1.03.99 - Ciência da Computação

CLASSIFICADOR AUTOMATIZADO DE TEXTOS DE DRAFTS E RFCS: UMA NOVA FERRAMENTA DE BUSCA PARA A COMUNIDADE DA IETF E IRTF

Lucas Andrade[1]*, Marcelo Santos [2]
1. Estudante técnico de Informatica do IF Sertão-PE
2. IF Sertão-PE - coordenação de informática / Orientador

Resumo

A IETF e IRTF são duas das comunidades mais importantes no que se refere a padronização na Internet. Os padrões desenvolvidos são denominados RFCs (Request for Comments). Nesse contexto, a IETF e IRTF formam uma comunidade internacional que funciona através da organização de grupos de trabalho compostos por milhares de pessoas de várias partes do mundo. No entanto, embora os grupos de trabalho sejam organizados por áreas de interesse específicos, não raramente há discussões sobre o mesmo tópico entre diferentes grupos de trabalho que não se comunicam entre si. Além disso, há uma dificuldade de acompanhar temas específicos dentre as dezenas de grupos de trabalhos disponíveis. Assim, neste artigo propomos uma ferramenta de classificação de texto que classifica os diversos Drafts e RCFs dentro da IETF e IRTF de forma automatizada.
Intel

Palavras-chave: Processamento de Texto; IETF/IRTF; RFCs

Apoio financeiro: Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano (IF Sertão PE).

Trabalho selecionado para a JNIC pela instituição: IF Sertão PE

Introdução

A IETF (*Internet Engineering Task Force*) pode ser definida como uma entidade padronizadora de protocolos e boas práticas na Internet. Seu funcionamento ocorre baseado em listas de e-mails organizadas por temas ou grupos de trabalhos (*Working Groups*). A partir das discussões geradas através da lista de e-mails ocorrem três reuniões anuais presenciais onde são definidos protocolos amplamente usados na Internet como o TCP (*Transmission Control Protocol*) e IP (*Internet Protocol*). Além da IETF, existem grupos de pesquisa dentro da IRTF (*Internet Research Task Force*) onde são discutidas padronizações de tecnologias a médio/longo prazo. Assim, dentro dessas comunidades há dezenas de grupos de pesquisa e trabalhos sobre diversas tecnologias.

Os padrões definidos dentro da IETF/IRTF são públicos e disponibilizados na Internet. Existem basicamente dois tipos de documentos dentro dessa comunidade: Drafts e RFCs (*Request For Comments*). Drafts são rascunhos de padronizações que podem vir a ser adotados dentro de um grupo de trabalho e então virar uma RFC caso haja consenso da maioria. Por outro lado, Drafts possuem um tempo de expiração, onde durante esse período decide-se pela adoção ou não do draft em questão. O site “datatracker.ietf.org” oferece uma busca por palavra-chave de todos os drafts e RFCs dentro de diferentes grupos de pesquisa e trabalho na IETF e IRTF. Muitas vezes ao buscar-se por uma palavra-chave encontramos diversos drafts espalhados por diversos grupos de trabalho, tornando difícil de acompanhar as discussões por haver uma grande fragmentação das informações e, além disso, não raramente há a discussão de um mesmo tópico em diferentes grupos de trabalho que não se comunicam entre si. O TAO da IETF disponível em https://www.ietf.org/tao.html dá uma visão detalhada de como essa comunidade funciona.

Nesse contexto, propomos uma interface web que tem como objetivo auxiliar pesquisador e membros da IETF/IRTF a buscarem drafts e RFCs não apenas por palavras-chaves, mas por categorias ligadas a uma dada palavra-chave, realizando uma busca semântica através da classificação do texto dos Drafts e RFCs disponíveis. Para a correta decisão de como classificar os documentos disponíveis alimentamos a ferramenta desenvolvida com artigos científicos relacionado a categoria que se deseja classificar, o que chamamos de base de treinamento.

Metodologia

O primeiro passo para o desenvolvimento do classificar foi realizar um levantamento das bibliotecas e softwares disponíveis. Assim, analisamos as ferramentas Lingpipe (LingPipe, 2017), WEKA (WEKA

2017), OpenLPE (OPENLNP, 2017) e NaiveBayes. A ferramenta escolhida para desenvolvimento foi a biblioteca NaiveBayes (NaiveBayes, 2017), pois o NaiveBayes apresentou maior facilidade de implementação e bom desempenho na classificação de um grande conjunto de dados. Decidimos utilizar a linguagem Java (JAVA, 2017).

Como estudo de caso desenvolvemos um classificador para os drafts e RFCs que possuem a plavra-chave SDN (*Software Defined Network*), pois é uma tecnologia recente em redes de computadores com uma ampla discussão na indústria e academia. Constatou-se que as existências de muitas categorias dificultam a correta classificação de um determinado documento, causando muitos casos de falsos-positivo. Por isso, definimos um pequeno número de categorias baseadas em artigos relevantes da área de estudo disponíveis em sites.google.com/site/sdnreadinglist. As categorias selecionadas foram: *Control Plane; Data Panle; Emulation and Simulation; General; Monitoring and Measurement; Wireless, Routing; Security e Testing*.

Abaixo temos uma tabela que sumariza as características das ferramentas analisadas:

Tabela 1. Analise das ferramentas

	Plataforma	**Open Source**	**Facilidade na criação da base e uso**
LingPipe	Java	Não	Sim
Weka	Java, Python.	Sim	Não
Opennlp	Java	Sim	Não
NaiveBayes	Java, Python,C.	Sim	Sim

Resultados e Discussão

Primeiramente foi desenvolvido um web crawler que coleta os arquivos a serem classificados de forma automática diretamente do site da EITF. É necessário apenas a especificação da palavra-chave a ser buscada. No exemplo analisado a palavra chave foi SDN (*Software Defined Networking*). Assim, temos uma lista atualizada de um conjunto de Drafts e RFCs que não precisa ser alimentada manualmente.

Um classificador tem como objetivo associar objetos de classe desconhecida a um conjunto pré-definido de classes ou categorias. Antes da criação do classificador estudamos como o Naivebayes realizava o processo de classificação, sendo este uns dos classificadores mais utilizados na área de aprendizagem da máquina. A biblioteca Naivebayes tem seu funcionamento baseado no cálculo de probabilidade de Thomas Bayes e necessita de uma base de treinamento para classificar corretamente o texto em categorias. Nesta etapa é fundamental possuir uma boa base treinamento.

Alimentamos a base de treinamento após a conversão de artigos em pdf disponíveis no link sites.google.com/site/sdnreadinglist para uma cadeia de caracteres em Java. Identificamos que a melhor forma de entrada para a base de treinamento foi utilizar apenas o resumo de cada artigo e não o texto do artigo inteiro. Desa forma, a precisão das classificações dos Drafts e RFCs analisados foi maior.

Para verificar a precisão das classificações realizadas foi necessário classificar manualmente uma lista de Drafts e RFCs entre as categorias estabelecidas para posterior comparação com os resultados gerados. Após a execução da ferramenta desenvolvida tivemos uma precisão de 71% na classificação dos documentos analisados, demonstrando assim um bom índice de acerto.

Conclusões

Neste artigo apresentamos uma ferramenta que realiza a classificação de Drafts e RFCs através da biblioteca NaiveBayes utilizando a linguagem Java. Foi desenvolvido ainda um web crawler que coleta automaticamente um conjunto de Drafts e RFCs de forma automática do site da IETF/IRTF, necessitando apenas a definição de uma palavra-chave para realização da coleta.

Definimos um estudo de caso baseados na palavra chave “SDN”. Com as categorias definidas juntamente com a base de treinamento conseguimos obter mais de 70% de precisão na classificação dos documentos analisados.

Como trabalhos futuros, pretendemos extender os testes adicionando mais categorias e outros casos de uso. Em seguida, após uma validação por um grupo de usuários experientes em relação ao trabalho

desenvolvido pela comunidade do IETF/IRTF, temos a intenção de disponibilizar a ferramenta desenvolvida através de uma interface web para que seja usada por diversos usuários.

Referências bibliográficas

OpenNLP (2017)– Site Oficial: <https://opennlp.apache.org/>. /;acessado em março de 2017.

Java (2017)- Site oficial: <https://www.java.com/pt_BR/>. /;acessado em março de 2017.

Python (2017)- Site oficial:<https://www.python.org/>. /;acessado em março de 2017.

LingPipe (2017)– Site Oficial: <https://alias-i.com/lingpipe/>. /;acessado em março de 2017.

WEKA (2017)– Site Oficial: < http://www.cs.waikato.ac.nz/ml/weka/>. /;acessado em março de 2017.

Yahoo (2017)– Site Oficial: < https://br.search.yahoo.com/>.

NaiveBayes (2017)– Site Oficial: http://scikitlearn.org/stable/modules//naive_bayes.html. /;acessado em março de 2017.