**UNIVERSIDADE DE SÃO PAULO**
**ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO**
DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E SISTEMAS DIGITAIS
CENTRO DE CIÊNCIA DE DADOS (C²D)

# NOVO MODELO DE RELATÓRIO DOS BOLSISTAS

09/2021

| **Dados da Bolsa** |
|---|
| **Tipo de Bolsa:** ( X ) IC ( ) TCC ( ) PqEP ( ) ME |
| **Nome do/a Orientador/a:** Anna Helena Reali Costa |
| **Nome do Projeto:** Chatbot Q&A multi-agente |
| **Período da Bolsa:** 01/02/2021 a 01/02/2022 |
| **Relatório:** ( ) Final ( X ) Parcial |
| **Período do Relatório: 25/08/2021** a **25/09/2021** |
| **Descrição das Atividades de Pesquisa do Projeto** |
| **Descrição das atividades acadêmicas:** 2° Semestre Letivo, em que o aluno está cursando as seguintes matérias:<br>● PTC3213 - Eletromagnetismo<br>● PCS3225 - Sistemas Digitais II<br>● PSI3213 - Circuitos Elétricos II<br>● PSI3502 - Realidade Virtual<br>● PSI3323 - Laboratório de Eletrônica I<br>● PSI3322 - Eletrônica II<br>● PSI3572 - Computação Visual<br>● PSI3214 - Laboratório de Instrumentação Elétrica |
| **Descrição das atividades planejadas para o relatório (repetir do relatório anterior):**<br>Para este relatório era planejado:<br>● Ter feito uma revisão bibliográfica das técnicas de sumarização automática,<br>● Ter implementado um algoritmo para aplicar os modelos de sumarização dos dados das Atas do Diário da Assembleia da República (DAR)<br>● Fazer a submissão do projeto de IC para o SIICUSP 2021 - 29º Simpósio Internacional de Iniciação Científica e Tecnológica da USP. |

**Descrição das atividades de pesquisa realizadas:**

A arquitetura do sistema, que inicialmente se pretendia para Q&A, agora está mais geral, visando interações diversas com o usuário, a fim de extrair informações de forma simples e visual de algum tema -- que, por falta de material acessível e abundante da Amazônia Azul (inicialmente proposto) foi trocado para Atas do Diário da Assembleia da República (DAR) e Iniciativas do Parlamento Português. O motivo foi haver texto na língua portuguesa em abundância (pré-requisito do projeto) e texto de fácil acesso (material em txt disponível na web). Assim, um corpus com este material foi montado para a realização de experimentos que fornecem funcionalidades (como agentes).

Uma das funcionalidades desenvolvidas foi a **identificação e contagem de palavras-chave** (exemplo: corrupção e derivados) nas manifestações dos parlamentares nas Declarações Políticas das Atas. O resultado destes desenvolvimentos foi submetido em artigo ao evento ENIAC 2021 e estamos aguardando as respostas.

Outra funcionalidade foi a **Sumarização Automática**, desenvolvida no período deste relatório.

A arquitetura completa foi apresentada no resumo submetido ao SIICUSP 2021.

**Sumarização Automática:** Durante esse período, o bolsista trabalhou para a criação de um algoritmo para criação do algoritmo para sumarizar, sem perda de sentido, as falas de deputados descritas nas Atas do DAR. Para isso foram utilizados alguns modelos para testes. O algoritmo final conta com duas etapas (custosas) de tradução automática e segue o fluxo abaixo:

Fala (pt) → Tradução (pt-en) → Sumarização (en-en) → Tradução (en-pt) → Resumo

Esta primeira etapa de tradução é bem custosa computacionalmente, então o bolsista planeja utilizar as máquinas do C²D para efetuar a tradução de todo o banco de dados a priori.

Nas etapas de tradução foram testados os modelos PTT5 e M2M100. O primeiro não foi muito bom e precisaria de *fine-tuning*. O segundo apresentou resultados razoáveis e foi o adotado neste projeto.

A etapa de Sumarização foi testada usando os modelos: BertExt, BART, Pegasus e Pegasus-XSUM. O que apresentou melhores resultados foi o modelo Pegasus. A avaliação feita foi subjetiva, utilizando apenas a avaliação humana dos resultados. Mesmo assim, os resultados foram expressivamente superiores aos outros.

O bolsista também ajudou na montagem da funcionalidade de **Modelagem de Tópicos**, que visa determinar os temas mais relevantes em um documento. A proposta inicial usa o BERTopic para uma primeira segmentação de palavras relacionadas a um mesmo tema e depois atribui o tema final utilizando o modelo roBERTa para Zero-Shot Classification. Os temas finais são definidos pelos projetistas e incluem “Saúde”, “Educação”, “Ciências e Tecnologia”, entre outros.

**Apresentações:** O bolsista irá apresentar o projeto de IC durante o 29º Simpósio Internacional de Iniciação Científica e Tecnológica da USP, marcado para ser a segunda apresentação do dia 30/09/2021 às 9h00 na Sala 10. A apresentação será transmitida por este link. Além disso, o projeto foi apresentado no dia 24/09 para a equipe portuguesa, para traçar os objetivos de visualizações e outputs do sistema.

**Artigos:** Durante este período tivemos um *Extended Abstract* aceito para o Seminário "Artificial Intelligence: Democracy and Social Impacts" (site do seminário) pelo C4AI.

**Descrição das próximas atividades:**

Para o próximo mês, pretende-se integrar as funcionalidades na arquitetura e aprimorar a interface. Planejamos expandir a base de dados e também convertê-la para inglês para também poder testar em modelos que são reportados como estado-da-arte, mas apenas funcionam para língua inglesa.

**Houve alteração significativa no tema ou prazo:** ( ) Sim ( X ) Não

**Justifique em caso positivo:** Houve somente mudança no tópico de **aplicação**: por falta de material em língua portuguesa (e também pouco material em txt para a língua inglesa deste tópico) para a Amazônia Azul, resolveu-se trabalhar com os textos da Assembleia Portuguesa (Atas do DAR, Iniciativas, etc), que são abundantes e disponíveis na web. Porém, o foco técnico (desenvolver funcionalidades em agentes de uma arquitetura que processa língua portuguesa) do projeto continua o mesmo e tem avançado muito bem.

**Apreciação Circunstanciada do/a Orientador/a sobre as Atividades da/o Bolsista**

**Etapa cumprida no relatório:** ( x ) Ótimo ( ) Bom ( ) Regular ( ) Fraco

**Programação para a próxima etapa:** ( x ) Ótimo ( ) Bom ( ) Regular ( ) Fraco

**Resultados em relação às expectativas iniciais:** ( x ) Acima ( ) Dentro ( ) Abaixo ( ) Muito abaixo

**Previsão de conclusão no prazo:** ( x ) Sim ( ) Não

**Justifique em caso negativo:**

**Apreciação do/a orientador/a:** Conforme justifiquei acima, após a mudança da aplicação em foco (antes era a Amazônia Azul e agora estamos trabalhando com as Atas da Assembleia Portuguesa e outros documentos disponíveis na web), o trabalho evoluiu bem rapidamente. O bolsista amadureceu bastante, tem feito um ótimo trabalho, discutindo de igual para igual com outros alunos de mestrado trabalhando em ferramentas similares. O aluno submeteu um resumo de seu trabalho no SIICUSP 2021 e o apresentará no dia 30/09 (das 9h00 - 12h00). Acredito que teremos um bom trabalho no final deste projeto de IC.

**Protocolo**

**UNIVERSIDADE DE SÃO PAULO**
**ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO**
DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E SISTEMAS DIGITAIS
CENTRO DE CIÊNCIA DE DADOS (C²D)

| |
|---|
| **Data:** 25/09/2021 |
| **Nome Completo da/o Bolsista:** Enzo Bustos Da Silva |