Resumo 5 – PLN

Como de costume, a aula de Processamento de Linguagem Natural (quinta aula, no caso) foi iniciada a partir de um resumo da aula anterior e abordou uma breve revisão dos conceitos já tratados. Logo em seguida, foram iniciadas as tratativas da aula em questão: os N-Gramas.

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.

Outros exemplos práticos foram extremamente importante para a explicação dos N-Gramas, como por exemplo, uma música da Alannis Morissete, a divisão da frase “I hate cigarettes with a passion” em N-gramas diferentes e, por fim, um projeto envolvendo a atribuição de avaliadores para projetos, cujo objetivo é relacionar, através de alguns parâmetros definidos, os melhores avaliadores para cada projeto.

Após mencionar e detalhar alguns conceitos probabilísticos, uma atividade prática foi realizada de modo a fixar o conhecimento da turma. A ideia principal foi definir a probabilidade de uma determinada frase ocorrer, de acordo com as relações entre as palavras e as formações de N-Gramas (cadeias de Markov).

Um ponto importante, entendido durante a atividade, relaciona a presença de um caractere que indica o início (<s>) e o fim (</s>) de uma frase. Esses caracteres devem ser considerados durante o cálculo de probabilidade para o corpus.

A atividade em questão relacionava a quantidade de unigramas, bigramas e trigramas possíveis em um conjunto de n palavras (corpus). Algumas regras foram mencionadas para se chegar ao número correto. Por fim, foi pedido ao aluno o cálculo da probabilidade de ocorrência da frase “I want chinese food” de acordo com uma tabela de probabilidades fornecidas.

Já no encerramento da aula, foram abordados alguns conceitos sobre a identificação de padrões. Exemplos como a quantidade de menções sobre os anos no decorrer dos anos (anos 2000, ou y2k na suspeita do bug do milênio, por exemplo) ou então a menção a presidentes dos Estados Unidos em épocas diferentes de suas respectivas atuações. Uma dica final considerada pelo professor foi a utilização do Log para uma melhor visualização dos resultados probabilísticos.