Empoderando vidas.
Fortalecendo nações.

Projeto BRA/12/018 - Desenvolvimento de Metodologias de Articulação e Gestão de Políticas Públicas para Promoção da Democracia Participativa

Produto 03 - Ferramentas assistidas de categorização de conteúdo

Com base em Processamento de Linguagem Natural e de Redes Complexas, adaptadas para o ambiente do portal de participação

Renato Fabbri

Secretaria Geral da Presidência da República

Produto 03 - Ferramentas assistidas de categorização de conteúdo

Contrato n. 2013/000566

Objeto da contratação: Aporte de conhecimentos e tecnologias para especificação de vocabulário e ferramentas assistidas que utilizam processamento de linguagem natural e análise de redes complexas para o conteúdo do portal da participação social.

Valor do produto: R$ 10,800 (dez mil e oitocentos reais)

Data de entrega: 27 Julho de 2014

Nome do consultor: Renato Fabbri

Nome do supervisor: Gabriella Vieira Oliveira Gonçalves

Secretaria Geral da Presidência da República

Fabbri, Renato

Ferramentas assistidas de categorização de conteúdo: Com base em Processamento de Linguagem Natural e de Redes Complexas, adaptadas para o ambiente do portal de participação / 2014.

Total de folhas: 15

Supervisor: Gabriella Vieira Oliveira Gonçalves

SG/PR

Secretaria Geral da Presidência da República

Palavras-chave: reconhecimento de padrões, redes complexas, processamento de linguagem natural, participação social.

Sumário

Resumo

Este documento descreve procedimentos selecionados para categorização de conteúdo do portal federal da participação social (Participa.br). O produto relacionado no termo de referência desta consultoria preve somente propostas de especificações e códigos. Dado o aspecto prático do trabalho, estão descritas o contexto e possibilidade consideradas, asism como implementações e códigos operantes. Parte deste trabalho é acessível online via http, como os scripts no IPython Notebook e o endpoint SparQL que serve os dados do Participa.br com critérios semânticos.

Palavras-chave: reconhecimento de padrões, redes complexas, processamento de linguagem natural, participação social.

1 Introdução

1.1 Contexto e importância da consultoria

Em confluência com o portal federal de participação social (Participa.br), e o Plano Nacinal de Participação Social (PNPS), esta consultoria propõe métodos de classificação e priorização de conteúdo, e formas de auto-regulação para o portal. O presente produto apresenta uma seleção de métodos para classificação de conteúdo. Dadas a pertinência para o contexto participativo e a simplicidade, são apresentadas a classificação via 1) conectividade dos autors e 2) características dos documentos.

1.2 Contexto e importância do Produto

- Este Produto, através da classificação de conteúdos, visa 1) facilitar a assimilação das informações pelos participantes; 2) explicitar propriedades do sistema considerado; 3) permitir observação de conteúdos produzidos por nichos ou características em comum.
- São esperadas a incorporação destes métodos no funcionamento do Participa.br e pelos participantes.
- A especialização conectiva dos agentes sociais, e do texto produzido por indivíduos e grupos, é um fenômeno plenamente reconhecido. O aproveitamento destas diferenciações é uma realidade, mesmo ainda restrito a empresas e acadêmicos. A entrega prática destes conhecimentos ao poder federal e à sociedade capacita a democracia participativa.

2 Desenvolvimento

2.1 Etapas de desenvolvimento

2.1.1 Estudo ontológico e triplificação dos dados para API de acesso

Para viabilizar a classificação de conteúdos do portal participativo, em confluência com as propostas de web semântica desta consultoria e do Participa.br, foi necessário uma abordagem ontológica dos aspectos envolvidos no participa.br, assim como a triplificação dos dados. Para isso, a OPS (Ontologia de Participação Social) foi revisada, com melhoras substanciais, também a OPA (Ontologia do Participa.br) foi criada[1,2]. Já a representação dos dados do Participa.br em triplas RDF envolveu o uso destas e diversas outras ontologias[3].

2.1.2 Instanciação de um Fuseki/Jena e um IPython Notebook

Os dados triplificados podem ser usados diretamente em algum aplicativo ou programa. A forma padrão de disponibilizar dados em RDF, porém, é através de um endpoint, que prepara os dados na RAM para buscas especificadas via SparQL. Está online um endpoint Jena para consultas SparQL via HTTP. Também uma seleção dos scripts em Python estão disponíveis através de navegadores comuns, como o Firefox ou o Chrom(e,um). **Veja os anexos**.

2.1.3 Classificação dos textos (mineração de texto / processamento de linguagem natural)

As possibilidades de classificação de conteúdo com base nos textos são inúmeras. Nesta subsubseção, são apontados alguns dos caminhos considerados.

- Através de uso de textos previamente classificados, pode-se treinar classificadores automatizados. Este é o chamado "aprendizado supervisionado" de máquina. As técnicas atualmente em uso são inúmeras (redes neurais, algorítmos genéticos, etc). Para exemplificação, foi implementado uma aprendizagem Bayesiana. **Veja nos anexos**.

- A classificação de objetos não classificados previamente, com base nas propriedades dos objetos somente, é conhecido como "aprendizado não supervisionado". Pode-se impor a existencia de 2 classes (com base no balanço estrutural[easley]), ou mais classes, de forma a maximizar a dispersão inter-classe e diminuir a dispersão intra-classe. Este processo pode ser útil para observar nichos nas atividades, mesmo sem um conjunto de mensagens classificadas de antemão.

- Classificação de mensagens similares às escolhidas. Esta distância pode ser euclidiana no espaço de contagem de palavras, ou calculada via redes semânticas (e.g. wordnet).

- Classificação via contexto similar da palavra ou via simples incidência da palavra. Como buscadores usuais, com capacidades mais amplas para lidar com contexto (outras palavras, tipo de autor, classificação da mensagem).

- Ranqueamentos para mensagens, autores e palavras:

  - Mais adjetivos, mais substantivos, mais pontuações, etc.
  - Maior tamanho médio das palavras ou variedade de tamahos (desvio padrão). **Ver nos anexos**.
  - Frases mais longas em caracteres ou em palavras, variedade de tamanhos (desvio padrão).

– Uso de limiares para o rankeamento, p.ex.: os participantes que mais usam adjetivos (ou escrevem mensagens de mobilização) dentre os que possuem mais de 10 mensagens. **Ver nos anexos**.

–

2.1.4 Classificação dos agentes pela conectividade (Redes Complexas)

- Em geral, as redes formadas com rastros de atividade digital são: redes de interação ou redes de relações. No participa, há, em especial, a rede de amizades entre os usuários (relações) e redes de interação: quem responde quem, etc. **Ver nos anexos**.
- Pode-se classificar os usuários por comunidades detectadas nas redes. **Ver nos anexos**.
- Ranqueamento por centralidade é um dos recursos mais comuns. Há medidas de centralidade com base da conectividade (grau), intermediação (betweeness) proximidade (closeness) e outras medidas. **Ver nos anexos**.
- As redes sociais, por serem em geral "livres de escala", possuem especialização dos agentes, canonicamente pensado em "hubs", "intermediários" e "periféricos". Estes setores podem ser obtidos com mais propriedade comparando o histograma de conectividade da rede real com uma Erdös-Renyi com o mesmo número de vértices e arestas. **Ver nos anexos**.

2.1.5 Combinação de medidas de RC, PLN e outras

- As medidas de redes e de texto podem ser combinadas para melhorar a qualidade dos classificadores de mensagem. As estabilidades nestas medidas sugerem que hajam outliers[]]].
- Medidas de uso do sítio e do perfil do participante podem enriquecer os classificadores.

2.1.6 Aquisição de dados classificados

Para o aprendizado supervisionado (etiquetação automática, análise de semtimento, etc), é utilizado um conjunto de dados etiquetados de antemão, para "treinar" o classificador. Nas áreas de comunicação e monitoramento, são etiquetadas à mão as mensagens como positivas, negativas e neutras e em outras classes de interesse (e.g. geolocalizações, assuntos). Os autores são classificados em personas (autor masculino, feminino, ativista, militante, curioso, etc). Esta classificação manual pode servir para treinar um classificador, especialmente se revisada por uma ou mais pessoas.

2.2 Justificativa do método

- Classificações mais fundamentais: os métodos utilizados (bag-of-words, aprendizado bayesiano, medidas de grau e betweenness) são as mais usuais, além de facilitar a comparação e estabelecimento de benchmarks, possuem eficiência conhecida e significados mais facilmente compartilhados.

- Amadurecimento com equipe do Participa.br: há outros consultores e integrantes da SG/PR, e da sociedade civil, que compõem ou se comunicam com a equipe do Participa.br. Neste contexto, foram propostas e amadurecidas diversas possibilidades de classificação de conteúdos. Neste processo, foi decantado esta seleção, apresentada neste Produto.

2.3 Justificativa das fontes

- Pesquisa científica: o autor é pesquisador nas áreas relacionadas com produção bibliográfica em revistas internacionais e em circulação nacional.

- Os frameworks computacionais utilizados (nltk, networkx, rdflib, jena, etc) são amadurecidos no mundo todo, em desenvolvimento aberto, com comunidades em constante e pública discussão.

- A equipe do Participa.br é a equipe da SG/PR voltada para a participação social. Desta equipe provém boa parte dos avanços na participação social.

2.4 Confronto entre os resultados esperados e os alcançados

Este Produto preve "propostas de especificações e códigos" de classificação de conteúdo do Participa.br. Este Produto compreende estas propostas. Há, além disso, alguns resultados alcançados a mais:

- Propostas operantes em códigos online, já integrado aos dados semânticos e disponibilização via endpoint SparQL.

- Interfaces/frontends já estudadas para gráficos, reatividade e streaming (meteor+d3) [][][].

- Entrega, através dos resultados dos scripts, de uma breve análise do Participa.br em termos dos rastros digitais, dos conteúdos e dos usuários. **Ver anexos**.

Este documento e os scripts foram reunidos em um repositório git público usual[].

3 Usos dos resultados

O próximo Produto desta mesma consultoria possui foco na utilização destas classificações. Exemplos de usos estão topificados abaixo.

- Navegação dos conteúdos do portal: facilitar a aquisição das informações de interesse; permite observar o conteúdo com base em características dos participantes (e.g. hub, periférico, intermediarios) ou dos conteúdos (e.g. fração de adjetivos ou classificada com rótulos X ou Y).
- Enriquecimento do legado semântico do Participa.br e outros portais: boa parte dos cálculos, necessário para obtenção das estatísticas e classificações, requerem recursos computacionais poderosos e técnicas nada triviais. Assim, os resultados podem ser disponibilizados junto aos dados, em RDF.
- Atribuição de função: através das estatísticas dos grupos, pode-se recompensar atores ou convidá-los para atividades ou funções especiais.
- Resumos: usualmente dashboards, redes ou relações de palavras, visões gerais da entidade de interesse. A entidade pode ser um portal, uma comunidade, um usuário, uma trilha participativa. Estes resumos são bastante úteis para valorizar as instâncias e orientar os participantes.
- Coleta destas informações para usos/ações: difusão de mídia, consultas, propostas, estudos, etc.

4 Conclusão

A categorização de conteúdo do Participa.br pode ser feita de forma distribuída. Os dados, servidos por um endpoint SparQL, podem ser analisados por frontends, como um IPython Notebook, um ScrapperWiki ou um Meteor+d3 para visualizações interativas. Foi testado um intermediário em Flask para servir os dados já formatados para o frontend. Funcionou com serviços gratuitos do Heroku, Meteor e Mongo Labs, embora com limitações e alguns impasses para desenvolvimento em nuvem. Os anexos ao final deste documento[], e o repositório git[] são os resumos principais do Produto.

4.1 Comentários, sugestões, recomendações

- Para boas aplicações de classificadores de conteúdo, é necessário uma quantidade grande de conteúdo classificado previamente, geralmente à mão. Assim, é pertinente a etiqeutação das

mensagens com os parceiros da comunicação, para liberação junto aos dados semânticos e treino de classificadores.

4.2 Impacto do Produto para a elaboração, gestão e/ou avaliação de políticas públicas de participação social

- Facilita a apropriação dos processos participativos através da categorização de conteúdos e observação de suas características.
- Explicita a entrega das informações para a população, para observação distribuída.
- Entrega em tecnologias livres destes conhecimentos. Compostos por tecnologias livres e publicados em licença livre.
- Entrega de instâncias operantes de acesso aos dados do Participa.br, em formato RDF e enriquecidos.
- Semanticamente (OWL, OPS, OPA, FOAF, Dublin Core, etc).[]
- Entrega destes algoritmos em forma executável em navegadores HTTP comuns, como Firefox ou Chrome(um).[]

4.3 Como o Produto deverá impactar o público-alvo das políticas públicas a que se refere

- Permitindo navegação seletiva pelos conteúdos disponíveis.
- Valorizando as instâncias e as tornando mais informativas, com resumos estatísticos e visuais.
- Explicitando propriedades dos processos participativos e usos destas propriedades.
- Integrando o portal federal de participação social (Participa.br) ao legado humano de dados linkados (via critérios semânticos).
- Permitindo critérios funcionais para atribuição de papéis para participantes. Por exemplo, a construção de um manifesto ou resumo final pode ser feito requisitando: de periféricos, os substantivos; de hubs, os adjetivos; e de intermediários, que formem o texto com aquelas palavras. Outra possibilidade é a remuneração de hubs pela participação efetuada ou a convocação de periféricos para oxigenar o processo participativo.

- Aproximando perfis técnicos pela qualidade das tecnologias utilizadas e da relevância dos dados sobre participação social.

Referências

[1] "Single sign-on - wikipedia," http://en.wikipedia.org/wiki/Single_sign-on - Acessado em 22 de Maio de 2014.

Abreviações e jargão

OPS: Ontologia de participação Social

OPA: Ontologia do Participa.br

MMISSA: Monitoramento Massivo e Interativo da Sociedade pela Sociedade para Aproveitamento

AARS: A Análise de Redes Sociais

PNPS: Plano Nacional de Participação Social

RDF: Resource Description Framework

HTTP: Hypertext Transfer Protocol

SPARQL: Simple Protocol and RDF Query Language

endpoint SPARQL: ponto de acesso, geralmente HTTP, a dados em RDF via buscas em SPARQL.

Participa.br: Portal federal de participação social.

IPython Notebook: instância online para rodar scritps Python

Mateor: arcabouço para páginas reativas e com funcionamento distribuído.

D3js: biblioteca de visualização de dados.

Anexos

1. Exemplo em código computacional de classificação de conteúdo via texto
2. Seleção por rankeamento (tamanho de palavra) e limiar (número mínimo de palavras)
3. Criação de redes de amizade e de conteúdo do Participa.br
4. Deteção de comunidades
5. Ordenação (ranking) por centralidade
6. Setores da ordenação (ranking)
7. Exemplo em código computacional de classificação de conteúdo via conectividade dos participantes.
8. Script para testar o tempo de resposta do endpoint SparQL como conexão local e remota. OK
9. Experimentos online.