REPRESENTAÇÃO DE TEXTO

Tópicos especiais em gestão de TI.

PROF. DIEGO ROSSI

REPRESENTAÇÃO COMPUTACIONAIS DE TEXTO

- Computadores são capazes de processar apenas números
- Possuem uma certa dificuldade em tratar informações não estruturadas.
- Portanto, é preciso criar formas de representação de texto que os computadores sejam capazes de processar, ou seja, transformar informação não estruturada em estruturada.

FEATURE EXTRACTION

- Para usarmos um modelo estatístico ou de deep learning em NLP, precisamos de features: informações mensuráveis acerca de algum fenômeno, ou seja, uma forma estruturada de armazenar informações.
- Porém, textos são um tipo de dado não estruturado (não organizado de uma maneira pré-definida, fixa), assim, é difícil para o computador entendê-los e analisá-los.
- Por isso, realizamos a chamada feature extraction, ou seja, transformamos o texto em uma informação numérica de modo que seja possível utilizá-lo para alimentar um modelo.

BAG OF WORDS (BOW)

- BoW é uma forma de representar o texto de acordo com a ocorrência das palavras nele.
- Traduzindo para o português, o "saco de palavras" recebe esse nome porque não leva em conta a ordem ou a estrutura das palavras no texto, apenas se ela aparece ou a frequência com que aparece nele.
- Portanto, BoW pode ser um ótimo método para determinar as palavras significativas de um texto com base no número de vezes que ela é usada.
- Problemas:
  - Documentos longos possuem um peso maior do que documentos menores
  - Palavras muito frequentes "dominam" o documento e podem não representar tanta informação.

BAG OF WORDS (BOW)

- Basicamente, para gerar um modelo de bag of words precisamos realizar três passos:
  - 1) Selecionar os dados
  - 2) Gerar o vocabulário
  - 3) Formar vetores a partir do documento

BAG OF WORDS (BOW)

- Exemplo
  - Este é o primeiro documento
  - Este documento é o segundo documento

este	é	o	primeiro	documento	segundo
1	1	1	1	1	0
1	1	1	0	2	1

TF-IDF (TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY)

- A ideia é reajustar o peso das palavras a medida que elas aparecem em todos os documentos.
- Term Frequency (TF): Frequência da palavra no documento atual.
  - TF= (n° de vezes que o termo aparece no documento) / (número de termos no documento)
- Inverse document Frequency (IDF): quão rara é a palavra nos documentos.
  - IDF = log(n° de documentos / n° de documentos que possuem o termo)

TF-IDF (TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY)

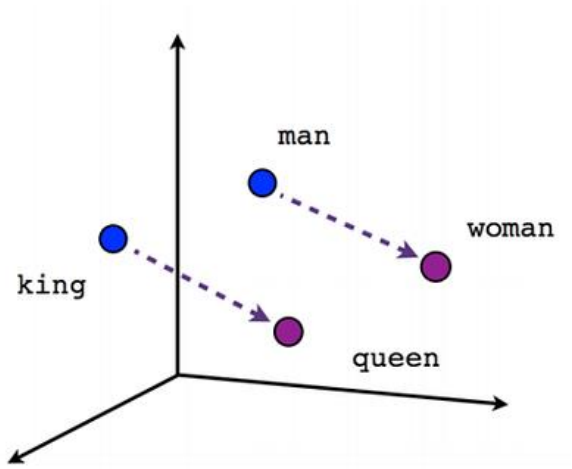
- Exemplo:
- Considerando que um documento contém 100 palavras e a palavra cachorro aparece 5 vezes.
- TF = 5 / 100 = 0,05
- No total de 100 documentos a palavra documento aparece em 20 desses documentos
- IDF = log(100/20) = 0,69
- TF-IDF = 0,05 * 0,69 = 0,034

- Portanto, quanto maior o valor do peso, mais raro é o termo;

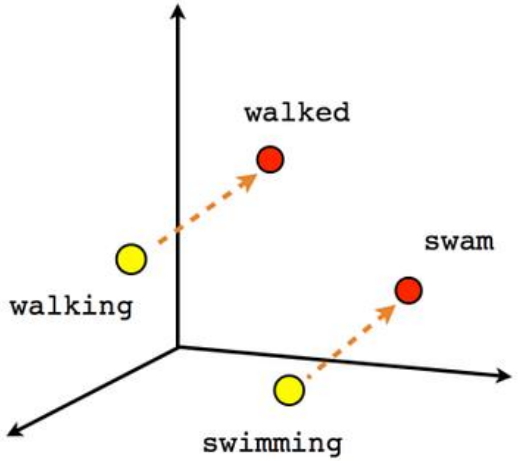
WORD EMBEDDINGS

- One Hot Encoding – Maneira primitiva de representar palavras, onde são utilizados símbolos discretos.
  - Exemplo: Hotel = [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]
  - Pousada = [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
- Semântica distributiva – Representar as palavras através do contexto.
- Uma maneira de representar palavras em um espaço vetorial.
  - Ex.: Alemanha = [0.286 0.792 -0.177 -0.107 0.109 -0.542 0.349 0.271]
- Frameworks para aprender word embeddings – Word2vec, Glove, BERT

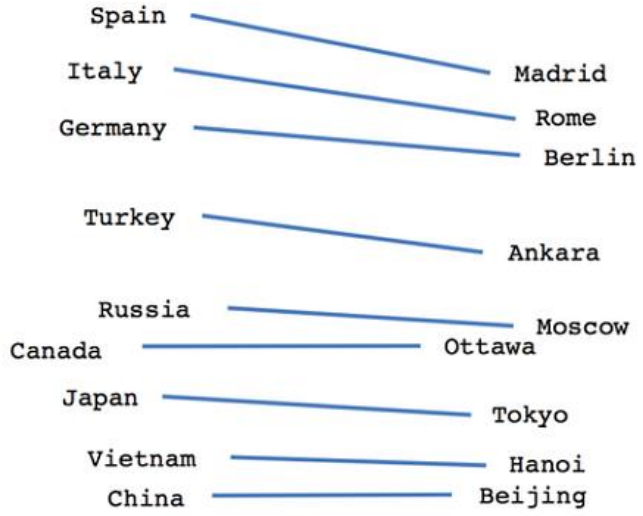
WORD EMBEDDINGS



Male-Female



Verb tense



Country-Capital