# Principais dúvidas e questionamentos:

* Quais bibliotecas utilizar? Utilizar modelos pré treinados?
* Fazer um modelo capaz de verificar redações apenas do Enem ou generalizar para outros vestibulares/provas?
* Um modelo pode funcionar bem para diferentes temas e gêneros de redação ou precisa ser ajustado por tipo?
* Modelo será capaz de se adaptar ou ideal fazer um modelo para cada tipo de texto e prova?
* Tokenização, utilizar lematização ou stemming?
* Como lidar com erros gramaticais específicos do português, modelos em outras línguas ajudariam?
* Quais tipos de métricas de avaliação de desempenho são mais importantes para modelos de correção de redação?
* Como dar feedback para o aluno sem simplesmente apontar erros? Algum algoritmo pode sugerir melhorias no texto?

# Plano de solução

Plano de solução utilizando Processamento de Língua Natural (PLN) para a correção de provas de redação do Exame Nacional do Ensino Médio (ENEM). O processo proposto será feito pensando em atender 3 critérios: detecção de desvios nos textos, atribuição de notas, global ou por critérios de avaliação e elaboração de um feedback para o aluno.

As etapas para o desenvolvimento do projeto são descritas a seguir:

**Definição do Escopo e Objetivos**

Identificar as três etapas principais: detecção de desvios, atribuição de nota e feedback ao aluno, seguindo o modelo completo de Correção Automática de Redação (CAR).

**Coleta e Preparação dos Dados**

Reunir redações do ENEM já corrigidas e anotadas. Limpar e padronizar o texto, removendo ruídos e formatando para uma análise eficiente.

Entender o conjunto de redações, identificando a estrutura dos textos e critérios de avaliação do ENEM, incluindo a análise descritiva e estatísticas iniciais de cada competência para melhor entendimento do comportamento dos dados.

**Pré-processamento dos Textos**

Simplificar o texto utilizando técnicas como tokenização, lematização e remoção stopwords, visando uma estrutura adequada para extração de features e treinamento do modelo. Extrair métricas básicas, como contagem de palavras e frases, e identificar desvios ortográficos e gramaticais usando ferramentas como CoGroo e LanguageTool. Ferramentas como SpaCy e NLTK poderiam ser utilizadas para realizar o pré-processamento, tokenização, lematização, e extração de padrões no texto.

**Detecção de Desvios**

Implementar uma abordagem simbólica baseada em regras para identificar desvios ortográficos, gramaticais e de uso de vocabulário. Extrair features como contagem de palavras, frases, média de conectivos, entre outras. Ferramentas como LanguageTool e BERT poderiam ser utilizadas para detecção de erros ortográficos e gramaticais e verificações contextuais.

**Extração de Features e Engenharia de Atributos**

Criar features que representem as competências específicas do ENEM: Língua Portuguesa, Tema e Gênero, Coerência, Coesão e Proposta de Intervenção. Adicionar contagem de conectivos e variedade no uso de vocabulário para avaliar a construção textual.

**Treinamento e Avaliação do Modelo de Atribuição de Nota**

Utilizar modelos de redes neurais para prever as notas de cada competência. Implementar um modelo separado ou multitarefa para atribuição de nota com avaliação de acurácia. Bibliotecas como Scikit-learn e Pytorch poderiam ser utilizadas para modelagem de algoritmos de regressão e para o desenvolvimento de redes neurais para análise semântica e atribuição de notas.

**Feedback Construtivo Automatizado**

Desenvolver um sistema que devolva feedbacks específicos baseados nos desvios identificados, incluindo sugestões de melhoria na estrutura e gramática.

**Validação e Testes**

Realizar validação cruzada e testes em conjunto com corretores humanos para refinar o sistema. Coletar feedback dos usuários para melhorias.