Resumo 14 – PLN

O professor Jesús deu início a décima quarta aula de Processamento de Linguagem Natural abordando a biblioteca NLTK e comentando sobre algumas análises a serem feitas nas próximas aulas. Recomendação de livro: NLTK Essentials.

Como primeiro exemplo, o professor mostrou um texto de exemplo e fez uma experiência com os alunos perguntando sobre o número de frases presentes em determinado parágrafo. Obviamente, temos o ponto final como separados de frases. Entretanto, existem algumas situações apartadas que podem ser confundidas com frases (como por exemplo, o ponto final em No. 754). A biblioteca NLTK automaticamente realiza a separação de frases a partir do método sent\_tokenize(string), dividindo o texto a partir de um ponto.

Uma segunda função chamada word\_tokenize(sentence), utiliza uma expressão regular e divide a sentença em palavras, incluindo sinais de pontuação, como ponto, vírgula, parênteses, entre outros.

A terceira função é a pos\_tag(tokens) que, por sua vez, recebe uma lista de palavras (tokens) e atribui um rótulo (label) a cada uma das palavras. Este método é um rotulador de palavras e um componente principal em análises de PLN. A tarefa de etiquetagem de POS implica simplesmente em rotular palavras com sua característica de linguagem (i.e. substantivo, verbo, adjetivo, advérbio, pronome). Uma questão importante realizada em sala mencionou a presença de palavras ambíguas, ou seja, de grafia igual mas com significados diferentes (por exemplo, Washington, Albert Einstein, entre outros). Para estes casos, o professor disse que o que tem sido visto é o contexto sendo considerado para essa definição. Um corpus famoso relacionado a essa classificação dos tokens é o Penn Treebank.

Voltando um pouco ao exemplo utilizado para a explicação dos métodos, o significado de alguns rótulos foi dito, além da dificuldade em realizar esse tipo de tokenização e classificação em idiomas árabes e chineses.

Até o momento, dividimos a frase, dividimos as palvras, e as palvras foram rotuladas. Dentre os rótulos mais comuns, temos People, Organization, Location(LOC), Geo-Political(CPE), Entity, Facility(FAC), Vehicles(VEH).

Na abordagem intuitiva, a sumarização é a classificação de frases em sua importância. Assim, é possível assumir que normalmente uma frase que tem mais substantivos e entidades tem maior relevância.

No final da aula, o professor comentou a respeito do que seria o Desafio 3 da matéria, sendo este contemplado pelo resumo de um artigo da revista CACM com requisito de que tal artigo tenha, no mínimo, duas ou mais páginas.