Disciplina Ciência de Dados Aplicada e

Ciência de Dados para Todos

Relatório 3 – Importação e análise de dados

Autor: Nur Corezzi Data: 27/05/2018

1. Introdução e Contextualização

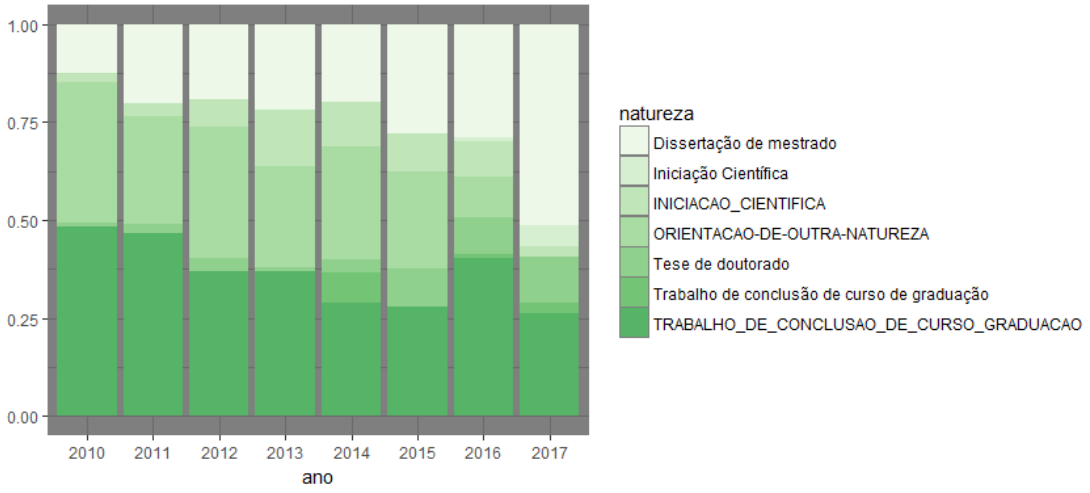
O trabalho em questão visa analisar o contexto de um dos programas de pós-graduação (PPG) realizados na UnB, para que sejam expostas de maneira organizada as informações levantadas pelos *datasets* disponibilizados pela plataforma do E-Lattes.

O programa escolhido para ser avaliado foi o programa de pós-graduação em ensino de ciências, que tem como objetivo qualificar os professores de nível básico em matérias relativas a ciências e afins. O foco é dado no desenvolvimento de conteúdos de ciências, o que envolve aspectos teóricos, metodológicos e epistemológicos relativo ao ensino de ciências além do desenvolvimento e aprendizado de novas tecnologias que permitam a evolução do ensino.

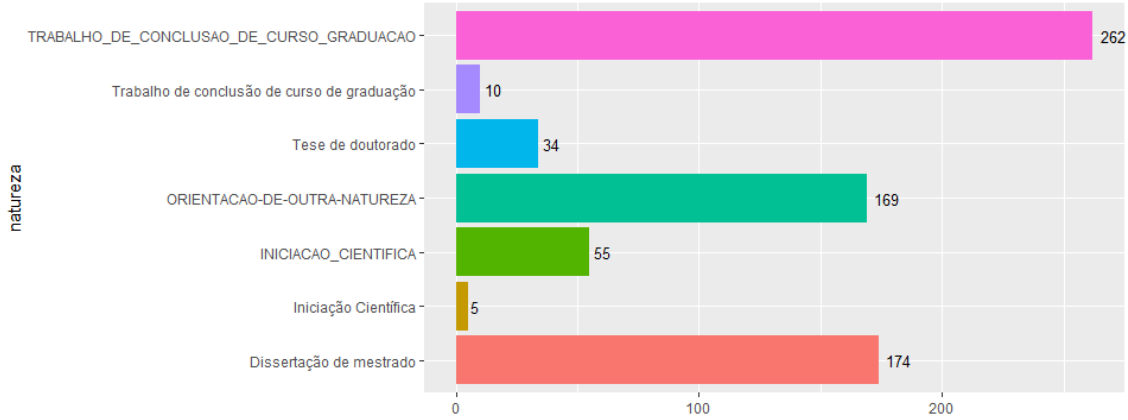
O programa engloba professores com as mais diversas áreas de atuação e de diferentes cantos da universidade. Podemos citar como áreas a física, bioquímica, ecologia, história, química, genética, medicina, psicologia entre outras, com destaques para a participação do Instituto de Física, Química e Ciências Biológicas. Os programas englobam três linhas de pesquisa sendo elas: Ensino-Aprendizagem em Ciências da Natureza em seus mutilplos aspectos; Formação de Professores de Ciências da Natureza e Educação Científica e Cidadania.

A implementação do Programa de Pós-Graduação em Ensino de Ciências na UnB, se deu inicialmente em 2007 e desde 2010, tem sido ampliado seu quadro de docentes por meio do credenciamento de novos orientadores, principalmente a partir da aproximação com o curso de Licenciatura em Ciências Naturais da Faculdade UnB Planaltina.

O programa é composto por 19 professores que já realizaram cerca de 709 orientações ao longo dos anos de 2010 a 2017 com cerca de 174 dissertações de mestrado concluidas até o momento. Podemos observar abaixo as distribuições de cada orientação realizada ao longo dos anos. É possível observar um grande aumento na quantidade(relativa) de orientações realizadas de mestrado, o que pode ser explicado pelo aumento do número de novos orientadores credenciados ao PPG.



Abaixo teremos as quantidades absolutas totais de orientações realizadas ao longo dos anos de 2010 a 2017 apenas para complementar a distribuição acima apresentada:



Em média os orientadores possuem nível de senioridade 7, tendo 5 como mínimo e 9 como máximo, o que indica que no geral os professores possuem bastante experiência no meio acadêmico, e pode ser relevante para explicar a quantidade de orientações realizadas pelo programa e também o número de publicações realizadas. No total os orientadores participam de 29 grupos de pesquisa e são em sua totalidade nascidos no brasil com uma idade média de 58 anos de idade. Entre os grupos aos quais os orientadores participam podemos citar os de Educação Científica e Cidadania, Ensino de Ciências e Mediação Pedagógica, Aprendizagem e mediação pedagógica, Ecologia e Comportamento e o grupo de Ensino de Química.

Atualmente estão sendo realizadas 74 orientações que incluem, mestrandos, doutorandos e PHDs. Deste conjunto existem alunos que vieram de diversas instituições, dentre elas Universidade Federal de Mato Grosso, Universidade Federal de Pernambuco, Universidade Federal de Sergipe e Centro Universitário Norte do Espírito Santo, porém nenhum estudante proveniente de instituições internacionais.

Tendo todo este conhecimento prévio sobre o programa a ser estudado iremos avaliar o contexto das publicações realizadas pelo departamento expondo alguns dados quantitativos e realizando algumas análises qualitativas com o objetivo de avaliar tendências e influências ao longo do tempo nas produções geradas pelo PPG.

2. Referencial

De acordo com Ricardo Barros Sampaio e Jorge Henrique Cabral Fernandes, “A ciência pode ser definida como o estudo metodicamente organizado de qualquer fenômeno que ocorre no universo, com finalidade de explicar e prever o comportamento e a estrutura de tal fenômeno”. O seguinte trabalho visa o estudo de tudo aquilo que é produzido cientificamente pela universidade (ciência da ciência) e qual a relevância destas descobertas para os próprios pesquisadores e alunos, com o intuito de validarem o método que utilizam e a qualidade daquilo que vem sendo criado.

O programa de pós-graduação sendo avaliado gerou ao longo dos anos diversos subprodutos que são consequências direta da atividade cientifica como por exemplo publicações e ou orientações acadêmicas que podem ser avaliadas para gerar um feedback qualitativo sobre o programa. Algumas questões devem ser levantadas ao realizar estes estudos, entre elas como definir uma métrica para o que está sendo avaliado? Sabemos que grande parte das bases de dados internacionais indexam apenas pesquisas realizadas em língua inglesa, portanto como é possível avaliar o impacto internacional do que é produzido no país?

Estas questões são discutidas e exploradas na ciência da ciência poís devemos ser capazes de medir e validar o impacto do que é produzido em nossas universidade e como é possível melhorar o que vem sendo produzido. Existem grupos de pesquisa como por exemplo o *Science and Evaluation Studies*(SES) que se dedicam exclusivamente ao estudo de métricas de avaliações de pesquisas cientificas.

A plataforma Lattes vem com o objetivo de auxiliar nos estudos acerca da produção realizada na universidade. Diversos conjuntos de dados acadêmicos são agregados com o intuito de gerar um material organizado e padronizado para que outros usuários possam disfrutar destas informações. Os estudos aqui realizados se baseiam nos conjuntos de dados fornecidos a respeito de publicações, orientações e perfis agregados pela própria plataforma do Lattes que se tornou bastante relevante no momento de definir nossas métricas de avaliação.

3. Metodologia

A metodologia utilizada tenta seguir o que é proposto pelo modelo CRISP-DM. Inicialmente o contexto do programa de pós-graduação foi explorado, buscando o máximo possível de informações sobre o PPG incluindo seu funcionamento e suas áreas de pesquisa. Os dados foram buscados e explorados a partir do website do programa com o intuito de compreender melhor o contexto que será trabalhado.

Uma vez compreendido do que se trata o programa os dados dispostos pelo E-Lattes foram explorados para um reconhecimento inicial. Pode-se notar que os dados vieram em formato JSON, englobando os universos de publicações, orientações, perfis e áreas de pesquisa dos orientadores do departamento, além dos dados referentes as bases do DGP e OASIS. Após importação devida os dados passaram por uma exploração inicial para entender sua formatação e se alguma modificação seria necessária para trabalhar com as ferramentas disponíveis framework da linguagem R.

Os dados constituem informações sobre os orientadores do departamento o que inclui, senioridade, produções, endereço profissional, produção bibliográfica e orientações que os mesmos realizaram. Também podemos identificar informações sobre as orientações realizadas pelo departamento que incluem informações sobre o aluno sendo orientado, como curso, instituição, nome dos orientadores, título da orientação, ano e qual estado se encontra a orientação, em andamento ou concluída. Por último temos os dados de publicações que englobam 6 tipos de publicações que serão detalhadas nos resultados obtidos.

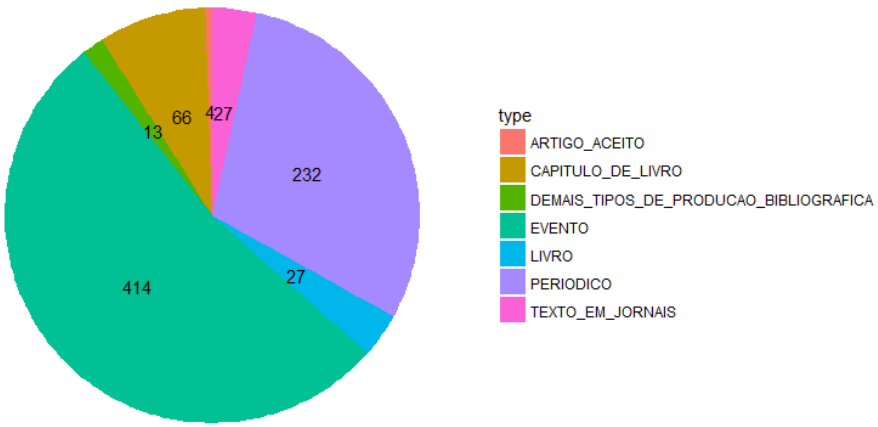
Em seguida foi identificado que alguns dados não estavam totalmente preparados para a realização de buscas, uma vez que é necessária a formatação em tabelas (e não em listas) para a realização de filtragens e *joins* de maneira clara prática. Portanto os arquivos de publicações e perfis foram modificados para que fossem removidas algumas listas internas a estruturam que dificultavam sua utilização.

Tendo os dados filtrados e organizados foram efetuadas consultas básicas com o intuito de ter uma visualização mais clara dos conjuntos de dados. As bibliotecas utilizadas para filtragem e visualização foram a *dplyr*, que permitem consultas semelhantes a consultas de bancos de dados, com dados organizadas em tabelas (em R conhecidas como *dataframes*), e a biblioteca *ggplot2* que possibilita a fácil criação de gráficos a parir de *dataframes*.

Após a realização das consultas foram geradas algumas conclusões que foram cruzadas com os dados iniciais obtidos a partir da avaliação do contexto do programa, apesar de o trabalho ter caráter expositivo foi gerado um modelo para tentar explicar algumas situações identificadas ao longo da pesquisa do PPG. O procedimento aqui demonstrado foi repetido até que fossem obtidos resultados satisfatórios assim como descreve o modelo CRISP-DM.

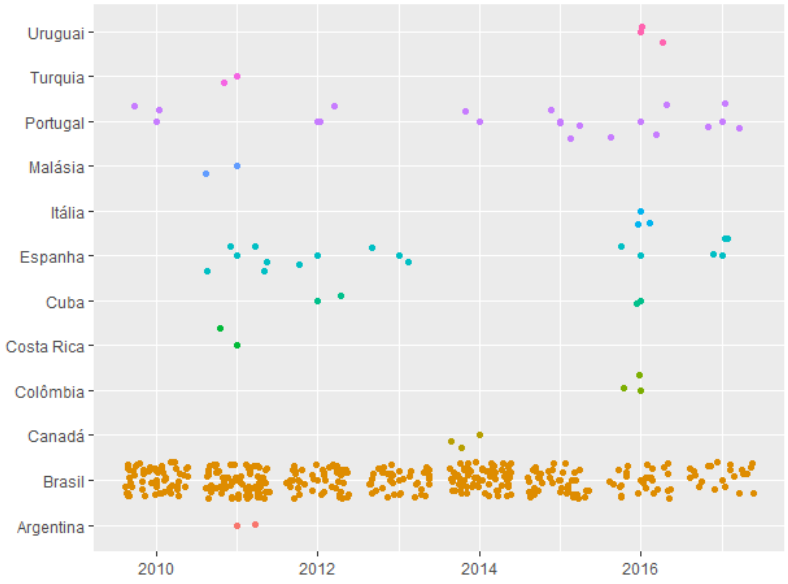
4. Resultados

Como foi dito o foco da análise será nas publicações realizadas pelo programa. Os tipos de publicações envolvem periódicos, livros, capítulos de livros, textos em jornais, eventos, artigos e outras demais publicações. Abaixo teremos as quantidades totais de registros no conjunto de dados obtidos:



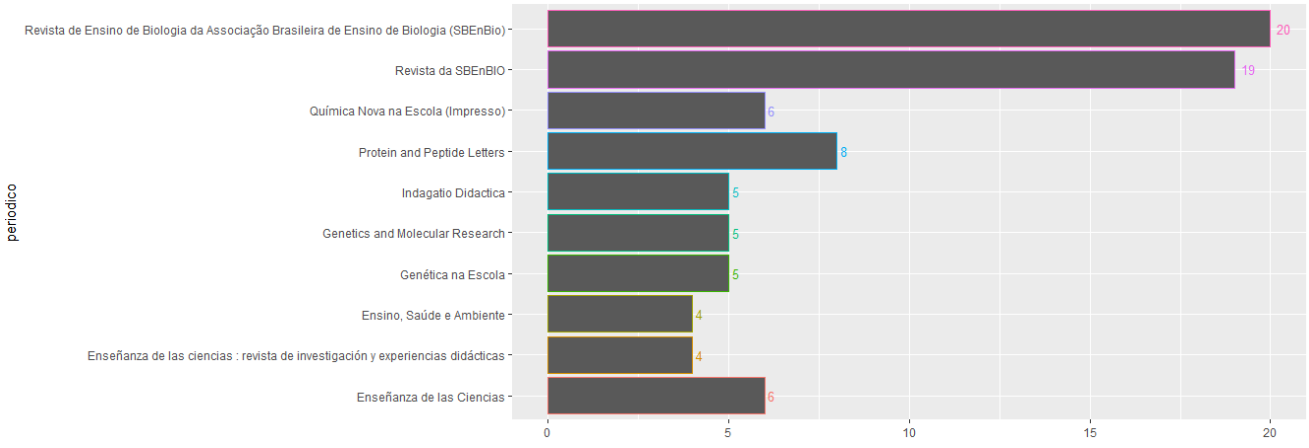
As maiores quantidades observadas são de participações em periódicos e de eventos sejam eles internacionais ou nacionais. Participações em eventos geralmente evidenciam algum tipo de engajamento dos pesquisadores com

as comunidades externas a universidade e podem trazer informações importantes sobre o quanto determinado departamento da UNB influencia ou é influenciado por outras comunidades. A seguir teremos as quantidades especificadas de participações em eventos:



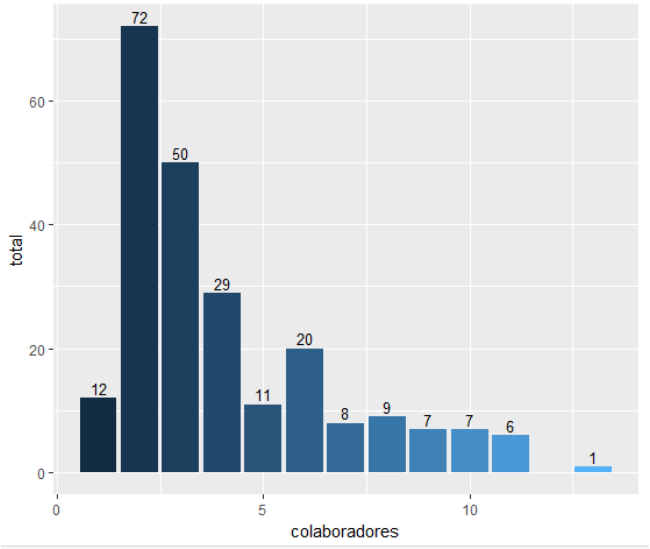
Como pode-se notar a maior quantidade de participações são de eventos nacionais, com destaque também para participações em Portugal e na Espanha, o que indica que o departamento tende a se relacionar majoritariamente com países que não possuam grandes barreiras línguisticas em relação ao português do Brasil.

Ao analisar as publicações realizadas pelo programa também podemos notar que em sua grande maioria são focadas em periódicos nacionais o que fortalece a ideia de que o departamento tende a manter sua influência apenas dentro do país. Isto poderia ser explicado pelo fato de que as áreas de pesquisa do programa são mais aplicáveis ou fazem mais sentido em um contexto brasileiro, talvez por questões culturais ou até mesmo porque o ensino sempre deve ser algo particular de uma população (o que é o foco do PPG). Abaixo podemos observar os dez periódicos em que ocorreram mais publicações pelo programa:



Apenas para medir a relevância dos periódicos publicados foram verificados os *qualis* a partir da plataforma sucupira e foram identificados para os periódicos acima citados as classificações B5, B2, B4, B1 e C que definem

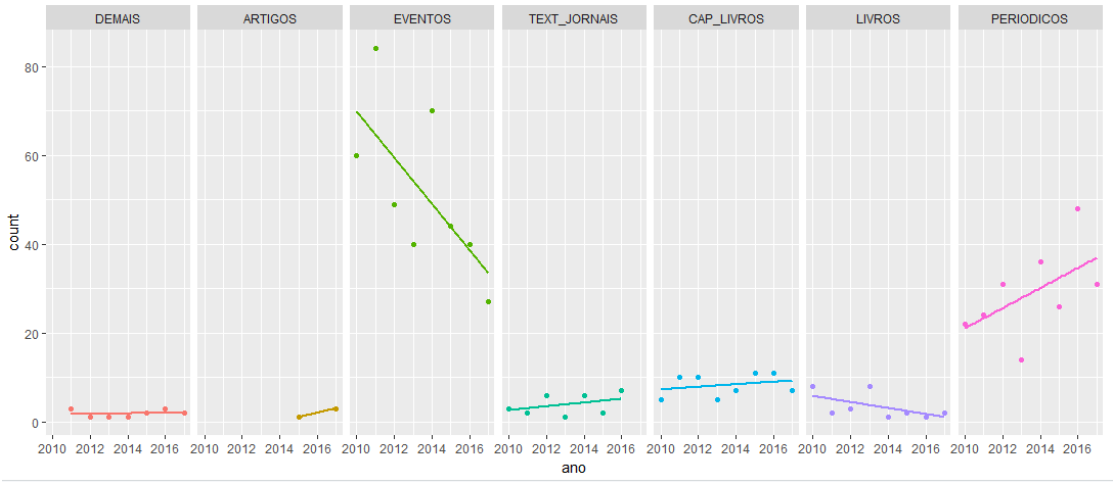
um nível de relevância médio, Outro aspecto importante de se avaliar são as parcerias que estes orientadores realizam, podemos avaliar por exemplo a quantidade de autores associada a uma mesma publicação de um periódico e observar se obtemos mais publicações solo ou publicações conjuntas e determinar o grau de colaboração entre os profissionais.



O gráfico acima mostra que em sua grande maioria as publicações em periódicos envolvem 2 ou 3 pesquisadores e a maior parte da distribuição se concentra em colaborações de 2 a 6 pesquisadores. Isto mostra que o programa possui um forte grau de colaboração entre seus pesquisadores e juntamente com a quantidade de grupos de pesquisa podemos dizer que o departamento possui um bom engajamento com relação a produção acadêmica sendo realizada na universidade.

Com relações aos outros tipos de publicações temos um volume de dados menor a serem considerados mas vale a pena também analisa-los de alguma forma. Com relação a publicação de livros, temos no total 27 publicações todas realizadas para o Brasil, o que evidencia a tendência observada anteriormente de que o foco do programa é em território nacional. No total existe uma média de 208 páginas por livro e as publicações são realizadas em colaborações em sua maioria de 2 e 8 escritores, o que também evidencia o caráter colaborativo do programa, sendo que grande parte das publicações foram realizadas em 2010 e 2013 com 7 publicações e 8 respectivamente.

Para propósitos mais gerais podemos analisar se existe alguma tendência em relação a quantidade de cada tipo de produção realizada ao longo do tempo pelo PPG. É de se esperar que as quantidades absolutas apresentem um aumento com relação ao tempo, uma vez que o departamento apresentou uma expansão de seu quadro de docentes em 2010. Também é importante que sejam analisadas as distribuições do que é produzido uma vez que isto pode revelar algum tipo de influência externa aos pesquisadores, como por exemplo um grande evento em que ocorrerão mais publicações ou algum período de crise em que as verbas de pesquisas foram cortadas. Abaixo seguem os modelos de regressão que apresentam as tendências para cada tipo de produção:



Como podemos notar, a quantidade de participações em eventos veio decrescendo nos últimos anos, talvez devido a cortes recentes em verbas acadêmicas que podem estar fortemente associadas a crise em que o país vem vivendo nos últimos anos. Já as participações em periódicos demostraram um aumento considerável o que pode também estar associado a diminuição de participações em eventos. Para os demais tipos de publicações não existe uma tendência clara uma vez que os mesmos se apresentaram relativamente constantes nos últimos 7 anos.

Portanto podemos observar que o programa de pós-graduação de ensino em ciências apresenta um baixo nível de internacionalização uma vez que grande parte de suas produções e orientações são voltadas a alunos de instituições nacionais e a periódicos nacionais em sua grande maioria publicados em língua portuguesa do Brasil.

É importante também notar que o programa possui um grande grau de colaboração entre seus participantes uma vez que a maior parte das publicações são realizadas em grupos, assim como as produções literárias e artigos produzidos. Também foram observadas mudanças nas produções do departamento ao longo dos últimos anos que podem ser analisadas mais a fundo, aqui foi proposta uma relação com cortes de fundos mas que devem ser avaliados traçando relações com informações sobre o cenário brasileiro.

Algumas informações sobre a mudança no quadro de funcionários também foi levantada durante a pesquisa e poderá também ser avaliada com mais profundidade para averiguar qual o real impacto disso no programa.