Nome do Aluno: Eduardo Ferrari Magalhães, Leonardo Pereira Medeiros

Nome do Orientador: Fabio José Ayres

Título: Desenvolvimento de um módulo python para processamento de textos com *tags*

Palavras Chave: *NLP, Text-Tagger, Text Processing, Business Intelligence*

1. Descrição do Problema

Analisar textos de acordo com suas caraterísticas, ou *tags*, é um dos maiores desafios de *Natural Language Processing* (NLP). Descobrir características de um tipo especifico de texto, encontrar relações entre diferentes tipos de texto ou até mesmo criar novos textos a partir de um conjunto de características são alguns dos exemplos de tarefas que podem ser realizadas a partir da combinação de um texto e sua classificação.

Nesse contexto análises desse tipo já possuem uma serie de recursos para serem realizadas, mas, ainda não existe um módulo ou toolbox popular para fazer as analises de texto-característica que seja simples e não demande conhecimentos avançados de NLP por parte do usuário. Como essas analises podem ser muito uteis para alguns negócios, se apresenta a oportunidade para esse tipo de biblioteca.

Alguns exemplos de analises dentro de negócios poderiam ser: Para um site de vendas descobrir se as descrições dos produtos tendem a um padrão dependendo da categoria de produto a que se referem, em um site de aluguel de apartamentos, descobrir se as descrições de apartamentos se referem aos apartamentos em si ou se as palavras usadas estão correlacionadas com o tamanho do imóvel, ou se por exemplo, em alguma região/local os comentários em rede social apresentam preferências em algum estabelecimento ou produto.

2. Objetivo

Desenvolver um módulo em Python que facilite a analise de textos com *tags* construindo relatórios das diferentes peculiaridades observadas em cada categoria, e seu nível de confiança.

3. Metodologia (Proposta)

Para atingir o objetivo o módulo desenvolvido tomará como input um *dataset* de textos e *tags*. As *tags* poderão ser de 2 tipos, *tags* cujo significado é independente de outras *tags* como tipos do texto, localização onde o texto foi escrito ou categoria de produto a que se refere (Absoluto) ou *tags* numéricas ordenáveis como longitudes e latitudes, tamanho de apartamentos, preço de produtos (Numéricas).

Esses dados serão então processados pelo *core* do módulo que utilizará algoritmos de NLP para realizar tratamento dos textos (remover caracteres especiais, remover *stopwords* e *stemming*) e criação de *embeddings* (vetores que representam os textos) ou de dicionários das diferentes palavras.

Após esse processo o usuário será capaz de utilizar as funções de criação de relatório do módulo para verificar resultados de diferentes análises. A princípio as análises serão:

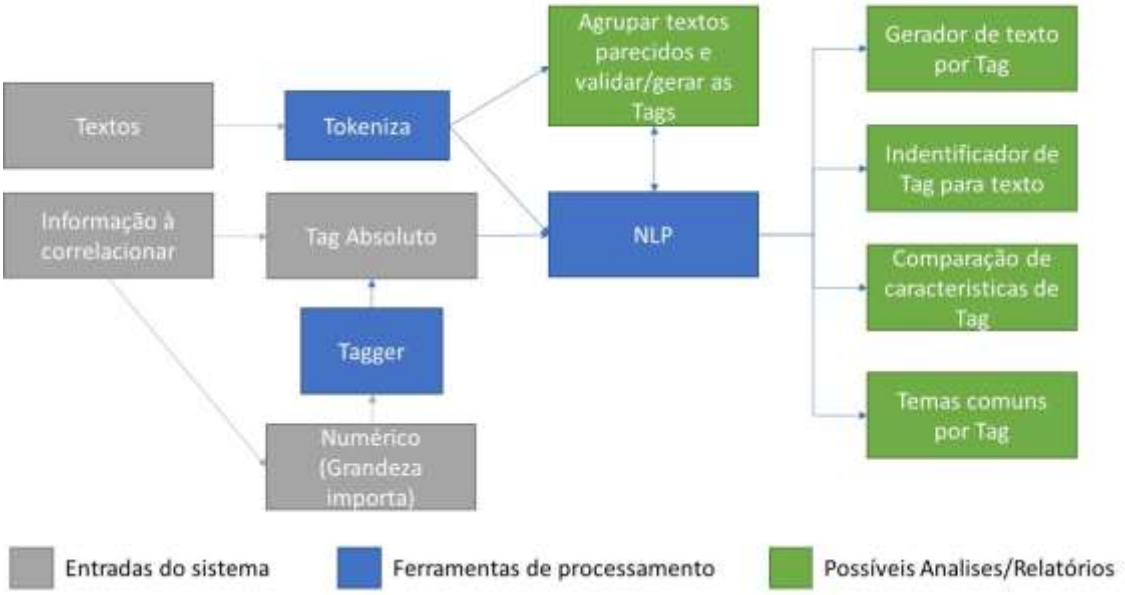
- Gerador de texto por *tag* – De acordo com uma *tag* gera um texto que possui caraterísticas semelhantes com os textos dessa *tag* do *dataset* original
- Identificador de *tag* por texto – Dado um novo texto tenta identificar a qual *tag* esse texto pertence, de acordo com as *tags* do *dataset* que estudou

- Comparação de características de *tags* – Verifica para as *tags* do *dataset* recebido se os textos de cada categoria possuem características semelhantes ou diferentes das características dos textos de outras categorias
- Temas Comuns por *tags* – Identificar principais palavras ou frases de cada conjunto de texto.

Além dos relatórios o módulo teria ainda outra função para criar *tags* para os textos do *dataset* através de algoritmos como clusterização, de forma que *datasets* que não possuem *tags* pré-existentes também pudessem ser analisados. Essa função também teria uma utilidade como utilizar algoritmos de classificação de texto para verificar se as *tags* definem textos de propriedades diferentes ou se algumas *tags* não possuem diferenças relevantes entre si.

Nesse contexto, o modulo poderia ser diagramado de acordo com a figura 1.

Figura 1: Diagrama do pipeline do módulo



4. Resultados Esperados

O resultado do projeto seria o módulo bem documentado com as funções previamente descritas com uma diversidade de algoritmos considerável (diferentes níveis de complexidade) para cada função e uma análise de exemplo para orientar pessoas que possam quere utilizar o módulo.

5. Referências Bibliográficas

Z. Cheng, J. Caverlee, K. Le – "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users" – 2010

C. Fink, C. Piatko, J. Mayfield, T. Finin, and J. Martineau. – "Geolocating blogs from their textual content." – 2009

Biblioteca de Python – NLTK – https://www.nltk.org/

Biblioteca de Python – Gensim – https://pypi.org/project/gensim/

Biblioteca de Python – skLearn – https://scikit-learn.org/stable/

Biblioteca de Python – Tensorflow – https://www.tensorflow.org/

Cronograma de atividades

Período: 2018-2

Atividades	Pré	1º semana	2º semana	3º semana	4º semana
Estudo do Problema	X	X			
Coleta de Dados	X				
Construção das funções		X	X	X	X
Analise dos dados			X	X	
Análise de Resultados			X	X	X
Criação da Apresentação				X	
Redação do Relatório					X