Atividades unidade 1

TERMINOLOGIA E CONCEITOS

TC.1.1. Selecione uma obra literária de domínio público (ex. livros tais como Vinte mil léguas submarinas (de Júlio Verne), a Bíblia, etc.) e ilustre a variedade de dados presente. Considere, por exemplo a construção de frases, orações etc. e compare com expressões de uso corrente. Para respaldar sua resposta, elabore um programa que contabilize, por exemplo, o número de palavras diferentes.
Hint: Utilizar obras mais antigas, textos infantis, etc. Pode ajudar você a ter insights relacionados a tal variedade. Considere textos de tamanho equivalente ou calcule um índice que permita dar a noção de variedade (ex: taxa de palavras distintas utilizadas em relação ao total de palavras do texto, para textos de tamanhos similares).

A obra literária de domínio público selecionada foi a Bíblia, que ilustra uma variedade de dados linguísticos, com uma gama de gêneros literários ao longo de seus capítulos, como narrativa histórica, poesia, cada um tendo formas gramaticais e expressões diferentes. Por exemplo, em Salmos usa metáforas, enquanto em Paulo, a argumentação.
A construção de frases são variadas, na parte do Antigo Testamento, as orações são mais formais e complexas, já no Novo Testamento, as frases são mais curtas, diretas, trazendo um entendimento acessível. A frase "E disse Deus: Haja luz. E houve luz.", comparando com expressões modernas, poderia ser algo como: “Deus falou para existir a luz, e a luz acendeu”, que é uma linguagem menos formal e arcaica, deixando mais fluido.
Para respaldo, o programa para contabilizar gerou o seguintes resultados:
Número de palavras distintas: 27408
Número total de palavras: 754097
Quantificando uma alta variedade linguística da Bíblia, demonstrando a diversidade de dados e amplitude do vocabulário, com sua riqueza e complexidade.

TC. 1.9. Para os sistemas abaixo, diga quais envolvem ou não PLN, justifique sua resposta:

a) Um sistema de triagem automatizado, via Whatsapp, utilizando em um hospital com pronto atendimento.

Sim, ao interpretar as mensagens, identificar os sintomas, processar a linguagem dos usuários, entender se é pergunta ou resposta, abreviações e mensagens escritas incorretamente.

b) Um sistema de diagnóstico de defeitos em um automóvel, baseado na descrição textual dos problemas relatados pelo motorista.

Sim, envolve PLN, ao interpretar as descrições, processar e extrair as informações.

c) Um sistema de geração automática de código em linguagem de programação a partir de um diagrama de blocos funcional.

Não, pois o diagrama de blocos não é uma linguagem natural, não exige a interpretação dela.

d) Um sistema de consulta a uma base de dados utilizando linguagem padrão SQL.

Não envolve PLN, pois a linguagem SQL é um conjunto de comandos predefinidos, de consultas estruturadas, de sintaxe formal e rígida, não de linguagem natural.

e) Um sistema de reconhecimento de fala utilizado pela Alexa.

Sim, a fala é um tipo de linguagem natural, envolve o reconhecimento de fala e a interpretação, realizando dessa forma a execução de tarefas e respondendo perguntas.

f) Um sistema de reconhecimento de gestos utilizando a língua de sinais.

Não envolve PLN, pois é utilizado a interpretação visual, o reconhecimento de gestos, que não envolve o processamento textual ou falado da língua natural.

g) Um sistema para conversão de linguagem de programação em Python para JS (Transpilador).

Não, pois utiliza a linguagem de programação, não a linguagem natural, não precisa de interpretação semântica para acontecer.

PRÁTICA DE PROGRAMAÇÃO

PP.1.5. Repita **PP.1.2.** considerando a língua espanhola.

PP.1.2. 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.

Stemização

Texto: "Las flores crecían en el jardín"

Las	**flor**es	**crec**ían	en	el	jardín
~~las~~	flor	crec	~~en~~	~~el~~	~~jardín~~

Eliminou os afixos apenas de duas palavras, flores e crecían, deixando respectivamente como flor e crec. Reduzindo para chegar em uma base, sem levar em conta se é possui significado ou correção gramatical.

Vetor ordenado com stems em ordem alfabética

["crec", "el", "en", "flor", "jardín", "las"]

Lematização

Texto: “Las flores crecieron en el jardín”

Las	flores	crecieron	en	el	jardín
la	**flor**	**crecer**	en	el	jardín

la ⇨ Artigo definido, singular

flor ⇨ Singular

crecer ⇨ Verbo no infinitivo

Reduziu as palavras para chegar na forma base se importando com a correção gramatical.

Vetor ordenado com lemas em ordem alfabética
["crecer", "el", "en", "flor", "jardín", "la"]

Textos diferentes conduzem a uma mesma saída através da lemmatization

Texto 1: “El gato corre rápido”

El	gato	corre	rápido
el	gato	**correr**	rápido

Texto 2: “El felino corrió rápidamente”

El	felino	corrió	rápidamente
el	felino	**correr**	**rápido**

Vetor ordenados com os lemas
[“correr”, “el”, “felino”, “gato”, “rápido”]

As palavras “felino” e “gato” não se lematizaram, pois foi levada em consideração a semântica e apareceram no vetor por serem substantivos sem modificações.