

Processamento de Linguagem Natural



Processamento de língua natural é uma subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais. Sistemas de geração de língua natural convertem informação de bancos de dados de computadores em linguagem compreensível ao ser humano e sistemas de compreensão de língua natural convertem ocorrências de linguagem humana em representações mais formais, mais facilmente manipuláveis por programas de computador. Alguns desafios do PLN são compreensão de língua natural, fazer com que computadores extraiam sentido de linguagem humana ou natural e geração de língua natural.



自然言語処理は、自動生成と自然な人間の言語の理解の問題を研究するコンピュータ科学、人工知能と言語学のサブ領域です。自然言語生成システムはコンピュータデータベースからの情報を人間が理解できる言語に変換し、自然言語理解システムは人間の言語の出現をより正式な表現に変換し、コンピュータプログラムによってより容易に操作される。 PLNの課題のいくつかは、自然言語の理解、コンピュータによる人間または自然言語からの意味の理解、および自然言語の生成です。

Para que fazer os computadores entender texto?

Mineração



30 bilhões de páginas

Google

bing™

Porque o Facebook comprou o WhatsApp?

“O WhatsApp está a caminho de conectar 1 bilhão de pessoas. Os serviços que alcançam essa marca são todos incrivelmente valiosos” – Mark Zuckerberg



Aplicações

Facebook

<https://www.globaldatinginsights.com/news/can-facebook-predict-the-start-of-a-relationship/>

Facebook

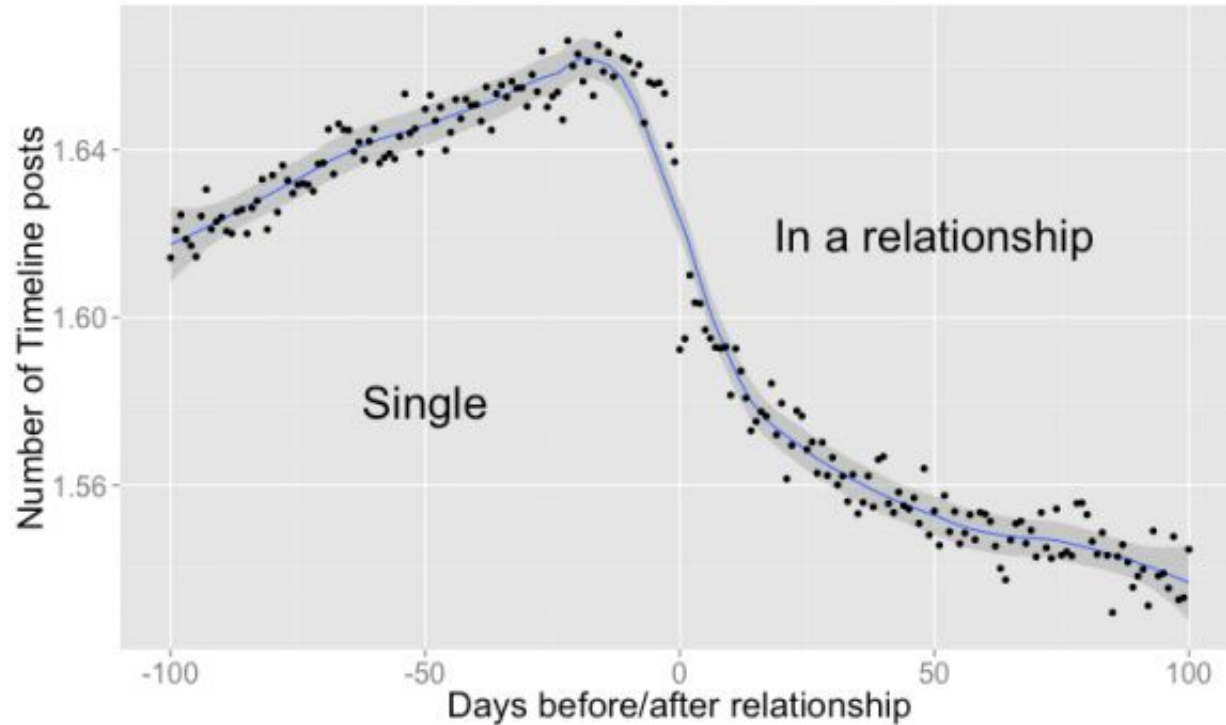
O Facebook sabe quando você vai começar a namorar

Apenas usando as interações dentro da rede

Só usando as informações da timeline



Facebook



Facebook

Depois de mudado o status as postagens tendem a diminuir, mas...

Palavras positivas como **feliz** e **amor** aumentam consideravelmente



Predição em textos jurídicos

Predição

IA prediz resultados de processos de direitos humanos

As decisões da corte europeia de direitos humanos foram preditas com uma acurácia de 79% usando métodos de inteligência artificial

<https://www.ucl.ac.uk/news/2016/oct/ai-predicts-outcomes-human-rights-trials>



Aplicações

Casos citados

Pessoas envolvidas - partes, advogados


Causa legal

Natureza do caso

Decisão

<https://medium.com/datadriveninvestor/where-are-we-in-legal-text-processing-with-artificial-intelligence-2eac562>

Ross Intelligence




- Search
- Document Analyzer
- Statutes & Regulations
- History
- Folders

Alex Abid

Search using... **Natural Language** · Terms & Connectors · Citation or Name All Federal & State

What is the distinction between independent contractors and employees?

Add... Motion Facts Exclude Terms Show Sample



IBM Watson

LegalMation

Helping legal teams draft high-quality litigation work in minutes and drive down costs by 80 percent

With intuitive IBM® Watson® offerings, LegalMation developed a first-of-its-kind AI platform to automate routine litigation tasks. Supported by the IBM Watson ecosystem, the company quickly launched its solution for drafting early phase response documents, helping legal teams save time, drive down costs and shift strategic focus.

Processamento de Linguagem Natural

Processamento de língua natural é uma subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais. Sistemas de geração de língua natural convertem informação de bancos de dados de computadores em linguagem compreensível ao ser humano e sistemas de compreensão de língua natural convertem ocorrências de linguagem humana em representações mais formais, mais facilmente manipuláveis por programas de computador. Alguns desafios do PLN são compreensão de língua natural, fazer com que computadores extraiam sentido de linguagem humana ou natural e geração de língua natural.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	8	0	9	4	2	5	2	5	8	2	4	7	1	3	4	7	7	4	3	3	3	6	2	0	1	8	9	7	2	1	3	4
2	3	5	6	3	2	1	9	8	8	2	1	1	9	0	4	5	2	6	1	8	2	7	5	1	2	6	2	7	1	0	9	5
3	1	3	3	0	6	3	3	1	3	7	5	3	9	8	9	3	8	7	3	8	6	8	1	5	1	3	3	8	8	5	4	3
4	3	6	6	5	0	0	1	6	2	2	4	3	6	4	3	2	4	7	9	6	6	0	9	5	5	2	0	3	1	6	2	0
5	7	6	5	0	5	9	2	5	5	5	9	8	7	3	1	1	2	1	8	2	4	6	4	5	3	5	3	0	5	5	8	9
6	4	4	9	0	5	4	1	7	9	7	2	7	6	1	5	3	5	9	0	1	4	8	7	8	9	9	8	0	9	8	7	7
7	6	5	4	5	9	1	0	4	9	3	1	8	8	5	1	9	7	5	3	7	2	7	8	5	9	3	7	3	2	4	4	5
8	3	6	2	6	5	9	9	5	1	2	1	5	9	7	5	3	9	2	2	3	5	6	5	8	2	9	4	4	2	8	9	9
9	4	6	6	5	4	8	2	0	7	5	5	4	0	6	1	2	9	6	8	3	4	2	5	1	9	1	3	8	1	7	0	9
10	6	4	9	8	7	5	1	9	0	4	7	4	7	8	1	8	6	6	3	2	9	6	8	3	9	8	7	2	4	0	9	0
11	6	7	2	2	9	8	6	9	9	3	6	1	7	6	7	5	4	8	8	3	1	3	1	5	9	6	7	9	8	8	3	4
12	9	7	4	8	5	9	3	2	5	1	1	5	2	7	2	1	0	0	3	3	9	3	0	3	9	7	1	3	4	0	1	2
13	5	6	4	1	1	4	1	7	1	4	1	9	7	4	3	4	8	1	6	5	7	3	6	8	1	2	1	8	5	0	3	9
14	7	4	4	4	9	2	0	0	8	8	4	0	5	8	8	2	4	3	9	8	3	9	0	4	9	1	9	9	9	3	3	6
15	8	2	7	9	3	0	1	9	4	6	7	2	3	7	4	3	3	9	7	9	4	6	8	9	9	0	2	1	6	9	9	0
16	0	1	6	1	7	6	1	7	1	0	2	4	2	3	8	7	2	8	9	1	6	6	7	7	1	5	8	5	2	4	8	2
17	7	3	8	8	9	7	5	9	7	5	5	5	6	6	2	4	9	9	7	7	2	0	0	8	5	5	9	6	9	7	4	0
18	7	8	3	0	4	7	1	4	3	6	9	6	2	9	1	9	1	8	0	4	4	0	4	4	1	0	3	4	2	5	9	7
19	9	8	8	7	4	2	1	6	6	5	2	6	4	5	3	5	8	4	3	0	5	2	7	0	9	6	0	5	0	7	8	8
20	1	2	6	1	2	5	1	6	8	5	6	9	2	3	1	0	9	9	9	9	6	7	0	3	9	8	4	1	0	3	5	3
21	3	9	4	7	4	9	3	7	7	6	3	4	2	5	4	3	6	2	3	9	7	4	5	5	2	0	5	5	7	7	9	5
22	4	5	5	0	8	1	0	3	1	2	5	0	2	3	0	4	1	1	3	8	9	7	8	8	9	1	4	4	4	5	2	6
23	1	3	4	4	9	6	9	7	2	3	8	3	6	9	7	6	6	2	5	1	4	2	0	1	2	0	3	6	6	5	5	2
24	8	9	7	6	5	8	2	3	8	4	8	7	0	4	5	0	3	1	0	6	9	1	6	5	2	7	1	7	7	6	0	1
25	7	7	1	0	9	9	4	3	6	9	7	8	6	2	7	3	9	7	1	4	9	7	0	0	1	5	6	6	2	8	6	9
26	8	9	5	9	6	0	0	8	8	4	4	2	2	2	8	2	1	5	2	4	2	5	1	7	5	8	1	8	0	0	8	1
27	7	9	4	1	2	3	1	2	2	4	3	1	6	7	0	2	9	9	8	4	3	4	6	9	3	0	2	5	4	7	6	2
28	2	2	8	4	0	8	9	6	9	1	0	7	5	5	4	2	7	3	1	9	3	7	8	2	1	0	8	9	5	7	4	4
29	9	5	9	4	7	4	1	6	9	3	6	5	6	0	4	5	1	1	6	3	5	9	1	6	9	5	9	9	1	1	4	3
30	4	6	1	3	8	5	4	9	6	3	6	9	3	2	0	8	5	1	0	9	9	6	8	0	1	1	6	8	6	1	3	3

Bag-of-Words (BoW)

Representa **o que** está sendo tratado no documento e não **como** está sendo tratado

Essa representação considera que as informações mais importantes do documento são as palavras estão sendo usadas

- Isso ignora, por exemplo, a ordem das palavras e informações de discurso



Bag-of-Words (BoW)

Cada documento é representado como um **vetor de palavras**, onde cada entrada é uma palavra e o valor é **quantas vezes** aquela palavra apareceu no documento

Exemplo:

- Eu gosto de gato
- Eu gosto de cachorro

Documento	Eu	gosto	gato	cachorro
1	1	1	1	0
2	1	1	0	1

Bag-of-Words (BoW)

Esses vetores de palavras podem ser usados para qualquer tipo de análise

- Regressão e classificação
- Agrupamento
- Visualização

Esses vetores são simples de interpretar, contudo:

- Esses vetores são grandes, então podem ser lentos
- Esses vetores são bem esparsos, podendo gerar overfitting



Limpeza de Dados

Análise ortográfica

Corrige erros ortográficos e palavras escritas de forma diferente

Funciona bem na maioria dos casos, mas pode trazer erros com nomes próprios e palavras de outros idiomas

A não ser que esteja se trabalhando com redes sociais, erros não deveriam ser tão comuns



Normalização

Converter o texto para maiúsculo ou minúsculo - em geral se usa minúsculo

O curso sobre PLN foi apresentado por Rafael.

o curso sobre pln foi apresentado por rafael.



Tokenização

Divide o texto em palavras

Em geral, pontuação também é considerada uma palavra

o curso sobre pln foi apresentado por rafael.

{o,curso,sobre,pln,foi,apresentado,por,rafael,.}



Remoção de stopwords

Palavras extremamente comuns e que não tem muito significado para o texto

Existem várias lista pré-definidas de stopwords em diferentes idiomas

Artigos e preposições também são consideradas stopwords

Palavras muito comuns para a coleção analisada

{o,curso,sobre,pln,foi,apresentado,por,rafael,.}

{curso,pln,foi,apresentado,rafael}



Stemming/Lemmatization

Remove a flexão da palavra - plural, verbo conjugado, entre outros

Muda a palavra para a forma mais primitiva

{curso,pln,foi,apresentado,rafael}

{curso,pln,ser,apresentar,rafael}



Mineração de texto

Mineração de texto

Encontrar documentos relevantes

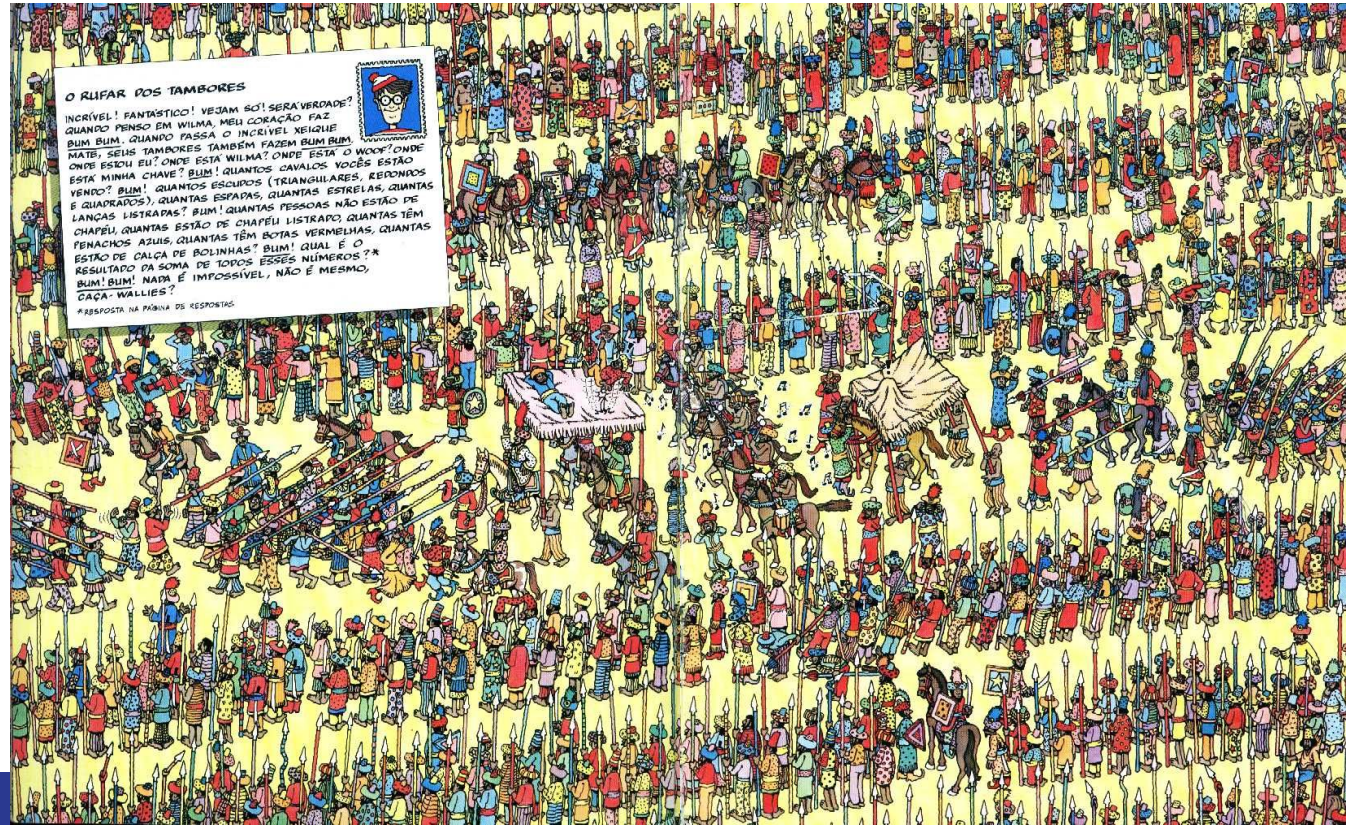
Categorizar documentos

Extrair informação específica dos textos

Sumarização de texto



Recuperação de informação



Categorização



Categorização

Europeus



Asiáticos



Americanos



Categorização

Desportivos



Não desportivos



Extração de informação

Texto Livre

4 de abril em Dallas – cedo na noite passada, um tornado varreu todo o noroeste da área de Dallas, causando extensos danos. Testemunhas confirmam que o ciclone passou sem advertência, aproximadamente às 7:15 da noite, e destruiu dois *motor-homes*. O posto Texaco, na Rua Principal, 102, Farmers Branch, TX, também foi severamente danificado, mas nenhuma morte foi informada. O valor total calculado dos danos é de U\$200.000.

Jornal



Sistema de Extração de Informações

Template

Evento: tornado
 Data: 4/4/2000
 Hora: 19:15
 Local: Farmers Branch : "noroeste de Dallas" : TX : USA
 Danos: "motor-homes" (2) : "Posto Texaco" (1)
 Perdas Estimadas: U\$200.000
 Mortes: nenhuma

Reconhecimento de entidade nomeadas

Reconhecimento de entidade nomeadas

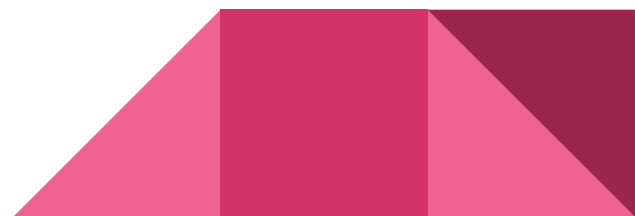
Neymar **PER** , principal jogador do **Brasil LOC** , atualmente defende o **PSG ORG** .

Reconhecimento de entidade nomeadas

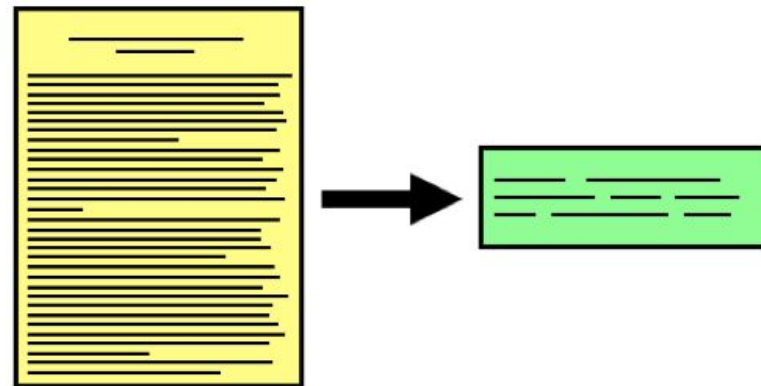
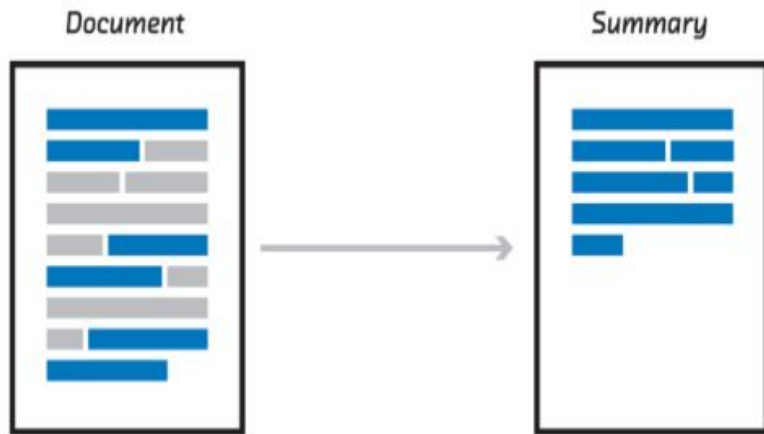
Os melhores métodos são aqueles que levam em consideração a classificação anterior

- HMM
- CRF
- LSTM

Principais classes:

- Pessoa
 - Lugar
 - Organização
 - Moeda
 - Data
- 

Sumarização de texto



Sumarização de texto

Declara e prova o autor, através de seu advogado constituído, que adquiriu passagem aérea da Cia Ré. Requer, nos termos do novo CPC que a acionada seja condenada a pagar a autora a quantia de R\$ 35.000,00. Requer ainda a condenação da Ré no pagamento de todas as despesas processuais e em honorários advocatícios. Dá-se à causa o valor de R\$ 35.000,00. Termos em que, Pede e espera Deferimento.

Declara e prova o autor, através de seu advogado constituído, que adquiriu passagem aérea da Cia Ré, (VOO 9179 – DE Recife para Salvador no dia 28.01.18) com horário de embarque para as 12:18h, conforme prova através do bilhete anexo. Ocorre que a passagem foi adquirida no horário transcrito, tendo em vista que o autor é Músico percussivo da banda "PARANGOLÉ", e realizaria um show na cidade de Taperoá-BA, conforme prova através do folder de divulgação anexado nos autos. Sucede que para sua surpresa, espanto e INDIGNAÇÃO o VOO atrasou por horas, sem qualquer justificativa. Dá-se à causa o valor de R\$ 35.000,00 (Trinta e cinco mil reais) Termos em que, Pede e espera Deferimento.

Sumarização de texto


[Artificial Intelligence Review](#)

March 2019, Volume 51, [Issue 3](#), pp 371–402 | [Cite as](#)

Text summarization from legal documents: a survey

Authors

[Authors and affiliations](#)

Ambedkar Kanapala , Sukomal Pal, Rajendra Pamula

Word Embeddings

Word Embeddings

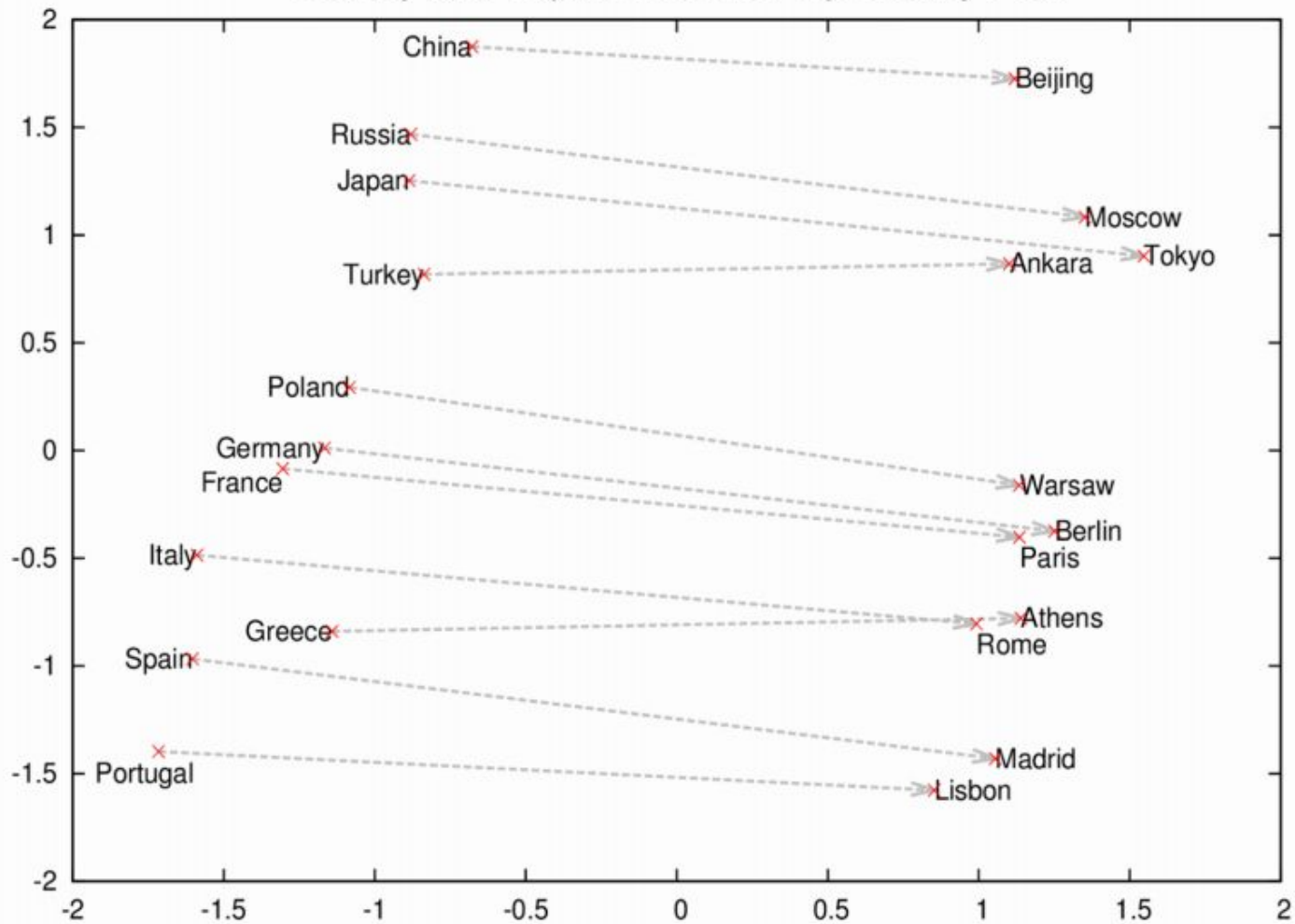
Técnica para transformar palavras ou sentenças em um vetor de números

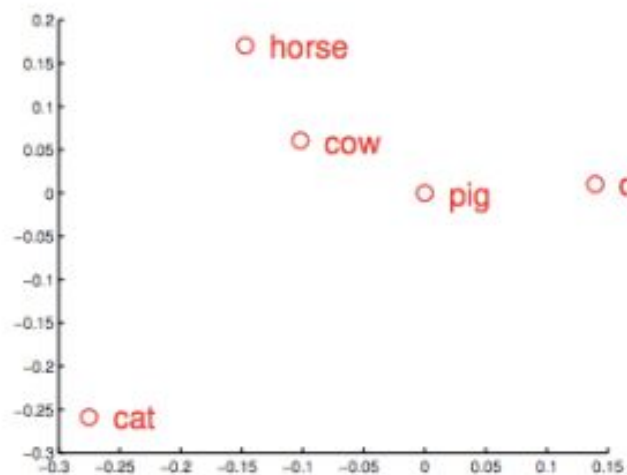
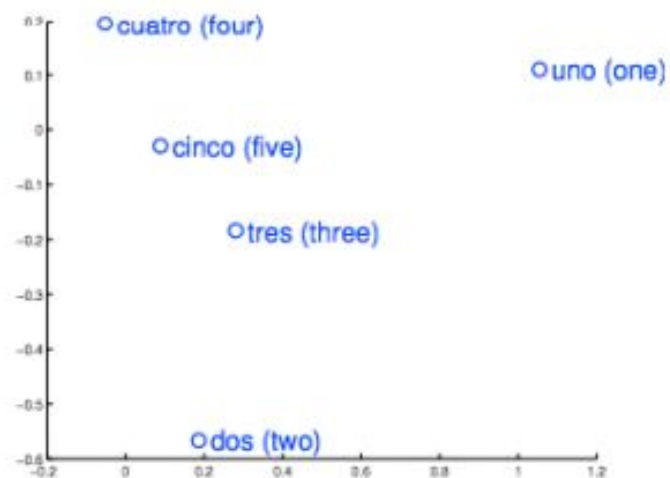
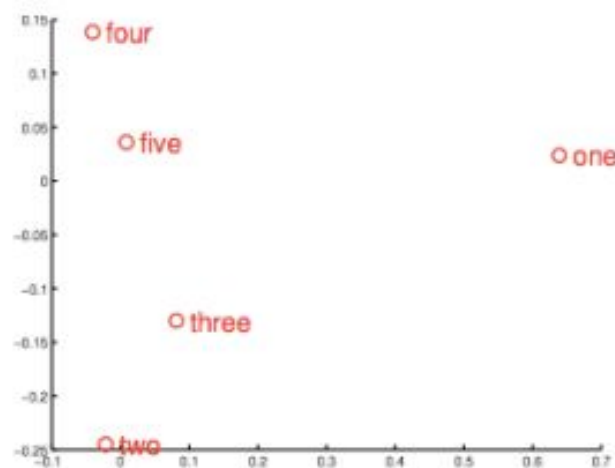
Geralmente diminui a dimensão de um vetor com uma grande quantidade de posições

Utiliza métodos de redes neurais, co-ocorrência de palavras, métodos probabilísticos



Country and Capital Vectors Projected by PCA





Processamento de Linguagem Natural