Proposta de TCC de Ciência da Computação

Tema
Automatização de coleta de informações da internet (técnicas de *web scraping*).

Título
“Construção de corpus a partir de textos de notícias em Português sobre a costa litorânea brasileira para aplicação em bases de conhecimento”

Problema
A coleta de conhecimento, a representação dele e o raciocínio envolvido em seu processamento estão entre os desafios fundamentais da inteligência artificial. Repositórios de conhecimento em grande escala sobre entidades, suas relações e abstrações são conhecidos como bases de conhecimento. A criação de uma base de conhecimento precisa, atualizada e completa é um desafio significativo, apesar dos esforços substanciais empregados em sua construção automatizada.

A palavra de origem latina *corpus*, segundo o dicionário Priberam, refere-se à coletânea acerca de um mesmo assunto ou, ainda, um conjunto de documentos que servem de base para a descrição ou o estudo de um fenômeno. Assim, um corpus de alta qualidade, limpo e abrangente (definições a seguir) sobre determinado assunto é fundamental para alimentar os códigos que visam à construção de bases de conhecimento.

Entretanto, para construir um corpus confiável, seja em termos de conteúdo ou formatação, muitos desafios devem ser contornados. A pesquisa específica de domínio (em inglês, *domain-specific search* (DSS)) é o primeiro passo a ser dado quando se decide construir um corpus, e ela, por si só, envolve obstáculos antigos, ainda a serem vencidos pelos cientistas de dados. Alguns deles, por exemplo, são a escalabilidade, custo e velocidade (análise/extração de conteúdo, alta memória aplicada); Erros e redirecionamentos ao visitar páginas; Conteúdo não indexado (deep web), páginas dinâmicas e rolagem infinita; Medidas contra-rastreamento *(login*, *captchas*, *traps*, falsos erros, banimento) e, por fim, atualização e desduplicação (identificar e rastrear novamente por novos conteúdos). No presente trabalho, serão endereçados três deles: análise, extração e atualização.

Citações bibliográficas:
"corpus", in Dicionário Priberam da Língua Portuguesa, 2008-2021. Disponível em: https://dicionario.priberam.org/corpus. Acesso em: 07-06-2021.

Automated Knowledge Base Construction Program, 2019. Disponível em: http://www.akbc.ws/2019/. Acesso em: 02/06/2021.

Nguyen, D.B., Abujabal, A., Tran, N.K., Theobald, M., Weikum, G. Query-Driven On-The-Fly Knowledge Base Construction, p.1, set. 2017. Disponível em: https://core.ac.uk/download/pdf/147014933.pdf. Acesso em: 04/06/2021.

Kejriwal, M., Knoblock, C., Szekely P. Constructing Domain-Specific Knowledge Graphs. Disponível em: https://usc-isi-i2.github.io/slides/2018-02-aaai-tutorial-introduction.pdf. Acesso em: 07/06/2021.

Relevância

As principais empresas de tecnologia e universidades têm somado esforços científicos substanciais na construção de bases de conhecimento. A complexidade dos assuntos envolvidos requer equipes multidisciplinares para lidar com todos os desafios que surgem na execução do projeto, como, por exemplo, estatísticos, cientistas da computação de diversos expertises (teoria computacional, inteligência artificial, sistemas, infra-estrutura etc.), além de profissionais do domínio específico que está sendo explorado, como biólogos, biomédicos e cientistas sociais.

As bases de conhecimento são usadas em aplicações estratégicas de aprendizado de máquina, recuperação e extração de informação, como predição de diagnósticos médicos e agentes conversacionais que imitam humanos. Sendo assim, um avanço nesse meio significa um posicionamento muito à frente da concorrência e exploração pioneira em determinada tecnologia inovadora.

Mas, para atingir esse nível de excelência, é necessário, também, que o software fruto desse imenso trabalho multidisciplinar seja alimentado com dados confiáveis, num formato que os computadores consigam manipular e interpretar sem oferecer (mais) obstáculos para os desenvolvedores da base de conhecimento. É, então, quando entra em ação o corpus.

Citações bibliográficas:
Automated Knowledge Base Construction Program, 2019. Disponível em: http://www.akbc.ws/2019/. Acesso em: 02/06/2021.

Objetivos

- Levantar informações para melhor compreender o problema de construção de corpus de textos jornalísticos, em português.
- Criar um código em python com uma compilação de técnicas de *web scraping* que ajudem na coleta e formatação desses textos, provindos de portais de notícias, jornais eletrônicos, sites de divulgação científica e governamentais, cujo tema principal é a costa litorânea brasileira;
- Construir um corpus bem formatado e confiável (definições a serem feitas para atingir esses critérios) que ajude no desenvolvimento de novas técnicas computacionais para construção de bases de conhecimento que levem agentes conversacionais a um novo nível, aumentando sua capacidade de argumentação, explicação e consistência temporal e factual.

Hipóteses

- As técnicas de *web scraping* de alguns livros (a serem citados) atingem os objetivos listados?
- A implementação delas em Python é viável?

Metodologia Científica

- Paradigma: **quantitativo**;
- Tipo de pesquisa quanto aos fins: **descritiva** (expõe especificidade de um determinado fenômeno, estabelece correlações entre variáveis, não tem compromisso de explicar os fenômenos que descreve), **metodológica** (construir um modelo para determinado fim) e **aplicada** (resolver problemas, finalidade prática motivada pela curiosidade intelectual do autor);
- Tipo de pesquisa quanto ao meio: **laboratorial** (investigação realizada em lugar circunscrito, no caso simulação computacional);
- Universo e amostra: Notícias ou artigos jornalísticos, em Português, sobre a costa litorânea brasileira, oriundos de portais de notícias, jornais eletrônicos, sites de divulgação científica e governamentais. A amostra será **não-probabilística**, ou seja, definida por tipicidade e acessibilidade. A coleta de dados será através de **pesquisa documental**.

Estrutura dos capítulos do TCC

1) Resumo e palavras-chaves;
2) Introdução (apresenta o tema, o problema de pesquisa, sua relevância, justificativa, objetivos, hipóteses/questões de estudo, delimitação do estudo e seu desenvolvimento);
3) Capítulo teórico (revisão bibliográfica sobre DSS, técnicas de scraping, construção de corpus e exemplos de aplicação (bases de conhecimento));
4) Capítulo metodológico (acima);
5) Capítulo de análise de dados (tratamento e apresentação);
6) Conclusão (lições aprendidas, possíveis desdobramentos e aplicações em outras áreas);
7) Referências bibliográficas;
8) Apêndices (docs produzidos pelo autor);
9) Anexos (docs produzidos por terceiros);
10) Glossário.