A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias

UFMG

Universiteit Leiden

Evandro Cunha
Gabriel Magno
Virgilio Almeida

Um tipo de conteúdo muito relevante para pesquisadores em humanidades digitais - inclusive linguistas - são os **comentários publicados por leitores de notícias** em portais online

comentários

Ver todos os comentários (140)

Caro leitor,

Termos e condições

para comentar, é preciso ser assinante da **Folha**. Caso já seja um, por favor entre em sua conta cadastrada. Se já é assinante mas não possui senha de acesso, cadastre-se.

Faça seu login | Cadastre-se | Assine

Lorenzo Frigerio *ontem às 12h38* 49 3 Denunciar

+ COMPARTILHAR

Blá blá blá blá blá blá blá blá... os tempos de militância estudantil acabaram faz tempo, senhora. Caia na real. A propósito, como anda o estoque de papel higiênico? Vai precisar dele.

O comentário não representa a opinião do jornal; a responsabilidade é do autor da mensagem

Responder

Renato Donati *ontem às 12h56* 45 0 Denunciar

+ COMPARTILHAR

Essa não é a Barraqueira que foi impedida de entrar na reunião do Mercosul?

O comentário não representa a opinião do jornal; a responsabilidade é do autor da mensagem

Responder

ANA CAROLINA AMORIM BARBOSA *ontem às 12h47* 42 3 Denunciar

+ COMPARTILHAR

Bom mesmo é viver na Venezuela!! Amiga, você tem problemas suficientes para resolver em casa: cuide deles.

n.br/tradesolution/TradeSolution/Plataformasnãoprofissionais/tabid/402/Default.as

comentários

Ver todos os comentários (140)

Termos e condições

Caro leitor,

para comentar, é preciso ser assinante da **Folha**. Caso já seja um, por favor entre em sua conta cadastrada. Se já é assinante mas não possui senha de acesso, cadastre-se.

Faça seu login | Cadastre-se | Assine

Lorenzo Frigerio *ontem às 12h38* 49 3 Denunciar

+ COMPARTILHAR

Blá blá blá blá blá blá blá blá... os tempos de militância estudantil acabaram faz tempo, senhora. Caia na real. A propósito, como anda o estoque de papel higiênico? Vai precisar dele.

O comentário não representa a opinião do jornal; a responsabilidade é do autor da mensagem

Responder

Renato Donati *ontem às 12h56* 45 0 Denunciar

+ COMPARTILHAR

Essa não é a Barraqueira que foi impedida de entrar na reunião do Mercosul?

O comentário não representa a opinião do jornal; a responsabilidade é do autor da mensagem

Responder

ANA CAROLINA AMORIM BARBOSA *ontem às 12h47* 42 3 Denunciar

+ COMPARTILHAR

Bom mesmo é viver na Venezuela!! Amiga, você tem problemas suficientes para resolver em casa: cuide deles.

A análise desses comentários permite o estudo de

questões linguísticas nos mais variados domínios:

lexical

morfossintático

pragmático

...

Além de uma série de **outras questões**

em diversas áreas do conhecimento:

sociologia

ciências políticas

comunicação social

...

Torna-se necessário o desenvolvimento de **ferramentas** capazes de auxiliar a **coletar** e a **organizar** esse material

Xereta

(a) um **coletor** de comentários de portais de notícias

(b) um **corpus** composto por comentários de leitores do portal UOL

(a) O **coletor** Xereta

> Desenvolvido em **código aberto** e **livre para uso**, **modificação** e **distribuição**

> **Funcionamento simples**, permitindo a utilização por usuários não-técnicos

> Coleta **todos** os comentários de **várias notícias** após simplesmente informar **lista de URLs**

> Disponível para **uso online** e **download**

(b) O **corpus** Xereta

> Corpus de comentários do **portal UOL**

> Comentários de notícias publicadas

em **várias seções** do site (política, esportes etc.)

> Versão **preliminar**: 25 mil comentários **já disponibilizados**

> Versão em **fase final** de coleta: **1 milhão** de comentários

Coletor para uso online,

coletor para download

e corpus disponíveis em

xereta.herokuapp.com

(ou www.dcc.ufmg.br/~evandrocunha)

Coletor: principal desafio

> As **estruturas** das páginas de cada site (UOL, Folha, Terra, G1 etc.) são **completamente diferentes** umas das outras

> As estruturas das seções **de um mesmo site** podem ser completamente diferentes

> As **informações disponíveis** para coleta também variam

Coletor: principal desafio

> As **estruturas** das páginas de cada site (UOL, Folha, Terra, G1 etc.) são **completamente diferentes** umas das outras

> As estruturas das seções **de um mesmo site** podem ser completamente diferentes

> As **informações disponíveis** para coleta também variam

Ou seja: os códigos não são reaproveitáveis!

Corpus: principal desafio

> Conseguir **listas de URLs**

para serem coletadas

> **Solução**: descobrimos listas “escondidas”

no próprio site do UOL

Próximos passos

> Tornar possível a utilização do coletor em outros portais **brasileiros** (G1, Terra etc.) e **estrangeiros** (NYT, Washington Post etc.)

> Disponibilizar a **segunda versão** do corpus, com cerca de **um milhão de comentários**

Lembrando que...

> ...o coletor Xereta é desenvolvido em **código aberto** e **livre para uso**, **modificação** e **distribuição**

> Criar uma **comunidade** de **desenvolvedores** e **usuários** interessados em melhorar a ferramenta

> Se tiver interesse, entre em contato:

evandrocunha@dcc.ufmg.br

UFMG

Obrigado!

Evandro Cunha

evandrocunha@dcc.ufmg.br
www.dcc.ufmg.br/~evandrocunha
xereta.herokuapp.com

Universiteit Leiden