



Recomendação de criação de artigos da Wikipédia

Aluno: Rennan de Lucena Gaio



Índice

- Apresentação do artigo
- Proposta do trabalho
- Obtenção dos dados
- Sistema de ranqueamento
- Resultados
- Problemas encontrados



Artigo selecionado

Nome: Growing Wikipedia Across Languages via Recommendation

Autores: Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec

Artigo realizado pela universidade de Stanford em parceria com a Wikimedia

Data de publicação: Abril de 2016

Conferência: World Wide Web Conference (WWW Conf.)



Proposta do artigo selecionado

- Sistema de recomendação de criação de artigos do Wikipédia para todos os idiomas, utilizando os top 50 idiomas como base de conteúdo já existente.
- Dados extraídos a princípio do Wikidata e Wikimedia.
- Ranqueamento dos artigos em relação a relevância de um determinado tópico com o idioma alvo.
- Utilização de técnicas de regressão para a avaliação.
- Recomendação de artigos não existentes para editores baseados em seus interesses, e sua relevância para o idioma.



Classes de dados do Wikidata

Classes de objetos ou conceitos eram do tipo Q(algum número inteiro)

Eles podiam representar conceitos como:

- Barack Obama (Q76)
- Matéria negra (Q79925)

Os conceitos eram independentes de idioma, e possuíam links para todas as páginas do wikipédia de cada idioma que representavam esse conceito.

E eles também podem ser interligados por propriedades, que são do tipo P(algum número inteiro), ex:

- Part of (P361)



Modelagem do artigo

- Cada par (S, T) era modelado como uma aresta de um grafo, e essas arestas eram ligadas pelas propriedades de cada conteúdo.
- S é o idioma que já possui conteúdo.
- T é o idioma que se quer criar o conteúdo.
- Eram retiradas possíveis palavras sinônimas, em que os ids eram diferentes, mas possuíam o mesmo significado para línguas diferentes.



Ranqueamento do artigo

Algoritmo com melhor performance utilizado por eles: Random Forest

É feita uma normalização nos dados.

Parâmetros:

- wikidata count
- page views
- geo page views
- source article length
- quality
- edit activity
- links
- topics



Matching de artigos





Proposta

Fazer um sistema reduzido para atender em específico a recomendação de artigos para o português.

Utilizado de base artigos em inglês (pois são os com maior volume de informação disponíveis)

O perfil dos autores será desconsiderado, recomendando apenas de forma a seguir os artigos que seriam mais relevantes para o idioma em geral.

Extração dos dados



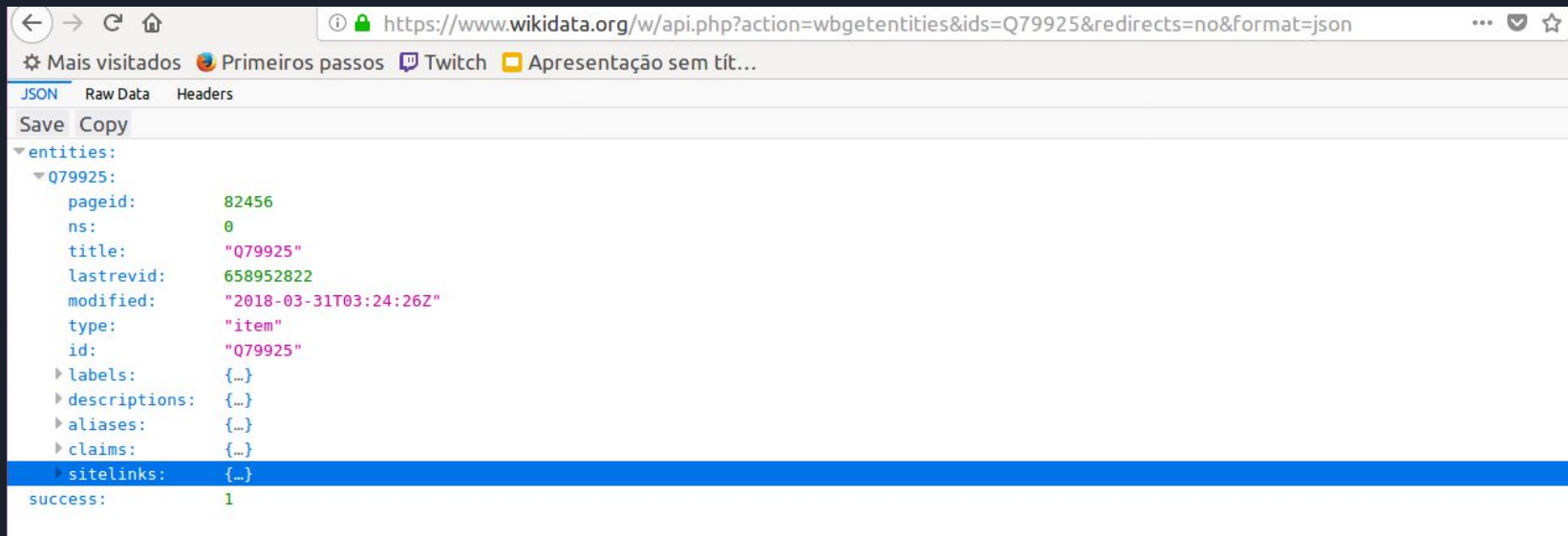


Extração dos dados wikidata

Formas possíveis de se fazer:

- Via API
- Via bots
- Via bibliotecas
- Baixando toda a base de dados

Formato dos dados extraídos



The screenshot shows a web browser window with the address bar displaying the URL: `https://www.wikidata.org/w/api.php?action=wbgetentities&ids=Q79925&redirects=no&format=json`. The browser's tab bar shows several tabs, including "Mais visitados", "Primeiros passos", "Twitch", and "Apresentação sem tít...". The main content area displays the JSON response from the Wikidata API, with tabs for "JSON", "Raw Data", and "Headers". The "JSON" tab is active, and the response is shown in a collapsible tree view. The root of the JSON object is "entities:", which contains a single entry for "Q79925:". This entry is a JSON object with the following properties: "pageid" (82456), "ns" (0), "title" ("Q79925"), "lastrevid" (658952822), "modified" ("2018-03-31T03:24:26Z"), "type" ("item"), "id" ("Q79925"), "labels" ({}), "descriptions" ({}), "aliases" ({}), "claims" ({}), and "sitelinks" ({}). The "sitelinks" property is currently selected and highlighted in blue. At the bottom of the JSON object, there is a "success:" property with the value "1".

```
{
  "entities": {
    "Q79925": {
      "pageid": 82456,
      "ns": 0,
      "title": "Q79925",
      "lastrevid": 658952822,
      "modified": "2018-03-31T03:24:26Z",
      "type": "item",
      "id": "Q79925",
      "labels": {},
      "descriptions": {},
      "aliases": {},
      "claims": {},
      "sitelinks": {}
    }
  },
  "success": 1
}
```




Obtenção das entidades

- Entidades se relacionam em forma de tripla:
 - Entidade - Propriedade - Entidade
- Não existe uma lista de Entidades disponível.
- Mas existe uma lista de Propriedades!
 - ex. P4466("Unified Astronomy Thesaurus ID"), P118 ("divisão esportiva")
- Obtenção das entidades a partir das propriedades utilizando o SPARQL!
- Selecionar os artigos que possuem página em inglês, separando os que possuem página em português, e os que não possuem ainda.



Estatísticas para ranqueamento

- Parte que gerou MUITOS problemas na execução do trabalho!
- Falta de disponibilidade de métricas.
- Utilizado somente quantidade de acessos a página.
- Dados puderam ser obtidos baixando a base de dados do wikimedia.
- Mais de 10GB de dados para apenas 1 dia.
- Dia escolhido: 11-05-2015 (Escolhi esse dia só por ser meu aniversário YESSSSS)
- Formato das tabelas baixadas:
 - origem do dado - wiki_name - page_views - tamanho do artigo



Armazenamento e processamento dos dados

- Dados salvos em arquivos csv
- Formato:
 - id - en_page_views - pt_page_views
- Um arquivo utilizado para criação do modelo de machine learning
- Outro arquivo utilizado para obtenção dos resultados



Classificadores

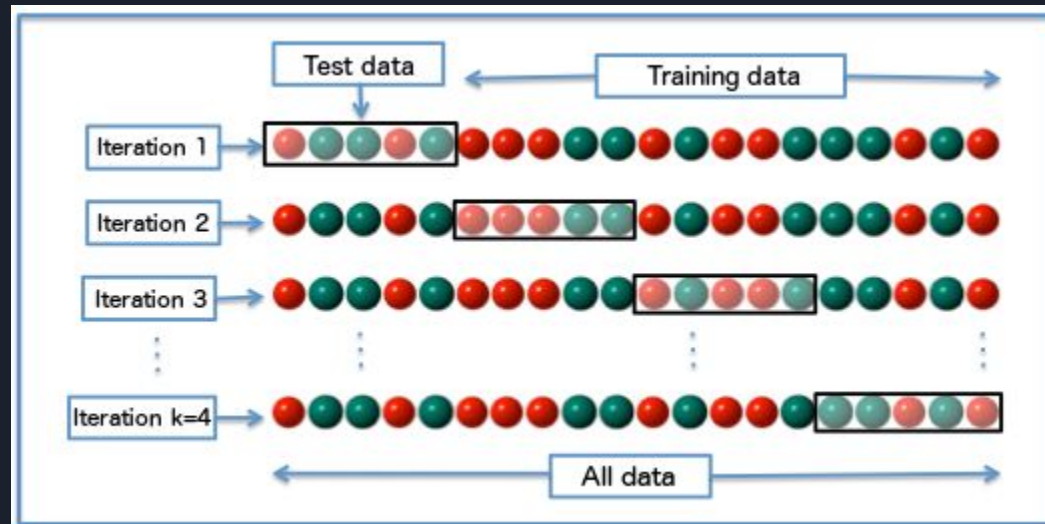
Foi proposto uma série de testes para poder escolher qual classificador obteria o melhor resultado para os dados que possuíamos.

Algoritmos de classificação utilizados (Utilizando a biblioteca do sklearn):

- Logistic Regression
- KNN
- Random Forest
- Naive Bayes
- XGBoost

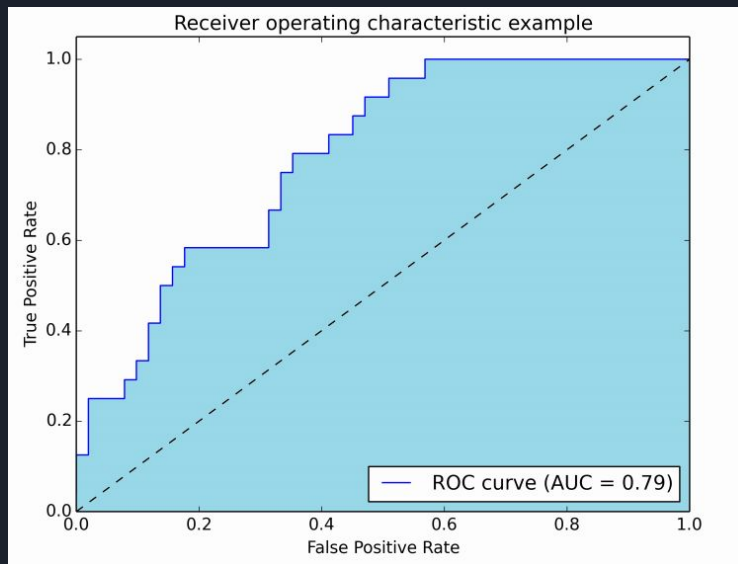
Separação dos dados

Os dados do csv completo foram divididos em conjunto de treino, e conjunto de teste utilizando a técnica de KFold.



Escolha de melhor classificador

Para escolher qual classificador foi mais performático, foi utilizada a métrica de AUC para a avaliação dos resultados de teste. (isso deve ser mudado, para serem usadas técnicas de regressão)





Resultados encontrados

Entãoooo..... meu notebook não aguentou a quantidade de dados (mesmo reduzida), devido a necessidade de cruzamento de dados de 2 datasets para a obtenção das features de avaliação (wikidata e wikimedia).



YOU
LOSE



Problemas encontrados

- Grande dificuldade em obtenção dos dados.
- Quantidade muito grande de dados.
- Descentralização dos dados.
- Despadronização dos dados.
- Baixo poder de processamento.

Dúvidas?

