Resumo 13 – PLN

A décima terceira aula de Processamento de Linguagem Natural teve início com um comunicado sobre trabalhar com a biblioteca nltk para algumas análises um pouco mais aprofundadas de comparadas com o que vimos até o momento.

Um dos grandes desafios dentro de Processamento de Linguagem Natural e das áreas relacionadas a Computação como um todo é a de identificação do tipo linguístico de uma determinada palavra. Por exemplo, é possível identificar se uma palavra é um verbo? Um sujeito? Respondendo essas perguntas adentramos ao assunto da aula: entidades nomeadas.

Um reconhecedor permite identificar e classificar um texto escrito em linguagem natural. Tal tarefa pode se realizada através de buscas em dicionários. Assim é possível entender, por exemplo, que uma data está relacionada a uma determinada pessoa em um determinado local. Em um trecho de exemplo utilizando uma breve descrição da Universidade Federal do ABC, o algoritmo pôde identificar palavras individuais em algumas categorias como, por exemplo, eventos, conceitos, lugares, organizações. Em outras palavras, o contexto foi levado em consideração.

Entre as principais aplicações práticas, as entidades nomeadas podem ser utilizadas para reconhecer e entender relações e associações entre outras atividades. Como exemplo, foi mostrada a estruturação de dados utilizada pelo Wikipedia, onde o conceito de entidades nomeadas é aplicado a composição de conteúdo das páginas para identificação de páginas relacionadas.

Um outro exemplo de código foi desenvolvido e mostrado pelo professor para aplicar, a um trecho textual, uma espécie de classificação entre palavras “Importantes” e “Não Importantes” baseada em uma regra que verifica se a primeira letra daquela palavra é maiúscula. Apesar de simples e aparentemente efetivo, alguns casos incoerentes surgem (principalmente levados por palavras em início de frase). Outras regras podem ser adicionadas ao código par a identificar se a palavra em questão encontra-se após um sinal de pontuação, evitando assim a identificação de palavras comuns em início de frase.

Em mais um exemplo, outro trecho foi utilizado para identificação de Localização, Moeda, Organização, Período e Pessoa. Dessa forma, algumas palavras desse trecho foram sinalizadas de acordo com seu significado dentro das categorias propostas. Para a realização de tudo isso, é necessário um dicionário. Alguns casos fora da curva são inevitáveis: JK se trata de que? Escola? Avenida? Pessoa? Governo? Mesmo os reconhecedores modernos e mais sofisticados enfrentam esse tipo de problema.

Para finalizar a aula, o professor comentou a respeito das entregas para o projeto final.