



Triplificação de tweets em português com análise de sentimentos

Matheus Feijoó

feijoo@ufrj.br

MAI715 - Organização do Conhecimento



AGENDA

- Artigo Selecionado
- Análise de Redes Sociais
- Twitter
- Quais dados devem ser analisados?
- TweetsKB
- Implementação da tradução em RDF com análise de sentimentos em português
- Execução e Testes
- Conclusão



TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets

- 15th Extended Semantic Web Conference (ESWC'18), Greece, June 3-7, 2018
- **Autores:** Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi e Stefan Dietze
- Fazem parte do centro de pesquisa **L3S** na Alemanha com foco no **desenvolvimento de métodos e tecnologias** para a mudança digital e **investiga os efeitos da digitalização** para obter opções de ação, recomendações e estratégias de inovação para negócios, política e sociedade.
- O artigo foi nomeado ao prêmio de **Best Resource Paper** da conferência.

Análise de Redes Sociais

- As redes sociais são **extremamente populares** nos dias de hoje.
- São gerados **quantidades absurdas de dados**.
- Brasil é o **segundo país que mais gasta tempo em redes sociais** (225 min média/dia)
- **500 milhões de tweets** são publicados a cada dia.
- Brasil é o **2º colocado em contas ativas no Twitter**.



blog.statusbrew.com/social-media-statistics-2018-for-business

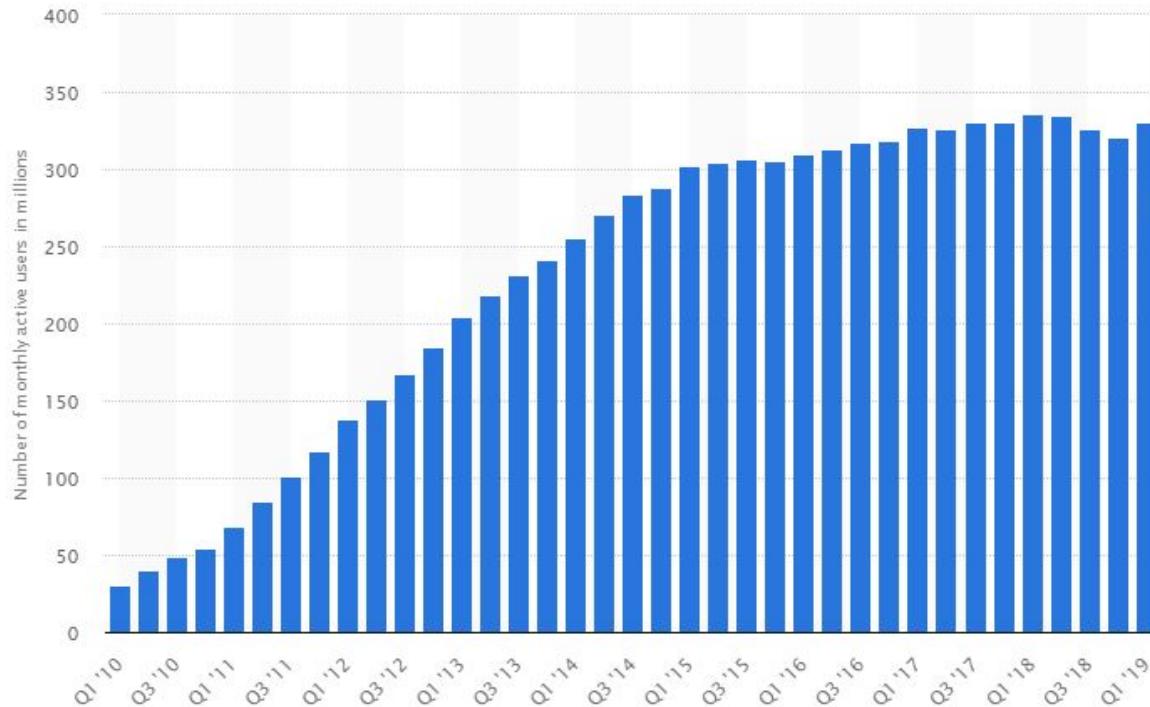
<https://f5.folha.uol.com.br/nerdices/2019/09/brasil-e-2o-em-ranking-de-paises-que-passam-mais-tempo-em-redes-sociais.shtml>

Twitter

- É uma das **maiores plataformas de microblogging**.
- Os dados gerados pelo Twitter são **altamente analisados por diversas áreas de pesquisa**: Ciência de Dados, Psicologia, Sociologia, Publicidade e etc.
- **Possui inúmeros nichos**: Esportes, Política, Música, Negócios, Estilo de Vida, Comportamento Social e etc.
- **Google Acadêmico retorna 7 milhões de resultados**. 65 mil só para 2019.
- Uma das únicas plataformas que possibilita uma **análise real de seus dados**.

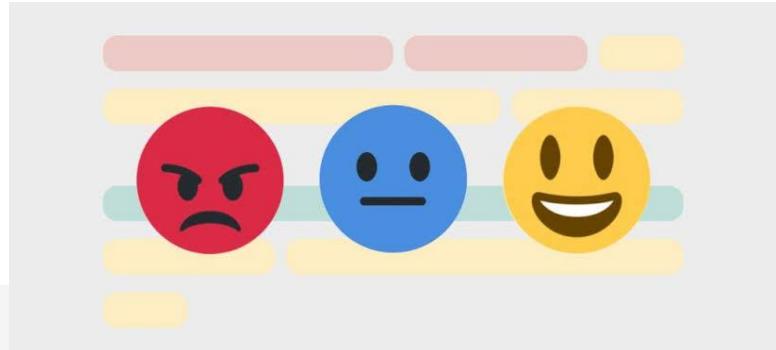
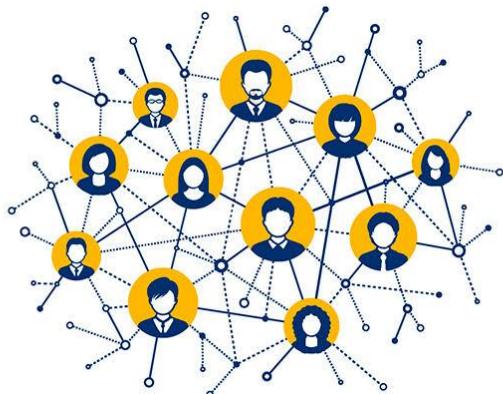
Twitter

(Usuários ativos por mês em milhões)



Quais dados devem ser analisados?

- Dados como seguidores de uma conta, número de curtidas e compartilhamento são informações extremamente importantes para realização de análises, contudo é possível extrair mais!
- Como coletar o sentimento de um tweet?
- Quais são as relações entre as entidades presentes em um tweet?



Como esses dados são analisados?

- Os autores citam exemplos atuais em que os **dados são analisados manualmente por pesquisadores não pertencentes a computação**, como no exemplo de uma análise jornalística de **14 mil tweets** feita pelo NY TIMES nas eleições americanas de 2016.
- **DeltaFolha** é criada em Maio de 2019 para foco exclusivo na análise de dados para fins jornalisticos.
- **Não havia de armazenamentos de grandes dados de redes sociais.**
- Como auxiliar análise de dados de tweets?



TweetsKB

- Nada mais é que um conjunto público de tweets anônimos em RDF.
- Contém 1.5 bilhões de tweets.
- Entre Jan-2013 até Mar-2018.
- Inclui análise de sentimentos e de entidades.
- É ideal para:
 - Noção de tempo e exploração centrada em entidades.
 - Integração de dados por bases já existentes.
 - Análise de vários aspectos e análise centrada em entidades.



TweetsKB

TweetsKB

[What](#) • [Why](#) • [Dataset](#) • [Stats](#) • [Data model](#) • [Examples](#) • [Contact](#) • [About](#)

what

TweetsKB is a public RDF corpus of anonymized data for a large collection of **annotated** tweets. The dataset currently contains data for more than **1.5 billion** tweets, spanning more than **5 years** (February 2013 - March 2018). **Metadata** information about the tweets as well as extracted **entities**, **sentiments**, **hashtags** and **user mentions** are exposed in RDF using established RDFS vocabularies. For the sake of privacy, we encrypt the tweet IDs and usernames, and we do not provide the text of the tweets.

More information is available at the following paper:

P. Fafalios, V. Iosifidis, E. Ntoutsi, and S. Dietze,
TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets,
15th Extended Semantic Web Conference (ESWC'18), Heraklion, Crete, Greece, June 3-7, 2018.
Nominated for the "Best Resource Paper" award!
[pdf](#) • [bib](#) • [slides](#)

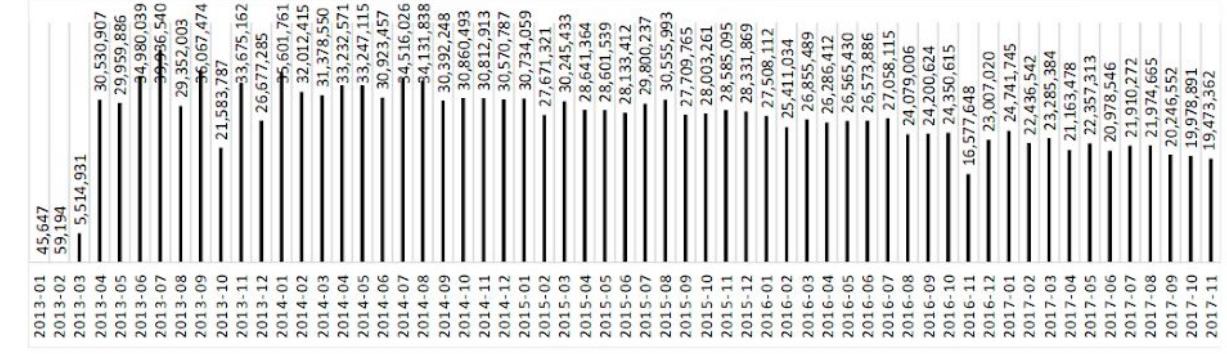
why

[top](#)

- For relieving data consumers from the computationally intensive process of extracting and processing tweets.
- For facilitating a variety of multi-aspect data consumption, exploration and analytics scenarios. These include:
 - time-aware and entity-centric exploration of the Twitter archive
 - data integration by directly exploiting existing knowledge bases (like DBpedia)
 - entity-centric analytics and knowledge discovery by inferring multi-aspect information related to one or more entities during certain time periods (like popularity, attitude or relations with other entities)

TweetsKB

- Iniciou-se com 6 bilhões de tweets. Foram retirados:
 - Retweets
 - Tweets que não estavam em inglês
 - Spams
- Foram adicionados metadados



TweetsKB : Análise de dados

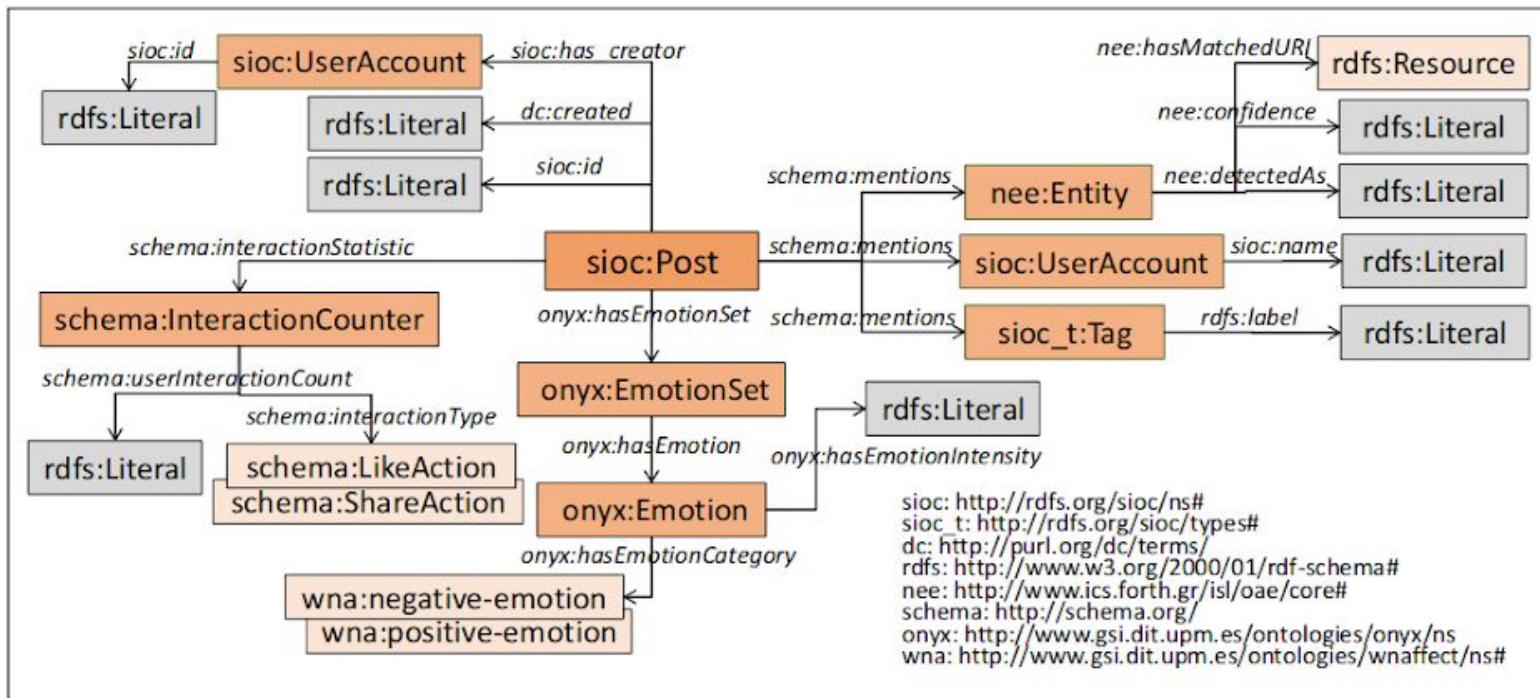
- **Conexão de Entidades:** Utilizou a ferramenta Yahoo FEL que faz a conexão com entidades da DBpedia, é desenvolvido para textos pequenos.
- **Análise de Sentimentos:** Utilizou o SentiStrength, uma ferramenta robusta para análise sentimental de dados sociais da Web. Realiza a pontuação negativa e positiva em relação ao texto, para assim analisar os dois sentimentos ao mesmo tempo.



TweetsKB : Modelo RDF

- Os tweets no modelo RDF/S foram associados com 6 tipos de elementos:
 - **Metadados:** *sioc:Post*→*tweet*, *sioc:UserAccount*→*user*
 - **Entidades mencionadas:** Para cada entidade foi armazenado o URI, pontuação de confiança (NEE) e a entidade no tweet.
 - **Menções do usuário:** *sioc:UserAccount*→*user*
 - **Hashtags mencionadas:** *sioc_t:Tag*
 - **Pontuação sentimental:** *onyx:EmotionSet*
 - **Estatísticas de interação:** *schema:LikeAction*→*favorite_count*, *schema:ShareAction*→*retweet_count*

TweetsKB : Modelo RDF



```
:t02243578068a7372090869a87a63f7d1 rdf:type sioc:Post ;
    dc:created "2013-02-06T23:17:49"^^xsd:dateTime ;
    sioc:id "02243578068a7372090869a87a63f7d1" ;
    sioc:has_creator _:u3fe01e294dd6f855db742c57000dd345 ;
    onyx:hasEmotionSet _:es02243578068a7372090869a87a63f7d1;
    schema:interactionStatistic _:i02243578068a7372090869a87a63f7d1_1,
                                _:i02243578068a7372090869a87a63f7d1_2 .

_:u3fe01e294dd6f855db742c57000dd345 rdf:type sioc:UserAccount ;
    sioc:id "3fe01e294dd6f855db742c57000dd345" .

:_102243578068a7372090869a87a63f7d1_1 rdf:type schema:InteractionCounter ;
    schema:interactionType schema:LikeAction ;
    schema:userInteractionCount "0"^^xsd:integer .

:_102243578068a7372090869a87a63f7d1_2 rdf:type schema:InteractionCounter ;
    schema:interactionType schema:ShareAction ;
    schema:userInteractionCount "3"^^xsd:integer .

_:es02243578068a7372090869a87a63f7d1 rdf:type onyx:EmotionSet ;
    onyx:hasEmotion _:em02243578068a7372090869a87a63f7d1Pos,
                                _:em02243578068a7372090869a87a63f7d1Neg .

_:em02243578068a7372090869a87a63f7d1Pos onyx:hasEmotionCategory wna:positive-emotion ;
    onyx:hasEmotionIntensity "0.0"^^xsd:double .

_:em02243578068a7372090869a87a63f7d1Neg onyx:hasEmotionCategory wna:negative-emotion ;
    onyx:hasEmotionIntensity "0.25"^^xsd:double .

:_t02243578068a7372090869a87a63f7d1 schema:mentions _:e02243578068a7372090869a87a63f7d1_0 .

_:e02243578068a7372090869a87a63f7d1_0 rdf:type nee:Entity ;
    nee:detectedAs "abortion" ;
    nee:hasMatchedURI <http://dbpedia.org/resource/Abortion> ;
    nee:confidence "-2.0047875069805587"^^xsd:double .

:_t02243578068a7372090869a87a63f7d1 schema:mentions _:h02243578068a7372090869a87a63f7d1_0 .

_:h02243578068a7372090869a87a63f7d1_0 rdf:type sioc:t:Tag ;
    rdfs:label "myThoughts" .
```

TweetsKB : Resultados

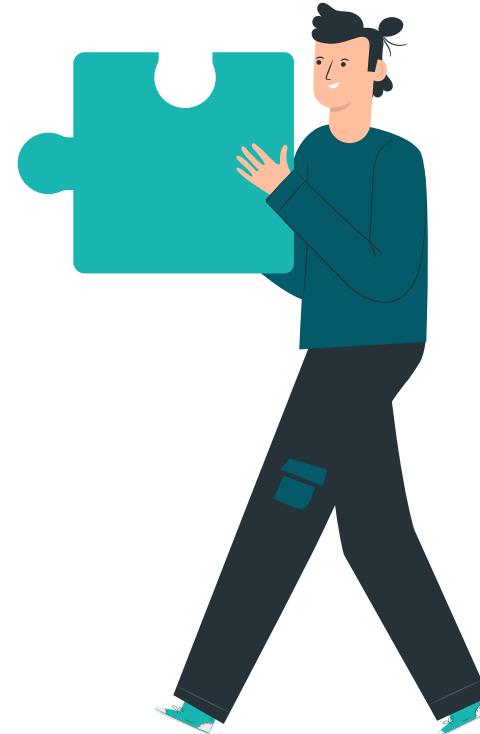
Table 2. Key statistics of *TweetKB*.

Number of tweets:	1,560,096,518
Number of distinct users:	125,104,569
Number of distinct hashtags:	40,815,854
Number of distinct user mentions:	81,238,852
Number of distinct entities:	1,428,236
Number of tweets with sentiment:	772,044,599
Number of RDF triples:	48,207,277,042

Implementação da tradução em RDF com análise de sentimentos em português



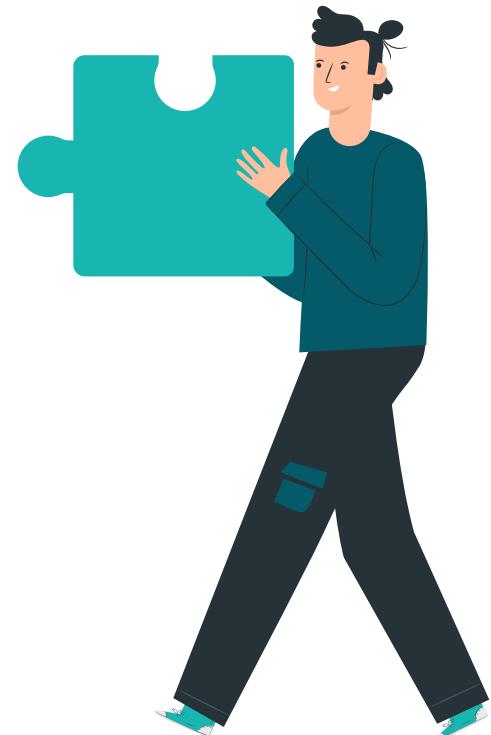
Metodologia

- 
- Analisar os dados anotados do TweetsKB.
 - Analisar a execução do Yahoo FEL para conexão de entidades e do SentiStrength para análise de sentimentos.
 - Levantar mais variáveis que podem ser analisadas e armazenadas via RDF.
 - Analisar se haveria mudanças a partir de tweets em português.
 - Conseguir uma base de dados em português
 - Desenvolver aplicação de triplificação.
- 

Conexão de Entidades

- Utilizou-se o pacote DBpedia Spotlight, por ser mais intuitiva, simples e com mais suporte que o Yahoo FEL.

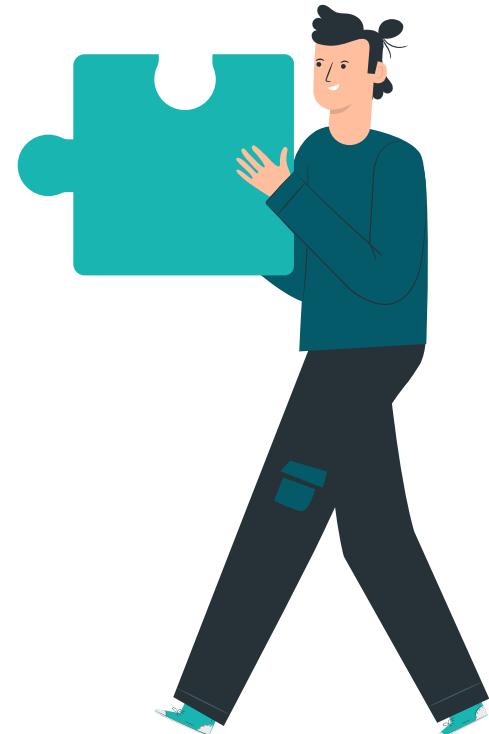
```
HOSTPT = 'http://api.dbpedia-spotlight.org/pt/annotate'
def get_annotationsPT(text):
    try:
        annotationsPT = spotlight.annotate(HOSTPT, text, confidence=0.35, support=10)
    except spotlight.SpotlightException:
        return None
    return annotationsPT
```



Análise de Sentimentos

- Utilizou-se o pacote `VaderSentiment` para a realização da análise de sentimento onde gera dados referentes a intensidade dos sentimentos positivo e negativo.

```
def do_sentAnalysis (text):  
    analyzer = SentimentIntensityAnalyzer()  
    scores = analyzer.polarity_scores(text)  
    return scores
```

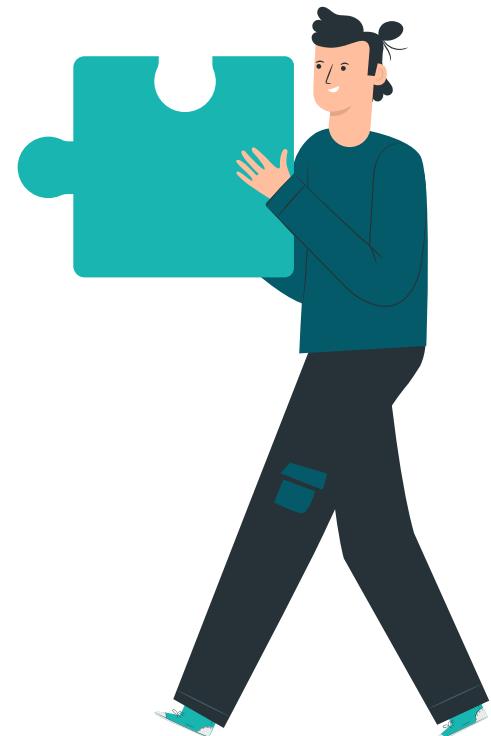


Análise de Sentimentos

- VaderSentiment só funciona em inglês, para isso o tweet em português foi traduzido utilizando o pacote GoogleTrans. Além disso os emoticons foram transformados em texto pelo pacote Emoji.

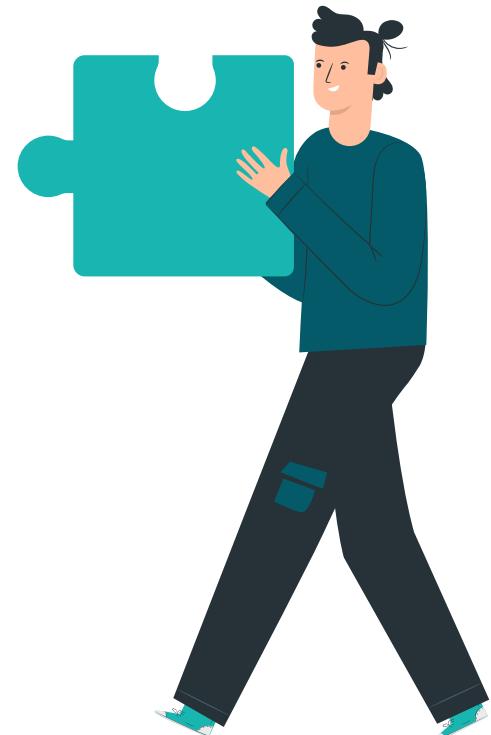
```
def do_translation(text):
    translator = Translator()
    ex = translator.translate(text)
    return ex

def extract_emojis(text):
    return emoji.demojize(text)
```



Armazenamento de Triplas

- O armazenamento foi feito baseado nas triplas do TweetsKB onde as mesmas foram transformadas em um modelo para que assim os tweets possam ser analisados, os dados e metadados coletados e armazenados no arquivo no formato .n3



```

with open("datasetBolsonaro.n3", "a") as f:
    #Base de todos os dados
    f.write(' \n' + ':t' + ' rdf:type sioc:Post ; dc:created "' + tempo + '"^^xsd:dateTime ; ' + 'sioc:id "' + codTweet + '" ; ' + 'sioc:has_creator :u' + codUser + ' ; onyx:hasEmotionSet :es' + codUser + '\n')
    f.write(':u' + codUser + ' rdf:type sioc:UserAccount ; ' + 'sioc:id "' + codUser + '" . ' + '\n')
    f.write(' :i' + codTweet + ' 1 rdf:type schema:InteractionCounter ; schema:interactionType schema:LikeAction ; schema:userInteractionCount "' + likes + '"^^xsd:integer .' + '\n')
    f.write(' :i' + codTweet + ' 2 rdf:type schema:InteractionCounter ; schema:interactionType schema:ShareAction ; schema:userInteractionCount "' + rt + '"^^xsd:integer .' + '\n')

    #Parte de emoções
    textoEN = do_translation(textoPT)
    traduzidoEN = re.search('text=(.*), pronunciation=', str(textoEN))
    var = traduzidoEN.group(1)
    senti = do_sentAnalysis(var)
    sentimentoPos = str(senti['pos'])
    sentimentoNeg = str(senti['neg'])
    f.write(' :es' + codTweet + ' rdf:type onyx:EmotionSet ; onyx:hasEmotion :em' + codTweet + 'Pos, :em' + codTweet + 'Neg .' + '\n')
    f.write(' :em' + codTweet + 'Pos onyx:hasEmotionCategory wna:positive-emotion ; onyx:hasEmotionIntensity "' + sentimentoPos + '"^^xsd:double .' + '\n')
    f.write(' :em' + codTweet + 'Neg onyx:hasEmotionCategory wna:negative-emotion ; onyx:hasEmotionIntensity "' + sentimentoNeg + '"^^xsd:double .' + '\n')

    spotlightPT = get_annotationsPT(textoPT)
    time.sleep(5)
    datasetSpolightPT = pandas.DataFrame(spotlightPT)
    #Se tiver alguma coisa da DBpedia
    if len(datasetSpolightPT) != 0:
        for varDatasetSpolightPT in range(len(datasetSpolightPT)):
            URI = datasetSpolightPT.loc[varDatasetSpolightPT,"URI"]
            pegou = datasetSpolightPT.loc[varDatasetSpolightPT,"surfaceForm"]
            score = datasetSpolightPT.loc[varDatasetSpolightPT,"similarityScore"]
            f.write(' :t' + codTweet + ' schema:mentions :e' + codTweet + ' .' + str(varDatasetSpolightPT) + ' .' + '\n')
            f.write(' :e' + codTweet + ' .' + str(varDatasetSpolightPT) + ' rdf:type nee:Entity ; nee:detectedAs "' + str(URI) + '" ; nee:confidence "' + str(score) + '" .\n')

    #Se tiver alguma hashtag
    hashtag = re.findall(r'#(.*)\ ', textoPT)
    if len(hashtag) != 0 :
        for varHashtag in range(len(hashtag)):
            f.write(' :t' + codTweet + ' schema:mentions :h' + codTweet + ' .' + str(varHashtag) + ' .' + '\n')
            f.write(' :h' + codTweet + ' .' + str(varHashtag) + ' rdf:type sioc_t:Tag ; rdfs:label "' + hashtag[varHashtag] + '" .' + '\n')

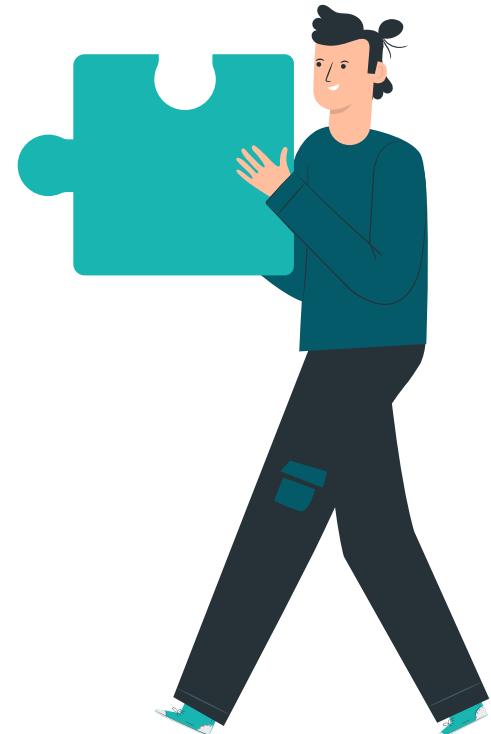
    #Se cita algum usuário
    arroba = re.findall('@(.*)\ ', textoPT)
    if len(arroba) != 0 :
        for varArroba in range(len(arroba)):
            f.write(' :t' + codTweet + ' schema:mentions :a' + codTweet + ' .' + str(varArroba) + ' .' + '\n')
            f.write(' :a' + codTweet + ' .' + str(varArroba) + ' rdf:type sioc_t:Microblog ; rdfs:label "' + arroba[varArroba] + '" .' + '\n')

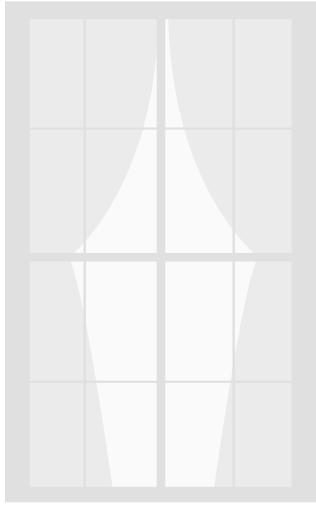
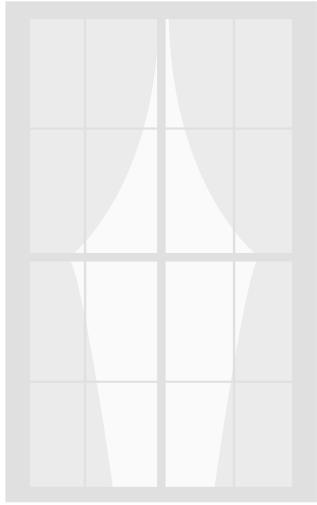
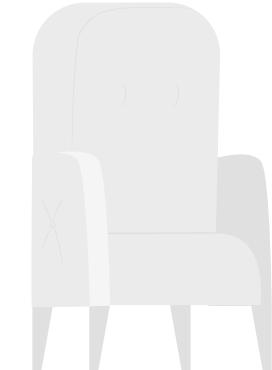
f.close()
return ('Feito')

```

Pontos a serem considerados

- Durante a análise dos dados armazenados do TweetsKB e do trabalho publicado não foi citado análises referentes aos sentimentos expressados por utilização de emoticons.
- Também não foi encontrado o armazenamento de usuários que são citados em cada um dos tweets, somente as hashtags.
- Essas duas questões possuem uma relevante importância, com isso as mesmas foram incluídas nesta implementação.





Execução e Testes

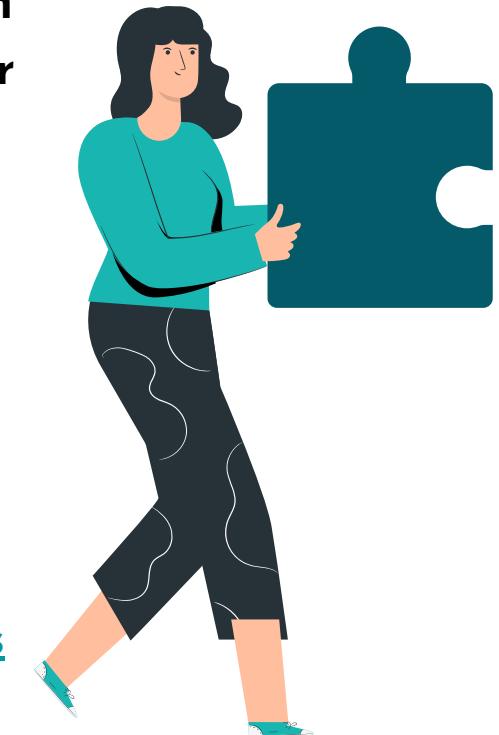


Implementação

- Para implementação foi utilizado um dataset em português contendo todos os tweets do Jair Bolsonaro desde do dia 1º de Janeiro de 2019.
- 2284 tweets em .csv contendo:
 - Id
 - Data e hora de criação
 - Texto do tweet
 - Nº de curtidas e de retweets

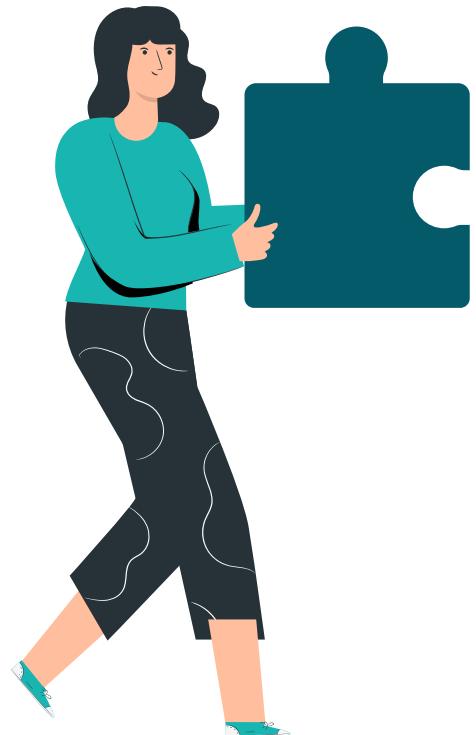
Fonte:

<https://www.kaggle.com/unanimad/bolsonaros-200-days-as-president-on-twitter>



Implementação

- **Tempo de execução do dataset: 2 horas**
- **Durante a execução aconteceram erros em relação a conexão com a DBpedia e com o Google Translate, por conta dos repetidos acessos simultâneos aos dois servidores.**
- **Houve um aumento exponencial do tamanho do arquivo .n3 em comparação com o arquivo em .csv**
- **Para realização de testes o arquivo gerado foi adicionado a ferramenta GraphDB para que seja possível a realização de consultas SPARQL.**



Testes: GraphDB

t1185960089271885824 

Source: <http://example.com/base/t1185960089271885824>

	subject	predicate	object	context	all	Explicit only	Show Blank Nodes	Download as	Visual graph
1	:t1185960089271885824	dc:created	"2019-10-20 16:44:54"^^xsd:dateTime						
2	:t1185960089271885824	sioc:has_creator	:u1						
3	:t1185960089271885824	sioc:id	1185960089271885824						
4	:t1185960089271885824	schema:interactionStatistic	:i1185960089271885824_1						
5	:t1185960089271885824	schema:interactionStatistic	:i1185960089271885824_2						
6	:t1185960089271885824	schema:mentions	:a1185960089271885824_0						
7	:t1185960089271885824	schema:mentions	:e1185960089271885824_0						
8	:t1185960089271885824	schema:mentions	:e1185960089271885824_1						
9	:t1185960089271885824	onyx:hasEmotionSet	:es1185960089271885824						
10	:t1185960089271885824	rdf:type	sioc:Post						

Testes: GraphDB

 twitterBolsonaro

e1185960089271885824_1 

Source: http://example.com/base/e1185960089271885824_1

	subject	predicate	object	context	all	Explicit only	Show Blank Nodes	Download as	Visual graph
1	:e1185960089271885824_1	nee:confidence	"0.9999991480465488""xsd:double						
2	:e1185960089271885824_1	nee:detectedAs	Brasil						
3	:e1185960089271885824_1	nee:hasMatchedURI	http://pt.dbpedia.org/resource/Brasil						
4	:e1185960089271885824_1	rdf:type	nee:Entity						

Testes: GraphDB (Entidades mais citadas)

```
SELECT ?Entidade (count(?Entidade) AS ?Vezes_Citado)
WHERE{
    ?tweet rdf:type sioc:Post.
    ?tweet schema:mentions ?mention.
    ?mention rdf:type nee:Entity.
    ?mention nee:hasMatchedURI ?Entidade.
}
GROUP BY (?Entidade)
ORDER BY DESC (?Vezes_Citado) LIMIT 10
```

	Entidade	
1	http://pt.dbpedia.org/resource/Brasil	"337"^^xsd:integer
2	http://pt.dbpedia.org/resource/Fernando_Henrique_Cardoso	"52"^^xsd:integer
3	http://pt.dbpedia.org/resource/Economia	"51"^^xsd:integer
4	http://pt.dbpedia.org/resource/Jair_Bolsonaro	"37"^^xsd:integer
5	http://pt.dbpedia.org/resource/Deus	"30"^^xsd:integer
6	http://pt.dbpedia.org/resource/Estados_Unidos	"30"^^xsd:integer
7	http://pt.dbpedia.org/resource/Deputado	"29"^^xsd:integer
8	http://pt.dbpedia.org/resource/S%C3%A3o_Paulo	"25"^^xsd:integer
9	http://pt.dbpedia.org/resource/Nordeste	"24"^^xsd:integer
10	http://pt.dbpedia.org/resource/Paran%C3%A1	"24"^^xsd:integer

Testes: GraphDB (Usuários mais citados)

```
SELECT ?Usuario (count(?Usuario) AS ?Vezes_Citado)
WHERE{
    ?tweet schema:mentions ?arroba.
    ?arroba rdf:type sioc_t:Microblog.
    ?arroba rdfs:label ?Usuario.
}
GROUP BY (?Usuario)
ORDER BY DESC (?Vezes_Citado) LIMIT 10
```

	Usuario	?
1	MInfraestrutura	"23"^^xsd:integer
2	MinEconomia	"18"^^xsd:integer
3	tarcisiogdf	"18"^^xsd:integer
4	SF_Moro:	"16"^^xsd:integer
5	CarlosBolsonaro:	"14"^^xsd:integer
6	SF_Moro	"14"^^xsd:integer
7	tarcisiogdf:	"14"^^xsd:integer
8	MEC_Comunicacao	"13"^^xsd:integer
9	exercitooficial	"13"^^xsd:integer
10	mdregional_br	"13"^^xsd:integer

Testes: GraphDB (Média de interações)

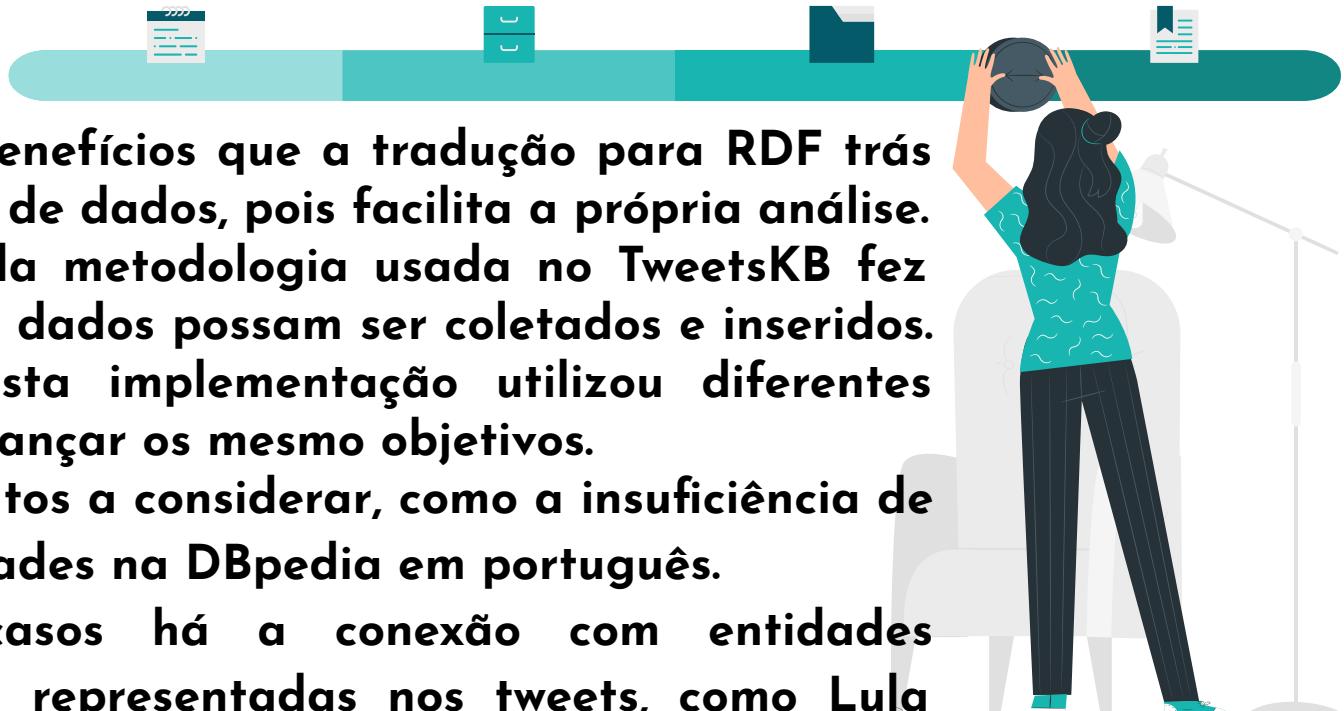
```
SELECT (avg(?likes) AS ?Media_de_Likes)
WHERE{
    ?tweet rdf:type sioc:Post.
    ?tweet schema:interactionStatistic ?interacao.
    ?interacao schema:interactionType schema:LikeAction.
    ?interacao schema:userInteractionCount ?likes.
}
```

1	"32221977568740955137481910275""xsd:decimal
---	---

```
SELECT (avg(?RTs) AS ?Media_de_RTs)
WHERE{
    ?tweet rdf:type sioc:Post.
    ?tweet schema:interactionStatistic ?interacao.
    ?interacao schema:interactionType schema:ShareAction.
    ?interacao schema:userInteractionCount ?RTs.
}
```

1	"5287.740231548480463096960926""xsd:decimal
---	---

CONCLUSÃO



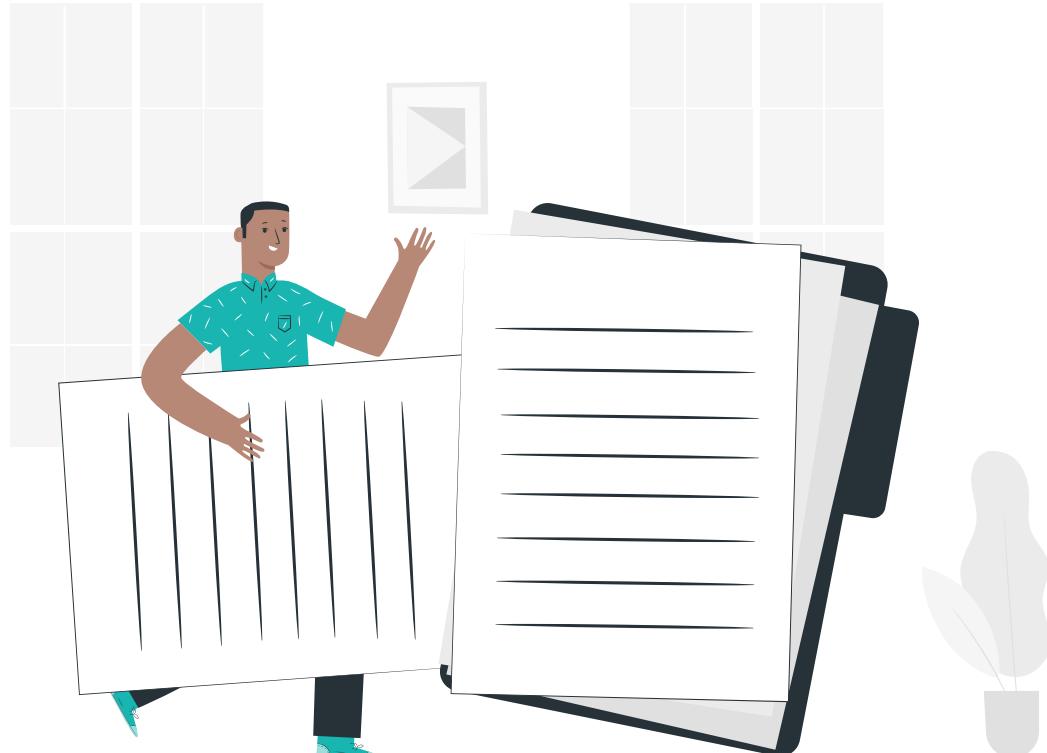
- É notório os benefícios que a tradução para RDF trás para a análise de dados, pois facilita a própria análise.
- A replicação da metodologia usada no TweetsKB fez com que novos dados possam ser coletados e inseridos. Além disso, esta implementação utilizou diferentes meios para alcançar os mesmo objetivos.
- Contudo a pontos a considerar, como a insuficiência de algumas entidades na DBpedia em português.
- Em alguns casos há a conexão com entidades diferentes das representadas nos tweets, como Lula sendo conectado ao animal e não a pessoa.

Obrigado

Alguma pergunta?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Please keep this slide for attribution.





Triplificação de tweets em português com análise de sentimentos

Matheus Feijoó

feijoo@ufrj.br

MAI715 - Organização do Conhecimento

