag of words

- Representação mais clássica dos textos
- Representação mais simples, de fácil compreensão e implementação.
- Método depende de um vocabulário de palavras e a frequência das palavras.
- Similar ao one-hot-encoding

Bag of words

- Construção do modelo:
  - Possui um vocabulário com palavras únicas.
  - É criada uma matriz onde cada coluna representa uma palavra do vocabulário e cada linha é uma frase (sentença, texto)
  - Se a palavra **x** está na sentença **y** o campo da coluna **x** e linha **y** é preenchido com a frequência de quantas vezes a palavra aparece na sentença.

Bag of words

- Exemplos de frases:
  Este filme é muito assustador e longo
  Este filme não é assustador e é lento
  Este filme é assustador e bom
  Este filme é show
  Gosto de pizza

Vocabulário: {este, filme, é, muito, assustador, e, longo, não, lento, bom}

Bag of words

	este	filme	é	muito	assustador	e	longo	não	lento	bom
Este filme é muito assustador e longo	1	1	1	1	1	1	1	0	0	0
Este filme não é assustador e é lento	1	1	2	0	1	1	0	1	1	0
Este filme é assustador e bom	1	1	1	0	1	1	0	0	0	1
Este filme é show	1	1	1	0	0	0	0	0	0	0
Gosto de pizza	0	0	0	0	0	0	0	0	0	0

Bag of words

Prós

- Simples de entender
- Textos com as mesmas palavras terão representações parecidas
- Frequência das palavras é considerado.

Contra

- O tamanho da representação depende do vocabulário
- Não capta sinônimos
- Não é capaz de associar palavras novas
- Não traz informação semântica

TF-IDF

TF-IDF

- TF: term frequency - frequência do termo
- IDF: inverse document frequency - inverso da frequência do documento
- É uma métrica que compara a frequência de um termo em um texto em relação aos demais.
- Avalia a importância da palavra no termo.

TF

- Frequência do termo:
  - Mede a relevância de um termo em um texto.

TF(t) = Nº de vezes que o termo t aparece no texto/Nº de termos no texto

TF

- Exemplos de frases:
  Este filme é muito assustador e longo
  Este filme não é assustador e é lento
  Este filme é assustador e bom

TF

Termo	C1	C2	C3	TF C1	TF C2	TF C3
este	1	1	1	1/7	1/8	1/6
filme	1	1	1	1/7	1/8	1/6
é	1	2	1	1/7	2/8	1/6
muito	1	0	0	1/7	0/8	0/6
assustador	1	1	1	1/7	1/8	1/6
e	1	1	1	1/7	1/8	1/6
longo	1	0	0	1/7	0/8	0/6
não	0	1	0	0/7	1/8	0/6
lento	0	1	0	0/7	1/8	0/6
bom	0	0	1	0/7	0/8	1/6

IDF

- Frequência inversa dos documentos
  - Mede o que rara uma palavra é entre os textos

IDF(t) = log (Total de textos/Nº de textos com o termo t)

Se um termo é frequente em todos os textos o IDF será de log(1) = 0
Se um termo é frequente em apenas um texto o IDF será de log(n).

IDF

Termo	C1	C2	C3	IDF
este	1	1	1	log(3/3) = 0
filme	1	1	1	log(3/3) = 0
é	1	2	1	log(3/3) = 0
muito	1	0	0	log(3/1) = 0,48
assustador	1	1	1	log(3/3) = 0
e	1	1	1	log(3/3) = 0
longo	1	0	0	log(3/1) = 0,48
não	0	1	0	log(3/1) = 0,48
lento	0	1	0	log(3/1) = 0,48
bom	0	0	1	log(3/1) = 0,48

IDF

Termo	C1	C2	C3	TF C1	TF C2	TF C3	IDF	TF-IDF C1	TF-IDF C2	TF-IDF C3
este	1	1	1	1/7	1/8	1/6	log(3/3) = 0	0,000	0,000	0,000
filme	1	1	1	1/7	1/8	1/6	log(3/3) = 0	0,000	0,000	0,000
é	1	2	1	1/7	2/8	1/6	log(3/3) = 0	0,000	0,000	0,000
muito	1	0	0	1/7	0/8	0/6	log(3/1) = 0,48	0,069	0,000	0,000
assustador	1	1	1	1/7	1/8	1/6	log(3/3) = 0	0,000	0,000	0,000
e	1	1	1	1/7	1/8	1/6	log(3/3) = 0	0,000	0,000	0,000
longo	1	0	0	1/7	0/8	0/6	log(3/1) = 0,48	0,069	0,000	0,000
não	0	1	0	0/7	1/8	0/6	log(3/1) = 0,48	0,000	0,060	0,000
lento	0	1	0	0/7	1/8	0/6	log(3/1) = 0,48	0,000	0,060	0,000
bom	0	0	1	0/7	0/8	1/6	log(3/1) = 0,48	0,000	0,000	0,080

TF-IDF

Prós

- Traz a importância da palavra para o texto.
- Palavras repetitivas ou muito frequentes terão menor peso.
- Palavras específicas terão um peso maior.

Contra

- O tamanho da representação depende do vocabulário
- Não capta sinônimos
- Não é capaz de associar palavras novas
- Não traz informação semântica

Embedding

Embedding

- Representação vetorial que possui o contexto das palavras
- Os números dos vetores são contínuos.
- Vetores similares representam palavras com significado ou função semelhantes.
- Necessita de um volume grande de textos para treinar o modelo
- Métodos mais conhecidos:
  - Word2Vec, GloVe, Fasttext, **Bert**

Embedding

Prós

- Palavras parecidas tem representação parecidas.
- Capaz de realizar cálculos entre as palavras.
- Representação do vetor é fixa.

Contra

- **Não é capaz de associar palavras novas**
- Não diferencia a palavra em contextos diferentes.

Word2Vec

Word2Vec

- Representação de embedding construída utilizando redes neurais.
- Para cada palavra do vocabulário é criada uma representação vetorial.
- Essa representação é uma camada intermediária de uma rede neural treinada de duas possíveis formas conhecidas como CBOW (continuos bag of words) e SkipGram.

Word2Vec

CBOW

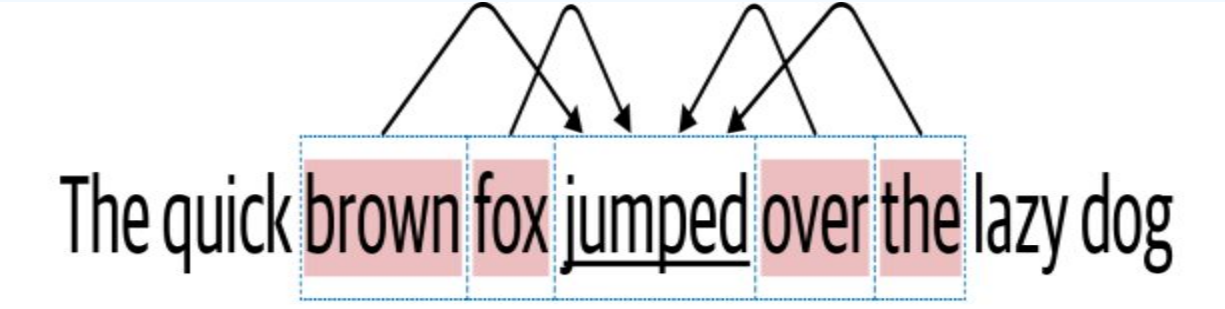
- Prevê uma palavra dado o contexto.
- Mais rápido para treinar
- Mais preciso para palavras mais frequentes.

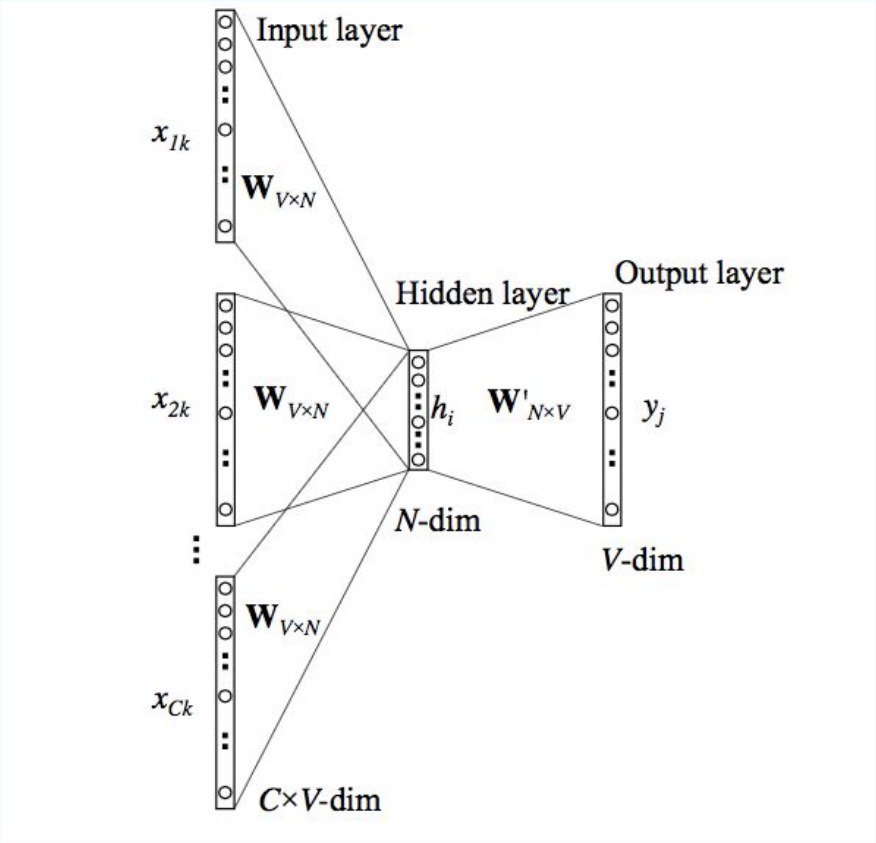
Skip-gram

- Prevê um contexto dado uma palavra.
- Funciona bem com poucos dados
- Representa melhor palavras raras

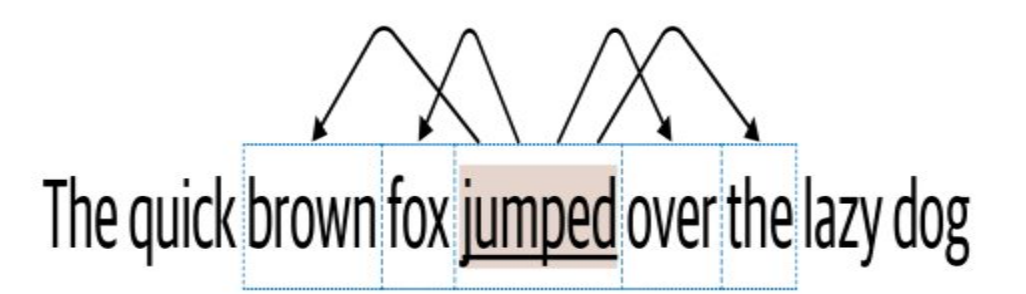
PUC Minas

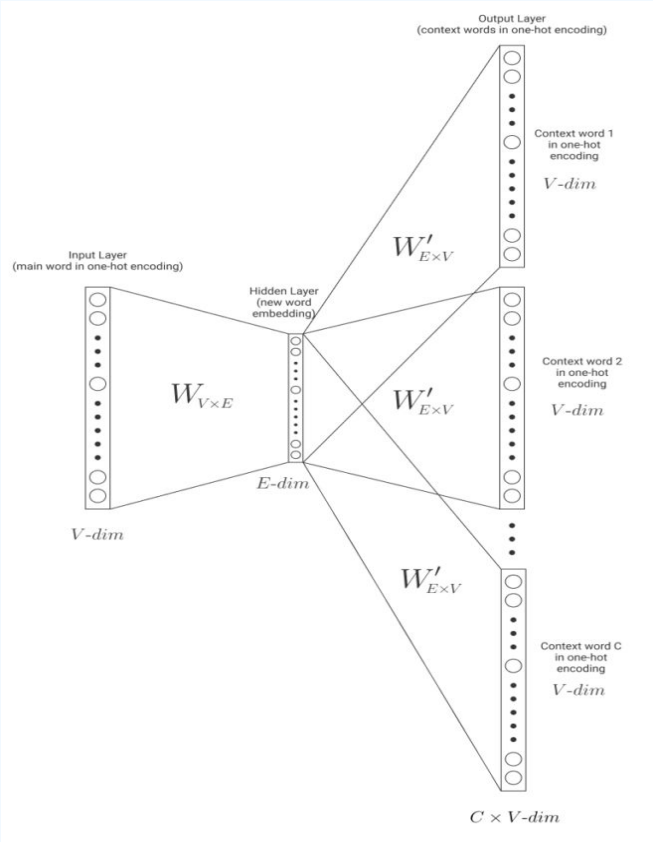
CBOW




Input layer
Hidden layer
Output layer
x_{1k}
x_{2k}
x_{Ck}
$\mathbf{W}_{V\times N}$
$\mathbf{W}_{V\times N}$
$\mathbf{W}_{V\times N}$
h_i
$\mathbf{W}'_{N\times V}$
y_j
N-dim
V-dim
$C\times V$-dim

Skip-gram




Output Layer
(context words in one-hot encoding)
Context word 1
in one-hot
encoding
V-dim
Input Layer
(main word in one-hot encoding)
Hidden Layer
(new word
embedding)
$W'_{E\times V}$
$W_{V\times E}$
$W'_{E\times V}$
Context word 2
in one-hot
encoding
V-dim
E-dim
V-dim
$W'_{E\times V}$
Context word C
in one-hot
encoding
V-dim
C × V-dim

GloVe

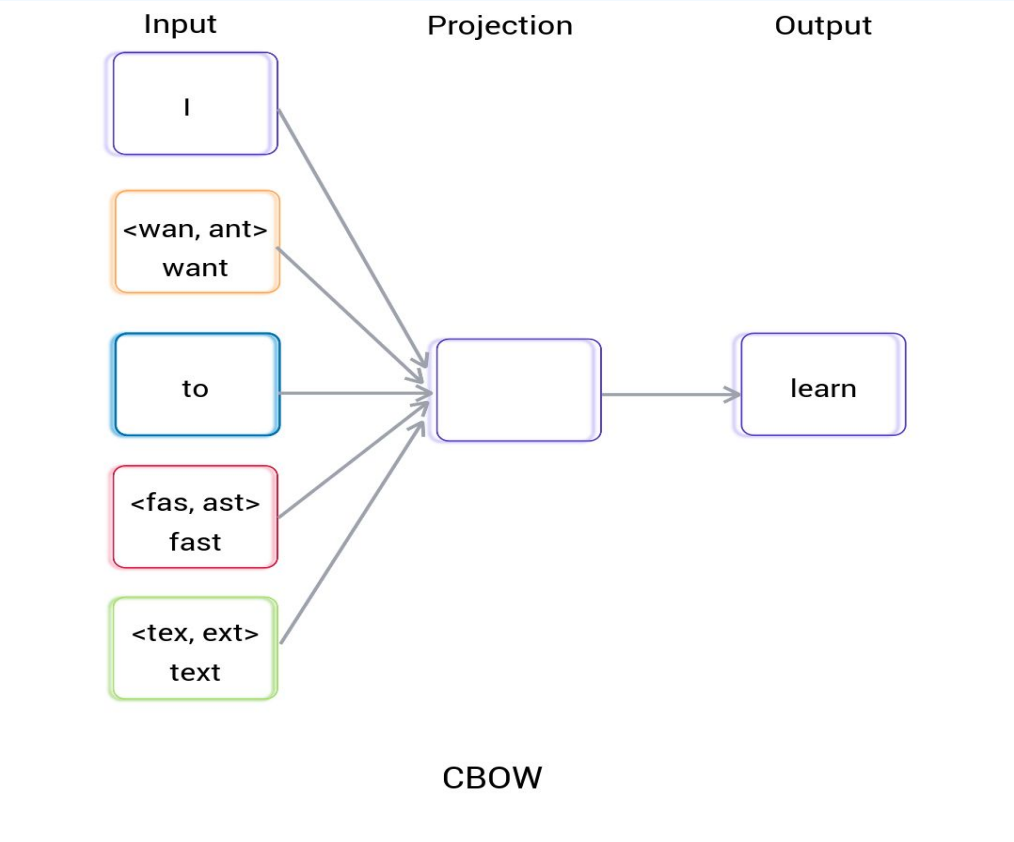
GloVe

- Representação de embedding através de decomposição de matriz de co-ocorrência.
- Para cada palavra do vocabulário é criada uma representação vetorial.
- Essa representação é construída calculando a probabilidade de duas palavras aparecem juntas.

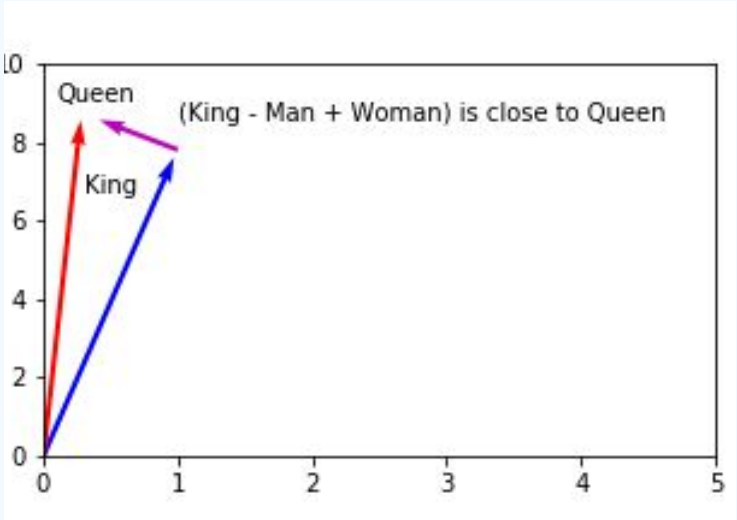
Fasttext

FastText

- Desenvolvido pelo time de IA do Facebook
- Considera a morfologia da palavra para a construção do embedding
- Funciona bem com palavras fora do vocabulário


Input
Projection
Output
I
<wan, ant>
want
to
<fas, ast>
fast
<tex, ext>
text
learn
CBOW

Operações


Queen
(King - Man + Woman) is close to Queen
King
10
8
6
4
2
0
0
1
2
3
4
5

Visualizações

Visualizações

- Possível criar visualizações usando PCA (análise de componentes principais) ou t-SNE para criar gráficos 3D que representam as palavras.
- Pode ser usado para mensurar por exemplo o quanto textos(documentos) similares possuem a mesma representação.
- Exemplo: https://projector.tensorflow.org/

Medidas de similaridade

Medidas de similaridades

- Similaridade entre palavras.
  - Uso: corretor ortográfico, agrupamento de palavras.
- Similaridade entre frases.
  - Uso: classificação de frases, agrupamento de frases.
- Similaridade entre documentos.
  - Uso: classificação de documentos.

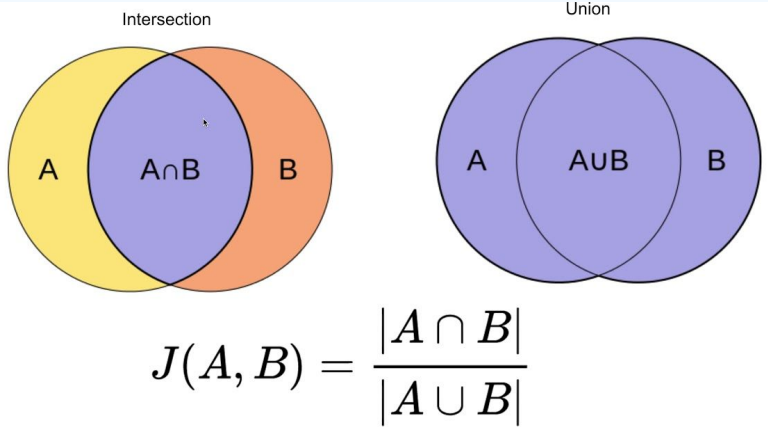
Distância de Jaccard

Cálculo:

Nº de palavras em comum nos textos

Nº de palavras no vocabulário dos dois textos.

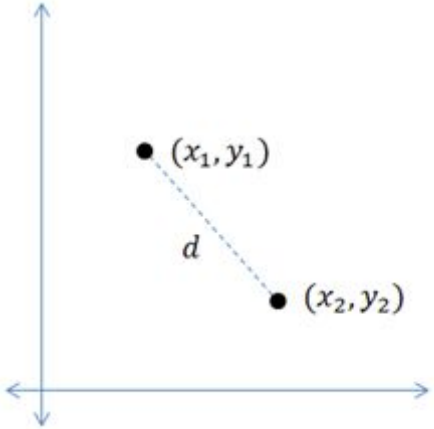
- Pode ser calculado usando a representação de one-hot-encoding
- Valores entre 0 e 1



PUC Minas

Distância Euclidiana

- Quanto mais similares forem as representações mais próximo de zero.
- Muito utilizado em representação como embedding, BoW ou TF-IDF
- Um ponto ruim é que o valor máximo tende a infinito.



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

PUC Minas

Similaridade Cosseno

- Quanto mais similares forem as representações mais próximo de 1.
- Quanto menos similares forem as representações mais próximo de -1.
- Muito utilizado em representação como embedding, BoW ou TF-IDF
- Possui máximo e mínimo definidos.