



<b>Evento</b>	Salão UFRGS 2024: SIC - XXXVI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
<b>Ano</b>	2024
<b>Local</b>	Virtual
<b>Título</b>	Extração de conteúdo linguístico em textos médicos do século XVIII: um ensaio com ferramentas computacionais
<b>Autor</b>	BIBIANA ZIM
<b>Orientador</b>	MARIA JOSE BOCORNY FINATTO

## Extração de conteúdo linguístico em textos médicos do século XVIII: um ensaio com ferramentas computacionais

Avalia-se o rendimento de diferentes sistemas de reconhecimento e categorização automáticos de palavras que lidam com arquivos de texto em português: Voyant Tools (2024), de consórcio acadêmico internacional (<https://voyant-tools.org/>) e Sobek (2024), produzido na UFRGS (<http://sobek.educacao.ws>). São testados manuais médicos impressos no século XVIII, que oferecem problemas para o processamento via ortografia e semântica diferentes da atual. Examinam-se os recursos de listagem e os tipos de categorização de palavras, lidando-se com a ortografia original da obra digitalizada *Observações Medicas Doutrinaes (1707)*, do médico português João Curvo Semedo (1635-1719), contendo 101 relatos de casos. São testados 3 relatos nos aplicativos de análise e representação de conteúdo, sendo uma vez em cada aplicativo com o relato normalizado na ortografia atual e também com a ortografia original, além de mais 2 testes apenas no Voyant Tools, aplicando duas listas de *stop-words* diferentes, gerando 6 resultados de cada relato. Ambos sistemas indicam elementos mais e menos recorrentes e geram representações de conteúdo textual com mapas e nuvens de palavras, sendo que o Sobek permite interação com um recurso de IA Generativa. O trabalho serve para que seja possível visualizar semelhanças entre a ortografia normalizada atual e a original da obra de Semedo, observando-se de que maneira cada relato se comporta em cada um dos aplicativos usados. Os resultados preliminares indicam que quanto mais específica for a lista de stop-words, melhor vai ser a aproximação da análise da ortografia original com o mesmo texto normalizado. Isso ocorreu tanto no Sobek quanto no Voyant Tools em 2 dos relatos considerados, e no relato restante observou-se que ambos aplicativos não foram capazes de ler o j (s longo) da mesma forma, separando palavras que seriam importantes para a análise e comparação final.