**Resumo Apresentações 22/08/19**

Dando início ao primeiro dia de apresentações dos projetos de Processamento de Linguagem Natural, tivemos os seguintes trabalhos: “Arquitetura para um Sistema de Perguntas e Respostas Usando Ontologias”, “Implicit Emoticon”, “Suicide Notes”, “Vocabulário dos Artistas Musicais Brasileiros”.

No projeto “Arquitetura para um Sistema de Perguntas e Respostas Usando Ontologias” foi utilizado o artigo “Architecture of na Ontology-Based Domain Specific Natural Language Question Answering System” publicado em Novembro de 2013 e proposto por Athura P.M., Sreeja M e P.C.Reghuraj. Com 19 citações, 11 referências e 10 páginas, a ideia geral do artigo e do projeto é desenvolver uma arquitetura para um sistema de perguntas e respostas para análise da ontologia de perguntas.

Dispostos de tags e utilizando como base o sistema de resposta automática do Stack Overflow, o grupo utilizou uma base de dados do Kaggle (StackSample: 10% of Stack Overflow Q&A). Dessa forma, com essa base de tags, o grupo treinou modelos para classificar automaticamente uma nova pergunta recebida.

Como análise e preparação do corpus, o grupo utilizou a distância de Levenshtein para obter uma aproximação entre todas as palavras das frases e depois a distância do cosseno para comparar as frases inteiras, somando os valores no final. O modelo Naive Bayes obteve a melhor performance.

No projeto “A Rule-Based Approach to Implicit Emotion Detection in Text”, foi utilizada Análise de Sentimentos para detecção de emoção de acordo com um texto recebido na entrada. Para preparação dos dados, um arquivo inicial foi processado, obtendo assim as sentenças juntamente com as palavras. Após isso, foi aplicada a técnica de POS tagging, Dependency Parsing, análise de tempo verbal e, por fim, cálculos de polaridade. O algoritmo utilizado para o treinamento foi o Naive Bayes. Como conclusão, o método baseado em regras alcançou melhores resultados que a abordagem por matching léxico, resultando em uma performance competitiva com o método supervisionado. Entretanto, a necessidade humana a cada iteração é um ponto a ser levado em consideração.

No projeto “Suicide Note Classification Using Natural Language Processing: A Content Analysis”, o artigo, de mesmo nome, foi proposto por John Pestian, Henry Nasrallah, Aurora Bennett e Antoon Leenaars. Tendo o suicídio como tema principal, o grupo trouxe alguns números importantes relacionando a gravidade desse assunto ao redor do mundo. O objetivo do projeto foi utilizar técnicas de PLN para identificar textos que pudessem indicar tendências suicidas.

No pipeline de pré-processamento dos dados, o grupo utilizou ferramentas da biblioteca NLTK, quebra de palavras (tokenização), remoção de pontuação, remoção de stopwords, população do dicionário global e criação dos dicionários de cada documento. Como resultado, o grupo alcançou melhores resultados com o modelo SVM.

No projeto “Vocabulário dos Artistas Musicais Brasileiros”, o grupo utilizou dados de músicas coletados através do portal Vagalume para realizar análises envolvendo técnicas de PLN. Assim, o grupo coletou os dados, realizou a limpeza e a preparação, o cálculo do vocabulário e, por fim uma análise dos resultados. O grupo utilizou um esquema de tradução de músicas utilizando uma biblioteca que acessa o grupo tradutor.

Os resultados são interessantes: artistas como Racionais MC’s, Hungria, Emicida, Projota e Anitta são aqueles com maior número de palavras. Já Midian Lima, Raça Negra, Fernandinho, Amado Batista e Belo são aqueles com o menos número de palavras. Com relação a taxa de palavras (vocabulário/total de palavras), Tio Carreiro e Pardinho, Zé Ramalho, Milionário e José Rico e Maneva são aqueles com maiores índices.