Mineração de Dados não Estruturados

Aula 03

Limitações da *Bag-of-Words*
Exemplo com Classificação
Word Embeddings
Exemplos com *Word Embeddings*

Prof. Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

LABIC Laboratório de Inteligência Computacional ICMC.USP SÃO CARLOS

ICMC USP SÃO CARLOS

Mineração de Textos

- Na aula anterior, introduzimos pré-processamento de textos com *bag-of-words*

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

Mineração de Textos

- Na aula anterior, introduzimos pré-processamento de textos com *bag-of-words*

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

Cada linha é um documento textual

Mineração de Textos

- Na aula anterior, introduzimos pré-processamento de textos com *bag-of-words*

	t_1	t_2	t_3	...	t_M
d_1	a_{11}	a_{12}	a_{13}	$\cdots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\cdots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\cdots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\cdots$	a_{NM}

Cada coluna é um atributo (e.g. termo extraído dos textos)

Mineração de Textos

- Na aula anterior, introduzimos pré-processamento de textos com *bag-of-words*

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	$\boxed{a_{23}}$	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

Células indicam a relevância do termo no documento, como binário (ausência/presença), frequência, TF-IDF, etc...

Mineração de Textos

- Na aula anterior, introduzimos pré-processamento de textos com *bag-of-words*

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

O documento é representado por um vetor, ou seja, representado em um modelo espaço-vetorial).

Mineração de Textos

- Quais as limitações da *bag-of-words*?

	t_1	t_2	t_3	…	t_M
d_1	a_{11}	a_{12}	a_{13}	…	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	…	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	…	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	…	a_{NM}

Mineração de Textos

- Quais as limitações da *bag-of-words*?

  1. Alta dimensionalidade

Métodos de extração de padrões são prejudicados na presença de alta dimensionalidade

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

Mineração de Textos

- Quais as limitações da *bag-of-words*?

  2. Semântica

Pobreza semântica do modelo.
Bag-of-words ignora ordem de ocorrência das palavras.

	t_1	t_2	t_3	…	t_M
d_1	a_{11}	a_{12}	a_{13}	…	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	…	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	…	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	…	a_{NM}

Mineração de Textos

- Quais as limitações da *bag-of-words*?

  2. Semântica

  Pobreza semântica do modelo. *Bag-of-words* ignora ordem de ocorrência das palavras.

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

  - Exemplo: qual a representação para os textos abaixo?
    - D1: *Company X acquired company Y.*
    - D2: *Company Y acquired company X.*

- Quais as limitações da *bag-of-words*?

  2. Semântica

  Pobreza semântica do modelo. Dificuldade em lidar com sinônimos ou proximidade semântica.

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

Mineração de Textos

- Quais as limitações da *bag-of-words*?

    2. Semântica

    Pobreza semântica do modelo. Dificuldade em lidar com sinônimos ou proximidade semântica.

	t_1	t_2	t_3	$\dots$	t_M
d_1	a_{11}	a_{12}	a_{13}	$\dots$	a_{1M}
d_2	a_{21}	a_{22}	a_{23}	$\dots$	a_{2M}
d_3	a_{31}	a_{32}	a_{33}	$\dots$	a_{3M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	a_{N3}	$\dots$	a_{NM}

- Exemplo:
    - *D1: Obama speaks to the media in Illinois.*
    - *D2: The President greets the press in Chicago.*

O Processo de Mineração de Textos

- Vamos testar a *bag-of-words* para classificação

(Rezende et al., 2003)

Classificação k-NN e *Bag-of-Words*

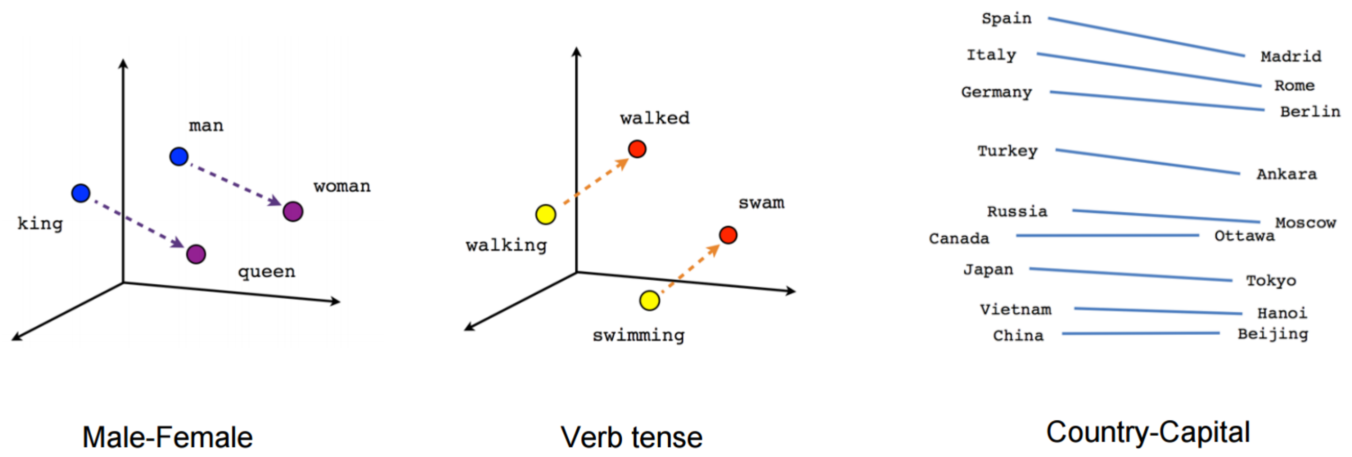
- Exemplo prático em aula...

Mineração de Textos

- Conhecemos as limitações da *bag-of-words*
- Quais as alternativas?

- Conhecemos as limitações da *bag-of-words*
- Quais as alternativas?

Introdução à *Word Embeddings*



Mineração de Textos

- Hipótese Distribucional
  - Palavras podem ser caracterizadas pelo contexto em que aparecem
  - Duas palavras que possuem contextos similares, são consideradas similares

Mineração de Textos

- Hipótese Distribucional
  - Palavras podem ser caracterizadas pelo contexto em que aparecem
  - Duas palavras que possuem contextos similares, são consideradas similares

O que é biroiskau?

Mineração de Textos

- Hipótese Distribucional
  - Palavras podem ser caracterizadas pelo contexto em que aparecem
  - Duas palavras que possuem contextos similares, são consideradas similares

> Uma garrafa de biroiskau está na geladeira
> Todos gostam de biroiskau
> Cuidado, pois biroiskau te deixa bêbado
> Fazemos biroiskau de malte e cevada

Mineração de Textos

- Hipótese Distribucional

  - Podemos inferir o significado de *biroiskau* pelo contexto e nosso conhecimento de mundo

> Uma garrafa de biroiskau está na geladeira
> Todos gostam de biroiskau
> Cuidado, pois biroiskau te deixa bêbado
> Fazemos biroiskau de malte e cevada

Mineração de Textos

- Hipótese Distribucional
  - Palavras podem ser caracterizadas pelo contexto em que aparecem
  - Duas palavras que possuem contextos similares, são consideradas similares

> Meu sapato está muito apertado e
> o meu _____ está doendo!

Mineração de Textos

- Hipótese Distribucional
- Matematicamente, queremos computar:

$$p(\text{ word } \mid \text{ context })$$

ou

$$p(\text{ context } \mid \text{ word })$$

Mineração de Textos

- Hipótese Distribucional
- O que é contexto?

$$\text{corpus:} \quad w_1 \mid w_2 \mid w_3 \mid \quad \cdots \quad w_t \quad \cdots \quad \mid w_{T-2} \mid w_{T-1} \mid w_T$$

$$\downarrow$$

$$\text{context of } w_t\text{:} \quad [w_{t-n} \ \cdots \ w_{t-1}] \ w_t \ [w_{t+1} \ \cdots \ w_{t+n}]$$

- O contexto da palavra w_t é formado por palavras próximas de w_t em um corpus de treinamento...

Mineração de Textos

- Hipótese Distribucional
- Dado um corpus, encontrar um conjunto de parâmetros θ para maximizar $f(\theta)$

$$p(\text{corpus};\theta) = \prod_{w_t} \prod_{w_c \in C_t} p(w_c | w_t;\theta).$$

$$f(\theta) = p(\text{corpus};\theta)$$

Mineração de Textos

- Hipótese Distribucional
- Dado um corpus, encontrar um conjunto de parâmetros θ para maximizar $f(\theta)$

$$p(\text{corpus}; \theta) = \prod_{w_t} \prod_{w_c \in C_t} p(w_c | w_t; \theta).$$

$$f(\theta) = p(\text{corpus}; \theta)$$

- Cada palavra é representada por um vetor proveniente de θ $\quad \text{word}_i = \langle \theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,300} \rangle$

Mineração de Textos

- Exemplo https://projector.tensorflow.org/

Como calcular θ ?

- Dado um corpus, encontrar um conjunto de parâmetros θ para maximizar *f(θ)*

$$p(\text{corpus}; \theta) = \prod_{w_t} \prod_{w_c \in C_t} p(w_c | w_t; \theta).$$

$$f(\theta) = p(\text{corpus}; \theta)$$

- Cada palavra é representada por um vetor proveniente de θ $\quad \text{word}_i = \langle \theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,300} \rangle$

Mineração de Textos

Como calcular θ ?

$$p(\text{corpus}; \theta) = \prod_{w_t} \prod_{w_c \in C_t} p(w_c | w_t; \theta).$$

$$f(\theta) = p(\text{corpus}; \theta)$$

- Nessa aula vamos apresentar apenas a ideia intuitiva de uma forma para calcular os parâmetros θ
- É necessária uma base de redes neurais (mais detalhes na próxima aula)

Visão geral de *Word Embeddings*

- Word Embeddings e Skip-Gram ficou conhecido em 2013 com o *word2vec*

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

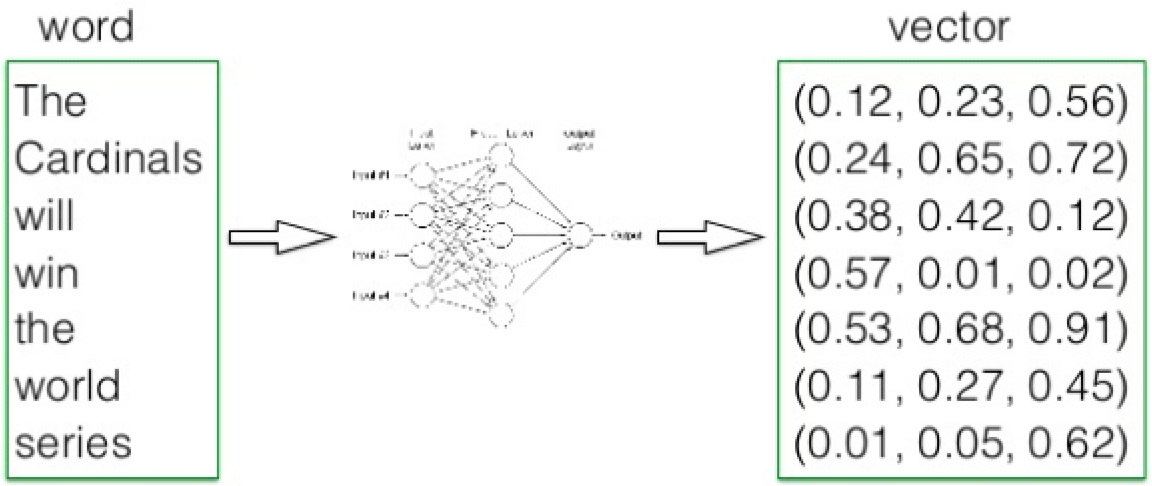
Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Visão geral de *Word Embeddings*

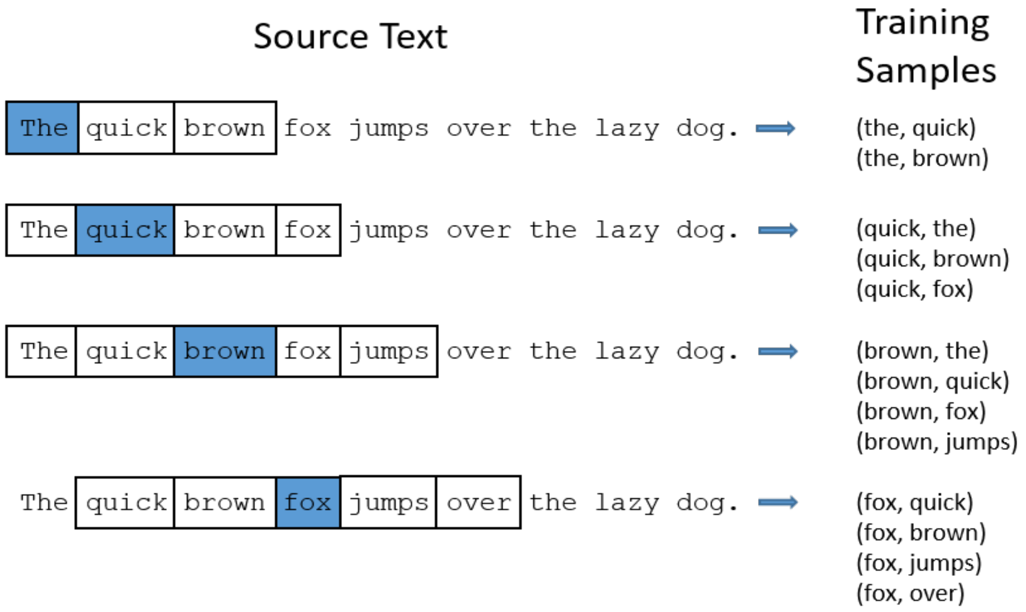
- O modelo Skip-Gram do Word2Vec
  - O objetivo é aprender “word vectors” para cada termo da coleção textual



  - No final, termos que coocorrem de forma significativa dado um contexto, terão “word vectors” similares

Visão geral de *Word Embeddings*

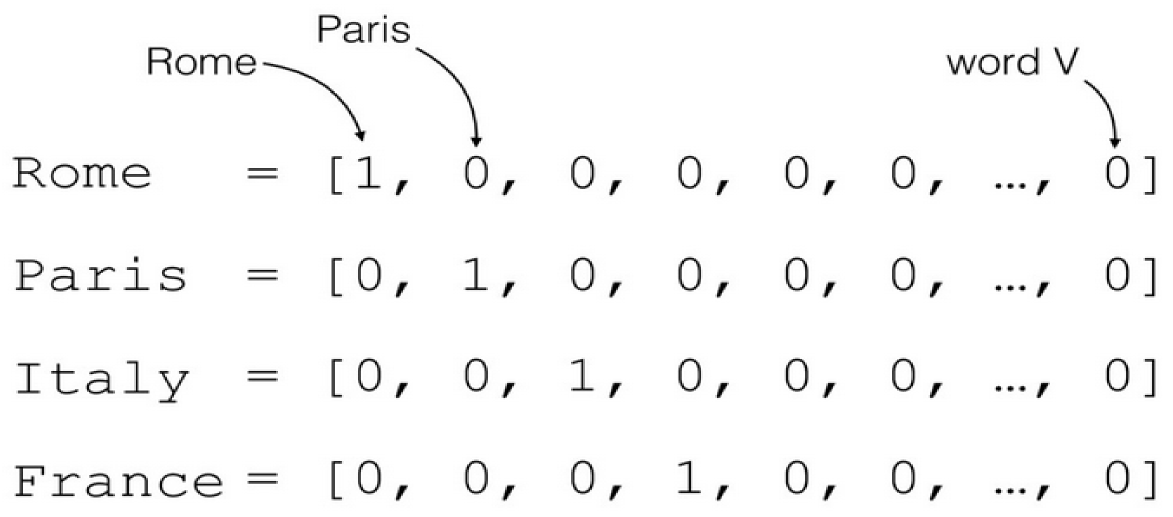
- Conjunto de treinamento com palavras que coocorrem em um determinado contexto.
  - Palavras alvo e palavras de contexto
- O contexto geralmente é uma janela deslizante no texto.
  - Exemplo com janela de tamanho 2.



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Visão geral de *Word Embeddings*

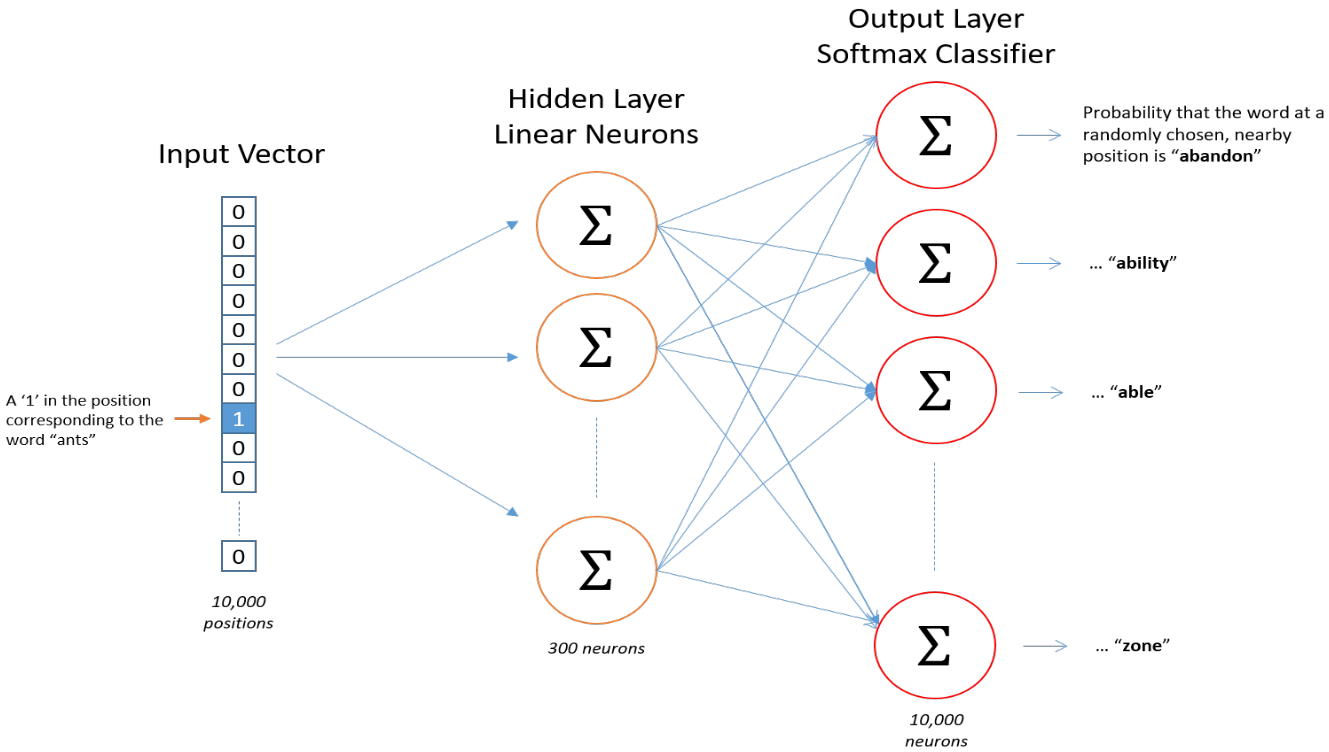
- O conjunto de treino precisa ser codificado para ser utilizada em uma rede neural
- Estratégia "*one-hot vector*".
  - Se há V palavras na coleção, a dimensão do *one-hot* será V



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

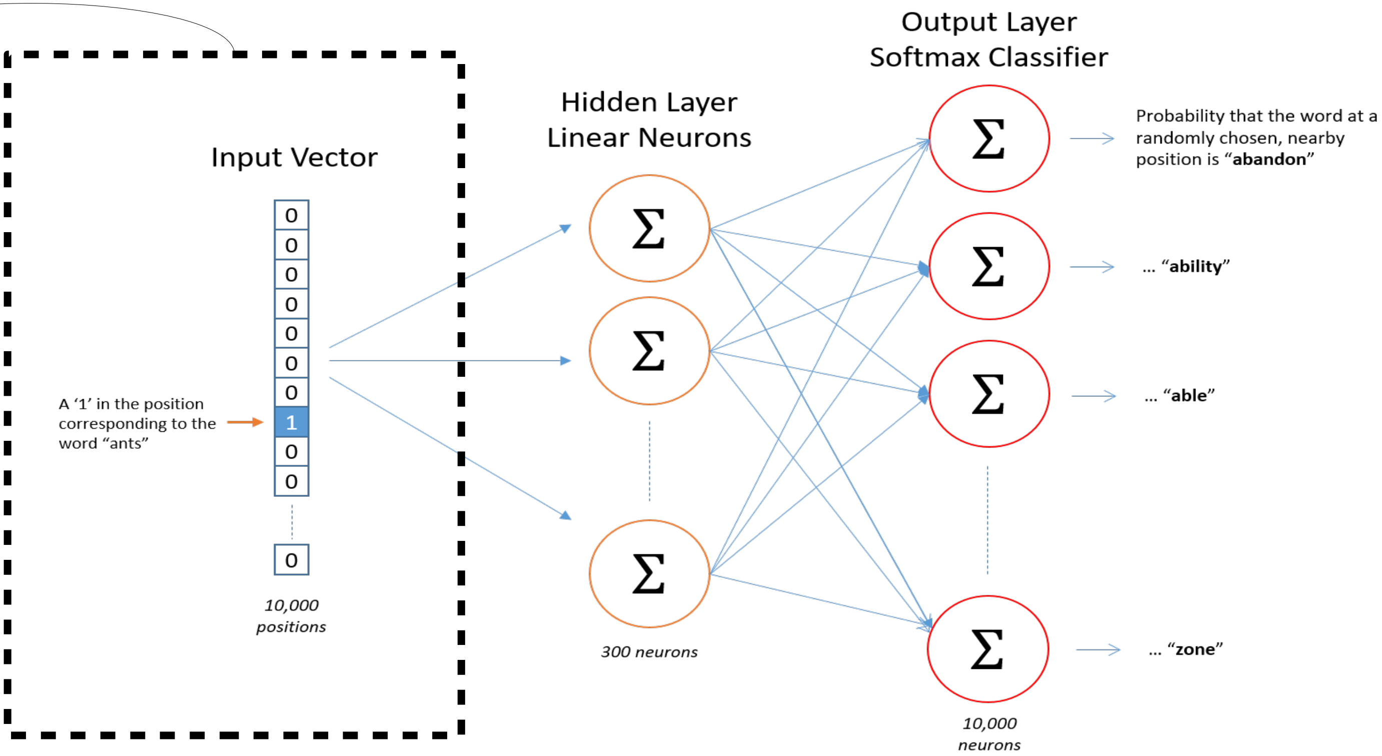


http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

Camada de entrada é de mesma dimensão dos one-hot vectors (número de termos)

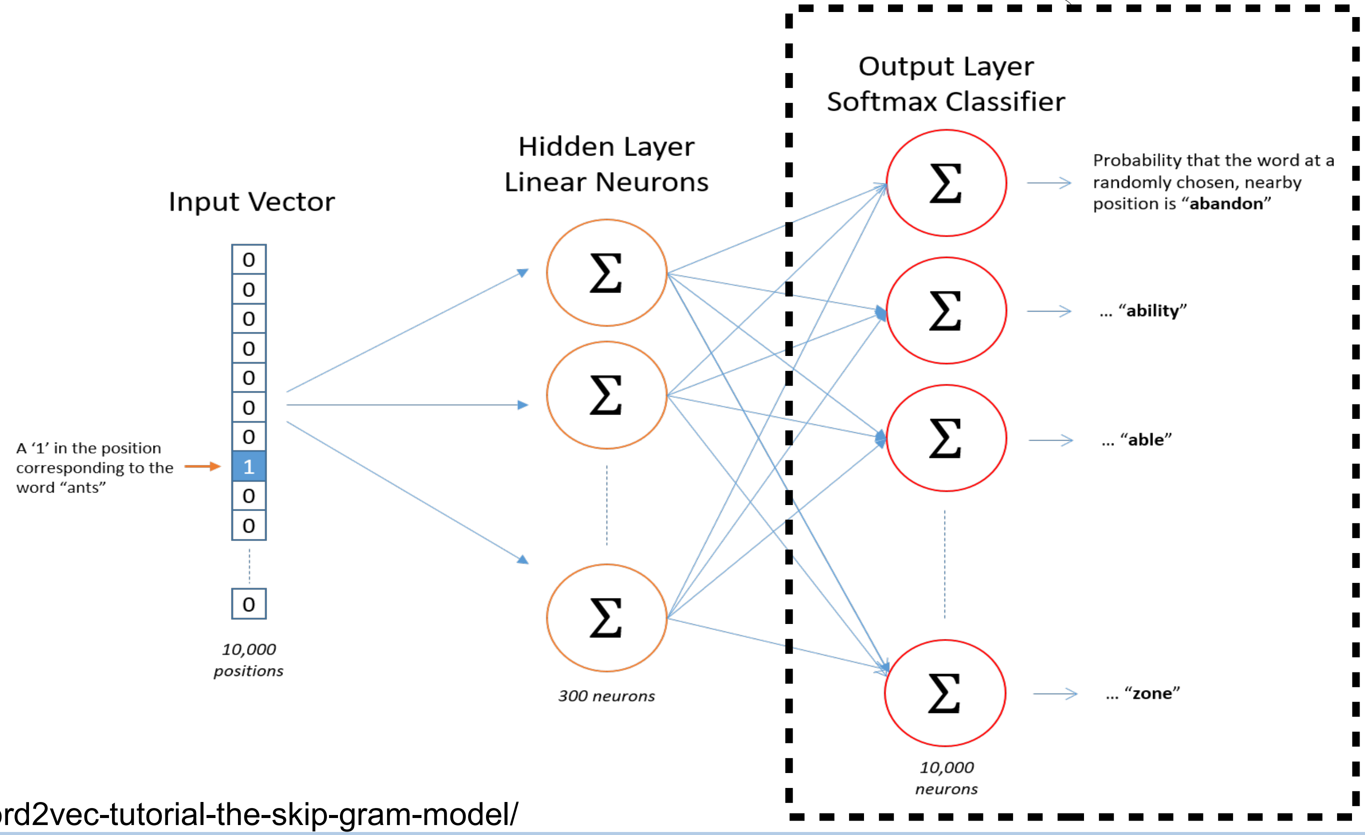


http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

Camada de saída possui a mesma quantidade de neurônios da dimensão do one-hot vector



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

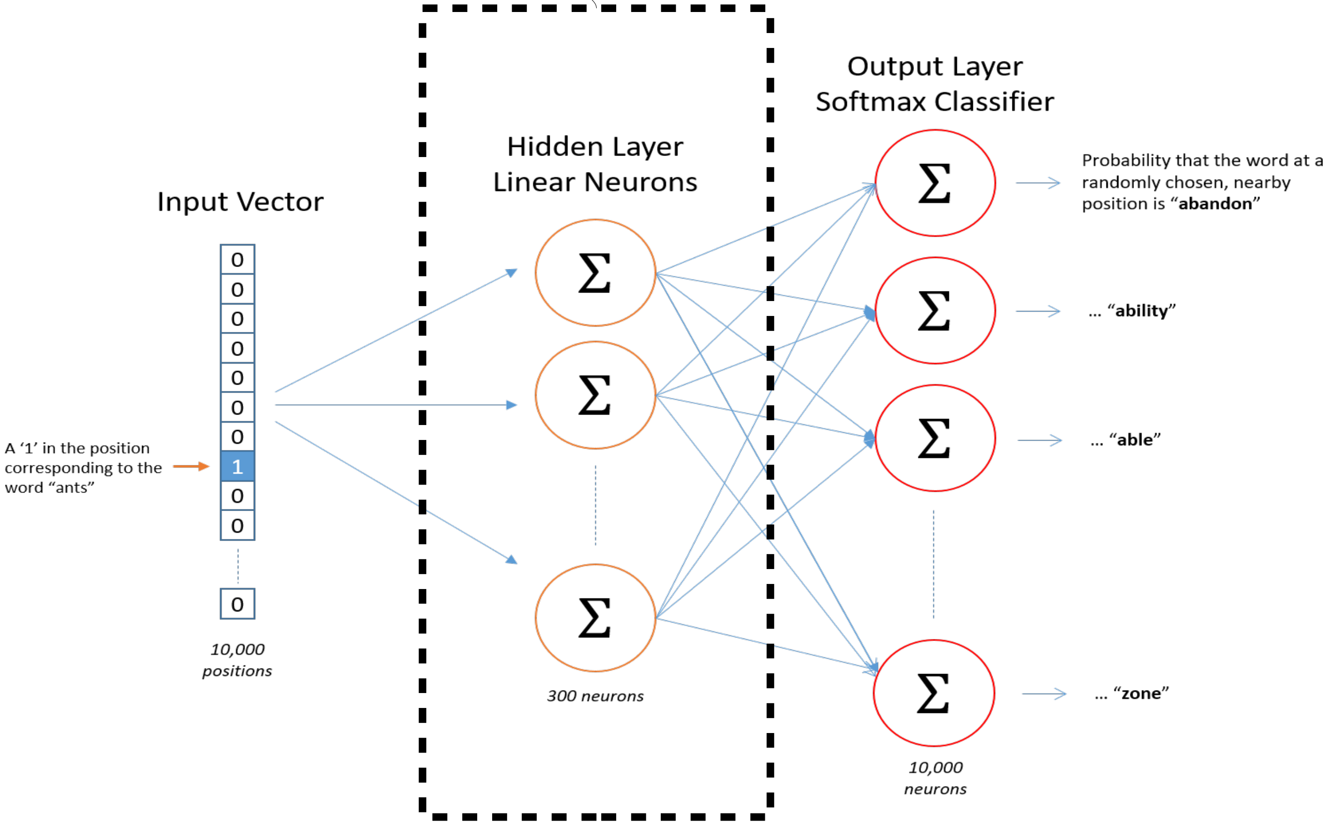
Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

Camada intermediária:

O número de neurônios determina a dimensão dos word vectors.

Word2Vec utiliza 300. Pode ser calibrado...



Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

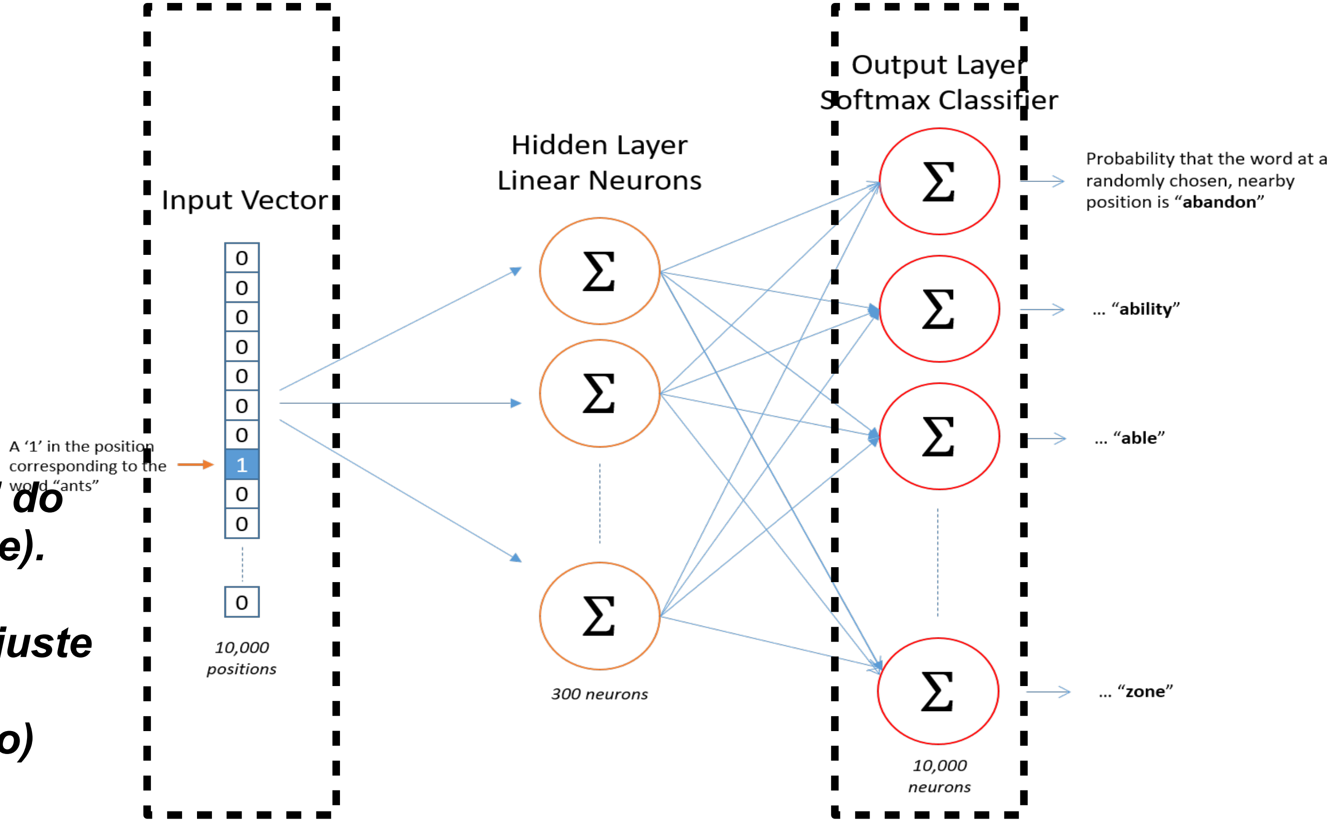
O que a rede aprende?

Ao apresentar um par (termo1,termo2).

Camada de entrada:
One-hot do termo1.

Camada de saída:
Estimar o "one-hot" do termo2 (em probabilidade).

Se errou: é realizado o ajuste dos pesos (treinamento/aprendizado)



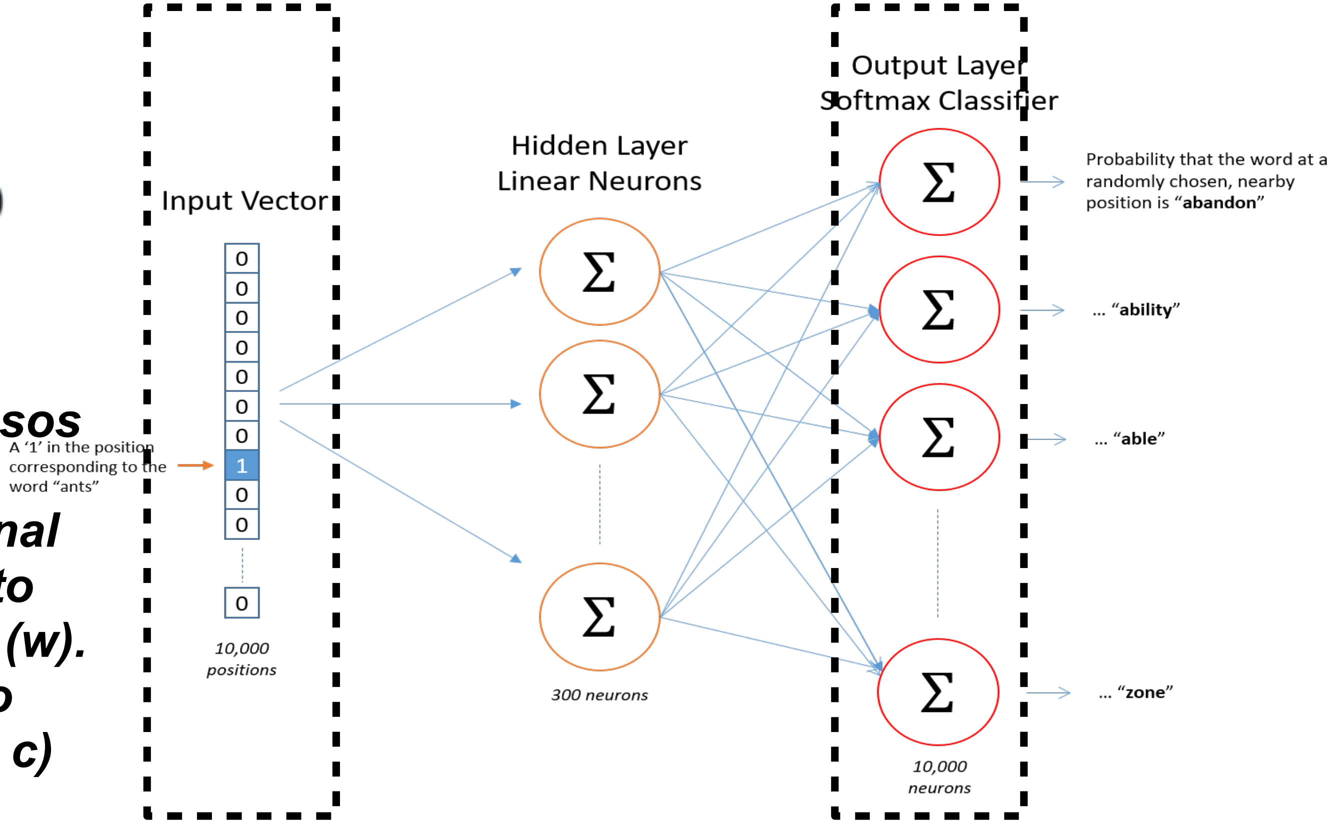
Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

O que a rede aprende?

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta)$$

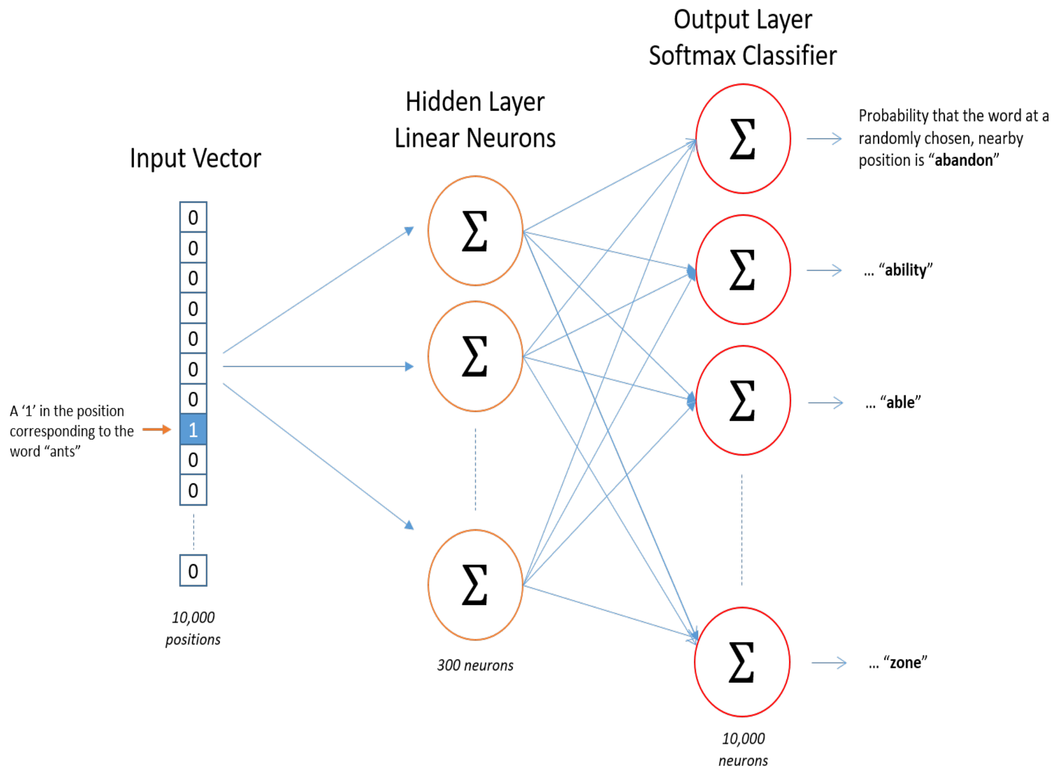
Otimizar parâmetros/pesos (Θ) que maximiza a probabilidade condicional de um termo de contexto (c) dado um termo alvo (w). D é o conjunto de treino para todos os pares (w, c)



Visão geral de *Word Embeddings*

- Arquitetura de uma Rede Neural para o modelo Skip-Gram considerando 10 mil termos e aprendizado de *word vectors* com 300 dimensões

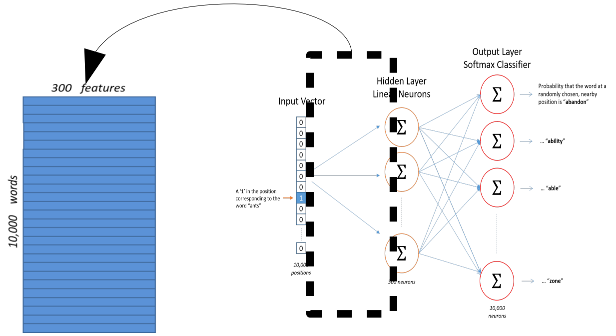
Onde estão os word vectors?



Visão geral de *Word Embeddings*

- O que podemos fazer com os *word vectors (word embeddings)?*
  - Relacionamento mais semântico entre termos (coocorrência em um contexto)

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Visão geral de *Word Embeddings*

- O que podemos fazer com os *word vectors (word embeddings)?*
  - Relacionamento mais semântico entre termos (coocorrência em um contexto)
  - Também é uma informa de incluir informação externa no dataset
    - Word embeddings são treinados em grandes bases de dados como Wikipedia

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

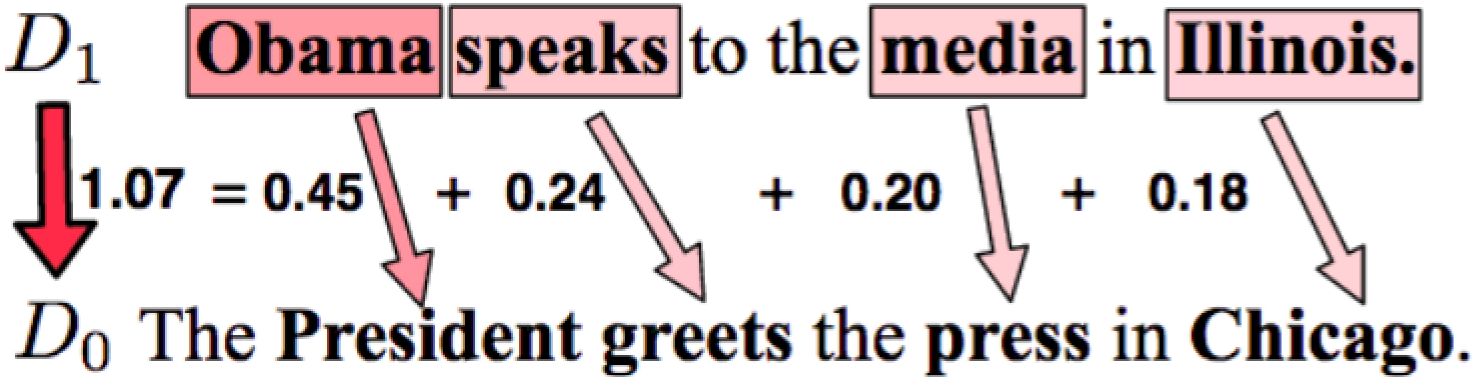
Visão geral de *Word Embeddings*

- O que podemos fazer com os *word vectors (word embeddings)?*
  - Relacionamento mais semântico entre termos (coocorrência em um contexto)
  - Também é uma informa de incluir informação externa no dataset
    - Word embeddings são treinados em grandes bases de dados como Wikipedia
  - Representar sentenças e documentos usando seus *word vectors*.
    - Basta calcular a média a média das *word vectors* da sentença ou do documento.
    - Similaridade entre sentenças e documentos...

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Visão geral de *Word Embeddings*

- O que podemos fazer com os *word vectors (word embeddings)?*
  - Alinhamento “semântico” entre sentenças com Word Movers.
  - Similaridade entre textos de forma mais robusta.



Visão geral de *Word Embeddings*

- Exemplo prático com *Word2Vec* e modelo pré-treinado...
- Testando *word mover distances*

Visão geral de *Word Embeddings*

- Exemplo prático
- Treinando a própria *word embeddings* usando *word2vec*

Visão geral de *Word Embeddings*

- FastText
  - Capacidade de lidar com *subwords*
  - *Out of Vocabulary(OOV) Words*
  - Tratamento da morfologia das palavras
  - Exemplo: <eating>
    - 3-grams = <ea, eat, ati, tin, ing, ng>

Visão geral de *Word Embeddings*

- FastText
  - Capacidade de lidar com *subwords*
  - *Out of Vocabulary(OOV) Words*
  - Tratamento da morfologia das palavras
  - Exemplo: <eating>
    - 3-grams = <ea, eat, ati, tin, ing, ng>

Vamos usar o FastText como
motivação para redes neurais e modelos
de linguagem na próxima aula...