Word2Vector

Clebson C. Alves de Sá
clebsonc@dcc.ufmg.br

Word2Vector

- Models used to create “Word Embeddings”
  - Neural Networks
  - Dimensionality Reduction
  - Probabilistic models
  - …
- Word Embeddings
  - Word or phrases from a vocabulary mapped to a vector of real numbers

The Skip-gram Neural Network

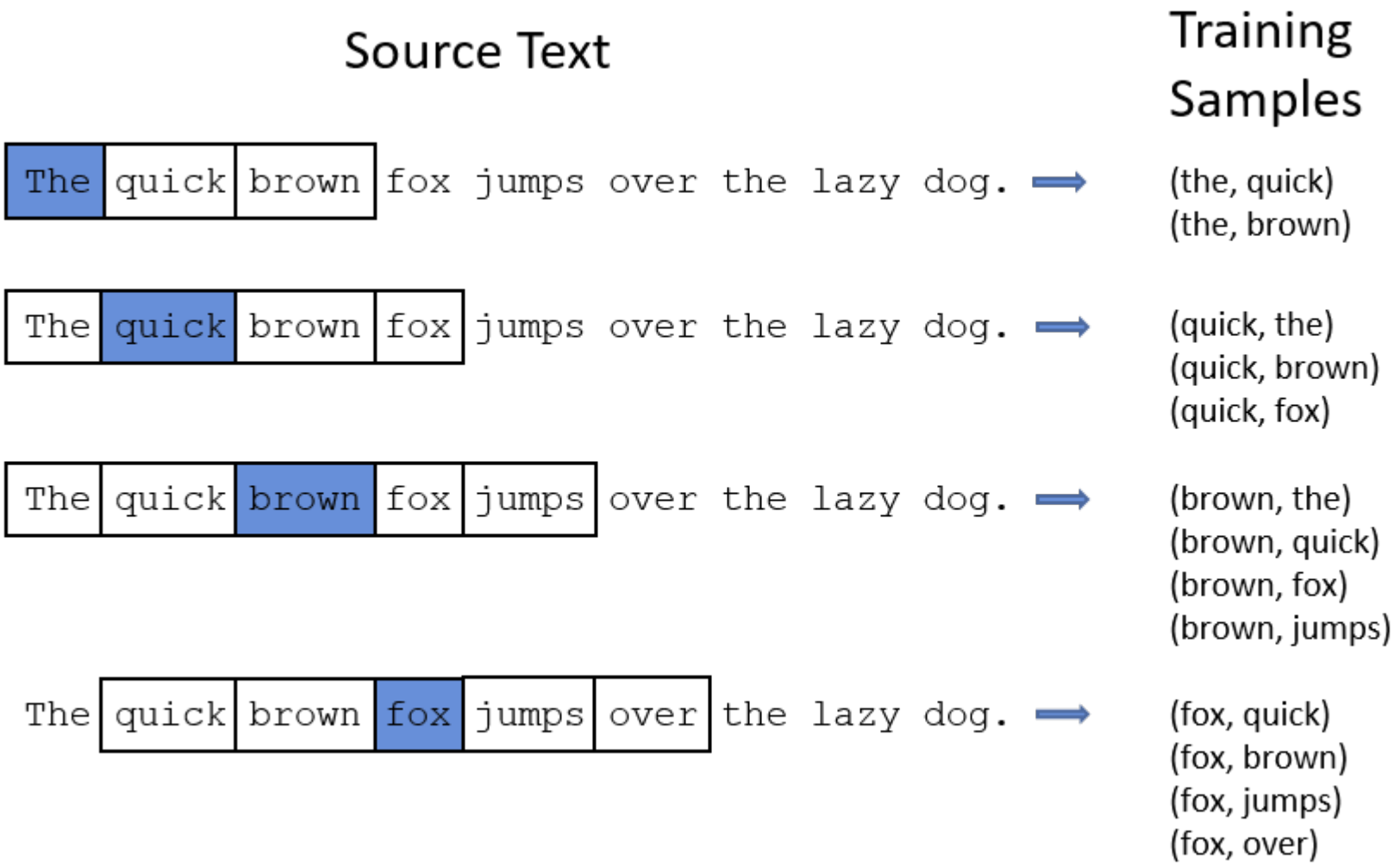
- Train a simple neural network with a single hidden layer to perform a certain task

- This model is actually not used anywhere
  - Just learn the weights of the hidden layer
  - These weights are actually the “Word vectors”

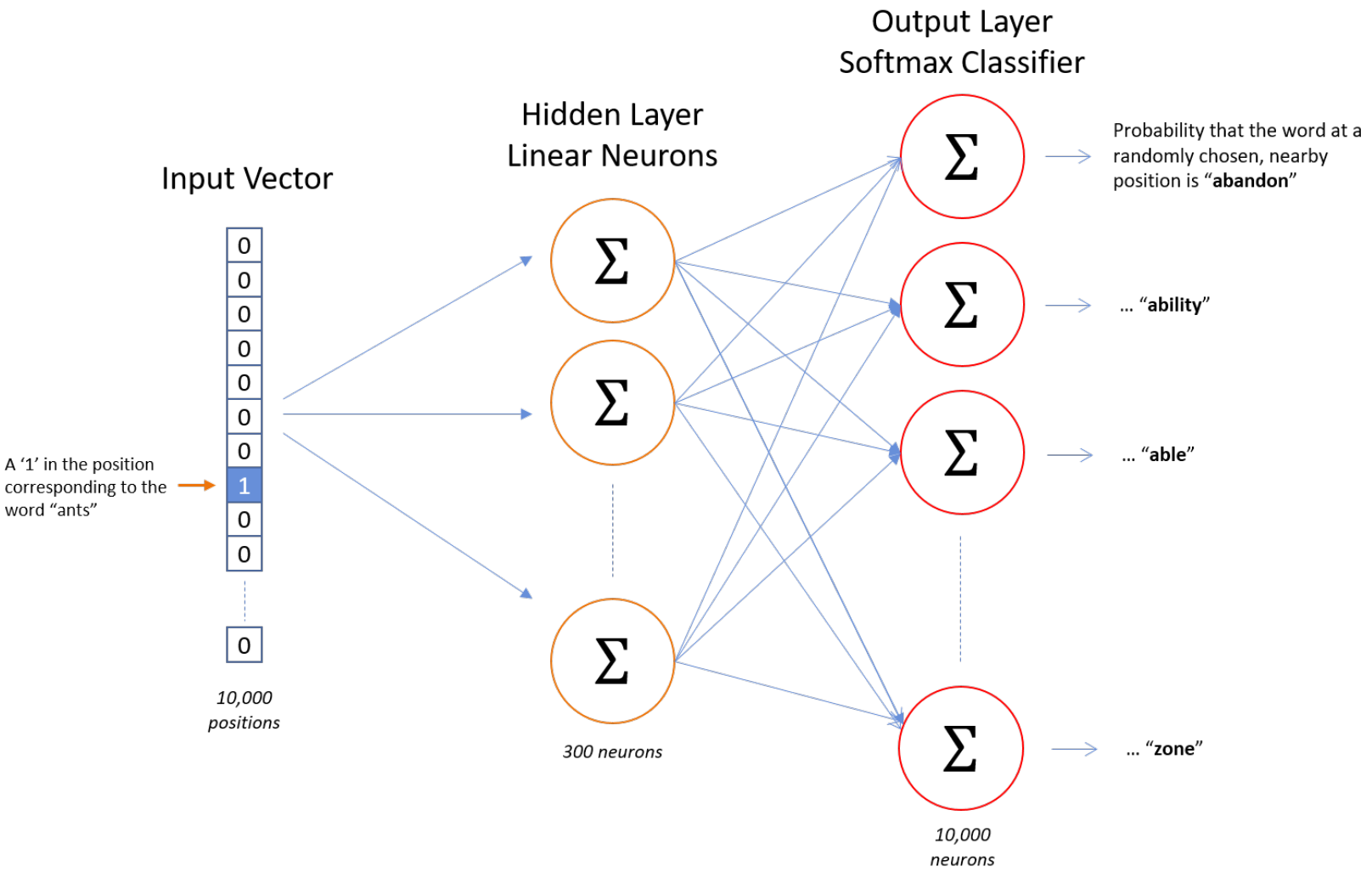
The Skip-gram Neural Network

- What the neural network does?
  - Given a specific word in the middle of a sentence (the input word), look at the words nearby and pick one at random.
  - The network is going to tell us the probability for every word in our vocabulary of being the “nearby word” that we chose.
    - Nearby in the sense that there’s a window size

Window Size


Source Text
Training Samples
The quick brown fox jumps over the lazy dog.
(the, quick)
(the, brown)
The quick brown fox jumps over the lazy dog.
(quick, the)
(quick, brown)
(quick, fox)
The quick brown fox jumps over the lazy dog.
(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)
The quick brown fox jumps over the lazy dog.
(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

Model Architecture



The Skip-gram Neural Network

- This implies that the model is trained to predict the probability of how semantically connected the words are. For example:

Input	Output
England	Queen
Soviet	Union, Russia
Kangaroo	Australian
Doll	Annabelle
Engine	Transmission

Google News

Notícias

Edição Brasil

Personalizar

- Notícias principais
- Belo Horizonte, Stat...
- Sugerido para você
- Mundo
- Brasil
- Negócios
- Ciência/Tecnologia
- Entretenimento
- Esportes
- Saúde

Notícias principais

Globo.com

'Não cabe ao TSE resolver crise política', diz Gilmar Mendes

Globo.com - há 3 horas

Para ministro, pedir vista no julgamento da chapa Dilma-Temer seria algo 'absolutamente normal'. Para ele, 'o Brasil parece que se transformou numa grande Organização Tabajara'. Facebook · Twitter ...

Monstro da judicialização Consultor Jurídico

Brasil vive 'normalidade institucional', diz Gilmar Mendes Epoca Negócios

Cobertura em tempo real

TV Gazeta - SP

Epoca Neg...

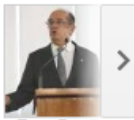
Ultimo Segu...

Consultor J...

Estado de M...

VEJA.com

Tribuna da ...



Terra Brasil

Globo.com

Presas em BH reclamam que Andrea Neves tem recebido regalias na cadeia

Globo.com - há 3 horas

Irmã do senador afastado Aécio Neves (PSDB) está presa desde o dia 18 de maio. De acordo com a PF, há suspeitas de que ela tenha pedido dinheiro ao empresário Joesley Batista, dono do grupo JBS, em nome do irmão. Facebook · Twitter ...

EBC

Com 44 mil pessoas afetadas, municípios de PE contabilizam prejuízos após chuvas

EBC - há 1 hora

Pernambuco registra 23 cidades afetadas pelas chuvas. O governador Paulo Câmara sobrevoou hoje as regiões atingidasAluísio Moreira/Governo de Pernambuco. Passado o período mais intenso das chuvas que atingiram a Zona da Mata Sul e parte do ...

Terra Brasil

Temer parte para ofensiva em última chance de salvar mandato

Terra Brasil - há 7 horas

Michel Temer tomou neste domingo (28/05) a ação mais explícita para salvar seu mandato. Em meio à crise provocada pela delação da JBS, o presidente decidiu tirar Osmar Serraglio do Ministério da Justiça e colocar no seu lugar Torquato Jardim, que ...

Mantega reconhece que tinha US$ 600 mil em conta não declarada na Suíça, mas

Tempo para Belo Horizonte, State of Minas Gerais

Hoje	ter	qua	qui
25° 11°	24° 12°	25° 14°	27° 16°

Belo Horizonte, State o... - Alterar local

Região Nordeste
Super Notícia - há 14 horas

Belotur prorroga prazo para cadastramento de festas públicas e privadas do Arraial de Belo Horizonte
Globo.com - há 7 horas

Blitz da saúde oferece exames gratuitos na Praça da Liberdade, em ...
Estado de Minas - 27 de mai de 2017

Sugestões dos editores

Observatório de Cinema

Jurassic World 2 | Atriz diz que primeiras cenas podem ser divulgadas em...
Ygor Palopoli

Fragmentado 2 | Diretor dá mais detalhes sobre o roteiro do filme

Google News as a Graph

- 3 million words and phrases
  - Each word is a node
  - Words are connected if there is a semantic meaning in the two words.

Gensim

```
>>> from gensim import corpora, models, similarities
>>>
>>> # Load corpus iterator from a Matrix Market file on disk.
>>> corpus = corpora.MmCorpus('/path/to/corpus.mm')
>>>
>>> # Initialize Latent Semantic Indexing with 200 dimensions.
>>> lsi = models.LsiModel(corpus, num_topics=200)
>>>
>>> # Convert another corpus to the latent space and index it.
>>> index = similarities.MatrixSimilarity(lsi[another_corpus])
>>>
>>> # Compute similarity of a query vs. indexed documents
>>> sims = index[query]
```

Gensim is a FREE Python library

- Scalable statistical semantics
- Analyze plain-text documents for semantic structure
- Retrieve semantically similar documents

Proposal

- Characterize the network structure applying the network metrics that we have seen so far in the discipline
- I’m going to use the following paper as a guide:

On the Various Semantics of Similarity in Word Embedding Models

Ábel Elekes
Karlsruhe Institute of Technology
Karlsruhe, Germany
abel.elekes@kit.edu

Martin Schäler
Karlsruhe Institute of Technology
Karlsruhe, Germany
martin.schaeler@kit.edu

Klemens Böhm
Karlsruhe Institute of Technology
Karlsruhe, Germany
klemens.boehm@kit.edu

Word 2 Vector Model

Clebson C. Alves de Sá