Agrupamento de Textos e suas Aplicações em Inteligência Analítica

Agrupamento de Textos (Extra)
Word Embeddings e BERTopic

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

Cursos de Extensão – Difusão de Conhecimento – Dezembro de 2021

ICMC SÃO CARLOS USP

LABIC Laboratório de Inteligência Computacional ICMC.USP SÃO CARLOS

WebSensors

Representação dos Textos

- Pré-processamento – Informação Textual
  - Considerar apenas a *Bag-of-words* para representar informação textual pode falhar em alguns casos na Mineração de Eventos
  - Exemplo:
    - *D1: Obama speaks to the media in Illinois.*
    - *D2: The President greets the press in Chicago.*

D1 e *D2* representam eventos relacionados.
Não possuem atributos em comum!

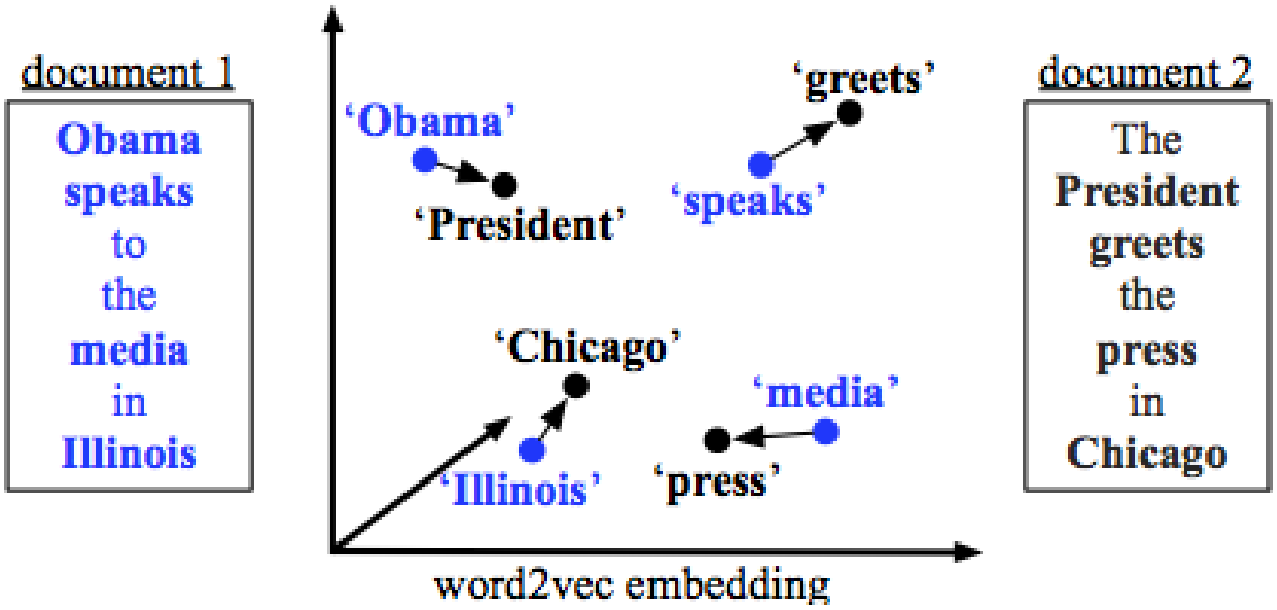
Representação dos Textos

- Pré-processamento – Informação Textual
  - Considerar apenas a *Bag-of-words* para representar informação textual pode falhar em alguns casos na Mineração de Eventos
  - Exemplo:
    - *D1: Obama speaks to the media in Illinois.*
    - *D2: The President greets the press in Chicago.*

Uma solução para este problema:
Word Embedding Models

Representação dos Textos

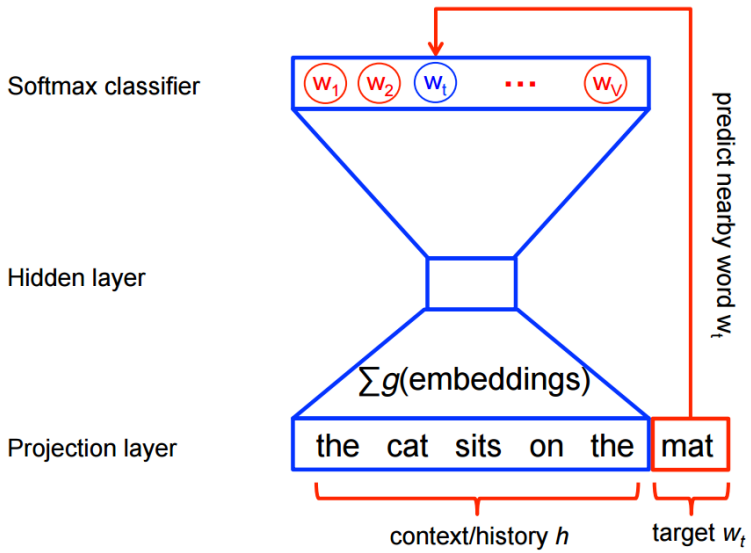
- Pré-processamento – Informação Textual
  - *Word Embedding Models*



MIKOLOV, Tomas et al. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111-3119.

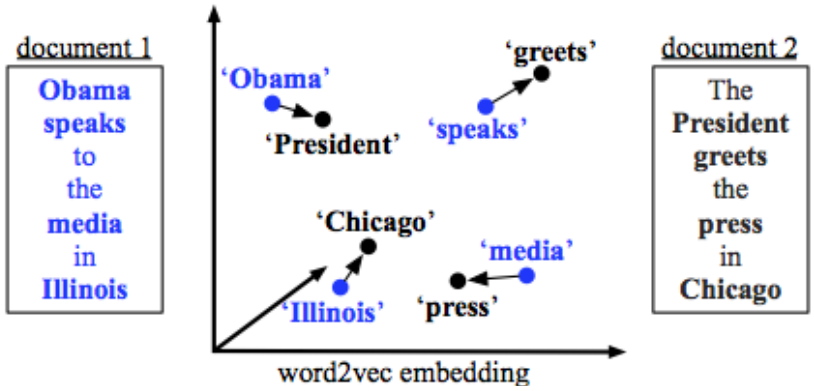
Representação dos Textos

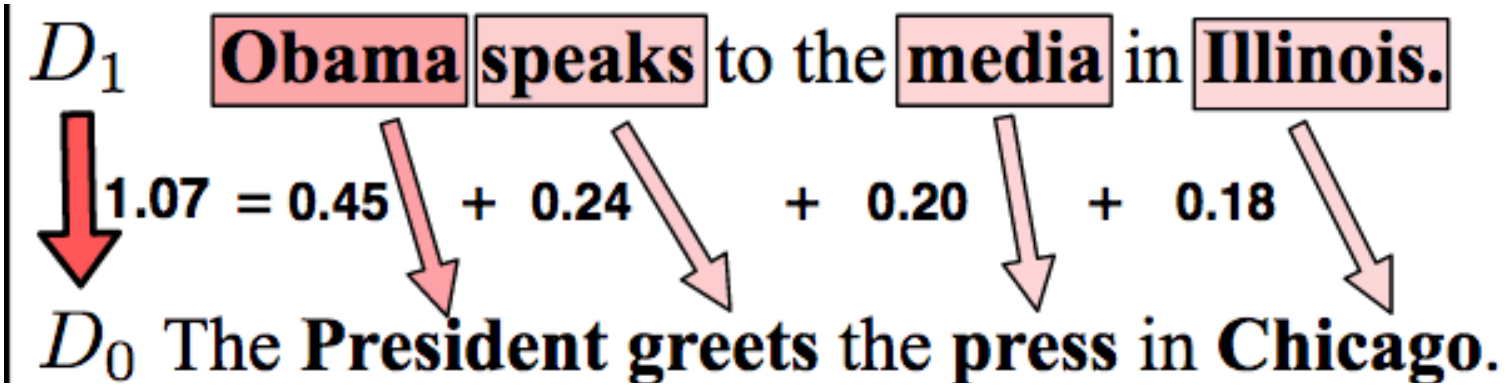
- Pré-processamento – Informação Textual
  - *Word Embedding Models*
  - Alguns modelos pré-treinados (públicos)
    - Word2Vec (Google)
      - Gogle News dataset (100 bilhões de tokens)
    - Glove (Stanford)
      - Wikipedia (6 bilhões de tokens)
    - Fasttext (Facebook)
      - Wikipedia e Statmt News (16 bilhões de tokens)
      - 157 línguas



Representação dos Textos

- Pré-processamento – Informação Textual
  - *Word Embedding Models*

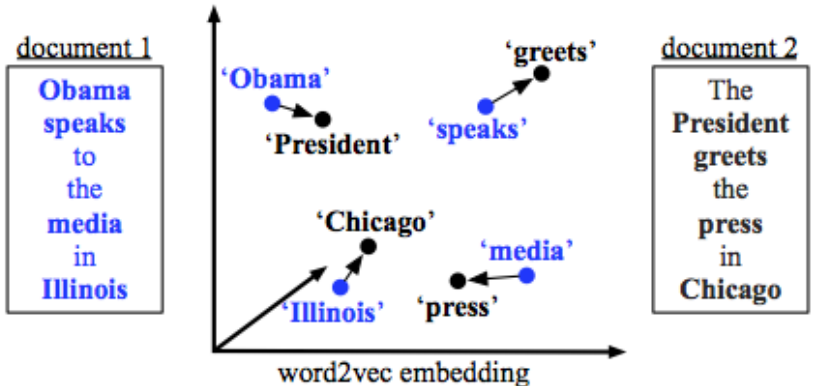


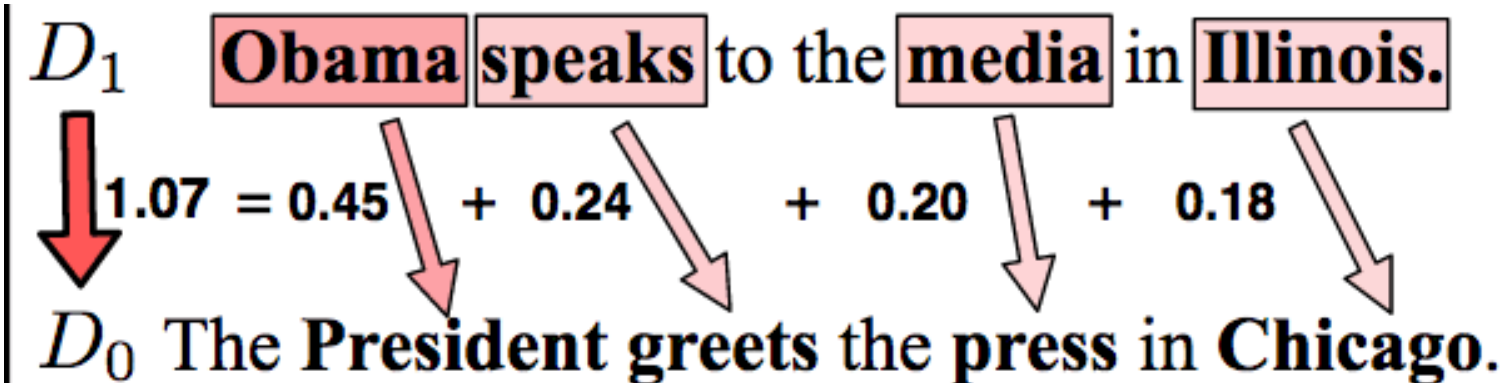


KUSNER, Matt et al. From word embeddings to document distances. In: International conference on machine learning. 2015. p. 957-966.

Representação dos Textos

- Pré-processamento – Informação Textual
  - *Word Embedding Models*





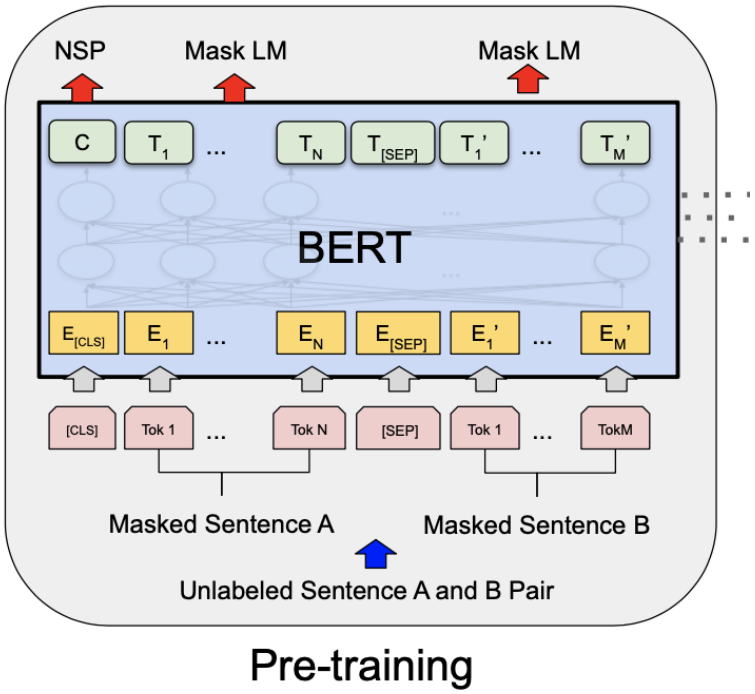
KUSNER, Matt et al. From word embeddings to document distances. In: International conference on machine learning. 2015. p. 957-966.

Representação dos Textos

- Pré-processamento – Informação Textual
  - *Word Embedding Models Livres de Contexto*
    - *Word2Vec*, *FastText*, *Glove*
    - Após o treinamento, os word vectors são estáticos
    - Exemplo
      - *Eu sentei no banco da praça*
      - *Eu fui no banco conferir o saldo*
  - Mais recentemente
    - *Word Embeddings* Contextuais
    - Exemplo: BERT (Bidirectional Encoder Representations from Transformers)

Representação dos Textos

- Pré-processamento – Informação Textual
  - *Word Embedding Models Contextuais*



Pre-training

- *Masked Language Model*
  - *Durante o treinamento, ~15% das palavras são mascaradas. O modelo tenta prever tais palavras.*
- *Next Sentence Prediction*
  - *Prever quando uma sentença é sucessora de outra sentença*

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Agrupamento de Textos

- Agrupamento com Word Embeddings
  - Exemplo prático
  - BERTopic (https://github.com/MaartenGr/BERTopic)