# USO DE MINERAÇÃO DE TEXTOS PARA ANÁLISE DE CARACTERÍSTICAS DA PRODUÇÃO CIENTÍFICA NACIONAL

**Vitor Yudi Kano, Luciano Antonio Digiampietri**

*Escola de Artes, Ciências e Humanidades, USP*

## Objetivos

Este trabalho tem como objetivo utilizar técnicas de mineração de texto para analisar a produção científica nacional. Em especial, até o momento, foi realizada a identificação automática dos idiomas dos títulos das publicações dos bolsistas produtividade, bem como foi feita uma análise de tendências considerando a frequência de palavras-chave nos títulos ano a ano e nas diferentes grandes áreas do conhecimento.

## Métodos/Procedimentos

A metodologia foi iniciada com a realização de uma revisão bibliográfica sobre o assunto do projeto e estudo sobre a estrutura dos dados dos currículos da Plataforma Lattes. A partir do estudo realizado, foram implementadas duas ferramentas na linguagem de programação JAVA. A primeira ferramenta combina a contagem de *stop words* e a presença dos radicais das palavras presentes nos dicionários disponíveis publicamente na Web para identificar o idioma dos títulos das publicações em periódicos presentes nos currículos Lattes. A segunda ferramenta utiliza a frequência relativa das palavras dos títulos para identificar tendências nas publicações considerando as diferentes grandes áreas do conhecimento.

## Resultados

Para a realização dos testes experimentos foram extraídos e processados 335.172 títulos de artigos publicados por 9.737 pesquisadores bolsistas produtividade em 2010 e cadastrados na Plataforma Lattes com apenas uma grande área de atuação. Todos os títulos tiveram seus idiomas identificados automaticamente e a distribuição dos títulos por área e por idioma está presente na Figura 1 (considerando os 3 principais idiomas da base).

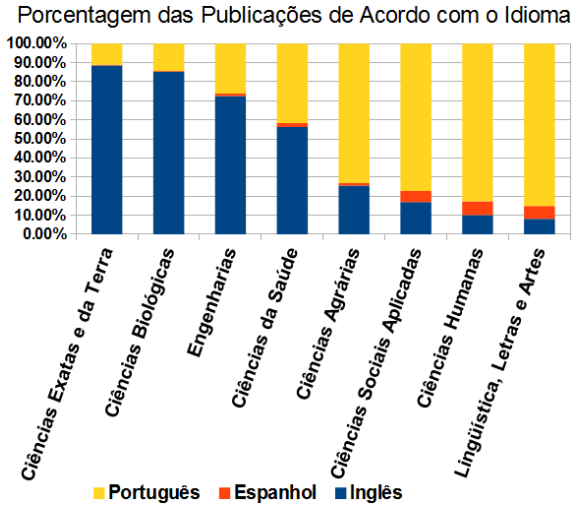


**Figura 1 - Distribuição dos Títulos por Idioma**

Neste projeto também está sendo feita a análise de tendências de palavras-chave em títulos de artigos (usando o mesmo conjunto de dados). A Figura 2 contém a frequência relativa da palavra *guerra* em cada uma das áreas.

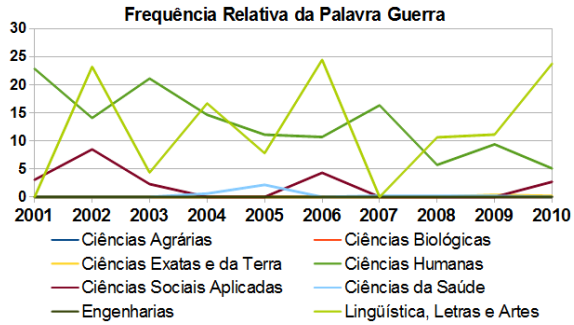


**Figura 2 - Frequência Relativa da Palavra Guerra**

## Conclusões

Este projeto está desenvolvendo ferramentas baseadas em mineração de texto para análise da produção nacional. Até o momento obtivemos resultados bastante promissores na identificação de idiomas e verificação de tendências em títulos de artigos.