MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA

RELATÓRIO -2019- Mineração de Dados **Manaus, 14 de novembro de 2019.**

1. DEMANDA

As atividades realizadas por este grupo na Operação Ricardo Franco 2019 tiveram como objetivo inicial realizar uma mineração de dados no software C2 em Combate. Estes dados, armazenados em texto corrido pelos militares que operam o software, referem-se a operações realizadas em todo área contida nos limites territoriais do CMA. Ao longo dos dias, devido à dificuldade encontrada com a primeira proposta de projeto, algumas metas foram redefinidas, mas manteve-se o propósito geral, idealizado originalmente, de otimizar o processo de obtenção de informações a partir de um tratamento e estruturação destes dados presentes no software em questão. Esta é uma demanda do COp do CMA.

2. OBJETO

Para a realização do projeto proposto foram abordados principalmente assuntos relativos à mineração de dados, visando explorar a grande quantidade de informações presentes no software para identificação de padrões e associações. Como o campo do banco em que estes dados estão inseridos é preenchido por meio da língua da portuguesa, o grupo teve contato com diversas ferramentas de processamento de linguagem natural, em especial a biblioteca em python spaCy, cujo modelo apresenta-se mais eficiente para a língua inglesa. Além disso, visando o armazenamento dos dados estruturados e a visualização deles por meio de uma interface gráfica, outras duas ferramentas foram estudadas e empregadas no trabalho: a plataforma de análise e visualização de dados Kibana e o servidor de buscas Elasticsearch, que permite trabalhar com números volumosos de dados.

3. EQUIPE

- Amon Rhaniery Brito Machado (1º Ten, Engenharia da Computação);

- Leandro de Mattos Ferreira (Maj, Orientador de Engenharia da Computação);

- Lincoln de Queiroz Vieira (1º Ten, Engenharia da Computação);

- Roberto Tadeu Abrantes de Araújo (1º Ten, Engenharia da Computação);

4. COLABORADORES EVENTUAIS

Cap Campana, Ch Seç Gerência e Captação do 4º CTA: junto ao comando do 4º CTA, disponibilizou uma sala dedicada à ORF equipada com diversos computadores para a confecção do projeto. Além disso, acompanhou de perto o desenvolvimento das atividades, fornecendo apoio técnico e sanando diversas dúvidas surgidas ao longo do período de trabalho. Proveu ainda o contato entre a equipe e os clientes do projeto, aqueles militares cujos trabalhos serão diretamente afetados pelo produto gerado.

2º Sgt Sheldon, Aux Seç Info Geográficos CCOp: representando os clientes do projeto, disponibilizou-se a todo momento para eventuais retiradas de dúvidas mais específicas relacionadas ao software C2Cmb. Detalhou o problema da forma que impacta atualmente as atividades do CMA e elaborou junto à equipe alternativas de soluções para os problemas encontrados. Realizou também alguns encontros em seu ambiente de trabalho a fim de mostrar ao grupo o banco de dados do software em questão e toda sua complexa estrutura.

5. HOSPEDAGEM E TRANSPORTES

Hospedados no 2º Grupamento de Engenharia, da 12ª região militar, durante todo o período da Operação Ricardo Franco;

Transportados de avião comercial até Manaus, bem como na volta ao Rio de Janeiro.

6. SÍNTESE DESCRITIVA DAS ATIVIDADES

a. 1º Dia – 11 de novembro de 2019

Durante a parte da manhã, nos dirigimos pela primeira vez ao 4º Centro de Telemática de Área e, assim que chegamos, fomos recepcionados pelo Capitão Campana, que nos levou à sala destinada ao desenvolvimento dos projetos. Logo em seguida, dividimos-nos entre os grupos, as ferramentas e computadores disponibilizados foram alocados aos que solicitaram e o capitão fez uma breve explanação acerca dos projetos, acrescentando algumas solicitações repentinas e sanando algumas dúvidas que surgiram sobre os temas dos trabalhos. Neste tempo também foram realizadas configurações preliminares ao início do desenvolvimento, como à rede de internet da OM.

No período da tarde, a equipe buscou aprofundar o entendimento do trabalho e, conforme havia planejado em um momento anterior ao início da operação, começou a trabalhar com a biblioteca em python spaCy, voltada ao processamento de linguagem natural. Assim, após baixá-la, começamos a analisar as associações que o seu modelo gerava quando fornecíamos como input algumas amostras dos campos textuais sobre os quais vamos trabalhar. De acordo com os resultados obtidos, levando em conta padrões identificados e relações entre as entidades, tentamos implementar algumas alterações no seu modelo de forma a otimizar o seu resultado para o nosso problema em específico.

b. 2º Dia – 12 de novembro de 2019

Durante o período da manhã, continuamos tentando melhorar o resultado do modelo acrescentando algumas funcionalidades a fim de se adaptar ao problema particular que queremos resolver. Devido à grande dificuldade encontrada, característica ao lidar-se com processamento de linguagem natural e potencializada pelo fato dos campos serem particulares da carreira militar e preenchidos ao livre arbítrio do operador, o grupo conseguiu resultados pouco satisfatórios, alcançando taxas de acerto baixas para o modelo. Vale ressaltar que outras duas ferramentas que trabalham com processamento de linguagem natural foram testadas: as bibliotecas de python NLTK e UDPipe, tendo em vista que possuíam modelos diferentes.

Logo após o almoço, o Cap Campana trouxe ao ambiente de trabalho dois militares que possuem grande contato com o banco de dados do software C2Cmb em suas funções diárias e, assim, poderiam auxiliar para uma melhor compreensão da sua arquitetura, bem como ajudar a entender os principais problemas enfrentados pelos militares que desejam extrair deles informações mais elaboradas. A conversa entre os integrantes do grupo, o Cap Campana e estes dois militares foi bem produtiva e ajudou a tornar mais clara a demanda real do cliente.

No final da tarde, entretanto, ao considerar-se a pouca evolução que a equipe alcançou até o momento e o consenso chegado com os professores orientadores acerca da dificuldade associada ao problema original, o objetivo do projeto foi reformulado pela primeira vez. Propôs-se que, em vez de implementar um modelo geral, que tente minerar todas as informações presentes nos campos textuais, o modelo seja orientado a uma solicitação do cliente. Ou seja, o militar entraria com uma palavra sobre a qual deseja obter-se alguma informação e o script retornaria todas as vezes em que ela apareceu nos campos como também o contexto associado a ela nessas aparições.

c. 3º Dia – 13 de novembro de 2019

Já que o projeto tinha ficado um pouco menos complexo, o grupo decidiu por voltar a usar a spaCy, ferramenta que até então tinha apresentado os melhores resultados. Além disso, o 2º Sgt Sheldon, um dos dois militares que vieram conversar sobre o projeto a pedido do Cap Campana, levou-nos ao CCOp e nos mostrou, na sua máquina de trabalho, exatamente como as tabelas do banco de dados do C2Cmb estão organizadas, detalhando de forma minuciosa o seu esquema e onde estão inseridos os campos textuais que contém os dados que nos interessa.

Seguimos com adaptações do modelo para alguns casos particulares de palavras chaves e os resultados, embora tenham melhorados consideravelmente quando comparado aos da primeira proposta de projeto, continuaram aquém do esperado. Desta forma, o Cap Campana em conjunto com o Sgt Sheldon, decidiram por reformular o objetivo do trabalho pela segunda vez. A terceira (e definitiva) proposta de projeto trabalhará em cima de somente alguns campos textuais. Estes apresentam informações de algumas operações específicas e caracterizam-se por terem sido preenchidos de forma padronizada, facilitando a separação dos dados e um posterior tratamento desta informação. Além disso, ficou como responsabilidade do grupo criar um local de armazenamento para estes dados estruturados e uma interface gráfica para apresentá-

los.

d. 4º Dia – 14 de novembro de 2019

Com o projeto enfim decidido, destinamos o período da manhã para a parte do script responsável por obter os dados e separá-los de forma estruturada conforme seus atributos. Como já tínhamos nos aprofundado bastante sobre o tema nos dias anteriores quando os desafios eram maiores, esta etapa foi relativamente rápida e tranquila. Logo em seguida, começamos a pesquisar ferramentas para a segunda etapa, relativa ao armazenamento do dado estruturado e a posterior visualização deles.

As duas ferramentas escolhidas foram o Elasticsearch e o Kibana, de modo que levamos o final da manhã e boa parte do período da tarde para instalarmos em nossas máquinas e aprendermos boa parte de sua sintaxe e funcionalidades.

Ao final do dia, tivemos um ligeiro avanço nesta etapa posterior, mapeando exatamente os próximos passos a serem realizados no dia seguinte (último dia) e, assim, tendo praticamente encaminhado a conclusão do projeto.

e. 5º Dia – 15 de novembro de 2019

Por ser feriado e pelo fato do projeto estar muito bem encaminhado, trabalhamos apenas meio expediente. Assim, conseguimos concluir as etapas faltantes nesta manhã, finalizando os últimos detalhes relativos à visualização de dados por meio do Kibana.

7. COLETA DE DEMANDAS FUTURAS

- Ferramenta envolvendo aprendizado de máquina para mineração dos dados presentes nos campos textuais, juntamente com um trabalho de reconhecimento previamente realizado;
- Modificação do software C2Cmb: padronização dos dados a serem inseridos;
- Projeto de fim de curso objetivando pesquisar e desenvolver soluções alternativas para o problema do processamento de linguagem natural, focando nas características particulares deste problema.

Possíveis contatos para demandas futuras:

- Cap Campana (Ch Seç Gerência e Captação do 4º CTA);
- 2°Sgt Sheldon (Aux Seç Info Geográficas CCop do CMA)

8. CONCLUSÃO

Durante a realização da Operação Ricardo Franco, o objetivo do projeto foi reformulado duas vezes. Nos dois primeiros escopos, a eficiência do modelo que

desenvolvemos ficou aquém de um limite para garantirmos a confiança do resultado retornado. Diversos foram os fatores que contribuíram para isto:

- A complexidade do problema proposto: processamento de linguagem natural, em especial tratando-se da língua portuguesa, ainda está em uma fase embrionária no mundo da computação, sendo as ferramentas existentes ainda muito pouco confiáveis, obtendo taxas de acertos para o contexto militar bastante aquém do aceitável;
- Dados inseridos sem nenhum tipo de padrão: além de ser um problema difícil, a particularidade dos campos textuais em que faríamos a mineração dificulta ainda mais a situação por serem preenchidos de forma quase aleatória, sem possibilitar a identificação de qualquer padrão existente e com a utilização de diversas siglas e abreviações, que contribuem ainda mais para o aumento da complexidade;
- Tempo destinado à equipe: como último fator, somado a tudo já falado anteriormente, os somente cinco dias de operação caracterizam um período curtíssimo para desenvolver uma solução que requer grande número de testes de eficácia. É, de fato, um problema bastante complexo para apenas uma semana de trabalho.

Tendo em vista o último objetivo proposto ao grupo, foi concluído com sucesso a estruturação dos dados da Operação Acolhida presentes no banco de dados C2Cmb. Estes dados, em específico, estavam conforme um padrão e foi possível realizar uma extração das informações contidas em seus textos corridos para o formato .CSV, arquivo de entrada para ferramentas de visualizações como Kibana.

9. ANEXOS

- Relatório de Informações Técnicas durante as atividades no 4° CTA.
- Script desenvolvido e documentação.

_____

1° TEN LINCOLN
Representante do Grupo de Trabalho