Unyleya
EDUCACIONAL

Processamento de linguagem natural

Natural Language Toolkit

O python conta com uma biblioteca especial para facilitar o processamento de linguagem natural, ou seja, o processamento de texto.

A biblioteca é chamada de NLTK - Natural Language Toolkit, dentro dela são disponibilizadas muitas interfaces, bibliotecas para classificação de texto, tokenização, stemming, análise e outras funcionalidades.

Natural Language Toolkit

São objetivos da biblioteca:

- **Simplicidade:** fornecer recursos simples de serem utilizados, permitindo que os usuários adquiram conhecimento prático sobre o processamento de linguagem natural. Reduz as dificuldades normalmente enfrentadas pelos usuários no processamento de linguagem natural;

- **Consistência:** ser um framework com muitos recursos unificados, oferecendo interfaces, estruturas para dados e métodos com nomes amigáveis e próximos da realidade;

- **Extensibilidade:** oferecer uma estrutura que permite que novos módulos, com novas funcionalidades, possam ser facilmente incorporados, permitindo até versões diferentes para uma mesma abordagem;

- **Modularidade:** permitir que o usuário baixe e utilize apenas aqueles componentes que lhe interessa.

Processamento de linguagem natural

Etapas do processamento de linguagem natural:

1. Tokenização;
2. Análise léxica;
3. Análise sintática;
4. Análise semântica;
5. Análise pragmática.

Tokenização

É a segmentação de um texto, identificando onde cada palavra começa e termina para realizar a separação corretamente e permitir o trabalho com as palavras.

Depois de separadas, para o processamento de linguagem natural cada palavra pode ser chamada de token.

Tokenização

Para linguagens artificiais, como as de programação, os tokens são bem definidos, todas as possibilidades de palavras.

Nas linguagens naturais isso não é tão simples, porque as pessoas utilizam abreviações neologismos que dificultam a identificação das palavras.

Tokenização

Existem linguagens em que as palavras são separadas por espaços, portanto, chamadas de linguagens segmentadas.

E outras linguagens que não realizam a separação de palavras por espaço e, por isso, são chamadas de linguagens não segmentadas. Na China e Tailândia as palavras são escritas sem qualquer separação para indicar início em fim.

Cada uma dessas linguagens exige uma abordagem diferente para realização da tokenização.

Tokenização

Nas linguagens segmentadas basta informar que o elemento para separação a ser utilizado é o espaço em branco.

Em linguagens sem quaisquer identificadores para separação das palavras, o processo de tokenização precisa ser feito utilizando a compreensão semântica e não estrutural.

Natural Language Toolkit

Para que suas funcionalidades possam ser utilizadas na construção de soluções de processamento de texto, primeiro é necessário instalar e importar os seus recursos.

```
import nltk
nltk.download()
frase = "As pessoas estão muito satisfeitas com o seu produto."
tokens = nltk.word_tokenize(frase)
tokens
['As', 'pessoas', 'estão', 'muito', 'satisfeitas', 'com', 'o', 'seu', 'produto', '.']
```

Análise léxica

Depois de serem separadas, as palavras precisam ser compreendidas, então começa-se a análise léxica. As palavras passam a ser enxergadas de modo diferente.

O verbo observar, por exemplo, é visto como *lemma* de outras palavras como: observador, observa, observado.

Análise léxica

Durante a análise léxica cada uma das palavras é relacionada com o seu devido *lemma*, que também são chamados de formas canônicas.

Para que seja possível fazer isso, é preciso que se conheça todos os *lemmas* da língua em que o texto estiver escrito.

Análise Sintática

Para o processamento de linguagem natural, a frase é uma unidade básica em termos de significado.

Portanto, depois de separadas e com os seus *lemmas* identificados, parte-se para a compreensão do significado de cada uma das frases.

Análise Sintática

Para estruturar frases corretas do ponto de vista da sintaxe da língua é fundamental que se tenha o conhecimento sintático a respeito dela.

A sintaxe diz respeito à forma como as palavras devem estar dispostas ao longo de uma frase para que elas estejam corretas do ponto de vista linguístico e que expressem o sentido desejado.

Análise Semântica

A análise semântica é referente aos significados das palavras, que para a PNL devem ser construídos por frases.

Aqui está a grande dificuldade do processamento de linguagem natural que é tratar as ambiguidades e possíveis interpretações de uma frase em função de seu contexto.

Análise Semântica

O sucesso da análise semântica está atrelado diretamente a qualidade dos procedimentos anteriores, que permitiram a devida composição da frase para que ela pudesse ser interpretada corretamente.

Análise Pragmática

A análise pragmática diz respeito principalmente ao tratamento do uso dos pronomes identificando o papel que eles cumprem em cada uma das frases.

Para definição do pronome é preciso analisar o sujeito, contar com a análise sintática para saber onde posicioná-lo ou não.

Obrigada!

Ana Laurentino