Atividade 03 - Projeto

Tema

Análise dos discursos de diferentes nações no Debate Geral da Assembleia Geral da ONU.

Equipe

Nome: *DataMinds*
Repositório GitLab da equipe: *https://gitlab.com/EricFukuyama/IntroCD2*
Membros:

- Anderson Nogueira Silva, 2126516, Andxyz8, andxyz8, andersonnogueira@alunos.utfpr.edu.br, Eng. Comp., UTFPR
- Eric Yutaka Fukuyama, 2126567, EricFukuyama, ericfukuyama, ericfukuyama@alunos.utfpr.edu.br, Eng. Comp., UTFPR
- Erick Jose Teles de Andrade, 2126575, @erickandrade1, Erick Andrade#3830, erickandrade@alunos.utfpr.edu.br, Eng. Comp., UTFPR

Objetivos

- Identificar quais tópicos globais foram mais discutidos por cada bloco (majoritariamente G7 e BRICS) em diferentes períodos.
- Analisar como mudanças geopolíticas, crises econômicas ou políticas impactam a retórica de cada grupo.
- Mapear os principais temas econômicos ao longo do tempo, destacando as semelhanças e contrastes para diferentes blocos econômicos.
- Analisar se os sentimentos de discursos de países aliados convergem ou divergem ao longo do tempo.

Dados e Modelos

Dados Utilizados

- **Corpus:** UN General Debate Corpus (UNGDC) [1], contendo 10.760 discursos de 202 países, abrangendo o período de 1946 a 2023.
- **Metadados:** Informações sobre país, ano, sessão, orador e cargo.
- **Tokenização, lematização e remoção de *stopwords*:** Todos os discursos foram previamente pré-processados para análise.

Modelos Selecionados

- **LDA Over Time (Latent Drichlet Allocation Over Time)**
    - **Descrição**: Este modelo será utilizado para identificar os tópicos mais relevantes nos discursos dos países do G7, BRICS e globalmente ao longo das décadas, permitindo analisar sua evolução temporal.
    - **Implementação**: A biblioteca utilizada será aquela referenciada em [2], lda-over-time, dividindo os discursos em intervalos de 4 anos.
- **Análise de Similaridade Semântica**
    - Utilizaremos *cosine similarity* entre vetores de tópicos para identificar convergências e divergências entre países de um mesmo bloco, ou países que acreditamos que tenham influência sobre outros.
- **Temporal Word Embeddings (TWE)**
    - **Descrição:** Para entender como palavras associadas à economia evoluíram em significado, utilizaremos modelos de embeddings temporais para comparar vetores semânticos em diferentes períodos.
    - **Foco:** Comparar palavras-chave como *economy*, *trade*, *development* e seus contextos em diferentes décadas.
- **Modelos de Análise de Sentimentos**
    - **Modelo Inicial**: Será usado um modelo baseado em Transformers uma vez que o TextBlob e o Flair não tiveram bons resultados. Inicialmente será utilizado o BERT. Todavia, a equipe resolveu deixar em aberto caso seja encontrado outro modelo que possua um melhor desempenho para a análise de sentimentos em discursos da ONU;

- **Estudo dos Resultados da Análise de Sentimentos**
    - **Motivação**: Explorar anos ou países com comportamento anômalo em termos da análise de sentimentos se comparado à década ou ao grupo em que pertencem;
    - **Ferramentas**: Plotagem de boxplots e análise contextual qualitativa dos outliers.

Como os Modelos Serão Usados

- **Identificar Tópicos nos Discursos**
    - O modelo LDA Over Time será aplicado ao corpus completo e aos subconjuntos de blocos econômicos (majoritariamente G7 e BRICS) para mapear a evolução dos temas discutidos.
    - Uma análise será feita comparativamente entre os tópicos identificados para cada bloco, com o intuito de encontrar tópicos que surgiram em épocas semelhantes, para averiguar a influência de um bloco econômico sobre outro.

- **Impacto de eventos Geopolíticos**
    - Esses dados de tópicos mapeados serão utilizados em conjunto com uma linha do tempo de eventos políticos, associados direta ou indiretamente ao bloco econômico em si, para observar mudanças na distribuição dos tópicos.

- **Análise de Semelhança de Tópicos**
    - Serão comparados os vetores de tópicos dos países dentro de cada bloco ou entre países que tenham influência e os influenciados para verificarmos se há semelhanças ou divergências.

- **TWE para Evolução Econômica**
    - Será explorado como o significado de palavras-chave relacionadas à economia mudaram ao longo das décadas para os blocos econômicos, buscando evidenciar as diferenças ou semelhanças.

- **Análise Sentimentos**
    - A partir dos dados limpos, será aplicado o modelo baseado em transformers para cada discurso dos blocos estudados para todos os anos. Após, será feito um agrupamento por período e por bloco através das médias e será usado o box plot para facilitar a visualização. Dessa forma, será possível encontrar outliers e a quantidade de dados próximos através do achatamento do boxplot. Observando padrões de convergência ou divergência entre países aliados.

Cronograma

1. Identificar palavras-chave para cada tópico obtido e categorizá-las.
2. Encontrar e analisar a similaridade dos tópicos encontrados por blocos econômicos e globalmente.
3. Identificar períodos marcados por crises, tratados ou mudanças econômicas significativas (e.g., Fim da URSS, 11 de setembro, etc.).
4. Aplicar análise de sentimentos por discurso e calcular métricas agregadas (positividade, negatividade, neutralidade) para cada bloco econômico.
5. Relacionar os outliers com eventos geopolíticos ou crises econômicas específicas.
6. Comparar dispersões entre blocos econômicos e interpretar resultados.
7. Contextualizar as variações em tópicos, sentimentos com esses eventos.
8. Relacionar resultados quantitativos com interpretações qualitativas baseadas no contexto histórico.
9. Consolidação e revisão dos resultados obtidos e análise final dos gráficos e modelos trabalhados.
10. Preparação da apresentação e relatório final.

Referências

1. **HARVARD DATAVERSE.** UN General Debate Corpus (UNGDC), 1946-2023. Atualizado em 08/2024. Disponível em: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y. Acesso em: 24 out. 2024.

2. OZAKO, Willian. *LDA Over Time*, 2024. Disponível em: https://github.com/lda-over-time/lda-over-time. Acesso em: 24 nov. 2024.