TRABALHO FINAL

BI

Análise de Sentimentos

(Hotéis)

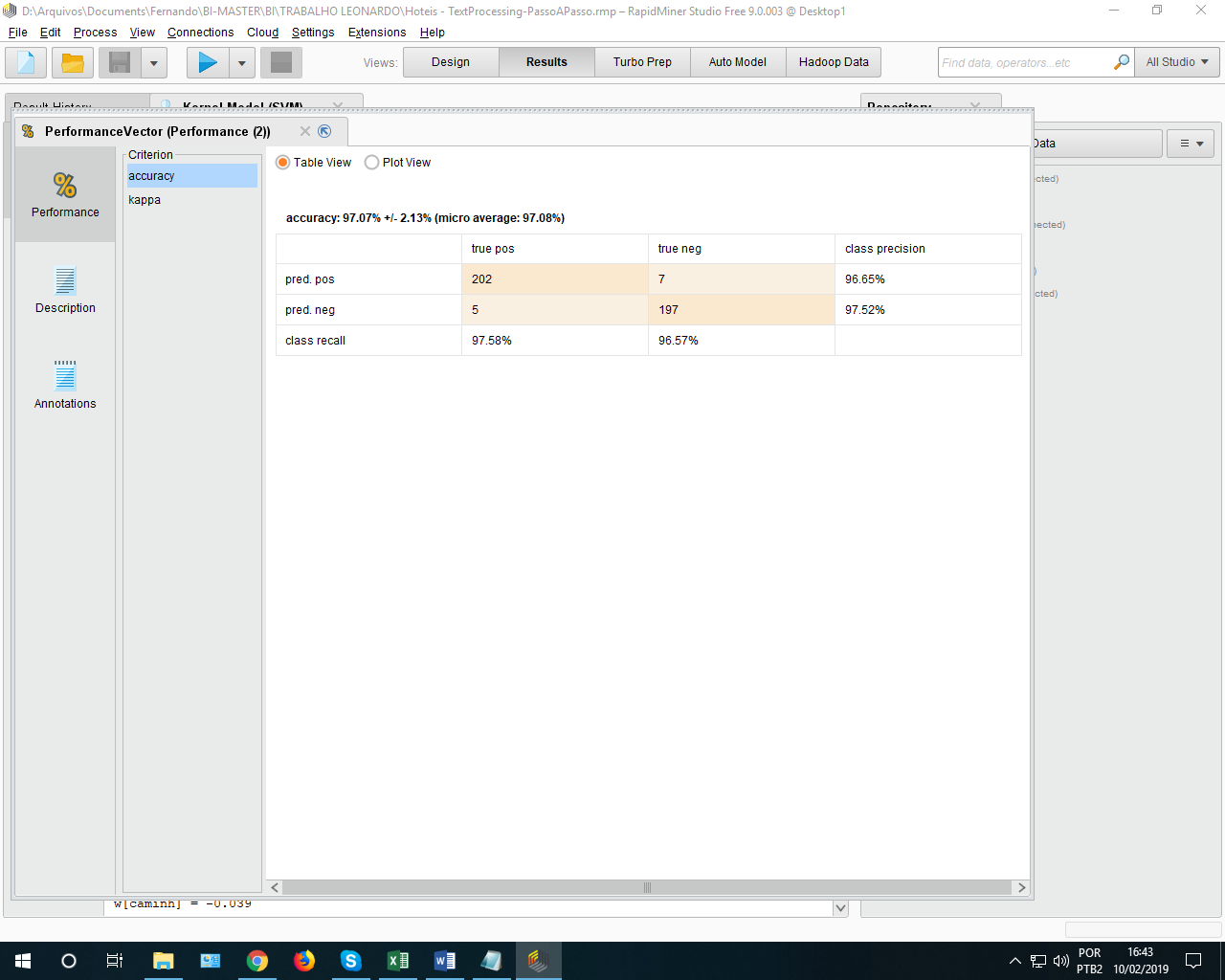
Prof. Dr. Leonardo Forero Mendoza

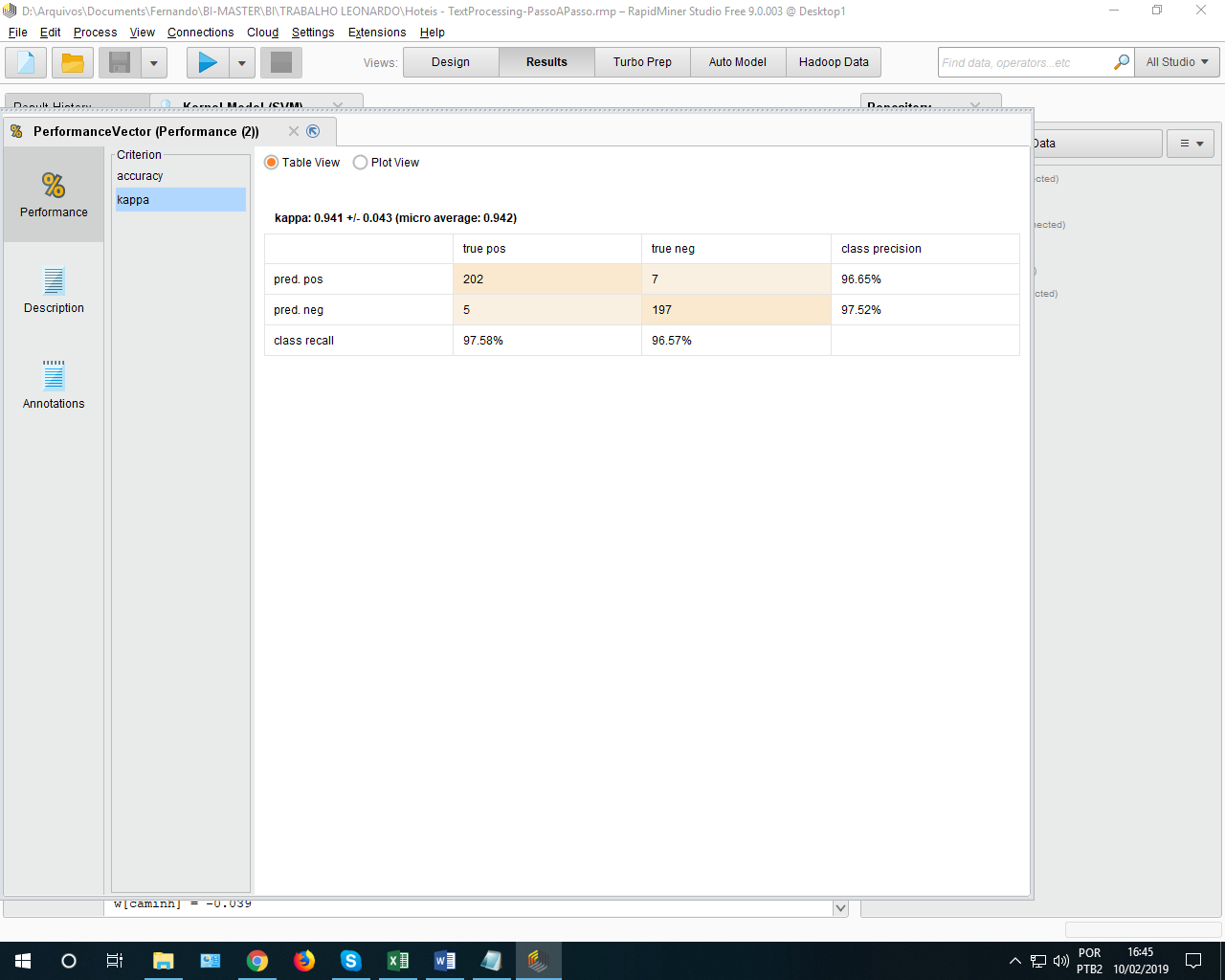
**ALUNO: FERNANDO MAURO BULEO BARBOSA**

**(buleo\_1021@esp.puc-rio.br)**

**RESUMO:**

* O trabalho foi realizado usando o **RapidMiner**
* Dentre as simulações realizadas a que proporcionou o melhor desempenho conta com a aplicação das seguintes técnicas:
* Tokenização
* Filtragem de StopWords
* Stemização (SnowBall)
* TF-IDF com aplicação de Método Prune
* Seleção de Atributos pelo Peso da Correlação deles com o Atributo Label (resultado esperado)
* Aplicação do Modelo SVM
* O melhor desempenho obtido foi **Acurácia** de **97,07%** e **Kappa** de **0,941**





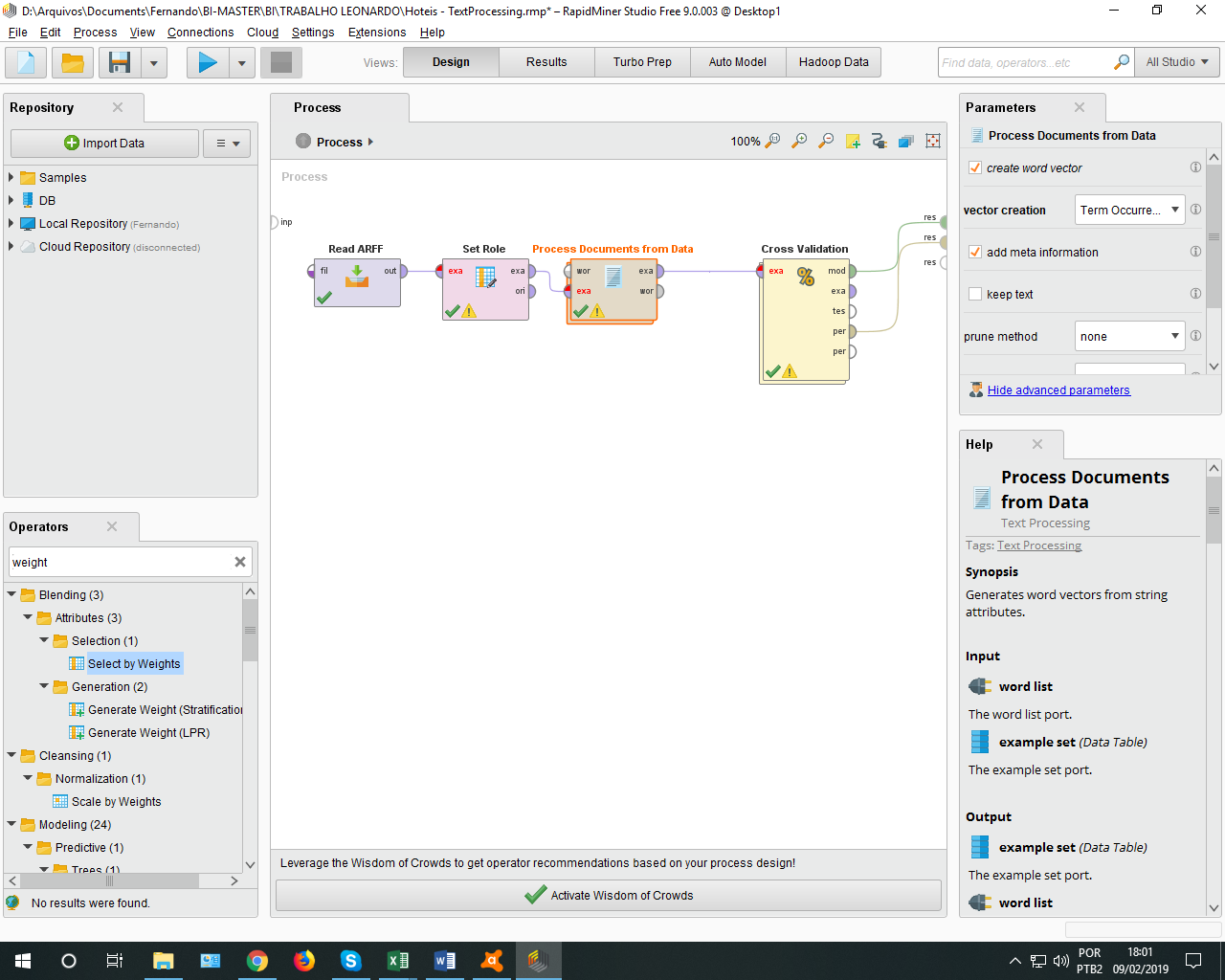
* Apresento nas próximas páginas o relatório detalhando os experimentos realizados
* Ao fim desse documento apresento os seguintes Anexos:
* Relação dos principais experimentos realizados e respectivos resultados
* Processos implementados no RapidMiner
* Dicionário de StopWords

**Trabalho FINAL – Enunciado:** *Utilize o exercício da segunda aula sobre analise de sentimento em hotéis e através das técnicas vistas melhore a acurácia obtida em aula.*

**Conforme apresentado presencialmente e no vídeo da aula, a acurácia obtida em aula foi de 80%.**

1. **Ponto de partida**

O ponto de partida foi o processo disponibilizado no curso (Aula 10 do Curso de BI – arquivo “Hoteis – TextProcessing.rmp”) para ser executado em RapidMiner, o qual executa tão somente a “Tokenização” do documento, sem aplicar as técnicas de StopWords, TF-IDF ou “Stemização”.



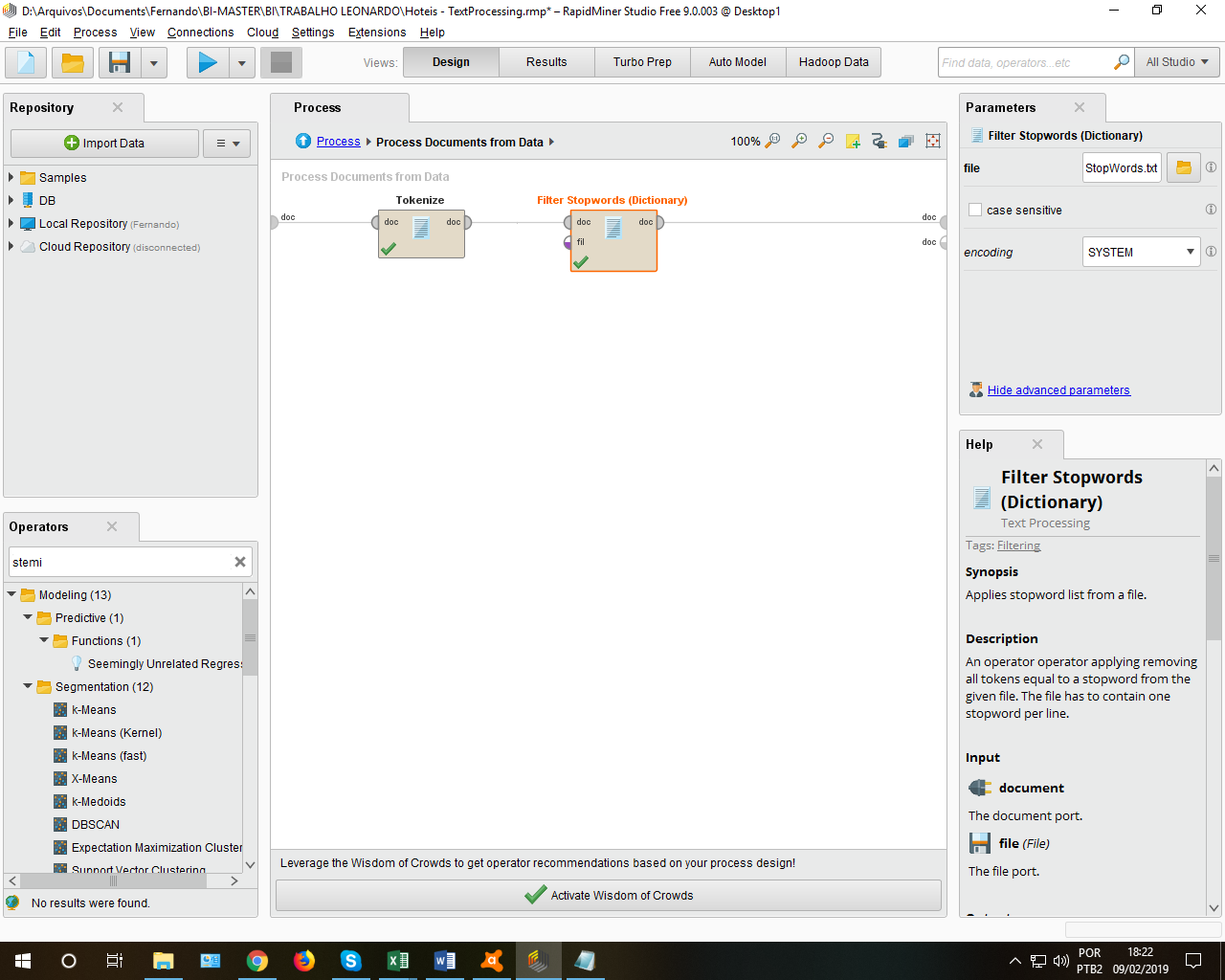
O resultado obtido com esse processo foi:

* Acurácia: 74,71%
* Kappa: 0,490

1. **STOP WORDS**

Analisei o texto fonte e os tokens gerados pelo processo e gerei um dicionário de StopWords. (O dicionário de StopWords encontra-se no Anexo 3 desse relatório).

Incluí o operador “Filter StopWords” do RapidMiner para excluir as StopWords do processo.



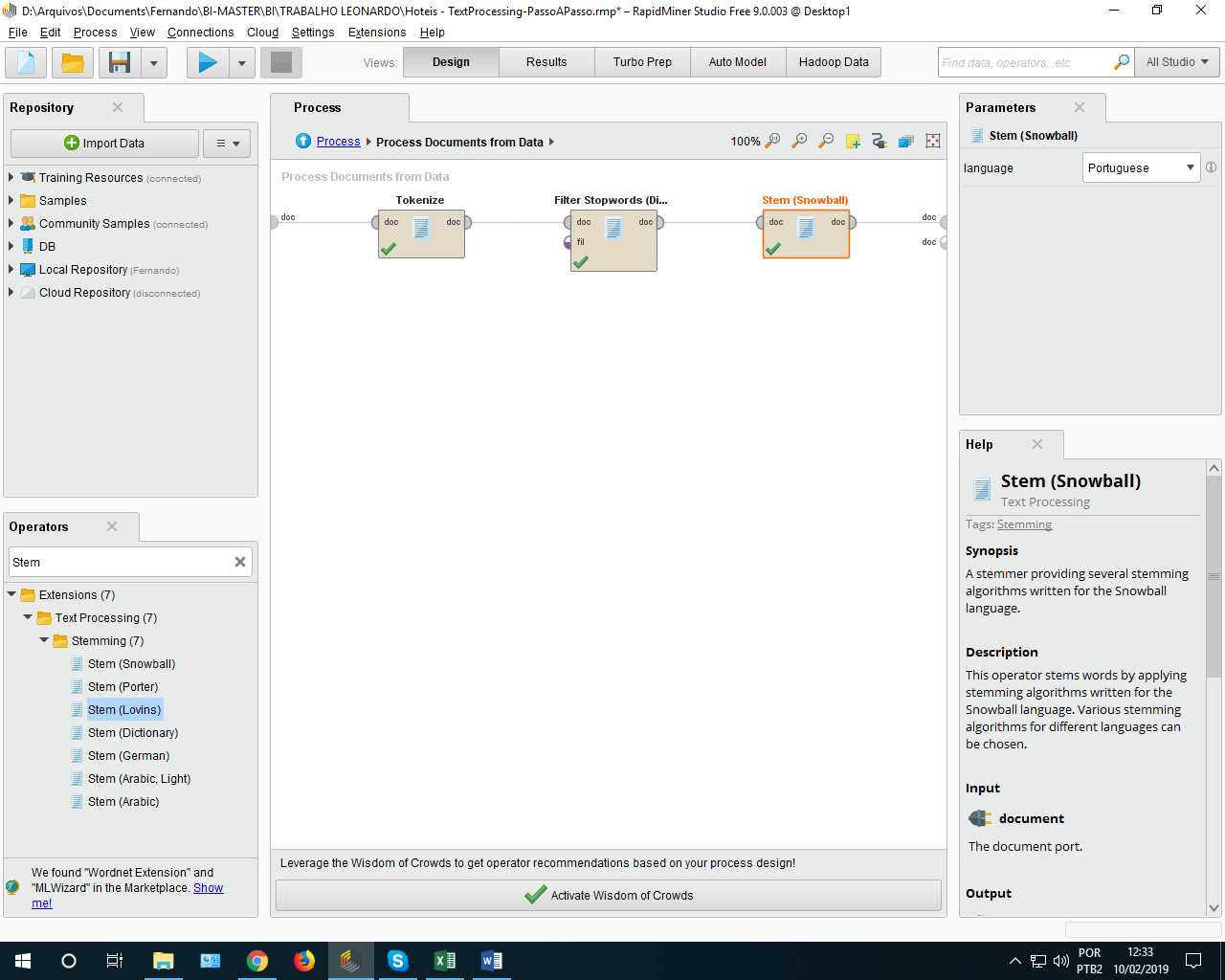
O resultado obtido com esse novo processo foi:

* Acurácia: 76,42%
* Kappa: 0,524

1. **“STEMIZAÇÃO”**

Dentre os operadores de Stemização disponíveis no RapidMiner, somente o SnowBall funciona com Português.

Incluí o operador “Stem (SnowBall)” do RapidMiner parametrizado para o português.

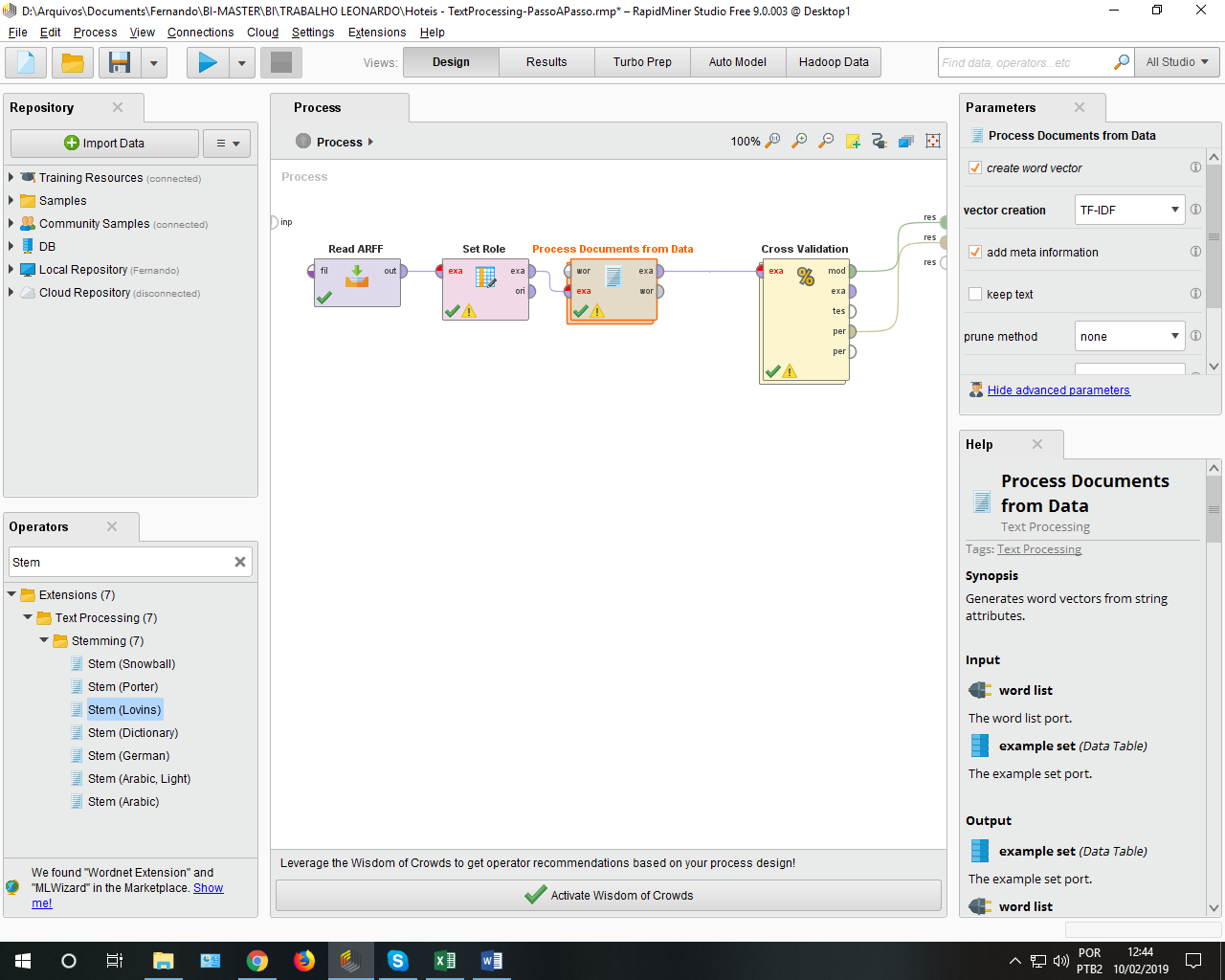


O resultado obtido com esse novo processo foi:

* Acurácia: 80,07%
* Kappa: 0,599

1. **TD-IDF**

O operador do RapidMiner “Process Documents from Data” pode ser parametrizado para aplicar a técnica de TD-IDF. Efetuei essa parametrização conforme abaixo:



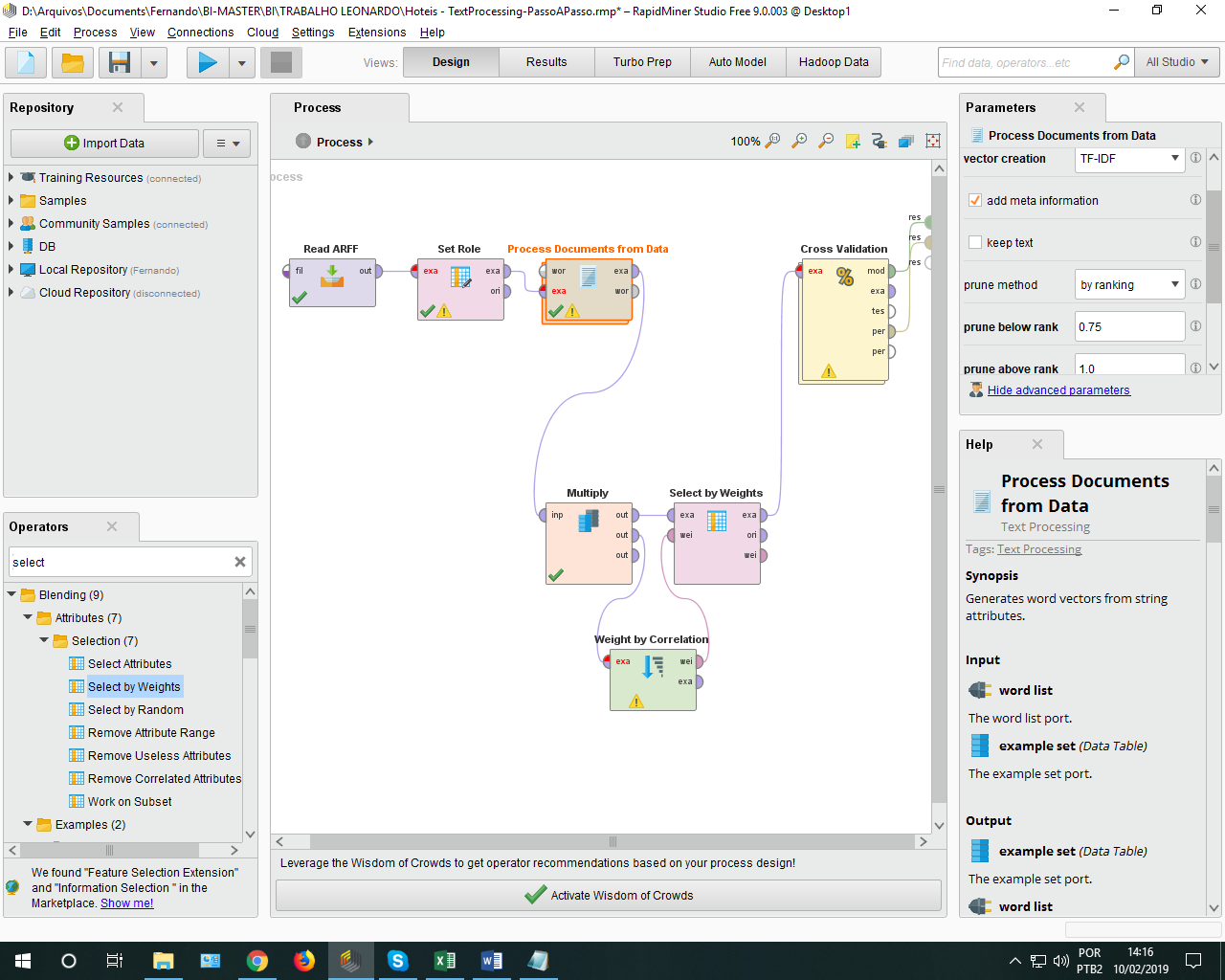
O resultado obtido com esse novo processo foi:

* Acurácia: 88,32%
* Kappa: 0,766

Busquei parametrizar o método TD-IDF usando o método “Prune”, o qual especifica se palavras frequentes ou infrequentes devem ser ignoradas de uma lista de palavras.

Aplicando o Prune by Ranking obtive o melhor resultado ao parametrizar a frequência mínima em 0,85. O resultado obtido foi:

* Acurácia: 93,67%
* Kappa: 0,873

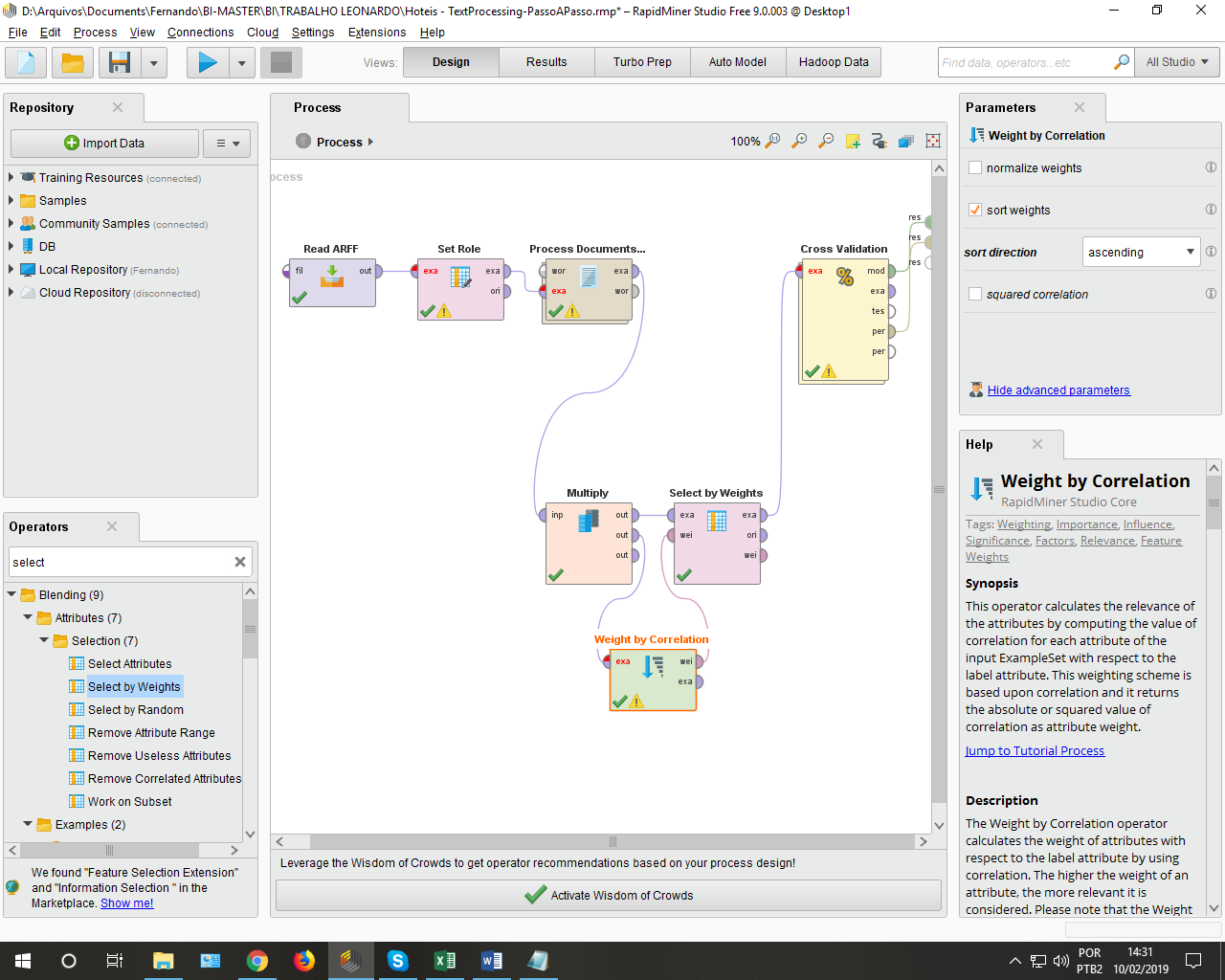


Aplicando o Prune by Percentual obtive o melhor resultado ao parametrizar o percentual mínimo em 5%. O resultado obtido foi:

* Acurácia: 91,97%
* Kappa: 0,839

1. **Select by Weight**

Avaliei que poderiam haver atributos (TOKENS) com uma correlação maior com o resultado final (campo LABEL) da análise de sentimento. Assim, partindo do modelo com Prune By Ranking superior a 0,85, incluí os Operadores “Weight by Correlation” e “Select by Weight” no processo.



Segue abaixo, tabela de atributos com maiores pesos (não representa a tabela completa de atributos e pesos, somente os maiores). Vemos que, com algumas exceções, em geral são palavras que, até pelo “bom senso”, salvo textos irônicos, devem estar associadas diretamente a contextos positivos OU negativos de forma excludente (bom, excelente, ótimo, velho, pior, confort, mal, bem, maravilh, mofo, sujo, novo, péssimo, limpo, perfeito, cheir). Observo, ainda, que as palavras (atributos) aparecem “truncados” pela aplicação do método de Stemização (Snowball).

|  |  |
| --- | --- |
| **Atributo** | **Peso** |
| bom | 0.3004789660786189 |
| excelent | 0.2937586207310673 |
| ótim | 0.273697492376431 |
| velh | 0.2722560962766664 |
| banheir | 0.26695392345979074 |
| ibis | 0.2643647092689376 |
| pior | 0.2508410032206436 |
| confort | 0.2496471334277191 |
| mal | 0.24149010942608717 |
| bem | 0.23713880677717092 |
| maravilh | 0.23545519278379387 |
| caf | 0.23223225281020068 |
| mof | 0.23192487992001673 |
| suj | 0.2311420262215225 |
| nov | 0.2277716407107158 |
| péssim | 0.22725873088228593 |
| limp | 0.2208757190995076 |
| copacaban | 0.21520176193828328 |
| restaur | 0.20889536972712192 |
| perfeit | 0.20405400678078164 |
| cheir | 0.2028776684793032 |

Utilizando peso superior a 0,057 como parâmetro de corte no operador “Select by Weight” obtive o melhor resultado nesse processo:

* Acurácia: 97,07%
* Kappa: 0,941

1. **Outros métodos preditivos**

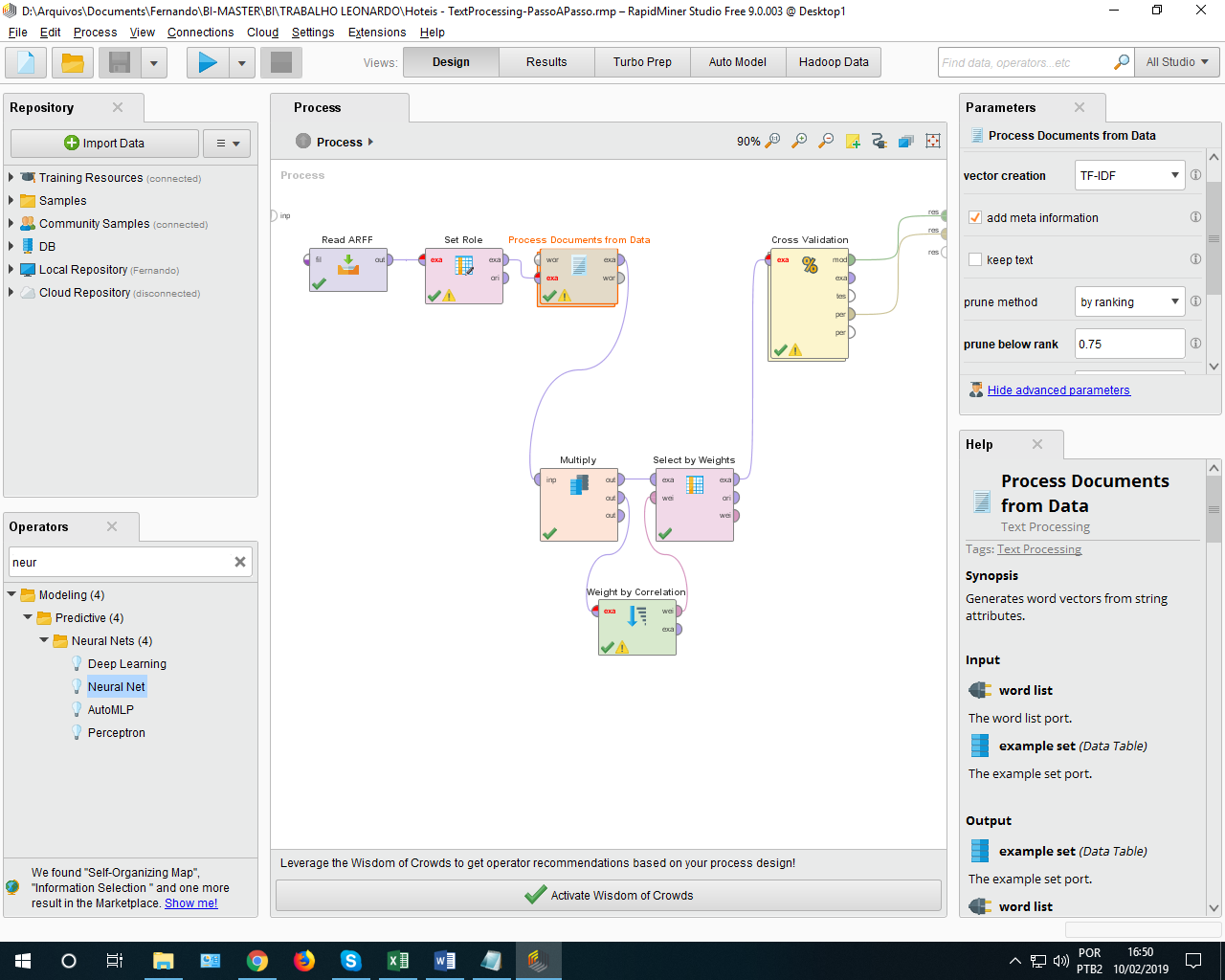
Substituí o SVM por outros métodos preditivos com parâmetros default do RapidMiner no processo: KNN, Árvore de Decisão, Random Forest, Rede Neural. O melhor resultado dentre esses foi com Random Forest (Acurácia de 91,99%), mesmo assim, abaixo do resultado com SVM. Então o modelo que proporcionou melhor resultado foi o SVM.

**ANEXO 1 - Relação dos principais experimentos realizados e respectivos resultados**

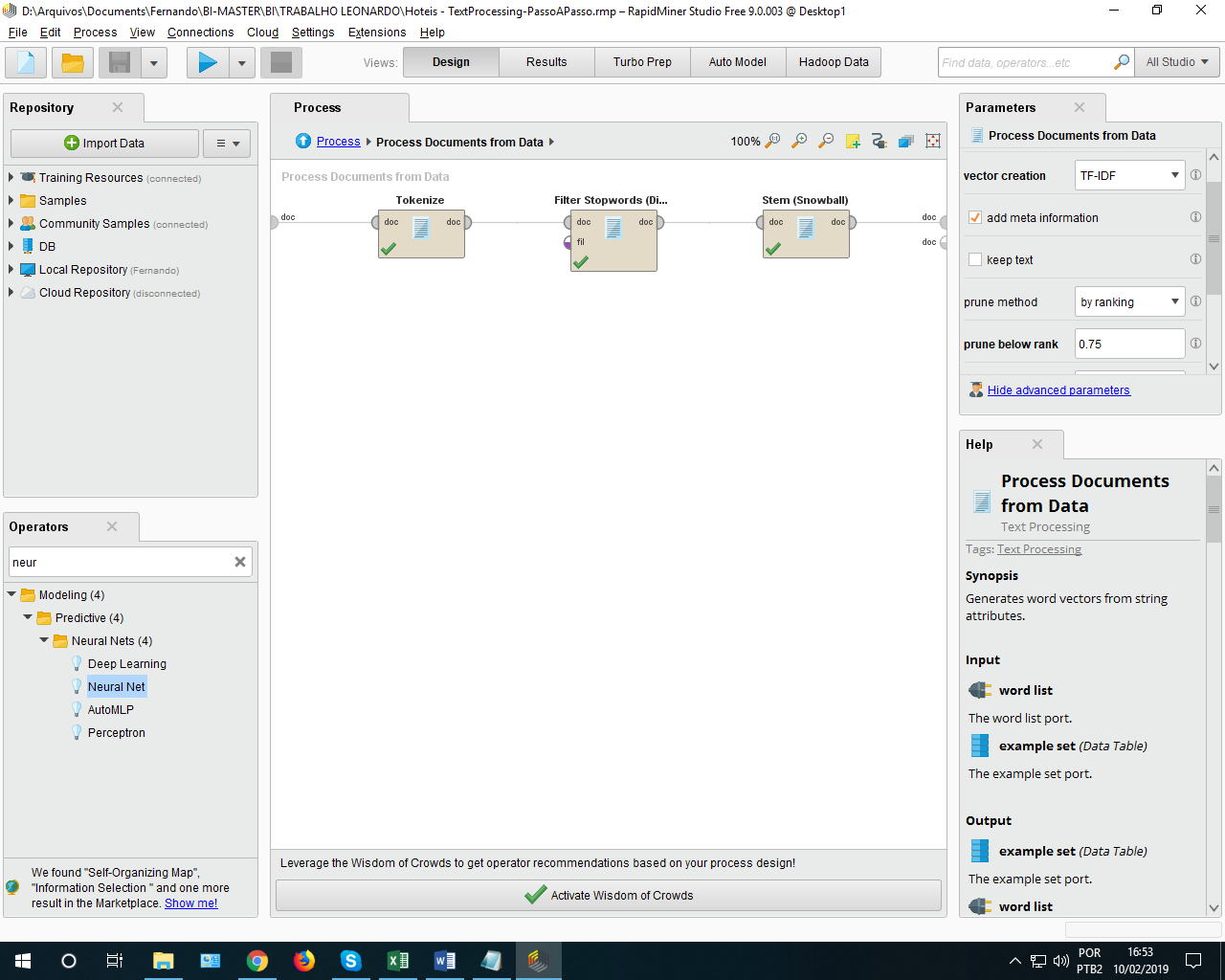


**ANEXO 2 - Processos implementados no RapidMiner**

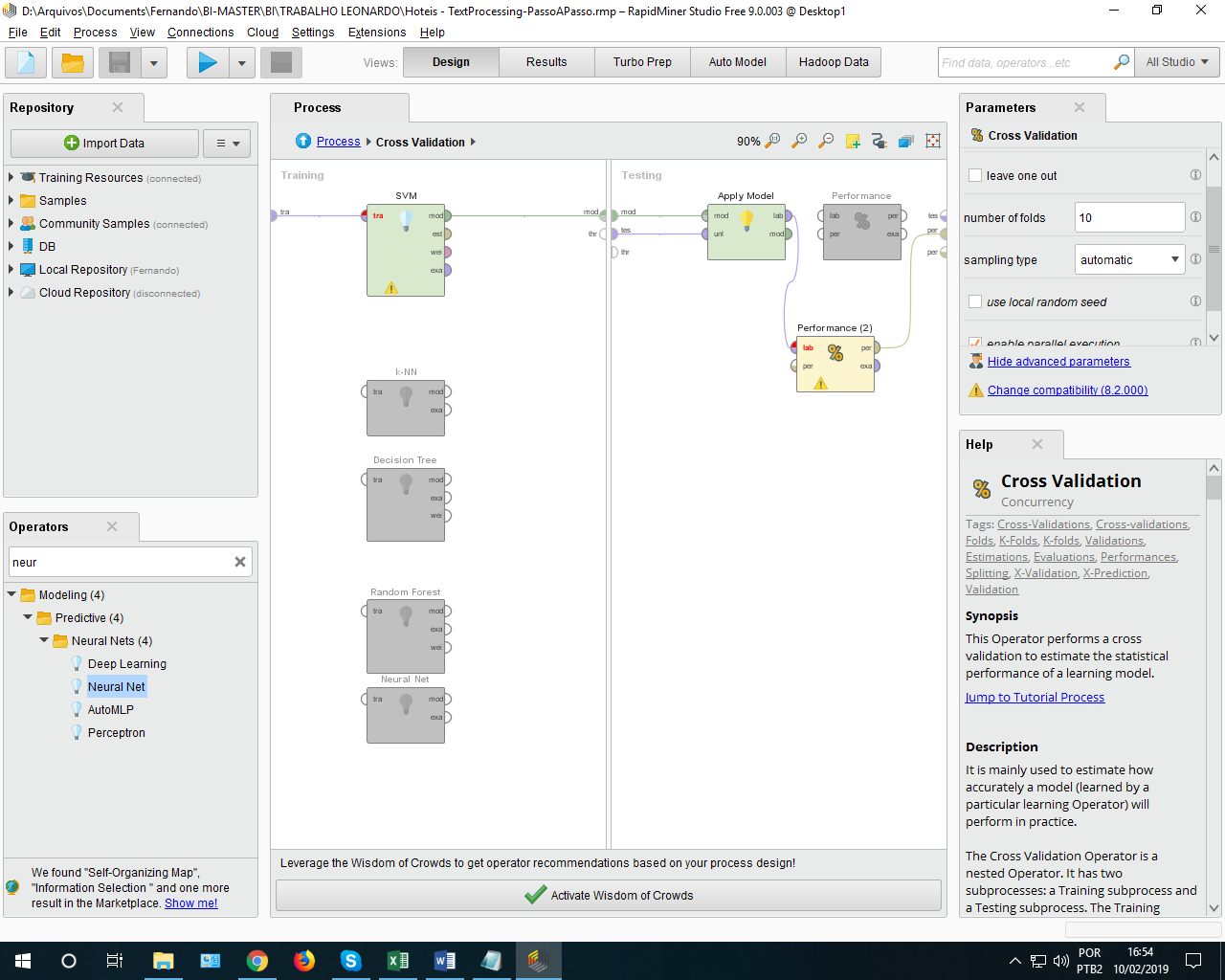
**Processo Principal:**



**Process Documents From Data:**



**Cross Validation:**



**ANEXO 3 – Dicionário de StopWords**

**a**

**A**

**à**

**abril**

**agora**

**Agora**

**Agosto**

**ainda**

**Ainda**

**algo**

**alguem**

**alguém**

**alguma**

**algumas**

**alguns**

**Alias**

**aliás**

**ano**

**Ano**

**ao**

**Ao**

**aos**

**aquela**

**aquele**

**aqueles**

**àqueles**

**aqui**

**as**

**As**

**ás**

**às**

**ate**

**até**

**Até**

**através**

**Av**

**cada**

**Cada**

**caso**

**Caso**

**casos**

**com**

**Com**

**como**

**Como**

**d**

**da**

**dá**

**das**

**de**

**desse**

**desta**

**deste**

**destes**

**disso**

**disto**

**do**

**Do**

**dos**

**Dos**

**e**

**E**

**é**

**É**

**eh**

**Eh**

**ela**

**Ela**

**ele**

**Ele**

**eles**

**Eles**

**em**

**Em**

**en**

**essa**

**esse**

**Esse**

**esses**

**esta**

**Esta**

**este**

**Este**

**etc**

**eu**

**Eu**

**isso**

**Isso**

**mas**

**Mas**

**MAS**

**me**

**meu**

**meus**

**mim**

**minha**

**Minha**

**minhas**

**Minhas**

**na**

**Na**

**nas**

**nele**

**nessa**

**nesse**

**nesta**

**neste**

**Neste**

**no**

**No**

**nos**

**Nos**

**nós**

**Nós**

**nossa**

**nossas**

**nosso**

**nossos**

**o**

**O**

**os**

**Os**

**ou**

**para**

**Para**

**pela**

**Pela**

**pelas**

**pois**

**Pois**

**por**

**Por**

**porem**

**porém**

**Porém**

**porque**

**Porque**

**portanto**

**que**

**quem**

**sao**

**são**

**se**

**Se**

**ser**

**serem**

**seria**

**seriam**

**seríamos**

**seu**

**seus**

**sua**

**tem**

**Tem**

**têm**

**tu**

**um**

**Um**

**uma**

**Uma**

**umas**

**voce**

**Voce**

**você**

**Você**

**voces**

**vocês**