

①

github.com/fabriicmarques/natural-langs

PANDAS e NLTK/SPACY

Introdução ao Processamento de Linguagem Natural

Fabrício Galende Marques de Carvalho

OBS: Esse texto foi produzido com o propósito de servir como um guia de estudo da disciplina Processamento de Linguagem Natural, ministrada por Fabrício Galende Marques de Carvalho, seja em cursos de graduação, pós-graduação ou extensão universitária. É permitida sua livre utilização e reprodução por parte dos alunos matriculados e que tenham concluído cursos dessa área que tenham sido integralmente ministrados pelo autor do material. Qualquer outro uso ou redistribuição deve ser feito somente mediante autorização, por escrito, fornecida pelo autor ao participante. Caso o conteúdo desse trabalho seja usado para subsidiar outras publicações científicas ou tecnológicas, esse deve ser citado da seguinte forma:

Carvalho, F. G. M. Notas de aula do curso Processamento de Linguagem Natural. São José dos Campos, 2025.

Resumo

Esse texto trata de aspectos introdutórios relacionados ao Processamento de Linguagem Natural. O texto começa com uma motivação ao estudo do tema, considerando diferentes aspectos tecnológicos pertinentes, seguido de diversos conceitos introdutórios relacionados ao PLN. Diversas aplicações relacionadas diretamente ao PLN e também aquelas relacionadas às análises de dados textuais (i.e., *Text Analytics*) são citadas. O texto é encerrado ilustrando-se os principais componentes de uma *pipeline* de processamento de linguagem natural. Para uma adequada compreensão, assume-se que o leitor possua familiaridade com aspectos básicos relacionados ao desenvolvimento de sistemas de software, bases de dados e inteligência artificial.

1 Motivação

Um dos principais motivadores para o estudo do PLN é o elevado volume de dados linguísticos disponíveis através dos sistemas informatizados. Exemplos de sistemas informatizados e serviços digitais que geram ou manipulam dados linguísticos textuais incluem:

- A World Wide Web.
- As redes sociais, tais como Facebook, Instagram e X.
- E-mails.
- Bibliotecas Digitais.
- Sites de conversa (*chats*).

Este fato é decorrente da característica pervasiva e ubíqua dos sistemas computacionais que passaram a fazer parte dos mais diversos negócios e do dia a dia das pessoas.

Quando se fala em volume de dados, três características devem ser analisadas: o volume de armazenamento, o volume de variedade e o volume de velocidade de geração

- **Volume de Armazenamento:** refere-se ao espaço físico ocupado pelos dados acessados, criados ou manipulados pelos sistemas computacionais. Bases de dados contendo dados de usuários, por exemplo, em geral são fisicamente volumosas. O mesmo vale para bases de dados de e-mails, registros de interações com chatbots, etc.
- **Volume de Variedade:** refere-se à grande quantidade de possibilidades para certas categorias de dados. Exemplos de dados com elevada variedade incluem aqueles presentes em textos de obras literárias, textos técnicos de diferentes domínios, etc.
- **Volume de Velocidade de Geração:** Possui relação com dados que são gerados com elevadas frequências. Exemplos de dados não linguísticos incluem dados gerados por sensores que executam tarefas periódicas (e.g., dados de temperatura por município). Exemplos de dados linguísticos incluem bases de dados de e-mails.

Se forem considerados dados em formato de linguagem natural (i.e., linguagem escrita, nesse caso), pode-se afirmar que esse geral elevado volume não só de armazenamento, mas também decorrente de variedade e também de velocidade de geração. Portanto, pode-se concluir que é interessante o desenvolvimento de ferramentas computacionais capazes de lidar com tais dados a tempo de maneira a gerar valor para o usuário de tais dados.

Alguns dos grandes desafios enfrentados nessa área incluem:

- Armazenamento adequado de maneira a tornar tanto os dados como as informações que eles transportam prontamente disponíveis para uso.
- Desenvolvimento de ferramentas capazes de lidar com elevados volumes de dados que são naturalmente desestruturados.
- Transformar as informações extraídas dos dados textuais em conhecimento diretamente aplicável a determinado domínio.

A área de Processamento de Linguagem Natural (PLN) trata justamente da busca de respostas a esses desafios, mais especificamente, é voltada ao desenvolvimento de estruturas, modelos, técnicas e algoritmos para processar dados linguísticos de maneira a torná-los úteis em um determinado contexto.

Uma das subáreas do PLN é a de Análise de Textos (do inglês, *Text Analytics* - TA), que foca na extração de informações relevantes e de valor a partir de dados de texto. A área de TA faz uso de ferramentas de estatística e de Inteligência Artificial (IA) para a análise de textos. TA inclui a mineração de texto (do inglês, *Text Mining*), que explora grandes quantidades de texto não estruturado, objetivando a extração de padrões e conhecimentos, e a aprendizagem de máquina (ML, do inglês, *Machine Learning*) a partir de texto.

2 Definições Fundamentais

Nesta seção, serão discutidos conceitos fundamentais envolvidos no estudo do Processamento de Linguagem Natural.

2.1 Processo Comunicativo

É o processo que envolve a transmissão de uma **mensagem** partindo do **transmissor** e chegando ao **receptor**.

O transmissor gera uma **referência**, utilizando um conjunto de **símbolos ou códigos**, que representam um **referente** (e.g. realidade, fato, etc.). O receptor recebe os símbolos e, através do processo cognitivo, consegue **decodificar** e inferir o referente.

2.2 Linguagem

É o meio de se expressar os sinais utilizados no processo comunicativo.

A linguagem pode ser falada, gestual, escrita, gráfica, etc.

Dependendo da linguagem utilizada, o processo comunicativo pode ser mais ou menos facilitado.

A ciência que estuda a linguagem é a **linguística**. Quando a linguagem é estudada sob o ponto de vista da computação, ou seja, técnicas de representação, compreensão geração e demais aspectos linguísticos através de computadores, tem-se a chamada **linguística computacional**



Apesar de muitas vezes serem tratados como sinônimos, os termos processamento de linguagem natural, *text analytics* e linguística computacional (do inglês, *computational linguistics*) referem-se a áreas distintas do conhecimento. Apesar disso, deve-se ter a compreensão de que estes três termos, com frequência, aparecem em trabalhos de PLN.

2.3 Linguagem Natural

É aquela que foi criada ao longo de vários anos e que evoluiu, gradativamente, através do uso diário pelos seres humanos.

A linguagem natural é dinâmica e temporal, ou seja, sofre mudanças ao longo do tempo e, em um dado instante de tempo, manifesta aquilo que é aceito pela maioria de seus usuários naquele instante.

A linguagem natural difere da linguagem artificial criada com um propósito específico (e.g.: SQL, linguagem de programação, etc.).

2.4 Processamento de Linguagem Natural

É um ramo especializado da computação/engenharia, originário da linguística computacional.

Tem relação direta com a área de interação humano-computador (IHC) pois foca no estudo e construção de sistemas que permitam a interação de máquinas com as linguagens naturais criadas pelos humanos.

As ferramentas e técnicas utilizadas pela disciplina de PLN, são baseadas em fundamentos matemáticos tais como probabilidade, estatística, álgebra linear e cálculo (e.g., modelos de aprendizado de máquina).

2.5 Sintaxe e Semântica

No processo comunicativo, **sintaxe** se refere às regras que regem a construção dos elementos básicos de uma determinada linguagem (palavras, orações, frases, etc.).

Apesar da área de PLN tratar de dados não estruturados até certo ponto, é requerido que os dados textuais sejam aderentes às regras básicas de sintaxe de uma determinada linguagem escrita.

Cabe também ressaltar que, apesar de simples para um ser humano, efetuar a análise sintática de uma linguagem não é tarefa fácil sob o ponto de vista computacional.

Ainda sob o ponto de vista da sintaxe, define-se como **léxico** o conjunto de palavras e expressões utilizados em uma língua. Se for considerado um subconjunto desse léxico tal como aquele que é utilizado por um grupo de pessoas, tem-se o que se chama de **vocabulário**.

A **morfologia** estuda as classes de palavras e, também, como estas são construídas a partir das suas menores unidades no contexto de uma língua. Essas unidades são denominadas de morfemas.

Exemplo: infeliz → morfema radical **feliz** + prefixo **in** morfema com sentido de “não”.

A morfologia das palavras é aplicada ao PNL em técnicas tais como lematização (em inglês *lemmatization*) e stemização (em inglês *stemming*).

Quando se consideram os sentidos atribuídos aos símbolos ou sinais de uma determinada linguagem tem-se o que se chama de **semântica**.

A análise semântica é uma das tarefas mais difíceis para sistemas computacionais de PLN, pois envolve modelar e executar tarefas que envolvem aspectos cognitivos.

No processo de rotulação de elementos constituintes de uma frase, escrita em uma determinada língua, a análise semântica e sintática geralmente acarretam a obtenção de rótulos (tags) para as palavras (e.g. Verbo, substantivo/nome, adjetivo, advérbio, artigo, etc.).

2.6 Fontes de Texto

As fontes de texto dizem respeito ao processo de obtenção do conteúdo linguístico textual utilizado pelo PLN.

Há diversas fontes possíveis, podendo ser citadas:

- Bibliotecas digitais.
- Sites de notícias da Internet.
- Redes sociais tais como Facebook, Instagram e X.

- Aplicações Web.
- E-mails.
- Chatbots utilizados em sistemas de autoatendimento.
- Logs de assistentes virtuais.
- Legendas de filmes.

Cabe ressaltar que cada tipo de fonte de conteúdo linguístico textual deve ser utilizada de acordo com a finalidade compatível com tal fonte (e.g., um log de um chatbot pode ser utilizado, por exemplo, para treinar outros chatbots para serem utilizados em contextos similares).

2.7 Documento, Corpus e Corpora

No contexto de PLN, um **documento** é uma unidade básica de dado textual que possui um significado atrelado. Como exemplo, pode-se citar uma revisão de produto presente em um site de comércio eletrônico (i.e., uma opinião de um consumidor acerca de um aparelho de telefonia celular).

Um **corpus** de texto refere-se a um conjunto de dados linguísticos reais, pertencentes a uma determinada língua, que pode ser utilizado pelo computador. Tipicamente um corpus é constituído por um conjunto de documentos. Um exemplo de corpus seria o conjunto de todas as revisões de clientes que adquiriram um determinado aparelho de telefonia celular no ano de 2025.

Dependendo do escopo de uso em PLN, os conceitos de documento e corpus podem ser considerados para um mesmo artefato. Por exemplo, um livro de Machado de Assis pode ser considerado como um único documento e um exemplo de uso em PLN seria a identificação da temática principal do livro. Por outro lado cada capítulo do livro pode ser considerado como um documento independente em que o tópico predominante pode ser identificado e, nesse caso, o livro completo seria tratado como um corpus de texto.

Alguns pontos importantes devem ser considerados quando se fala em corpus de texto:

- **Trata-se do resultado de um processo já realizado e não em realização**, ou seja, um corpus reflete algo que já começou e já terminou. Portanto, uma interação, por exemplo em um chatbot, não pode ser denominada de corpus de texto.
- Como consequência do aspecto anterior, um corpus é finito.
- Trata-se de um subconjunto das possibilidades linguísticas associadas a uma língua, ou seja, é um **recorte linguístico**.
- Não se deve incluir informações linguísticas em um corpus sob pena de se prejudicar as análises e resultados de PLN.

Quando se tem vários corpus de texto, tem-se um **corpora**. Corpora são utilizados para análise estatística e construção de sistemas de processamento de linguagem natural.

Dependendo da fonte de dados utilizada para a obtenção dos corpora, tem-se corpora:

CORPUS → CORPORA → ULM / VERIFICAR

- Oral: Obtidos a partir da língua falada (e.g., gravações telefônicas).
- Escrito: Obtidos a partir da língua escrita (e.g., bancos de dados de e-mail).

Um corpus (ou corpora) é dito ser **anotado** quando possui marcações ou **metadados** sobre os dados linguísticos que permitem que estes sejam classificados ou utilizados de alguma forma específica em PLN. Exemplos de anotações ou metados presentes em corpora incluem:

- Etiquetas (do inglês, *tags*) de categoria atreladas a frases ou documentos.
- Entidades nomeadas (e.g., Maria → pessoa).
- Partes do discurso, POS *tags* (do inglês, Part of Speech) (e.g., comer → Verbo).
- Relações semânticas (e.g., Thales nasceu em São José dos Campos → Thales, nasceu_em, São José dos Campos).
- Anotações de eventos (e.g, "Em agosto de 2025, uma picape capotou em uma região da cidade de São Paulo ocasionando um choque com outros 10 carros" → acidente de trânsito).
- *Stemming tags*, que correspondem à remoção de afixos de uma palavra sem necessariamente se chegar a um significado. Tem como foco a palavra base.
- *Lemma tags*, são similares aos tags de *stemming* stemming, mas consideram a redução a uma forma base preservando o significado.
- Tipos semânticos e papéis, que correspondem a anotações envolvendo tipicamente mais de um elemento constituinte de uma sentença. (e.g., na frase "Fabíola pegou o controle remoto da TV, Fabíola corresponde ao tipo semântico "pessoa" e possui o papel semântico de "agente", já controle remoto possui tipo semântico objeto com papel semântico "tema/paciente")

Alguns corpora especializados que contém formas avançadas de anotações sintáticas e semânticas são denominados de ***Tree Banks***.

Um exemplo de *tree bank* é a WordNet, criado pela Universidade de Princeton em 1985, focado na língua inglesa, que contém sinônimos (synsets), definição de palavras, relacionamentos, etc.

2.8 Lexema e Lema

Um **lexema** é um conjunto de palavras de uma mesma classe morfológica. A forma base dessas palavras é denominada de **lema** (em inglês, *lemma*)

Exemplo:

Lema	Lexema
	florescer
flor	florido
	floresce



A técnica de stemização, diferentemente da lematização, ao eliminar afixos de uma palavra, pode chegar a uma estrutura raiz que não necessariamente é provida de significado ou correção gramatical.

Exemplo:

Stem	Palavras originais
	bolacha
bol	boliche
	bolinha

→ 3 Áreas de Aplicação do PLN

O processamento de linguagem natural possui aplicabilidade em inúmeros problemas práticos envolvendo computação e grandes volumes de dados. A seguir, são ilustradas algumas dessas aplicações.

→ 3.1 Chatbots

Constituem uma das aplicações que mais fazem uso de PLN. Nesse caso, o objetivo é desenvolver um robô (*robot: bot*) capaz de interagir com um usuário de maneira a fornecer informações, efetuar triagens, etc.

Um dos grandes desafios presentes na criação de chatbots é a manutenção de uma interface conversacional capaz de comunicar adequadamente uma ideia ao usuário e, ao mesmo tempo, capaz de “compreender” aquilo que o usuário expressa, desconsiderando ou corrigindo erros de ortografia e interpretando adequadamente a semântica da frase.

Outras dificuldades no desenvolvimento desses sistemas incluem a manutenção de uma **base de conhecimento** que seja abrangente e suficiente para um determinado fim e, também, passível de consulta apropriada, além de prover um tempo de resposta baixo que seja adequado à aplicação.

A resolução e o reconhecimento contextual, ou seja, determinar a quem ou a que um determinado texto se refere e qual o sentido mais adequado da palavra, de acordo com o contexto (significado ou referente) são outros pontos de atenção a serem levados em consideração no projeto de chatbots.

Exemplo: “João chamou André. Ele estava ocupado.”

Nesse caso, o pronome “Ele” refere-se a quem? Notar que isso depende do contexto maior onde as frases estão inseridas e, sob o ponto de vista computacional essa determinação nem sempre é simples.

Apesar de já existirem chats de propósito geral (tais como o ChatGPT), a maioria dos chatbots possui especificidade com relação às bases de conhecimento (e.g., manutenção mecânica, saúde, etc.).

3.2 Tradução automática de texto

A tradução automatizada de textos também é uma das aplicações mais difundidas na Internet. Apresenta vários desafios, entre eles a existência de diferentes estruturas presentes em idiomas distintos, incluindo

regras de sintaxe e semântica. A seguir, ilustra-se um exemplo de dificuldade experimentada por um sistema automatizado de tradução:

Frase em português: "Que cachorro bonito!"

Frase em inglês: "What a beautiful dog!"

Claramente a inversão de ordem entre o substantivo (cachorro) e o adjetivo (bonito), quando se comparam os dois idiomas, ilustra uma dificuldade enfrentada pelo processo de tradução automatizada.



3.3 Geração de texto

A geração de texto vem ganhando significativa importância pois permite a entrega de informações ou conteúdos de maneira mais eficaz ao usuário.

Um exemplo prático de aplicação da geração de texto está na análise de tendências de séries temporais (i.e., variáveis observadas ao longo do tempo, tais como as cotações de uma ação na bolsa de valores ou o número de focos de calor em plantações). Nesse caso, a partir de valores numéricos gera-se informação útil ao usuário de modo a facilitar na tomada de decisão (e.g., se uma determinada cotação está em alta ou baixa ou se a quantidade de focos de calor indicam um aumento no número de queimadas.)

3.4 Reconhecimento de fala

Utilizados no projeto de assistentes virtuais, tais como o Alexa, da Amazon, ou naqueles que já vêm integrados aos painéis automotivos, que permitem ao motorista a execução de funções secundárias sem retirar o foco principal e as mãos do volante.

Esses sistemas geralmente são dependentes de domínio (i.e. um sistema que funciona bem para um pré-atendimento em uma oficina mecânica não necessariamente funcionará bem para uma triagem médica) e também são tipicamente customizados de acordo com as necessidades de um determinado usuário.

Aspectos relacionados ao sotaque regional, entonação, entre outros, constituem evidentes dificuldades experimentadas por esses tipos de sistemas.



3.5 Correção gramatical

Esses sistemas focam principalmente nos aspectos sintáticos de uma determinada linguagem. Nesse caso o objetivo é a verificação e correção de erros em relação aos aspectos da língua formal ou considerada como de referência.

Em sistemas de processamento de linguagem natural a correção gramatical muitas vezes precede os demais componentes da aplicação, pois “blinda” os demais de estruturas ou palavras irreconhecíveis.

3.6 Assistentes virtuais

Assistentes virtuais são sistemas computacionais que efetuam ações de acordo com os comandos fornecidos pelo usuário. Estão, portanto, intimamente ligados aos aspectos de IHC - Interação Humano-Computador.

Assistentes virtuais, em especial aqueles integrados ao reconhecimento de língua falada, são muito utilizados para aumento de acessibilidade ou para tornar certas tarefas mais intuitivas ou menos cansativas aos usuários.

3.7 Sumarização

Lida com a redução do conteúdo de modo apropriado para criar um sumário/resumo que contenha os principais pontos de um determinado documento ou corpus de texto.

A similaridade entre tópicos abordados ou sua elevada diversidade tornam essa tarefa bastante desafiadora.

Os sumarizadores de texto podem ser agrupados em duas grandes categorias:

1. **Sumarizadores extrativos:** são caracterizados por um subconjunto de palavras, frases, orações ou parágrafos que podem ser considerados representativos em relação ao texto original.
2. **Sumarizadores abstrativos:** operam através da reescrita do texto original de modo a torná-lo menos volumoso sem perder a representatividade associada ao significado. Tipicamente parafraseiam um texto original são caracterizados por algoritmos avançados de aprendizagem de máquina.



3.8 Máquinas de busca

Máquinas de busca com frequência precisam efetuar a indexação de grandes volumes de conteúdos textuais. O principal objetivo desse tipo de processo é permitir que o usuário encontre de modo fácil e rápido o conteúdo que mais lhe interessa.

3.9 Filtros de spam

São utilizados para filtrar mensagens de e-mail de acordo com o conteúdo. Geralmente fazem uso de estatísticas textuais associadas a conteúdos indesejável por parte do usuário.

3.10 Organização de conteúdo

Apesar de similar ao processo utilizado na indexação, o enfoque aqui é outro, mais voltado à organização. Um exemplo fácil de entender envolve a organização de portais de notícias. Nesse caso, uma vez que o conteúdo textual é gerado esse pode ser facilmente categorizado e publicado na seção adequada para divulgação.

3.11 Sistemas de recomendação

Sistemas de recomendação são aqueles que oferecem conteúdo textual, produtos ou outros itens que podem ser consumidos por um usuário. Nesse caso, baseado, por exemplo em uma opinião que um usuário forneceu sobre um determinado produto, o sistema pode ser capaz de oferecer outros produtos que estejam mais alinhados às expectativas desse.

3.12 Mineração de opiniões

A mineração de opiniões é conhecida como *text mining* e envolve a extração de informação ou insights qualitativos relacionados a um aspecto específico de um produto ou serviço fornecido ao usuário.

Seja, por exemplo, um conjunto de reviews de refrigeradores vendidos em um site de e-commerce. Após a execução da mineração de opiniões sobre um conjunto significativo de documentos, o vendedor pode concluir que o elevado preço é um ponto que está atrapalhando as vendas desse produto.

3.13 Análise de sentimentos

Similar à mineração de texto, a análise de sentimentos tem um enfoque mais quantitativo. nesse caso, considerando-se o mesmo exemplo anterior, o vendedor do produto pode concluir que mais de 80% dos consumidores estão satisfeitos com o produto, ou seja, manifestam opiniões que podem ser consideradas “positivas”.

4 Componentes Principais de uma Pipeline de PLN

Para facilitar o estudo e a implementação de sistemas computacionais que fazem uso de PLN, pode-se considerar que o fluxo completo de processamento é composto por etapas ou atividades menores, executadas em sequência e que, em conjunto, são denominadas de *pipeline de PLN*. Esse nome remete ao fato da similaridade com uma série de canos interconectados em um sistema de tubulação convencional, só que, ao invés de fluxo de algum fluido, o fluxo é de dados.

Em uma pipeline de PLN, cada atividade recebe os dados que foram processados pela atividade anterior e executa todas as operações sobre os dados antes de enviar para a atividade seguinte.

O diagrama de atividades típico de uma pipeline de PLN é ilustrado na Figura 1.

Na etapa de pré-processamento, os dados de linguagem natural são tratados de modo a se ter uma entrada livre de ruídos e relativamente padronizada.

A etapa de extração de características é a responsável por selecionar quais dados, da entrada padronizadas, serão utilizados na montagem do modelo.

O modelo de linguagem corresponde à representação numérico-matemática que será associada aos elementos da linguagem.

A etapa de treinamento de modelo de Inteligência artificial faz uso da representação do modelo de linguagem para treinar um modelo específico de IA voltado a uma certa tarefa.

Por fim, o modelo de IA, anteriormente treinado e devidamente salvo, é carregado e utilizado na execução da tarefa específica de PLN, tais como as citadas nas possíveis aplicações da seção anterior.

A discussão efetuada nesse texto não é exaustiva, mas sim introdutória. Para maiores detalhes e aprofundamentos, recomenda-se que o estudante consulte referências tais como (AGGARWAL, 2023), (SARKAR, 2019) e (FERREIRA; LOPES, 2025)

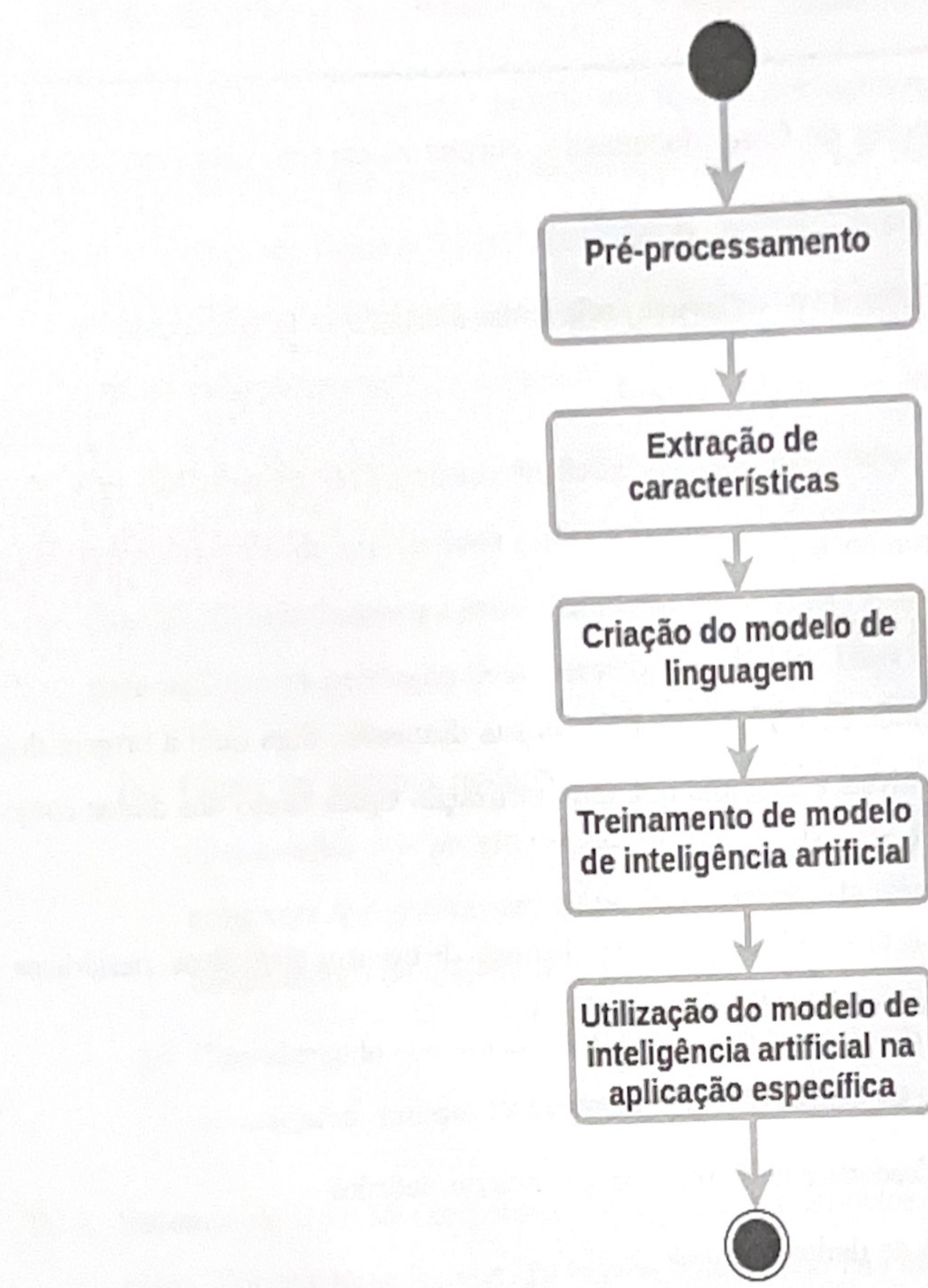


Figura 1: Uma pipeline típica de PLN.

TC.17. Cite e descreva, em alto nível e sem se preocupar com detalhes de implementação, os modelos de linguagem baseados em sacos de palavras e os modelos preditivos do tipo naive bayes. Adicionalmente, descreva os seguintes modelos de inteligência artificial: K-Means e Redes Neurais. Qual a diferença básica entre essas duas categorias de modelo (i.e., linguagem e IA)?

TC.18. As entradas para um sistema de PLN podem ser oriundas de diferentes fontes. Alguns exemplos incluem: obras literárias, textos jornalísticos, chats com usuários, entradas faladas, etc. Cite e exemplifique aplicações de PLN envolvendo diferentes fontes, discutindo potenciais dificuldades técnicas associadas, para o caso de:

- (a) Entradas através da língua falada.
- (b) Entrada através de sistemas de raspagem de dados.
- (c) Entrada através de interação direta e escrita com o usuário (ex. Chatbot).
- (d) Entrada oriunda de um tradutor automatizado de texto.

PC.1. Descreva as principais características, tais como origem do texto, metadados, ano dos textos, anos de produção, etc. e exemplifique o acesso computacional aos seguintes corpora utilizando a biblioteca NLTK.

- (a) Machado.
- (b) Mac-Morpho.
- (c) Floresta Sintáctica

PC.2. Exemplifique o acesso a dados textuais prontamente disponíveis na biblioteca Spacy. Adicionalmente, exemplifique o acesso ao corpora SUBTLEX-pt-br. Exemplifique, através de um pequeno trecho de código executável, o acesso aos dados e metadados do corpora aplicável a um problema simples de PLN.

PC.3. Discuta sobre os corpora disponíveis no projeto Gutenberg. Desenvolva um pequeno programa executável que acesse mais de um corpus do projeto Gutenberg. O seu programa deve acessar não somente os dados disponíveis nos corporas, mas também os metadados. Exemplifique, também, a utilidade prática da informação disponível no contexto de PLN.

* PC.4 Baseando-se no código-fonte fornecido pelo professor, exemplifique o carregamento da biblioteca NLTK, em Python e efetue a tokenização de um texto em português pertencente a alguma obra literária de domínio público. Efetue uma tokenização por sentenças e uma tokenização por palavras. Qual a diferença de saída entre cada um dos processos? Utilize um texto de pelo menos 2000 caracteres. Mostre o funcionamento do seu programa e ~~descreva~~ ao menos 5 POS tags.

PC.5. Desenvolva um pequeno exemplo de utilização da biblioteca Natural para PLN em JavaScript. Ilustre a operação de pelo menos dois conceitos abordados no material introdutório. Cite limitações dessa biblioteca.

PC.6. Desenvolva um pequeno exemplo de utilização da biblioteca Compromise para PLN em JavaScript. Ilustre a operação de pelo menos dois conceitos abordados no material introdutório e cite limitações dessa biblioteca.

PC.7. Se chamarmos de “riqueza lexical” a quantidade de palavras em um determinado documento, teremos como documentos mais ricos aqueles que possuem um número maior de palavras (i.e., maior vocabulário). Desenvolva um programa que compare a riqueza lexical considerando dois documentos de entrada. A saída deve ser a quantidade de palavras de cada um dos documentos e o percentual de palavras que o documento de menor riqueza tem com relação ao maior. (Obs: desconsidere caracteres de pontuação ao construir seu programa e faça uso de funções de bibliotecas tais como NLTK e Spacy).

PC.8. Desenvolva um programa que efetue uma análise estatística dos dados de um documento de texto. A entrada do seu programa deve ser um documento em formato.txt, a saída deve ser um gráfico de pizza que mostre os percentuais de palavras de acordo com suas classes gramaticais (i.e., artigos, substantivos, etc.). Faça a geração desses gráficos considerando pelo menos duas categorias distintas de documentos (e.g., obra literária infantil, texto técnico, notícia jornalística, etc.)

PC.9. A similaridade de Jaccard, entre dois documentos A e B, é definida como sendo $J(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|}$. Desenvolva um pequeno programa para detecção de plágio utilizando a definição de similaridade de Jaccard. Considere que um documento é considerado plágio de outro caso a similaridade seja superior a 50%. Para as operações de união e intersecção, considere o vocabulário de cada documento e faça uso de operações utilizando a biblioteca NLTK ou Spacy.

PC.10. Fazendo uso de tags do tipo POS, desenvolva um assistente virtual simples que identifique se o usuário está fazendo uma pergunta ou se é afirmação ou comando.

Hint: olhe para a primeira palavra da frase e seu POS.

PC.11. Repita o item PC.10. mas utilizando entradas por comando de voz. **Hint:** Utilize bibliotecas tais como pyaudio, portaudio e SpeechRecognition.

PC.12. Utilizando tags do tipo POS, desenvolva um pequeno summarizador extrativo. Seu summarizador deve gerar os principais adjetivos associados a um determinado conjunto de revisões. Considere como entrada uma lista de documentos representativos de revisão e como saída os adjetivos mais representativos das reviews (Obs: use contagem simples).

PC.13. Exemplifique a stemização e a lematização de um texto, em língua portuguesa. Ilustre um caso onde textos diferentes conduzem a uma mesma saída através do stemming ou lemmatization. Considere como saída um vetor ordenado contendo lemas e stems.

PC.14. Repita o problema PC.13. considerando a língua inglesa.

PC.15. Repita o problema PC.13. considerando a língua espanhola.