



FERRAMENTA DE BUSCA SEMÂNTICA DE CÓDIGO CATMAT E COLETA DE PREÇOS NO COMPRASNET: pesquisa em andamento

Yhasmin Justino¹; Tiago Possato²; Manassés Ribeiro³; Rodrigo Zuffo⁴

¹ Aluna do Instituto Federal Catarinense, Campus Videira. Curso Ciência da Computação. E-mail: ifcyhas@gmail.com

² Orientador do Instituto Federal Catarinense, Campus Videira. E-mail: tiago.possato@ifc.edu.br

³ Professor coorientador do Instituto Federal Catarinense, Campus Videira. E-mail: manasses.ribeiro@ifc.edu.br

⁴ Coorientador do Instituto Federal Catarinense, Campus Videira. E-mail: rodrigo.zuffo@ifc.edu.br

A pesquisa de preços é etapa essencial do planejamento das contratações públicas. A utilização de outras compras públicas é obrigatória para a formação de preços de referência. No âmbito federal, toda compra utiliza o CATMAT (Catálogo de Materiais), e os sistemas registram apenas essa descrição genérica, não a descrição do termo de referência: o CATMAT é a ponte entre a descrição do material e as contratações realizadas. Na prática, essa ponte é frágil, pois a descrição técnica serve de base para a escolha do código; entretanto, materiais idênticos são adquiridos com códigos distintos, enquanto um mesmo código pode ser associado a materiais diferentes. A literatura documenta falta de padronização, cadastros incompletos e anomalias de preços decorrentes dessas falhas. Consultar um único código pode deixar de fora outras compras semelhantes, comprometendo a qualidade da estimativa, a economicidade e a legalidade da contratação; a busca manual por vários códigos é ineficiente. Estudos recentes no domínio governamental brasileiro mostram que a busca semântica supera a busca léxica. Este trabalho investiga a hipótese de que recuperar um conjunto de códigos CATMAT semanticamente compatíveis é mais adequado do que utilizar apenas o código mais semelhante selecionado manualmente, empregando esses candidatos como pontos de entrada para a recuperação automática de compras. Para isso, está sendo desenvolvida uma ferramenta em Python que sanitiza a descrição original, extraindo palavras-chave e removendo expressões voltadas à compreensão humana; busca códigos CATMAT candidatos; classifica-os frente à descrição sanitizada, retém os melhores; e monta uma tabela com as compras públicas dos últimos doze meses, via API de Dados Abertos. São testadas alternativas em cada dimensão: sanitização com KeyBERT, com LLM ou desativada; classificação por similaridade de cosseno com sentence-transformers (modelo BGE-M3) ou por LLM; e fonte de dados nas APIs do CNBS ou em base offline com ChromaDB, totalizando nove configurações. Resultados preliminares indicam que o classificador LLM é o componente de maior impacto individual, que o score de cosseno isoladamente é um indicador frágil de relevância, e que as fontes CNBS e ChromaDB são complementares. A coleta de preços reuniu registros na maioria dos estados brasileiros, com variações expressivas para um mesmo item, evidenciando o potencial da ferramenta para a transparência e a economicidade nas compras públicas. Futuramente, o termo de referência de cada contratação será baixado e analisado, mantendo apenas os itens



compatíveis com a descrição pesquisada, para que a tabela não reproduza as falhas de uso do catálogo.

Palavras-chaves: CATMAT. busca semântica. pesquisa de preços.