Resumo 6 – PLN

A aula de número 6 da matéria de Processamento de Linguagem Natural foi iniciada a partir de um resumo da aula anterior, abordando conceitos de N-gramas. Como visto anteriormente, de acordo com a definição, um N-Grama é uma sequência contígua de N elementos, dados por caracteres, palavras, sílabas, fonemas, pares-base, entre outros, comumente obtidas a partir de um *corpus*. Para exemplificar, o professor Jesus Mena mostrou uma tabela com duas colunas relacionando o número de elementos e o número do N-Grama. Um N-Grama com 1 elemento é tido por Unigrama. Com 2 elementos, Bigrama. Com 3 elementos, Trigrama. Com 4 elementos, 4-grama e assim sucessivamente. Exemplos utilizando frases foram tratados para o melhor entendimento desse núcleo pela turma.

Após isso, foram abordados conceitos de probabilidade para identificar a probabilidade de uma palavra em uma frase. Logo em seguida, o pressuposto de Markov pôde ser definido para a turma como sendo a suposição que a probabilidade de uma palavra depende apenas da probabilidade de uma(s) palavra(s) anterior(es).

Alguns exemplos foram dados para fixar o conhecimento. As frases utilizadas foram: <s> I am Sam </s>, <s> Sam I am </s> e <s> I do not like green eggs and ham </s>. As probabilidades de ocorrência das frases foram calculadas utilizando os conceitos aprendidos.

Assim, foi possível mergulhar no verdadeiro objetivo da aula: propor cases práticos envolvendo os conceitos de N-gramas. O primeiro deles visa analisar bi-gramas em frases. Assim como em algumas aulas anteriores, a situação prática utilizada para exemplificar este conceito foi a análise realizada no livro de Machado de Assis. Além do exemplo prático, o professor Jesus Mena mostrou o código por trás do cálculo das probabilidades de ocorrência de frases, obtendo os resultados esperados para a mesma frase (“ele é uma pessoa”) utilizada como exemplo dentro da obra de Machado de Assis.

O segundo case demonstrado foi o de associação a próxima palavra de uma frase. Ainda utilizando a obra de Machado de Assis, o exemplo dado partia da inserção de uma frase e tinha, como saída, a próxima palavra de ocorrência mais provável dentro do contexto com sua respectiva probabilidade. Assim como no case anterior, o professor Jesús Mena mostrou o código em Python responsável por essa implementação.

O terceiro case exemplificou a análise de n-gramas em frases e, partindo dos modelos probabilísticos de unigramas, bigramas e trigramas, calculava a probabilidade de ocorrência de frases em diversos contextos. Como exemplo, esse case foi aplicado na obra A Semana de Machado de Assis e também em todas as suas obras, obtendo resultados diferentes para cada um dos contextos.

No quarto case, tivemos contato com a geração de frases a partir de palavras unidas em tuplas ou dicionários. Tendo em mãos diversas palavras, o algoritmo (também mostrado pelo professor) realizava o melhor agrupamento dentro de um contexto (obra A Semana de Machado de Assis). Nesse caso, foi possível aplicar o algoritmo utilizando unigramas, bigramas, trigramas e outros.

No case de número cinco, utilizamos algo semelhante ao case de número quatro, porém com sinais de pontuações e números. Os resultados foram demonstrados também na obra de Machado de Assis e as expressões regulares foram mostradas em sala.

Logo em seguida, falamos sobre correção ortográfica e sua real importância para sistemas de busca e processamento de linguagem natural, seja na internet ou na academia. Os tipos de erros foram elencados e uma atividade foi passada em sala para fixação do conhecimento. Conceitos de canal ruidoso (uma das questões do teste) foram abordados e a regra de Bayes foi mostrada com importância crucial dentro desta análise. A dica de pesquisa do professor para as aulas posteriores foi Peter Norvig.