CEPID NeuroMat
25/03/2021

A linguagem natural do voto:

modelagem com estruturação intencional de dados para a produção de verbetes sobre eleições municipais brasileiras na Wikipédia

Éder Porto • Érica Azzellini • João Peschanski

Érica Azzellini
erica@wmnobrasil.org

Introdução

- O jornalismo busca recursos de automatização de notícias sobre eleições, baseado em softwares de Natural Language Generation (NLG). Artigos criados por robôs, revisados por humanos e publicados.
- Proposta de um novo processo e ferramenta para indexação e difusão de dados legíveis por humanos sobre política

Introdução

Perguntas:

1. Qual é um processo eficiente para contar com narrativas em linguagem natural para a produção de conteúdo Wikimedia sobre eleições?
2. Qual é a eficácia e qual é o impacto social do conteúdo da Wikimedia produzido via NLG sobre eleições?

Discussão teórica

Jornalismo Computacional e softwares NLG

1. Mudanças e desafios no campo profissional do jornalismo

- Receio de substituição de humanos por máquinas **x** aumento na eficiência de redações e da satisfação profissional
- Integração de programação às práticas jornalísticas

Discussão teórica

Jornalismo Computacional e softwares NLG

2. Transparência e prestação de contas de governos

- Pressão para digitalização de governos
- Dados públicos, verificáveis por jornalistas e cidadãos
- Cultura open source: amplo acesso e engajamento comunitário

Discussão teórica

Jornalismo Computacional e softwares NLG

3. Camadas narrativas acionadas pelo uso de bancos de dados estruturados

- Automatic multimediality to the texts; language diversity; adaptability to local contexts; unprecedented velocity on real time coverage and data feeding from third party databases.
- Grande volume de dados: agenda jornalística, potenciais de cobertura e de formatos narrativos

Projetos Wikimedia

O ecossistema do conhecimento livre

- Atuam na disponibilização da soma de todo o conhecimento humano de forma livre e gratuita para todos
- Ambiente propício para experimentação e inovação aberta

Projetos Wikimedia

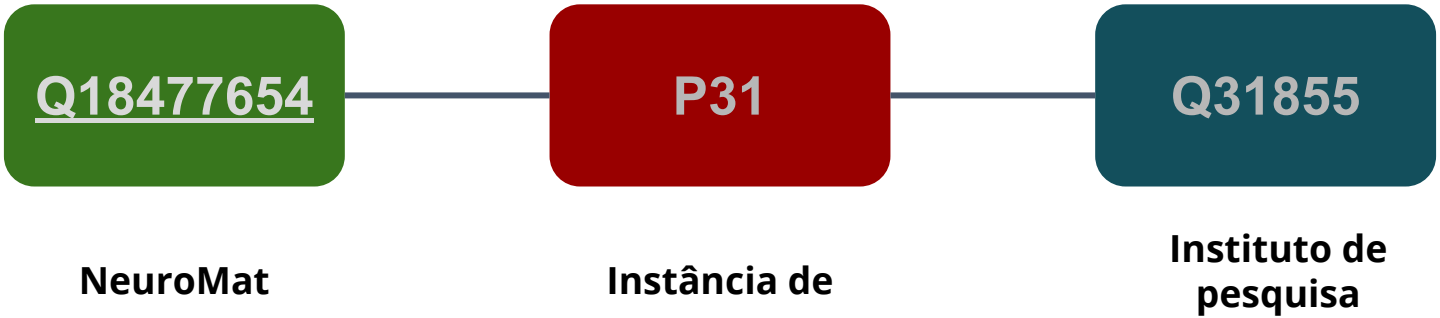
O ecossistema do conhecimento livre

- **Wikipédia:** a enciclopédia livre
- **Wikidata:** o banco de dados web semântico livre multilíngue
- Base de voluntários; toda informação deve ser referenciada; não comercial; possibilidades de integração entre as plataformas

Projetos Wikimedia

O ecossistema do conhecimento livre

- **Wikidata:** itens, propriedades, valores



Projetos Wikimedia

O ecossistema do conhecimento livre

- **Wikidata:** https://www.wikidata.org/wiki/Q61870761
- Eleição municipal de Cachoeiro de Itapemirim em 2016

Metodologia

- Abordagem metodológica de desenvolvimento de ferramenta orientada ao usuário
- Intersecção entre pesquisa e inovação
- Escopo das Humanidades Digitais

Metodologia

- **Diagnosis of the problem:** identification of pitfalls and caveats in the perspective of users of the application;
- **Solution proposal:** outline of the framework of the process and tool that can address the identified problem;
- **Pilot prototyping:** presentation of context, features and results of the developed tool in real life experimentation; and
- **Tool assessment:** critical evaluation of functionalities, scalability and impacts in the perspective of problem to be solved.

Diagnóstico do problema

- **Ignorância racional:** decisão individual de não adquirir informação política de qualidade porque os custos (tempo, esforço, dinheiro) geram pouco ou nenhum benefício
- **A Wikipédia pode ser um caminho:** custo zero e alta qualidade
- **A Wikipedia dilemma:** content on elections is of high relevance for internet users, yet its production has been associated to low engagement from editors

Solução proposta

- **Ferramenta Mbabel:** geração automática de rascunhos de verbetes na Wikipedia de eleições brasileiras baseada em dados do Wikidata
- Desenvolvimento durante pesquisa da bolsa de Jornalismo Científico FAPESP no CEPID NeuroMat em 2018
- Baseada em recurso criado por Richard Knipel, The Metropolitan Museum of Art

Narrativas estruturadas

O desenvolvimento de textos verbais, compreensíveis por humanos, automatizados a partir de arranjos pré-determinados que processam informação de bases de dados estruturados

WIKIDATA

MBABEL

WIKIPÉDIA

Prototipagem piloto

- Modelagem de dados no Wikidata sobre eleições municipais e estaduais

- Processo de transclusão em linguagem natural para a Wikipédia

Ferramenta Mbabel

Templates narrativos:

- Pré-determinados, genéricos e editáveis
- Obras de arte, museus, arquivos, bibliotecas, livros, filmes, jornais, sismos, **eleições municipais e estaduais**
- Introdução, sugestões de seções, infocaixa automática, referências, categorias, tabelas, Listeria

Ferramenta Mbabel

Templates narrativos:

- **eleições municipais e estaduais**

múltiplos Qids

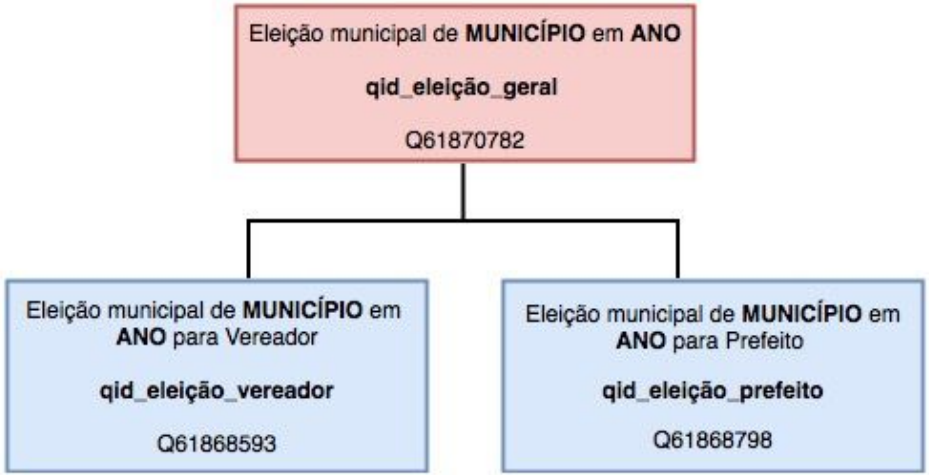
- Mais dados disponíveis
- Frases mais completas, complexas e versáteis
- Outros recursos de visualização da informação possíveis
- Aproximação de verbete completo
- Modelagem com estruturação intencional

Ferramenta Mbabel

Processo de elaboração de narrativas:

1. Dados raspados do Tribunal Superior Eleitoral (TSE)
2. Uso de Placar UOL Eleições ID como referência
3. Criação de itens para a eleição e para os candidatos

Eleição municipal de turno único

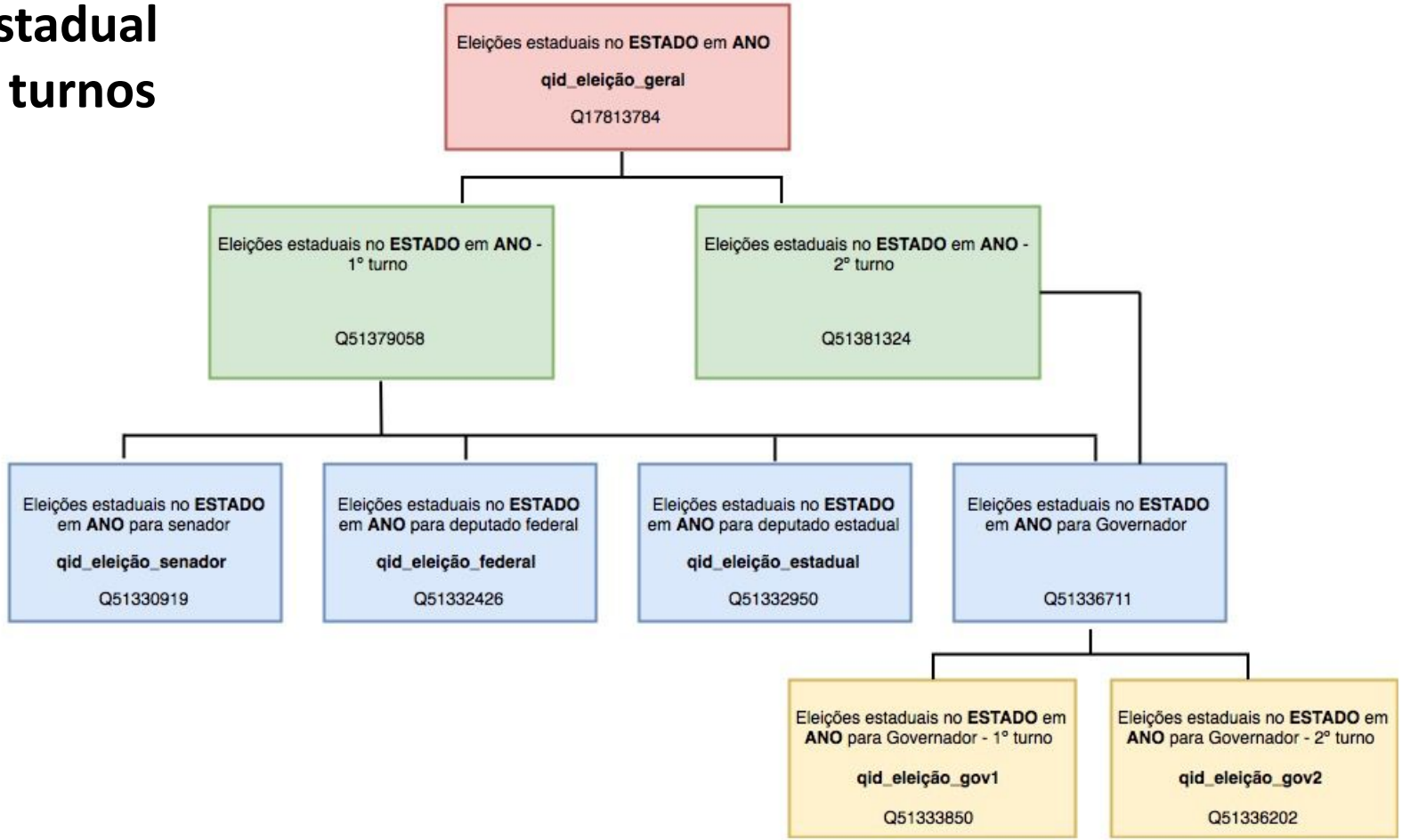


No Mbabel, 4 qids:

qid_município
qid_eleição_geral
qid_eleição_vereador
qid_eleição_prefeito

- Eleição municipal de Ilhéus em 2016 (Q61870746)
- Eleição municipal de Ilhéus em 2016 para Prefeito (Q61868765)
- Eleição municipal de Ilhéus em 2016 para Vereador (Q61868575)

Eleição estadual com dois turnos



Processo de elaboração de narrativas:

1. Observar verbetes (feitos por humanos) que já existem e seguir estrutura narrativa
2. Investigar as possibilidades do Módulo WikidataIB e criar funções, caso necessário
3. Antecipar possíveis variáveis, como turnos
4. Buscar uma linguagem neutra, considerando marcações de número/gênero
5. Cuidado com critérios de notoriedade
6. Testar recorrentemente

Avaliação da ferramenta

- Atividade em 2019 com alunos da matéria de Jornalismo Multimídia da Faculdade Cásper Líbero (FCL)
- **Três pontos:** introduzir os alunos ao conceito de Jornalismo Computacional; piloto para automatizar parcialmente a geração de conteúdo de eleições na Wikipédia; e produção de conteúdo de qualidade na Wikipédia em ampla escala pelo alunos

Avaliação da ferramenta

- 102 alunos no programa de educação
- Workshops de Wikidata e Wikipédia
- Sequência de passos: criação de páginas de teste na Wikipédia para ativação do bot; geração de página que une as páginas de teste; adição das seções “antecedentes”, “campanha” e “análise”, a partir de curadoria humana

Avaliação da ferramenta

- As a result of this activity, 102 entries on Brazilian elections were created on Wikipedia in Portuguese (REF). Students have made 2.36 thousand edits, adding 1.48 million characters to Wikimedia projects. Content that was produced was seen during the period of the specific activity in 2019 64,1 thousand times.

Avaliação da ferramenta

Number of views of Wikipedia articles about Brazilian elections between July, 2015 and October, 2020

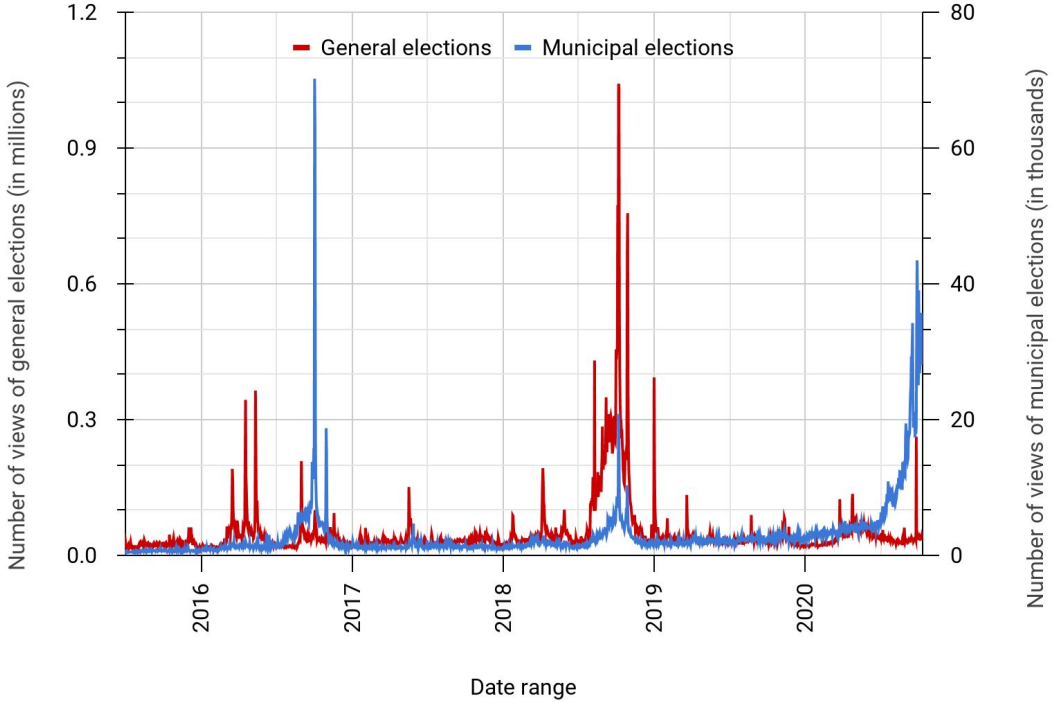


Gráfico por Éder Porto.

Avaliação da ferramenta

Cumulative number of views of Portuguse Wikipedia articles about Brazilian elections between July 2015 and October, 2020

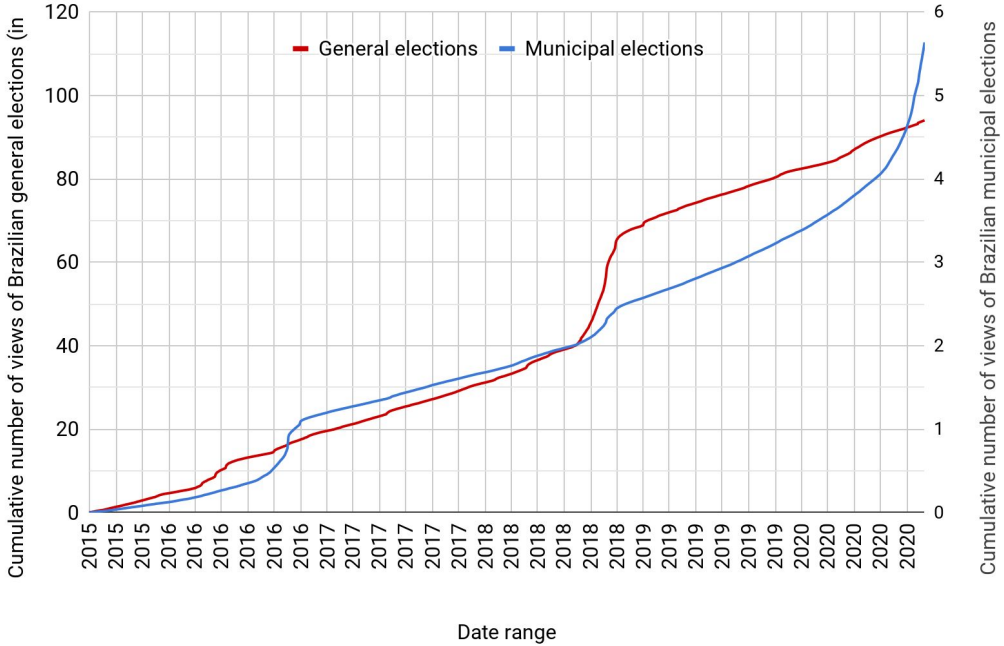


Gráfico por Éder Porto.

Considerações finais

A definir :)

1. Qual é um processo eficiente para contar com narrativas em linguagem natural para a produção de conteúdo Wikimedia sobre eleições?
2. Qual é a eficácia e qual é o impacto social do conteúdo da Wikimedia produzido via NLG sobre eleições?