Discussão parcial

Conseguimos realizar a extração bem sucedida de uma amostra de documentos de patentes, do qual pré processamos e criamos um corpora que poderá ser usado não somente para este trabalho como para outros trabalhos com documentos de patentes. Fizemos um primeiro levantamento dos tópicos usando o modelo LDA e obtivemos 20 tópicos que serão avaliados e rotulados corretamente.

Utilizando termos simples para rotular os tópicos neste primeiro momento percebemos que dois tópicos puderam ser removidos por serem semelhantes. Com os termos, expandimos com novos termos que são sinônimos, hiperônimos e hipônimos dos termos associados aos tópicos fazendo com que o nosso dicionário cubra uma área maior do assunto.

Realizamos a construção de um modelo a partir do Random Forest, onde obtivemos o valor de score de 0,56, este é um valor ruim, o que era esperado era algo acima de 0,70.

Para os próximos passos para a finalização do trabalho, esperamos aumentar o tamanho da base de dados, testar um conjunto de parâmetros que auxiliem ao modelo resultar em um score mais alto e testar o modelo Naive Bayes e SVM.