Dados da Bolsa

Tipo de Bolsa: (X) IC ( ) PqEP ( ) TCC ( ) ME ( ) DO ( ) PD

Nome do Orientador/Supervisor: 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: 25/04/2021 a **25/05/2021**

Descrição das Atividades de Pesquisa do Projeto

Descrição das atividades acadêmicas:

- PSI3211 - Circuitos Elétricos I
- PCS3115 - Sistemas Digitais I
- PSI3212 - Laboratório de Circuitos Elétricos
- PEF3208 - Fundamentos de Mecânica das Estruturas
- PSI3321 - Eletrônica I
- MAP3121 - Métodos Numéricos e Aplicações
- 4323301 - Física Experimental C
- SCC0633 - Processamento de Linguagem Natural

Descrição das atividades de pesquisa:

Augmented Democracy:

- Os *Dashboards* para a visualização de dados já estavam prontos, mas foram atualizados para futuramente serem disponibilizados em um notebook interativo no Google Colab de forma mais amigável para o usuário
- Para cada uma das falas dos políticos foi executada uma rotina de pré-processamento para poder serem extraídas as *features* mais relevantes para os nossos futuros modelos:
    - Remoção de *typos*: prováveis erros de digitação ou erros que ocorreram na hora da extração das atas.
    - Remoção de símbolos gráficos: queremos apenas as palavras, então removemos todos os sinais de pontuação e caracteres especiais.
    - Etapa de *Tokenização*: transformando o texto em um vetor de palavras, sensível a casos de nomes que devem ser um *token* único.
    - Remoção de *stopwords* "padrão" da língua portuguesa: palavras que acontecem com uma elevada frequência e não tem um grande carga de sentido semântico ("que", "de", "a" → geralmente artigos e preposições são *stopwords*).
    - Criação e remoção de *stopwords* "personalizada" ao vocabulário das Atas utilizando a Lei de Zipf: essa lista tem, não só, palavras específicas do vocabulário que se comportam como *stopwords* (como "Presidente"), mas também os *tokens* que aparecem uma única vez (por não agregarem muito semanticamente).
    - Aplicação de *Lematização* sobre os *tokens*: esse processo permite reduzir significativamente o tamanho do vocabulário, pois ele extrai o Lema das palavras. Isso corresponde à forma masculino singular para substantivos e adjetivos ("deputadas" → "deputado") e, para a forma no infinitivo para o caso de verbos ("fazendo" → "fazer").

- Conversão para vetores: a última etapa foi aplicar algoritmos de word2vec e doc2vec usando o vocabulário de todas as Atas, dessa forma poderíamos criar vetores numéricos a partir de cada uma das falas. O modelo doc2vec foi mais interessante, pois era capaz de colocar uma linha de fala inteira em um vetor. Foi estudada também a possibilidade de utilizar o BERTimbau para aplicar o *embedding*.

Próximos passos:

- Foi criada uma versão piloto de um algoritmo para delimitar continuidade ou interrupção de um assunto baseado nas falas do Presidente usando *Random Forest*.
- Queremos fazer um teste de ablação comparando diversos tipos de *embeddings* e tipos de modelos diferentes para ver como eles se saem na tarefa de classificar se uma dada fala do Presidente da DAR é uma interrupção ou continuação
- Para isso temos ainda que testar outros tipos de *sentence embeddings*, como *GloVe, Skip-Gram, CBOW e BERTimbau* específicos para o nosso vocabulário e também usando os *embeddings* de modelos já prontos, como os do NILC.
- Uso de *K-fold cross validation* para determinarmos as principais métricas dos modelos e também extrair a incerteza

Houve alteração no cronograma original: ( ) Sim (x) Não
Justifique em caso positivo:

Apreciação Circunstanciada do Orientador/Supervisor sobre as Atividades do Bolsista

Apreciação:

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

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

Resultados em relação às expectativas iniciais: ( ) Acima (x) Dentro ( ) Aquém ( ) Muito aquém

Previsão de conclusão no prazo: (x) Sim ( ) Não
Justifique em caso negativo:
Comentários: O bolsista de IC tem trabalhado com afinco, mostrando proatividade nas tarefas, muito bem integrado com a equipe do projeto. Tem contribuído substancialmente para o avanço do projeto. Acredito que vai terminar com sucesso seu IC.

Protocolo

Data: 25/05/2021

Nome Completo do Bolsista: Enzo Bustos Da Silva