

Introdução ao Processamento de Linguagem Natural com Python

Prof. Aline Paes - alinepaes@ic.uff.br

Jéssica Soares - jsoares@ic.uff.br

Paulo Mann - paulomann@id.uff.br

Introdução ao Processamento de Linguagem Natural com Python

Disclaimer

Images: collected from Google Images, no intention of
harming the author's rights, illustrative purposes only

Agenda

- Visão Geral
- Técnicas de Pré-processamento
- Tarefas de NLP
- Hands On (NLTK e SpaCy)

Processamento de Linguagem Natural (NLP)

- Meta: desenvolver sistemas que possam se **comunicar** com as pessoas usando a linguagem que usamos todo dia (“linguagem natural”)
 - Fazer computadores **processarem e entenderem** a linguagem natural para executar tarefas úteis
 - texto ou fala
- Combina diferentes áreas como: ciência da computação, inteligência artificial e linguística

Componentes de (NLP)

- Natural Language Understanding (NLU)
 - mapeamento da linguagem natural para uma representação computacional
- Natural Language Generation (NLG)
 - geração de texto (linguagem natural)
 - direção oposta do NLU

Um (pequeno) conjunto de aplicações de NLP

- Verificação gramatical, busca por palavras chave, encontrar sinônimos
- Detecção de emails que são spam
- Sistemas de geração de textos e diálogos
 - Chatbots
- Tradução automática
- Recomendação de anúncio/propaganda (baseado no histórico do usuário)

Um (pequeno) conjunto de aplicações de NLP

- Geração automática de legenda
- Extrair informações de documentos, textos e websites, como produtos, datas, locais, pessoas, relações entre entidades
- Resposta a consultas
- Prever a próxima palavra de um texto
- Análise de sentimentos
- Reconhecimento de fala

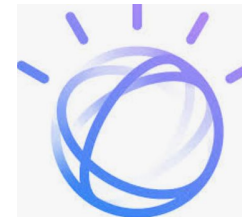
Aplicações de NLP na prática



Hi



TayTweets 
@TayandYou



IBM Watson



amazon alexa



Google Assistant



Aplicações de NLP por níveis de dificuldade

- Bem resolvidas
 - Detecção de Spam
 - "Taguear" partes da fala
 - Reconhecimento de Entidades Nomeadas

Aplicações de NLP por níveis de dificuldade

- Em um caminho de progresso
 - Desambiguação do sentido de palavras
 - Análise de sentimento
 - Tradução Automática
 - Geração de texto

Aplicações de NLP por níveis de dificuldade

- Ainda precisa de mais trabalho
 - Sumarização de texto
 - Diálogo
 - Inferência
 - Extração de Informação
 - Resposta a consultas

Técnicas de Pré-processamento

- Limpeza de dados
- Tokenização
- Stemming
- Lemmatization
- POS-Tagging
- Named Entity Recognition (NER)
- Chunking

Limpeza de dados

- Remoção de caracteres especiais, pontuação, números
- Remoção de Stopwords
 - Palavras sem significado semântico isoladamente

Ex: ['a', 'ao', 'aos', 'aquela', 'aquelas', 'aquele',
'aqueles', 'aquilo', 'as', 'é']

- A limpeza de dados deve levar em conta o problema que estamos querendo resolver

Tokenização

- Quebrar uma string de caracteres em pequenos pedaços (tokens)
 - palavras
 - frases
- Em algumas línguas escritas (e.g. Chinês), as palavras não são separadas por espaços

Tokenização

- N-grams
 - Unigrams: cada token é uma única palavra.
 - Bi-grams: cada token é composto por duas palavras.

Ex: Eu estou lendo um livro.

Unigrams: “Eu”, “estou”, “lendo”, “um”, “livro”, “.”

Bigrams: “Eu estou”, “estou lendo”, “lendo um”,

Normalização de palavras

- Redução ou simplificação ou radicalização de palavras
 - Stemming
 - Lemmatization

Stemming

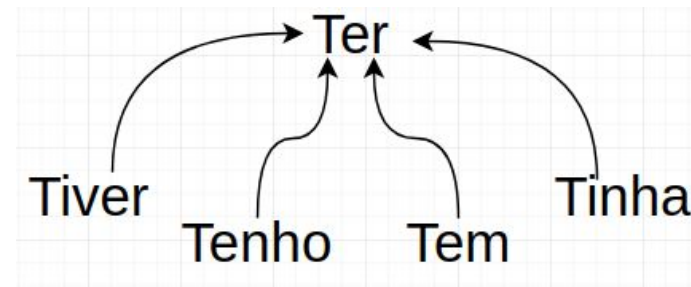
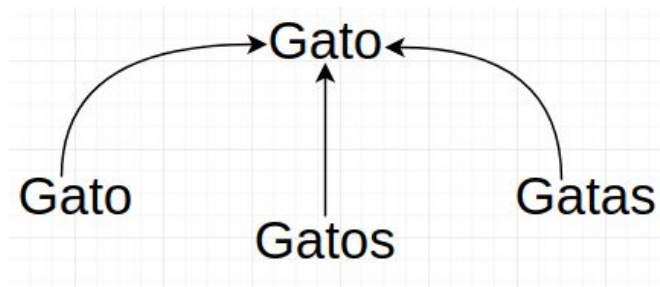
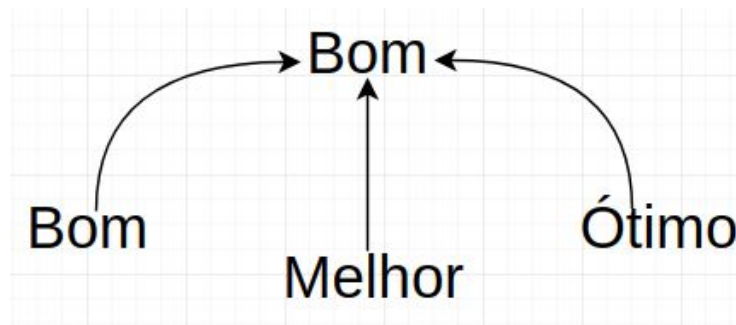
- Reduzir a palavra a seu radical (sem levar em conta a classe gramatical)



- Usa uma heurística ou um conjunto de regras que dependem da linguagem
- Geralmente, corta as extremidades (afixos)

Lematização

- Reduzir a palavra ao seu lema, que é a forma básica da palavra
- Utiliza dicionários e realiza a análise morfológica das palavras

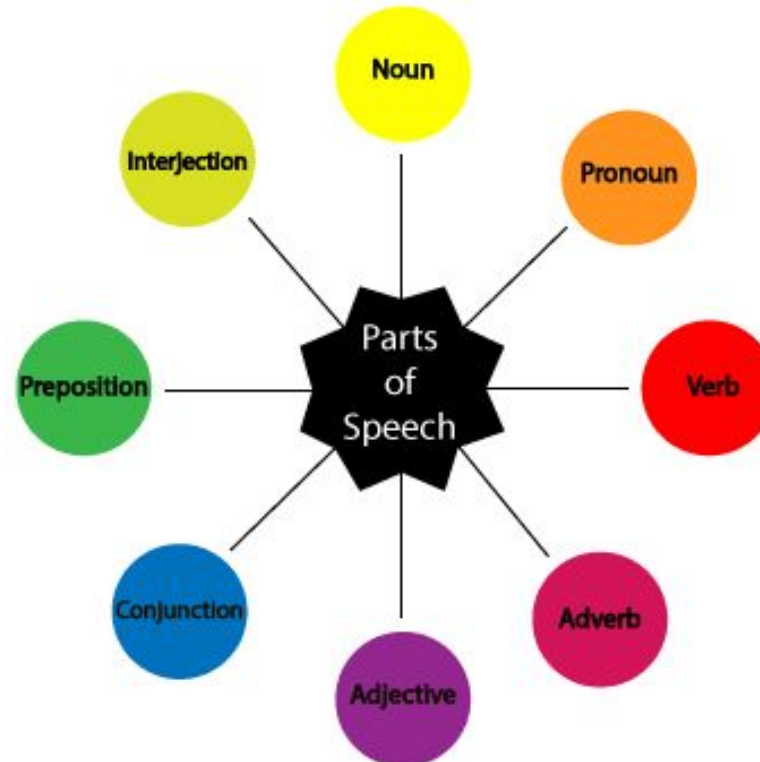


Stemming x Lematização

- O resultado do stemming não necessariamente é uma palavra válida na linguagem
 - ausência de significado
- O resultado da lematização é uma palavra válida na linguagem (existente no dicionário)
- Alguns buscadores tratam palavras com o mesmo radical/lema como sinônimos como um tipo de expansão de consulta na tentativa de retornar mais resultados relevantes

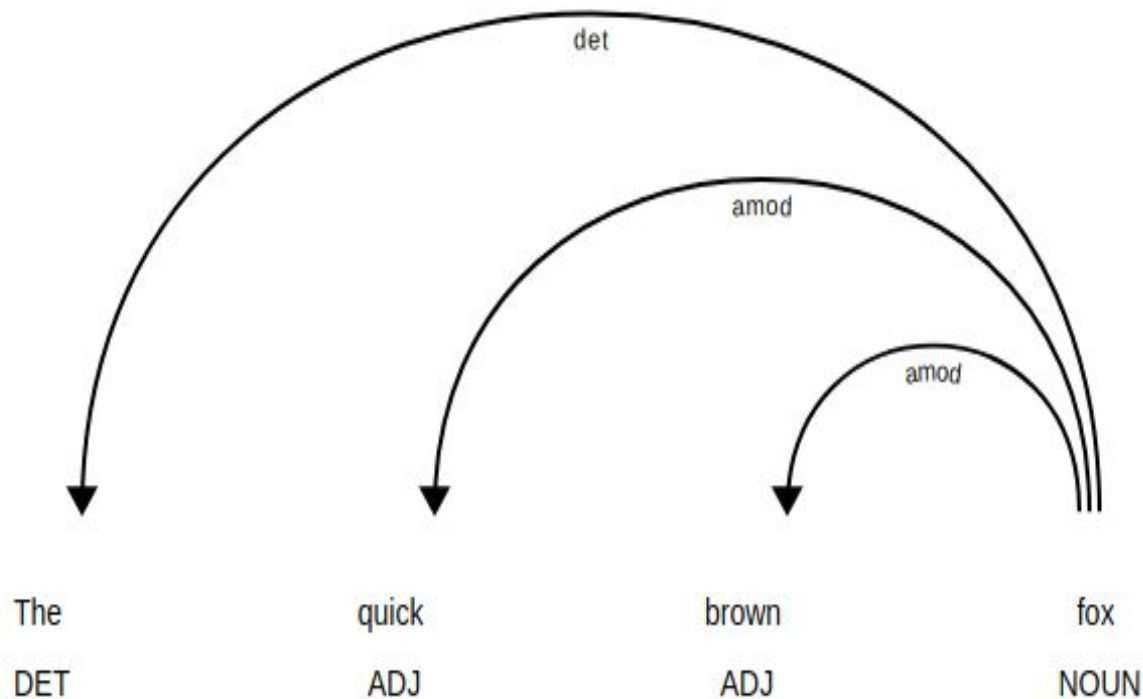
Part-of-speech (POS-tag)

- Anotar cada palavra com um tipo gramatical
 - verbo, substantivo, adjetivo, advérbio, artigo, etc.



Part-of-speech (POS-tag)

- Determina o papel da palavra em uma frase



Part-of-speech (POS-tag)

- Em alguns casos as POS tags são divididas em subclasses
- Uma palavra pode ter mais de um POS (dependendo do contexto)
 - Ex: “Google” something on the internet

Reconhecimento de Entidade Nomeadas (NER)

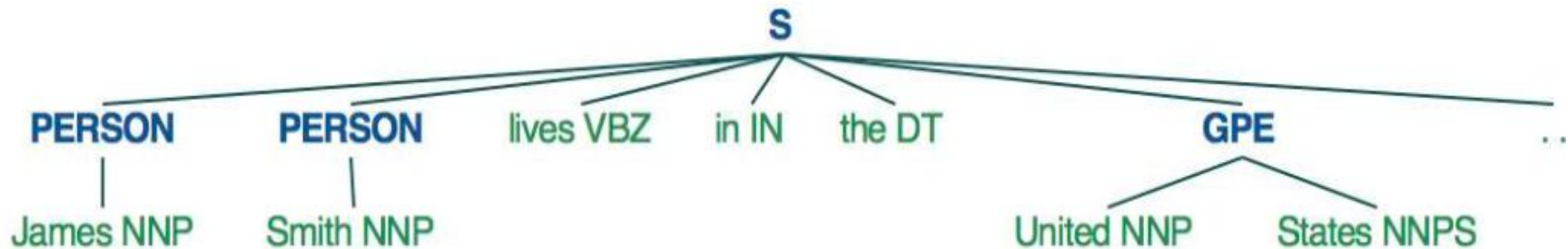
- Detectar entidades nomeadas no texto
 - pessoas
 - organizações
 - lugares
 - datas
 - ...

Ex: Michael Dell is the CEO of Dell Computer Corporation
and lives in Austin Texas.

people organizations places

Chunking

- agrupar tokens em chunks
- utiliza pos-tagging como entrada



Um pouco de história

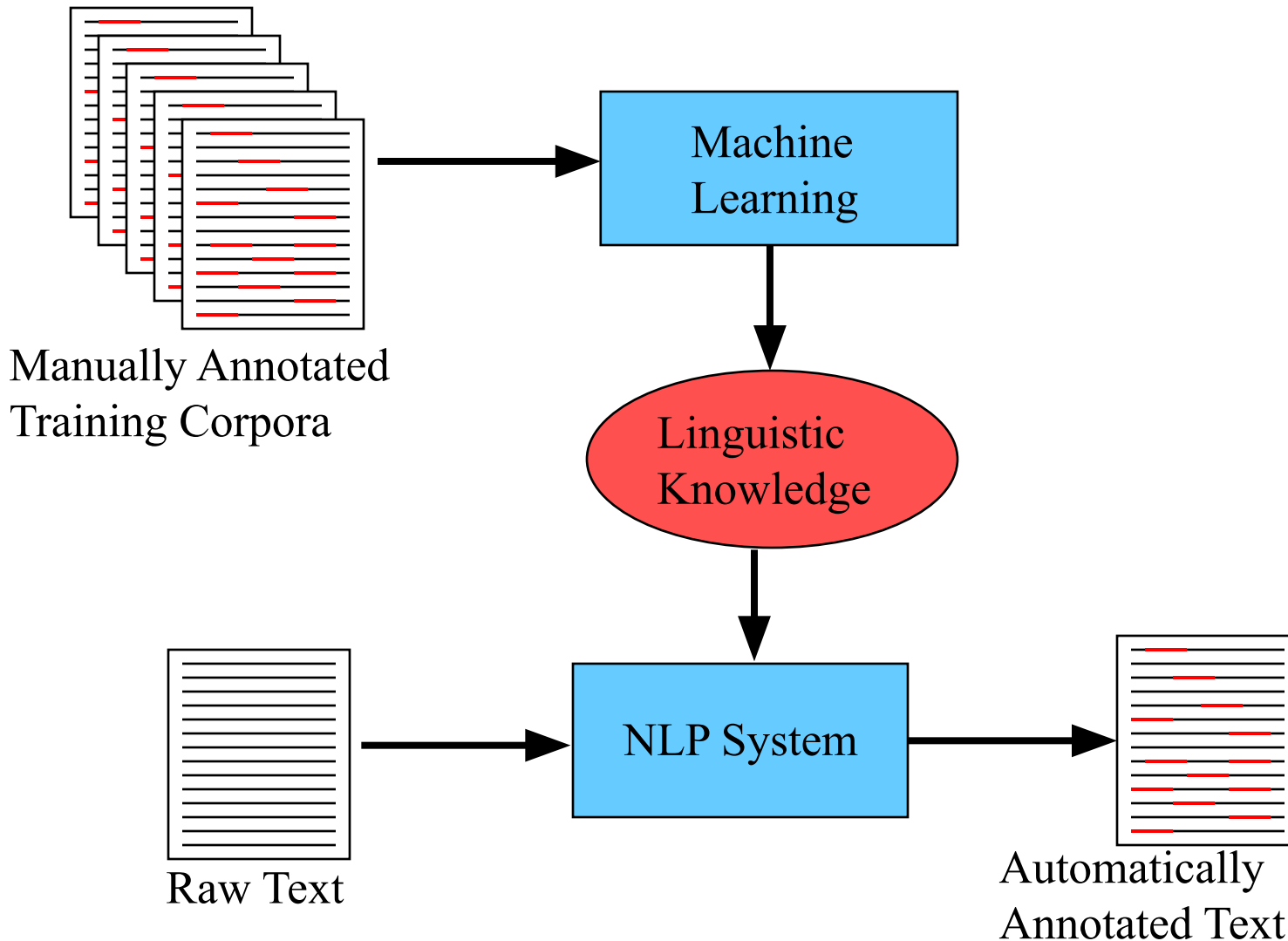
- Entre as décadas de 50 e 70, NLP consistia no uso de métodos linguísticos e regras predefinidas
- Métodos estatísticos se intensificaram a partir do fim da década de 70
- Na década de 90, abordagens de **aprendizado automático** se expandiram e dominaram quase todas as áreas de NLP
 - uso de corpora de textos anotados para adquirir de forma automática o conhecimento do problema.

Um pouco de história

- Rule Based Systems → Machine Learning → Deep Learning



Abordagem de Aprendizado



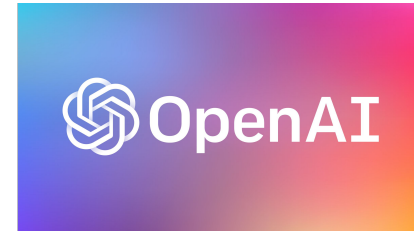
Aplicações de NLP

- Resolução de co-referência
 - Determinar que estruturas em um documento se referem a mesma entidade
 - John put the carrot on the plate and ate it.
 - Bush started the war in Iraq. But the president needed the consent of Congress.

Aplicações de NLP

- Sumarização de texto
 - Produzir um sumário a partir de um texto
 - **Article:** With a split decision in the final two primaries and a flurry of superdelegate endorsements, [Sen. Barack Obama](#) sealed the Democratic presidential nomination last night after a grueling and history-making campaign against [Sen. Hillary Rodham Clinton](#) that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against [Sen. John McCain](#), the presumptive Republican nominee....
 - **Summary:** Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Aplicações de NLP



- Geração de linguagem
 - Construir um texto a partir de poucas premissas ou de uma introdução

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

1º Parágrafo!

7º Parágrafo!

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Referência: <https://openai.com/blog/better-language-models/>

Aplicações de NLP

- Agentes de conversa / geração de diálogo

Baseline mutual information model (Li et al. 2015)	Proposed reinforcement learning model
<p>A: Where are you going? (1) B: I'm going to the restroom. (2) A: See you later. (3) B: See you later. (4) A: See you later. (5) B: See you later. (6) </p>	<p>A: Where are you going? (1) B: I'm going to the police station. (2) A: I'll come with you. (3) B: No, no, no, no, you're not going anywhere. (4) A: Why? (5) B: I need you to stay here. (6) A: I don't know what you are talking about. (7) ...</p>
<p>A: how old are you? (1) B: I'm 16. (2) A: 16? (3) B: I don't know what you are talking about. (4) A: You don't know what you are saying. (5) B: I don't know what you are talking about . (6) A: You don't know what you are saying. (7) ...</p>	<p>A: How old are you? (1) B: I'm 16. Why are you asking? (2) A I thought you were 12. (3) B: What made you think so? (4) A: I don't know what you are talking about. (5) B: You don't know what you are saying. (6) </p>

Aplicações de NLP

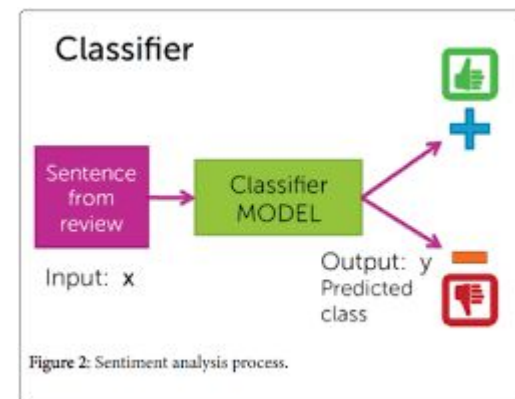
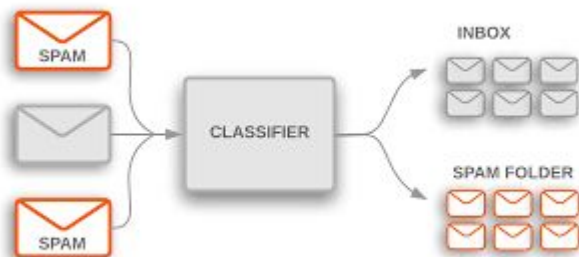
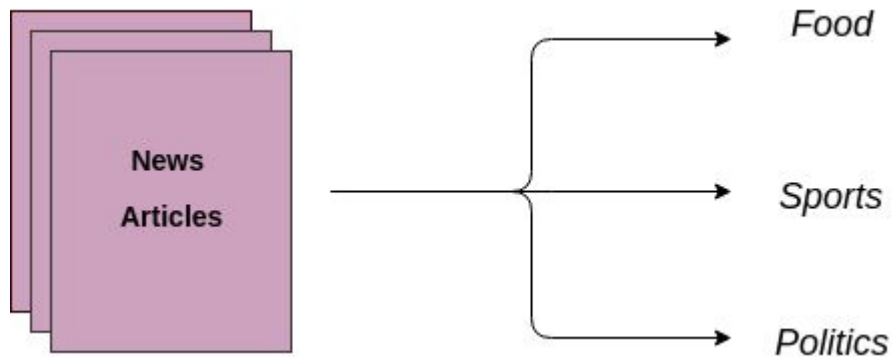
- Tradução automática
 - Traduzir uma sentença de uma linguagem para outra
- Hasta la vista, bebê ⇒
Until we see each other again, baby.

<i>Input sentence:</i>	<i>Translation (PBMT):</i>	<i>Translation (GNMT):</i>	<i>Translation (human):</i>
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Aplicações de NLP

- Tradução automática - História!
- **1990-2010:** Statistical Machine Translation (SMT)
- **2014:** Neural Machine Translation (NMT)
- **2016:** Google troca **SMT** por **NMT**
 - Mecanismo de Atenção!
- Vantagens x Desvantagens?

Tarefa com Aprendizado de Máquina: classificação de texto



Mas e as Features?

Aprendizado de Máquina Clássico X NLP

Aprendizado de Máquina: Tarefa supervisionada

Sepal length	Sepal width	Petal length	Petal width	Species
6.7	3.0	5.2	2.3	Virginica
6.4	2.8	5.6	2.1	Virginica
4.6	3.4	1.4	0.3	Setosa
6.9	3.1	4.9	1.5	Versicolor
4.4	2.9	1.4	0.2	Setosa
4.8	3.0	1.4	0.1	Setosa
5.9	3.0	5.1	1.8	Virginica
5.4	3.9	1.3	0.4	Setosa
4.9	3.0	1.4	0.2	Setosa
5.4	3.4	1.7	0.2	Setosa

Target

Aprendizado de Máquina: Tarefa supervisionada

Features 

Sepal length	Sepal width	Petal length	Petal width	Species
6.7	3.0	5.2	2.3	Virginica
6.4	2.8	5.6	2.1	Virginica
4.6	3.4	1.4	0.3	Setosa
6.9	3.1	4.9	1.5	Versicolor
4.4	2.9	1.4	0.2	Setosa
4.8	3.0	1.4	0.1	Setosa
5.9	3.0	5.1	1.8	Virginica
5.4	3.9	1.3	0.4	Setosa
4.9	3.0	1.4	0.2	Setosa
5.4	3.4	1.7	0.2	Setosa

Aprendizado de Máquina: Tarefa supervisionada

Examples →

Sepal length	Sepal width	Petal length	Petal width	Species
6.7	3.0	5.2	2.3	Virginica
6.4	2.8	5.6	2.1	Virginica
4.6	3.4	1.4	0.3	Setosa
6.9	3.1	4.9	1.5	Versicolor
4.4	2.9	1.4	0.2	Setosa
4.8	3.0	1.4	0.1	Setosa
5.9	3.0	5.1	1.8	Virginica
5.4	3.9	1.3	0.4	Setosa
4.9	3.0	1.4	0.2	Setosa
5.4	3.4	1.7	0.2	Setosa

Classificação de texto com Bag of Words

- Cada token ou chunking se torna uma feature
- O valor do atributo será
 - booleano
 - contagem de frequência absoluta
 - contagem ponderada
 - contagem normalizada

Vocabulary

0	0	0	0	0	0	0	0
are	cat	dog	is	now	on	table	the

Document 1

the dog is on the table

0	0	1	1	0	1	1	2
are	cat	dog	is	now	on	table	the



Document 2

the cat is on the table now

0	1	0	1	1	1	1	2
are	cat	dog	is	now	on	table	the

Classificação de texto com Bag of Words

Features!

Document	are	cat	dog	is	now	on	table	the
	0	0	1	1	0	1	1	2
	0	1	0	1	1	1	1	2

Problemas ao utilizar somente a frequência?

- A relevância de uma palavra != frequência
 - “os”, “as”, “do” ...
 - “indulgente”



Classificação texto - frequência do termo

- Termos mais frequentes em um documento são mais indicativos de um tópico
 - f_{ij} = frequência do termo i no documento j
- Frequência normalizada:
 - $tf_{ij} = f_{ij} / \max_i \{f_{ij}\}$

Classificação texto

- Termos que aparecem em muitos documentos distintos são menos indicativos do tópico
 - df_i = número de docs contendo termo i
 - $idf_i = \log_2(N/df_i)$

Classificação texto

- TF-IDF
- Combina a "importância" do termo com o inverso da frequência
 - $w_{ij} = tf_{ij}idf_i = tf_{ij} \log_2(N/df_i)$
- Um termo ocorrendo frequentemente no documento, mas raramente no resto da coleção, recebe peso mais alto

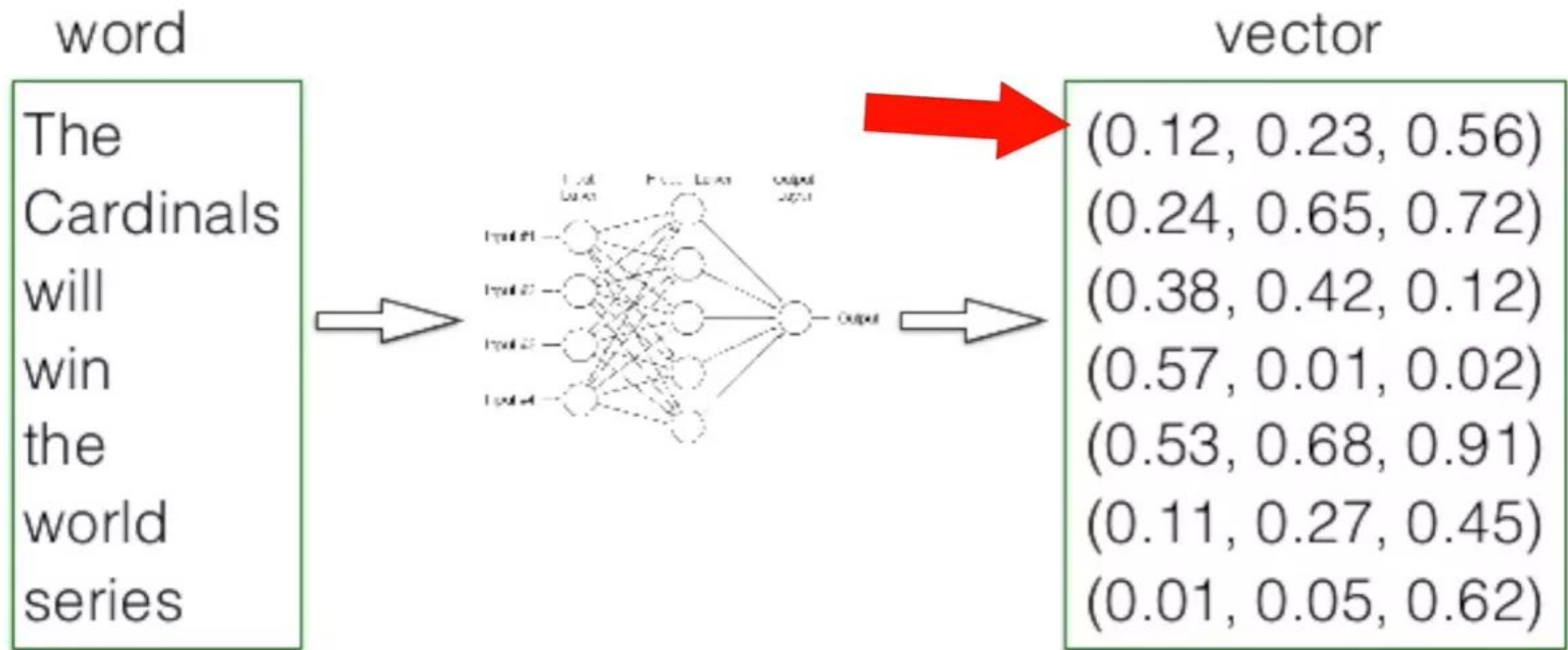
Classificação texto

Document	are	cat	dog	is	now	on	table	the
	0.0	0.0	0.47	0.33	0.0	0.33	0.33	0.67
	0.0	0.47	0.0	0.30	0.42	0.30	0.30	0.60

Deep NLP

- Combina ideias e metas de NLP com aprendizado de representações e métodos de Deep Learning
- Vários breakthroughs nos últimos anos com diferentes tarefas de NLP
 - Níveis: fala, palavras, sintaxe, semântica
 - Ferramentas: POS, NER, Parsing
 - Aplicações: tradução automática, análise de sentimento, agentes de diálogo, QA

Representações vetoriais

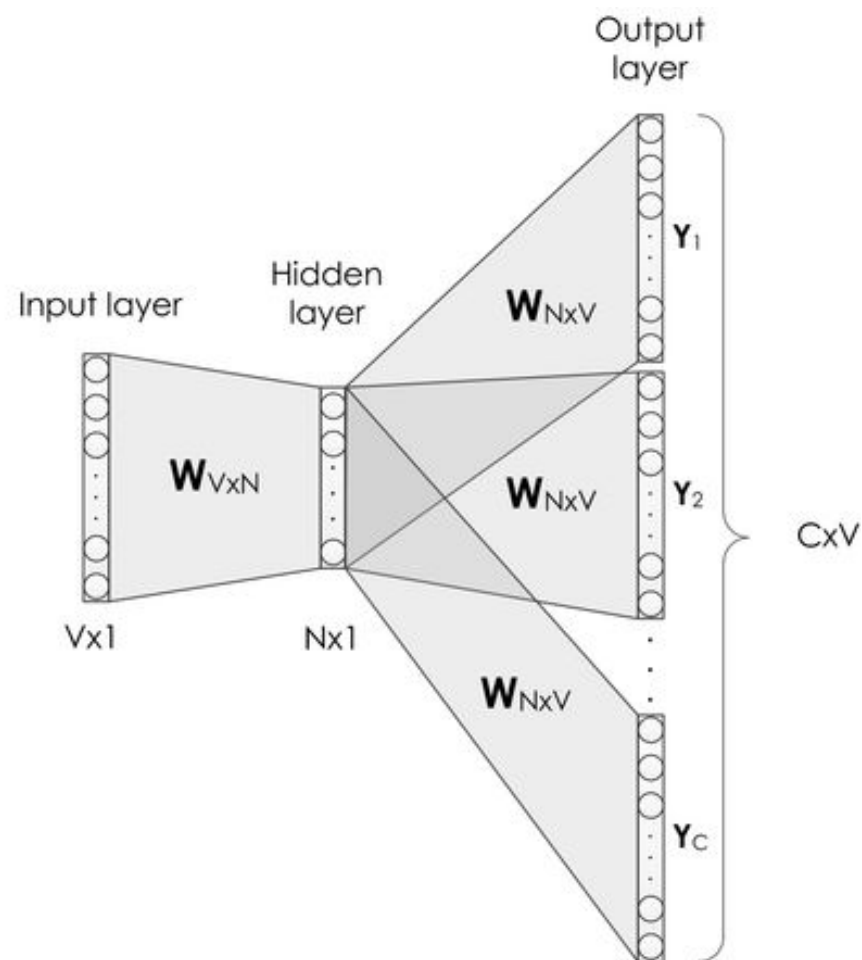
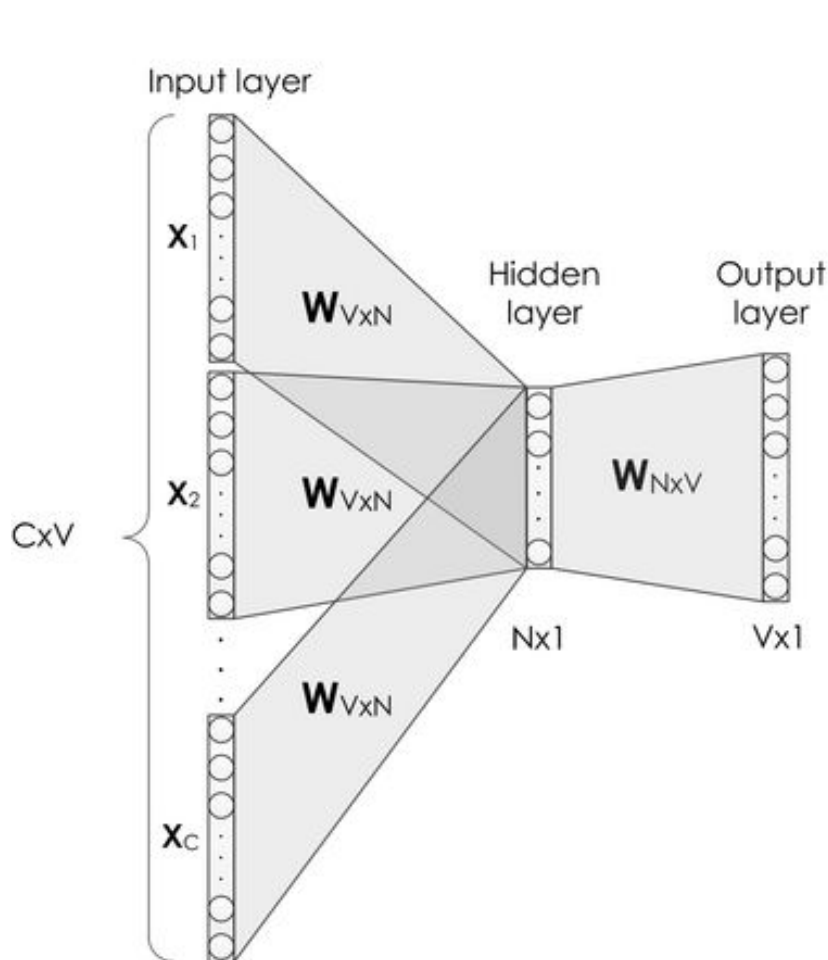


Word2Vec

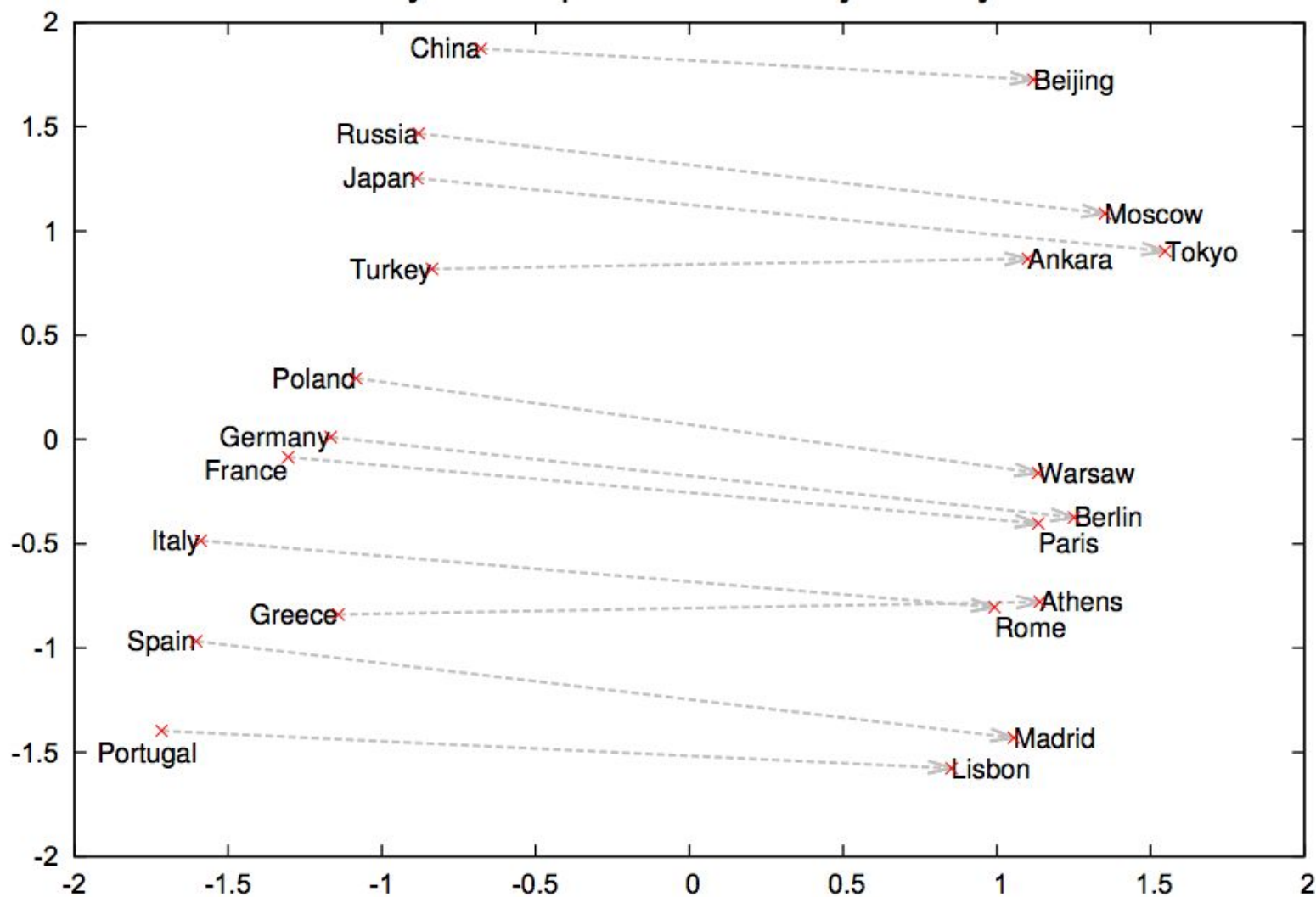
Mikolov et al. 2013

2. Sliding Window

#1	natural	language	processing	and	machine	learning	is	fun	and	exciting	#1
	X _k	Y(c=1)	Y(c=2)								
#2	natural	language	processing	and	machine	learning	is	fun	and	exciting	#2
	Y(c=1)	X _k	Y(c=2)	Y(c=3)							
#3	natural	language	processing	and	machine	learning	is	fun	and	exciting	#3
	Y(c=1)	Y(c=2)	X _k	Y(c=3)	Y(c=4)						
#4	natural	language	processing	and	machine	learning	is	fun	and	exciting	#4
		Y(c=1)	Y(c=2)	X _k	Y(c=3)	Y(c=4)					
#5	natural	language	processing	and	machine	learning	is	fun	and	exciting	#5
			Y(c=1)	Y(c=2)	X _k	Y(c=3)	Y(c=4)				
#6	natural	language	processing	and	machine	learning	is	fun	and	exciting	#6
				Y(c=1)	Y(c=2)	X _k	Y(c=3)	Y(c=4)			
#7	natural	language	processing	and	machine	learning	is	fun	and	exciting	#7
					Y(c=1)	Y(c=2)	X _k	Y(c=3)	Y(c=4)		
#8	natural	language	processing	and	machine	learning	is	fun	and	exciting	#8
						Y(c=1)	Y(c=2)	X _k	Y(c=3)	Y(c=4)	
#9	natural	language	processing	and	machine	learning	is	fun	and	exciting	#9
							Y(c=1)	Y(c=2)	X _k	Y(c=3)	
#10	natural	language	processing	and	machine	learning	is	fun	and	exciting	#10
								Y(c=1)	Y(c=2)	X _k	



Representações vetoriais



HandsOn: NLTK e SpaCy

Obrigada!

Aline Paes

alinepaes@ic.uff.br