**Resumo Apresentações 27/08/19**

Caminhando para o terceiro dia de apresentações dos projetos de Processamento de Linguagem Natural, o artigo “Understanding Semantic Change of Words Over Centuries”, publicado em 2011 na Conference of Information and Knowledge Management da ACM (ACM – CIKM) – Escócia, foi utilizado por um dos grupos para analisar as mudanças na semântica ao longo dos séculos.

Para a implementação, o grupo utilizou o dataset Google N-Gram com palavras em 5-grams distribuídas em 800 arquivos de 1GB cada. Para a análise, o grupo coletou aproximadamente 1,07 milhões de 5-grams relacionadas a “Mulheres” e aproximadamente 4,5 milhões de 5-grams relacionadas a “Homem”. O grupo utilizou LDA (Latent Dirichlet Allocation) para análise dos tópicos de cada palavra em cada documento “d”.

No processamento, o grupo selecionou, para cada um dos grupos (“Mulheres” e “Homens”) duas palavras anteriores e posteriores, criou um corpus para cada século, remoção de caracteres via RegEx, transformação em minúsculas, tokenização, remoção de stopwords, aplicação de Stemmer (Porter Stemmer), criação de bag-of-words, TD-IDF e, por fim, LDA. Alguns dos resultados do grupo mostrou as palavras mais comuns para “Mulheres” (old, Young, beautiful, elderly, older) e para “Homens” (old, honest, Young, woman, came).

No projeto “A Web of Hate: Tackling Hateful Speech in Online Social Spaces”, o grupo comparou diferentes abordagens para definição de discurso de ódio e propôs uma nova abordagem para detecção de discurso de ódio. Para tal, o grupo utilizou os classificadores Naive Bayes, SVM e Regressão Logística) em um dataset do twitter adaptado (“Hate Speech Identification”).

Seguindo com as apresentações, o projeto “Classificadores Naive Bayes Informados pelo Léxico de Sentimentos” teve como objetivo a aplicação de análises em críticas de restaurantes. O artigo utilizado foi o “Senti-lexicon and Improved Naive Bayes Algorithms for Sentiment Analysis of Restaurant Reviews” publicado por Hanhoon Kang, Seong Joon Yoo e Dongil Han em 2012 na Sejon University, Coreia do Sul.

Alguns dos problemas citados pelo grupo durante a implementação do projeto envolveram a utilização do SentiWordNet como um léxico de sentimentos generalista que não avalia bigramas, além da diferente de aproximadamente 10% entre a acurácia de classificação positiva para a acurácia de classificação negativa, reduzindo assim a acurácia média.

Durante a preparação, o grupo utilizou a tokenização das palavras para encontrar a classe gramatical das palavras na crítica. Assim, foi possível extrair o significado comum da palavra e selecionar os sentimentos positivo e negativo desse significado usando o SentiWordNet. O léxico nos sentimentos foi adicionado para substantivos, adjetivos e advérbios. Assim, o grupo pôde preparar os dados e utilizar classificadores para analisar o sentimento das críticas.

Para o projeto “Aplicação de Técnicas de PLN para Detecção de Plágio em Documentos” foi utilizado o artigo “Using Natural Language Processing for Automatic Detection of Plagiarism” apresentado em “Proceedings of the 4th International Plagiarism Conference 2010” com 34 citações até o momento. O dataset utilizado foi o corpus “Clough & Stevenson” de 2009 composto por pequenos trechos de texto em que alunos respondiam questões sobre textos extraídos da Wikipedia: 100 documentos (95 suspeitos de plágio e 5 originais).

As respostas foram classificadas em 4 classes: Near Copy (copy e paste do texto original), Light Revision (pequenas substituições do texto original), Heavy Revision (texto baseado no original realizando reestruturação e parafraseando) e Non-plagiarism (baseado apenas nos conhecimentos dos alunos).

Durante a preparação, o grupo aplicou a tokenização, normalização, remoção de StopWords, Stemming, Lemmatisation, Distância entre vetores por cossenos, calculando assim as similiaridades de cada documento com os documentos originais com o intuito de gerar features para a aplicação de um classificador de textos.

Como conclusão, o grupo obteve um desempenho menor com relação ao artigo que serviu como base para desenvolvimento da metodologia. No artigo original, os resultados obtidos para a acurácia ficaram entre 60 e 70%, entretanto não foram aplicadas todas as técnicas do artigo em nossa metodologia, justificando o valor da acurácia de 57,89%. Para melhoria do trabalho realizado e futuros projetos, pode-se combinar as técnicas utilizadas com os demais procedimentos de PLN mais avançados, como a modelagem com N-gramas e nomeação de entidades, para gerar melhor desempenho e melhores resultados na classificação de textos em diferentes níveis de plágio. Além disso, para melhorar a acurácia e o desempenho do classificador de Naive Bayes pode-se utilizar um corpus com maior massa de dados classificados, dado que não foram utilizados muitos dados para treino e teste do modelo.