Projeto 1: Corretor ortográfico

Este projeto será feito em duplas. Você deve submeter o link do github da dupla pelo Blackboard. Apenas um da dupla precisa mandar o link. Assegure-se que no readme.md do seu repositório você coloca o nome de ambos os membros da dupla!

Data de entrega: 29/03/2021

Introdução

Neste projeto vamos construir um corretor ortográfico. Parece simples, mas vamos pensar: como detectar se uma palavra está errada?

Uma opção é comparar com todas as palavras do vocabulário. Mas isso é complicado:

- Primeiro, temos que ter um vocabulário. Mesmo que a gente compre a lista de palavras de um dicionário oficial, como ficam os neologismos?
- Isso somente nos permite identificar as palavras corretas (quando encontra a palavra em questão no vocabulário) ou incorretas (quando não localiza a palavra). Como decidir qual a palavra correspondente a uma palavra errada?
- Para cada palavra do texto temos que procurar no vocabulário todo, isso não é viável!

E agora como proceder? Vamos tentar resolver esse problema!

Construção do vocabulário

Para construir um vocabulário, precisamos de um *corpus* da língua desejada para nosso corretor ortográfico. Neste projeto, vamos usar o dataset de documentos da Wikipedia em português, fornecido pelo professor.

Tarefa 1

Aprimore o limpador de documentos desenvolvido em sala de aula.

Tarefa 2

Selecione as 3000 palavras mais comuns da lingua portuguesa. Nosso corretor vai funcionar apenas para essas palavras.

Construção do corretor

Nosso corretor ortográfico vai se basear na ideia em https://norvig.com/spell-correct.html. Peter Norvig é um famoso pesquisador em inteligência artificial, autor de um dos livros-texto mais utilizados da área, e atualmente é diretor de pesquisa no Google.

Sendo um cara bastante “fora-da-curva”, ele resolveu se divertir escrevendo um corretor ortográfico “simples” no percurso de um vôo trans-continental ao invés de assistir filme de super-herói naquela telinha microscópica, como as pessoas de bem fazem. (Invejoso eu? ‘Magina, quiéisso!)

A ideia dele foi a seguinte:

- Para uma palavra dada (possivelmente escrita errado), gerar todas as variantes da palavra que estejam a uma pequena distância de edição desta (distância 1 ou 2)
- Consultar o dicionário de palavras corretas e verificar quais palavras da lista de variantes está no dicionário
- Se alguma variante faz parte do dicionário, eis aí a palavra corrigida!

Tarefa 3

Leia a página do Norvig para entender como o corretor funciona.

Tarefa 4 (entregável)

Escreva um programa que le uma *string* e imprime sua versão corrigida!

Rubrica de avaliação

Conceito	Definição
I	Não entregou ou entregou groselha
D	Entregou o programa que não é groselha, mas está gravemente bugado
C	Entregou o programa funcional, sem nenhum melhoramento acima do que foi feito em sala de aula.
B	Entregou o programa, apresenta melhoramentos (na limpeza dos documentos e criação de vocabulário) XOR (no corretor ortográfico).
A	Entregou o programa, apresenta melhoramentos (na limpeza dos documentos e criação de vocabulário) AND (no corretor ortográfico).