Tokenição e Anotação UD no *corpus* **DANTE** de *tweets* da bolsa de valores

Ariani Di Felippo, Caroline Postali, Gabriel Ceregatto, Laura Santos Gazana e **Norton Roman**

Coord. geral: Thiago A.S. Pardo

Projeto POetiSA

- C4AI (FAPESP-IBM-USP) → Frente de PLN → POetiSA (sintaxe/*parsing*)
- *Treebank* **multigênero** de tamanho significativo para o português
  - Filiado ao modelo *Universal Dependencies* (NIVRE, 2015; NIVRE *et al.,* 2020)
    - **Tweet** é um desses gêneros
    - DANTE - ***Dependency-Analised corpora of tweets***
    - Projeto que tem como objetivo construir e anotar vários *corpora* de tweets
      - *Corpus* do domínio "**mercado financeiro**" (*stock market*)

Cenário da anotação UD do DANTE

- ***Corpus* do mercado financeiro *(stock markets*)**
  - 4,517 *tweets* do mercado financeiro
  - Coletados de março a maio de 2014
- **Equipe**
  - 3 anotadores
- **Fase atual: treinamento dos anotadores** (até o final de abril)
  - Anotação de **PoS *tags*** dos primeiros 200 *tweets* pelos **3 anotadores** (Ariani)
    - Até o momento, 100 *tweets* anotados (50 por todos e 50 apenas por mim)
- **Material de suporte**
  - Todos os produzidos para a anotação do Folha-Kaggle (manual e listas de expressões e palavras ambíguas)
  - Diretrizes de Sanguinetti *et al* (2020)[1] e outras formuladas a partir do próprio DANTE
    - Diretrizes específicas serão um apêndice do manual principal do projeto POeTiSA
- **Editor ArboratorGrew**

[1] SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; ÇETINOĞLU, Ö.; CIGNARELLA, A. T.; LYNN, T.; REHBEIN, I.; RUPPENHOFER, J.; SEDDAH, D.; ZELDES, A. (2020) Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In: *Proceedings* of the 12th Language Resources and Evaluation Conference. (LREC 2020), p. 5240–5250, 11-16 May, Marseille, France.

Para o treinamento...

- Tokenização do *corpus* inteiro
  - *tokenizador próprio
- Geração da versão cnll do *corpus tokenizado* pelo UD-Pipe
  - Requisito do ArboratorGrew
- Anotação de PoS dos primeiros 50 *tweets* e posterior discussão em grupo

Particularidades de tokenização e PoS *tagging*

Problema	Tweet original	Tokenização com problemas (às vezes o UD-Pipe)
Nome de ação	• "vai, **oibr4**. um troux... ops... investidor precisa pagar as minhas férias."	• vai, **oibr 4** . um troux ... ops ... investidor precisa pagar as minhas férias .
Abreviações como p/	• Dilma ñ permitirá q se utilizem ações **p/** tentar	• Dilma ñ permitirá q sutilizem ações **p /** tentar
Símbolo de moedas	• Petrobras confirma precificação de títulos de **US$8,5** bilhões	• Petrobras confirma precificação de títulos de **US $** 8,5 bilhões
Emoticon	• pô #PETR4 ... faça o favor de furar os R$ 14,40 de uma vez ! **=)**	• pô #PETR4 ... faça o favor de furar os R$ 14,40 de uma vez ! **=)**
	• R$ 13 ... que ironia hein ? **;)** #PETR4	• R$ 13 ... que ironia hein ? **;)** #PETR4
Data	• no dia **05/02/14** uma Onda 3	• em o dia 05/02 **/ 14** uma Onda 3
Menção e hashtag	• **@webtraderx @BlackWizardX** Vou dar uma olhada	• **@ webtraderx @ BlackWizardX** Vou dar uma olhada
	• R$ 13 ... que ironia hein ? ;) **#PETR4**	• R$ 13 ... que ironia hein ? ;) **# PETR4**
	• **#petr4** Recompra : R$ 15,17 . Hoje LAVEI A ÉGUA !	• **# petr4** Recompra : R$ 15,17 . Hoje LAVEI A ÉGUA !
	• Estratégia de Análise Técnica para o pregão de hoje . **#analisetecnica #CIEL3** http://t.co/TgcPz0VU27	• Estratégia de Análise Técnica para o pregão de hoje . **# analisetecnica # CIEL3** http://t.co/TgcPz0VU27
Questões de terminologia	• @Rivanews BRKM5 15min : com formação de pivot de alta, após divergência de alta no estocástico **%k**	• @ Rivanews BRKM5 15min : com formação de pivot de alta, após divergência de alta em o estocástico **% k**
	• A LIGHT S.A. fechou o dia de hoje ao preço de R$ 18,65 (**-1,37**%)	• A LIGHT S.A. fechou o dia de hoje a o preço de R$ 18,65 (**-1,37** %) → **- 1,37**

Particularidades lexicais e PoS *tagging*

Questão	Exemplo
Corte de palavras	Vale diz que acesso de Valemax a a China é desejável , mas não necessário : Os navios , com **capac** ... http://t.co/DUDYluUL6A #infomoney #vale5 PROPN, VERB, SCONJ, NOUN, ADP, PROPN, ADP, DET, PROPN, AUX, ADJ, PUNCT, CCONJ, ADV, ADJ, PUNCT, DET, NOUN, PUNCT, ADP, **NOUN**, PUNCT, SYM, PROPN, PROPN 18 ações passam por ' ressaca ' após carnaval e fecham em queda , Vale recua 3 % : Mesmo com **proj** ... http://t.co/GNveU1iOgh #infomoney #vale5 NUM, NOUN, VERB, ADP, PUNCT, NOUN, PUNCT, ADP, NOUN, CCONJ, VERB, ADP, NOUN, PUNCT, PROPN, VERB, NUM, SYM, PUNCT, ADV, ADP, **NOUN**, PUNCT, SYM, PROPN, PROPN
Abreviações de palavras comuns (***"contractions"***)	FOI ! ! ! **OBJ** CUMPRIDO ! ! ! VERB, PUNCT, PUCNT, PUNCT, **NOUN**, VERB, PUNCT, PUNCT, PUNCT

Particularidades estruturais e PoS *tagging*

Questão	Exemplo
Sintaxe peculiar	**#petr4 Recompra** : R$ 15,17 . Hoje LAVEI A ÉGUA ! **PROPN**, **NOUN**, PUNCT, SYM, NUM, PUNCT, ADV, VERB, DET, NOUN PUNCT
	Quem lembra de o Post #Vale5 armando uma linda venda de diário? **pois** é 27,35 obj de a queda . PRON, VERB, ADP, DET, NOUN, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT, **CCONJ,** AUX, NUM, NOUN, ADP, DET, NOUN, PUNCT
Sintaxe?	#MRFG3 (mensagem : 953403) http://t.co/QZc5iN18om PROPN, PUNCT, NOUN, PUNCT, NUM, PUNCT, SYM
Elipses	**Petrobrás ordinárias** , + 6,82 % , **Preferenciais** , + 6,45 % **ERRO_DE_DIGITAÇAO**, **ADJ**, PUNCT, SYM, NUM, SYM, PUNCT, **ADJ**, PUNCT, SYM, NUM
	Rastreamento ações - Gráfico MENSAL - 11h . **NOUN, NOUN**, PUNCT, NOUN, ADJ, PUNCT, NOUN, PUNCT

Particularidades gráficas e PoS *tagging*

Questão	Exemplo
Capitalização aleatória	$CTIP3 - Cetip (ctip-nm) Ago / e 28/04/2014 10h00 Distribuicao De Dividendo http://t.co/8E9tQwMP72 PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT, CCONJ, NUM, NOUN, ERRO_DIGITAÇÃO, ADP, NOUN, SYM
	$LREN3 - Lojas Renner (lren-nm) - Declaracao E Pagamento De Juros Sobre O Capital Proprio http://t.co/flXNxtpbT PROPN, PUNCT, PROPN, PROPN, PUNCT, PROPN, PUNCT, PUNCT, ERRO_DIGITAÇÃO, CCONJ, NOUN, ADP, NOUN, ADP, DET, NOUN, ERRO_DIGITAÇÃO, SYM
	05/03/14 - 17:20 : Maiores Baixas : BRAP4 - 4,50 % R$ 20,35 , RSID3 - 4,09 % NUM, PUNCT, NUM, PUNCT, NOUN, ADJ, PUNCT, PROPN, SYM, NUM, SYM, SYM, NUM, PUNCT, PROPN, SYM, NUM, SYM
	Em a #PETR4 fizemos em a sexta passada uma Sub Onda 5 de fundo e em o dia 05/02/14 uma Onda 3 de fundo alvos 15,44 16,40 http://t.co/5DPBQCulWr ADP, DET, PROPN, VERB, ADP, DET, NOUN, ADJ, DET, ADJ, NOUN, NUM, ADP, NOUN, CCONJ, ADP, DET, NOUN, NUM, DET, NOUN, NUM, ADP, NOUN, NOUN, NUM, NUM, SYM
	Remem para o Money : O melhor (e o pior) de o Mercado Global em o primeiro ... Petr4 , vale5 , opções , bovespa , ogxp3 http://t.co/EpwT6TzAmc PROPN, ADP, DET, PROPN, PUNCT, DET, NOUN, PUNCT, CCONJ, DET, NOUN, PUNCT, ADP, DET, NOUN, ADJ, ADP, DET, NOUN, PUNCT, PROPN, PUNCT, PROPN, PUNCT, NOUN, PUNCT, PROPN, PUNCT, PROPN, SYM
	Rastreamento ações - Gráfico MENSAL - 11h . NOUN, NOUN, PUNCT, NOUN, ADJ, PUNCT, NOUN, PUNCT
	Quem lembra de o Post #Vale5 armando uma linda venda de diário? PRON, VERB, ADP, DET, NOUN, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT
	FOI ! ! ! OBJ CUMPRIDO ! ! ! VERB, PUNCT, PUCNT, PUNCT, NOUN, VERB, PUNCT, PUNCT, PUNCT

Particularidades gráficas e PoS

PROPN com inicial maiúscula

Questão	Exemplo
Capitalização aleatória	$CTIP3 - **Cetip** (**ctip-nm**) Ago / e 28/04/2014 10h00 **Distribuicao De Dividendo** http://t.co/8E9tQwMP72 PROPN, PUNCT, **PROPN**, PUNCT, **PROPN**, PUNCT, PROPN, PUNCT, CCONJ, NUM, NOUN, ERRO_DIGITAÇÃO, **ADP, NOUN**, SYM
	$LREN3 - **Lojas Renner** (**lren-nm**) - **Declaracao E Pagamento De Juros Sobre O Capital Proprio** http://t.co/flXNxtpbT PROPN, PUNCT, **PROPN**, **PROPN**, PUNCT, **PROPN**, PUNCT, PUNCT, ERRO_DIGITAÇÃO**, CCONJ, NOUN, ADP, NOUN, ADP, DET, NOUN,** ERRO_DIGITAÇÃO, SYM
	05/03/14 - 17:20 : **Maiores Baixas** : BRAP4 - 4,50 % R$ 20,35 , RSID3 - 4,09 % NUM, PUNCT, NUM, PUNCT, **NOUN, ADJ**, PUNCT, PROPN, SYM, NUM, SYM, SYM, NUM, PUNCT, PROPN, SYM, NUM, SYM
	Em a #PETR4 fizemos em a sexta passada uma **Sub Onda** 5 de fundo e em o dia 05/02/14 uma **Onda** 3 de fundo alvos 15,44 16,40 http://t.co/5DPBQCulWr ADP, DET, PROPN, VERB, ADP, DET, NOUN, ADJ, DET, **ADJ, NOUN**, NUM, ADP, NOUN, CCONJ, ADP, DET, NOUN, NUM, DET, **NOUN**, NUM, ADP, NOUN, NOUN, NUM, NUM, SYM
	Remem para o Money : O melhor (e o pior) de o **Mercado Global** em o primeiro ... Petr4 , **vale5** , opções , **bovespa** , **ogxp3** http://t.co/EpwT6TzAmc **PROPN, ADP, DET, PROPN**, PUNCT, DET, NOUN, PUNCT, CCONJ, DET, NOUN, PUNCT, ADP, DET, **NOUN, ADJ**, ADP, DET, NOUN, PUNCT, PROPN, PUNCT, **PROPN**, PUNCT, NOUN, PUNCT, **PROPN**, PUNCT, **PROPN**, SYM
	Rastreamento ações - **Gráfico MENSAL** - 11h . NOUN, NOUN, PUNCT, **NOUN, ADJ**, PUNCT, NOUN, PUNCT
	Quem lembra de o **Post** #Vale5 armando uma linda venda de diário? PRON, VERB, ADP, DET, **NOUN**, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT
	FOI ! ! ! **OBJ CUMPRIDO** ! ! ! **VERB**, PUNCT, PUCNT, PUNCT, **NOUN**, **VERB,** PUNCT, PUNCT, PUNCT

Particularidades gráficas e PoS

PROPN com inicial minúscula

Questão	Exemplo
Capitalização aleatória	$CTIP3 - **Cetip** (**ctip-nm**) Ago / e 28/04/2014 10h00 **Distribuicao De Dividendo** http://t.co/8E9tQwMP72 PROPN, PUNCT, **PROPN**, PUNCT, **PROPN**, PUNCT, PROPN, PUNCT, CCONJ, NUM, NOUN, ERRO_DIGITAÇÃO, **ADP, NOUN**, SYM
	$LREN3 - **Lojas Renner** (**lren-nm**) - **Declaracao E Pagamento De Juros Sobre O Capital Proprio** http://t.co/flXNxtpbT PROPN, PUNCT, **PROPN**, **PROPN**, PUNCT, **PROPN**, PUNCT, PUNCT, ERRO_DIGITAÇÃO, **CCONJ, NOUN, ADP, NOUN, ADP, DET, NOUN,** ERRO_DIGITAÇÃO, SYM
	05/03/14 - 17:20 : **Maiores Baixas** : BRAP4 - 4,50 % R$ 20,35 , RSID3 - 4,09 % NUM, PUNCT, NUM, PUNCT, **NOUN, ADJ**, PUNCT, PROPN, SYM, NUM, SYM, SYM, NUM, PUNCT, PROPN, SYM, NUM, SYM
	Em a #PETR4 fizemos em a sexta passada uma **Sub Onda** 5 de fundo e em o dia 05/02/14 uma **Onda** 3 de fundo alvos 15,44 16,40 http://t.co/5DPBQCulWr ADP, DET, PROPN, VERB, ADP, DET, NOUN, ADJ, DET, **ADJ, NOUN**, NUM, ADP, NOUN, CCONJ, ADP, DET, NOUN, NUM, DET, **NOUN**, NUM, ADP, NOUN, NOUN, NUM, NUM, SYM
	Remem para o Money : O melhor (e o pior) de o **Mercado Global** em o primeiro ... Petr4 , **vale5** , opções , **bovespa** , **ogxp3** http://t.co/EpwT6TzAmc **PROPN, ADP, DET, PROPN**, PUNCT, DET, NOUN, PUNCT, CCONJ, DET, NOUN, PUNCT, ADP, DET, **NOUN, ADJ**, ADP, DET, NOUN, PUNCT, PROPN, PUNCT, **PROPN**, PUNCT, NOUN, PUNCT, **PROPN**, PUNCT, **PROPN**, SYM
	Rastreamento ações - **Gráfico MENSAL** - 11h . NOUN, NOUN, PUNCT, **NOUN, ADJ**, PUNCT, NOUN, PUNCT
	Quem lembra de o **Post** #Vale5 armando uma linda venda de diário? PRON, VERB, ADP, DET, **NOUN**, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT
	FOI ! ! ! **OBJ CUMPRIDO** ! ! ! **VERB**, PUNCT, PUCNT, PUNCT, **NOUN**, **VERB,** PUNCT, PUNCT, PUNCT

Particularidades gráficas e PoS

NOUN com inicial maiúscula

Questão	Exemplo
Capitalização aleatória	$CTIP3 - **Cetip** (**ctip-nm**) Ago / e 28/04/2014 10h00 **Distribuicao De Dividendo** http://t.co/8E9tQwMP72 PROPN, PUNCT, **PROPN**, PUNCT, **PROPN**, PUNCT, PROPN, PUNCT, CCONJ, NUM, NOUN, ERRO_DIGITAÇÃO, **ADP, NOUN**, SYM
	$LREN3 - **Lojas Renner** (**lren-nm**) - **Declaracao E Pagamento De Juros Sobre O Capital Proprio** http://t.co/flXNxtpbT PROPN, PUNCT, **PROPN**, **PROPN**, PUNCT, **PROPN**, PUNCT, PUNCT, ERRO_DIGITAÇÃO**, CCONJ, NOUN, ADP, NOUN, ADP, DET, NOUN,** ERRO_DIGITAÇÃO, SYM
	05/03/14 - 17:20 : **Maiores Baixas** : BRAP4 - 4,50 % R$ 20,35 , RSID3 - 4,09 % NUM, PUNCT, NUM, PUNCT, **NOUN, ADJ**, PUNCT, PROPN, SYM, NUM, SYM, SYM, NUM, PUNCT, PROPN, SYM, NUM, SYM
	Em a #PETR4 fizemos em a sexta passada uma **Sub Onda** 5 de fundo e em o dia 05/02/14 uma **Onda** 3 de fundo alvos 15,44 16,40 http://t.co/5DPBQCulWr ADP, DET, PROPN, VERB, ADP, DET, NOUN, ADJ, DET, **ADJ, NOUN**, NUM, ADP, NOUN, CCONJ, ADP, DET, NOUN, NUM, DET, **NOUN**, NUM, ADP, NOUN, NOUN, NUM, NUM, SYM
	Remem para o Money : O melhor (e o pior) de o **Mercado Global** em o primeiro ... Petr4 , **vale5** , opções , **bovespa** , **ogxp3** http://t.co/EpwT6TzAmc **PROPN, ADP, DET, PROPN**, PUNCT, DET, NOUN, PUNCT, CCONJ, DET, NOUN, PUNCT, ADP, DET, **NOUN, ADJ**, ADP, DET, NOUN, PUNCT, PROPN, PUNCT, **PROPN**, PUNCT, NOUN, PUNCT, **PROPN**, PUNCT, **PROPN**, SYM
	Rastreamento ações - **Gráfico MENSAL** - 11h . NOUN, NOUN, PUNCT, **NOUN, ADJ**, PUNCT, NOUN, PUNCT
	Quem lembra de o **Post** #Vale5 armando uma linda venda de diário? PRON, VERB, ADP, DET, **NOUN**, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT
	FOI ! ! ! **OBJ CUMPRIDO** ! ! ! **VERB**, PUNCT, PUCNT, PUNCT, **NOUN**, **VERB,** PUNCT, PUNCT, PUNCT

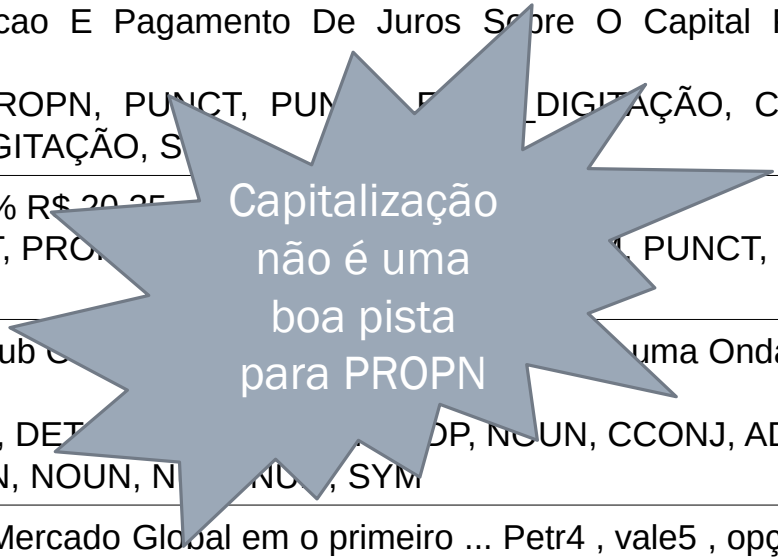
Particularidades gráficas e PoS

Caixa alta apenas como ênfase

Questão	Exemplo
Capitalização aleatória	$CTIP3 - **Cetip** (**ctip-nm**) Ago / e 28/04/2014 10h00 **Distribuicao De Dividendo** http://t.co/8E9tQwMP72 PROPN, PUNCT, **PROPN**, PUNCT, **PROPN**, PUNCT, PROPN, PUNCT, CCONJ, NUM, NOUN, ERRO_DIGITAÇÃO, **ADP, NOUN**, SYM
	$LREN3 - **Lojas Renner** (**lren-nm**) - **Declaracao E Pagamento De Juros Sobre O Capital Proprio** http://t.co/flXNxtpbT PROPN, PUNCT, **PROPN**, **PROPN**, PUNCT, **PROPN**, PUNCT, PUNCT, **ERRO_DIGITAÇÃO, CCONJ, NOUN, ADP, NOUN, ADP, DET, NOUN, ERRO_DIGITAÇÃO**, SYM
	05/03/14 - 17:20 : **Maiores Baixas** : BRAP4 - 4,50 % R$ 20,35 , RSID3 - 4,09 % NUM, PUNCT, NUM, PUNCT, **NOUN, ADJ**, PUNCT, PROPN, SYM, NUM, SYM, SYM, NUM, PUNCT, PROPN, SYM, NUM, SYM
	Em a #PETR4 fizemos em a sexta passada uma **Sub Onda** 5 de fundo e em o dia 05/02/14 uma **Onda** 3 de fundo alvos 15,44 16,40 http://t.co/5DPBQCulWr ADP, DET, PROPN, VERB, ADP, DET, NOUN, ADJ, DET, **ADJ, NOUN**, NUM, ADP, NOUN, CCONJ, ADP, DET, NOUN, NUM, DET, **NOUN**, NUM, ADP, NOUN, NOUN, NUM, NUM, SYM
	Remem para o Money : O melhor (e o pior) de o **Mercado Global** em o primeiro ... Petr4 , **vale5** , opções , **bovespa** , **ogxp3** http://t.co/EpwT6TzAmc **PROPN, ADP, DET, PROPN**, PUNCT, DET, NOUN, PUNCT, CCONJ, DET, NOUN, PUNCT, ADP, DET, **NOUN, ADJ**, ADP, DET, NOUN, PUNCT, PROPN, PUNCT, **PROPN**, PUNCT, NOUN, PUNCT, **PROPN**, PUNCT, **PROPN**, SYM
	Rastreamento ações - **Gráfico MENSAL** - 11h . NOUN, NOUN, PUNCT, **NOUN, ADJ**, PUNCT, NOUN, PUNCT
	Quem lembra de o **Post** #Vale5 armando uma linda venda de diário? PRON, VERB, ADP, DET, **NOUN**, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT
	FOI ! ! ! **OBJ CUMPRIDO** ! ! ! **VERB**, PUNCT, PUCNT, PUNCT, **NOUN**, **VERB**, PUNCT, PUNCT, PUNCT

Particularidades gráficas e PoS *tagging*

Questão	Exemplo
Capitalização aleatória	$CTIP3 - Cetip (ctip-nm) Ago / e 28/04/2014 10h00 Distribuicao De Dividendo http://t.co/8E9tQwMP72 PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT, CCONJ, NUM, NOUN, ERRO_DIGITAÇÃO, ADP, NOUN, SYM
	$LREN3 - Lojas Renner (lren-nm) - Declaracao E Pagamento De Juros Sobre O Capital Proprio http://t.co/flXNxtpbT PROPN, PUNCT, PROPN, PROPN, PUNCT, PROPN, PUNCT, PUN… …_DIGITAÇÃO, CCONJ, NOUN, ADP, NOUN, ADP, DET, NOUN, ERRO_DIGITAÇÃO, S…
	05/03/14 - 17:20 : Maiores Baixas : BRAP4 - 4,50 % R$ 20,35… NUM, PUNCT, NUM, PUNCT, NOUN, ADJ, PUNCT, PRO… …, PUNCT, PROPN, SYM, NUM, SYM
	Em a #PETR4 fizemos em a sexta passada uma Sub C… …uma Onda 3 de fundo alvos 15,44 16,40 http://t.co/5DPBQCulWr ADP, DET, PROPN, VERB, ADP, DET, NOUN, ADJ, DET… …DP, NOUN, CCONJ, ADP, DET, NOUN, NUM, DET, NOUN, NUM, ADP, NOUN, NOUN, N… …NUM, SYM
	Remem para o Money : O melhor (e o pior) de o Mercado Global em o primeiro ... Petr4 , vale5 , opções , bovespa , ogxp3 http://t.co/EpwT6TzAmc PROPN, ADP, DET, PROPN, PUNCT, DET, NOUN, PUNCT, CCONJ, DET, NOUN, PUNCT, ADP, DET, NOUN, ADJ, ADP, DET, NOUN, PUNCT, PROPN, PUNCT, PROPN, PUNCT, NOUN, PUNCT, PROPN, PUNCT, PROPN, SYM
	Rastreamento ações - Gráfico MENSAL - 11h . NOUN, NOUN, PUNCT, NOUN, ADJ, PUNCT, NOUN, PUNCT
	Quem lembra de o Post #Vale5 armando uma linda venda de diário? PRON, VERB, ADP, DET, NOUN, PROPN, VERB, DET, ADJ, NOUN, ADP, NOUN, PUNCT
	FOI ! ! ! OBJ CUMPRIDO ! ! ! VERB, PUNCT, PUCNT, PUNCT, NOUN, VERB, PUNCT, PUNCT, PUNCT



Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN, PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

@perfil

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN, PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN, PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

#assunto/tópico

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → **PROPN**

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN, PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN,** PROPN
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, PROPN
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo** **#EuApoioCPIdaPetrobras** SYM, **PROPN,** **PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! **PROPN**, PUNCT, **PROPN**, PUNCT, **PROPN**, PUNCT

$ticker

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN, PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

Algumas diretrizes de anotação de PoS

- Menções (@), hashtags (#) e cashtags ($) → PROPN

RT **@DenysonAnderson :** Petrobras confirma precificação de títulos de US$ 8,5 bilhões : SÃO PAULO , 10 Mar SYM, **PROPN**, PUNCT, PROPN, VERB, NOUN, ADP, NOUN, ADP, SYM, NUM, NOUN, PUNCT, PROPN, PROPN, PUNCT, NUM, NOUN
RT **@SouCalmo #EuApoioCPIdaPetrobras** SYM, **PROPN, PROPN**
Elet3 minha linda ! ! ! Hoje vou beber um **@vinho** pra comemorar ! ! ! PROPN, DET, NOUN, PUNCT, PUNCT, PUNCT, ADV, AUX, VERB, DET, **PROPN**, SCONJ, VERB, PUNCT, PUNCT, PUNCT
O tamanho de o iceberg que atingiu a **#Petrobras** é muito maior, esta refinaria é um cubo de gelo. **#PETR4** DET, NOUN, ADP, DET, NOUN, PRON, VERB, DET, **PROPN**, AUX, ADV, ADJ, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, NOUN, PUNCT, **PROPN**
#SortedoDia : ter perdido a alta de mais de 8 % em um único dia de a **$PETR4** **PROPN**, PUNCT, AUX, VERB, DET, NOUN, ADP, ADV, ADP, NUM, SYM, ADP, DET, ADJ, NOUN, ADP, DET, **PROPN**
#DEAL ! **#DEAL** ! **#DEAL** ! PROPN, PUNCT, PROPN, PUNCT, PROPN, PUNCT

Algumas diretrizes de anotação de PoS

- RT* e URL → SYM

RT @SouCalmo #EuApoioCPIdaPetrobras **SYM**, PROPN, PROPN
Confira a nova indicação agora em **http://t.co/kgt1YiTbF7** VERB, DET, ADJ, NOUN, ADV, ADP, **SYM** Vale diz que acesso de Valemax a a China é desejável , mas não necessário : Os navios , com capac ... **http://t.co/DUDYIuUL6A** #infomoney #vale5 PROPN, VERB, SCONJ, NOUN, ADP, PROPN, ADP, DET, PROPN, AUX, ADJ, PUNCT, CCONJ, ADV, ADJ, PUNCT, DET, NOUN, PUNCT, ADP, NOUN, PUNCT, **SYM**, PROPN, PROPN

Algumas diretrizes de anotação de PoS

- RT* e URL → SYM

Sanquinetti *et al.* (2020) propõem que URL com função sintática seja anotada com a PoS correspondente (p.ex.: NOUN)

RT @SouCalmo #EuApoioCPIdaPetrobras

SYM, PROPN, PROPN

Confira a nova indicação agora em **http://t.co/kgt1YiTbF7**

VERB, DET, ADJ, NOUN, ADV, ADP, **SYM**

Vale diz que acesso de Valemax a a China é desejável , mas não necessário : Os navios , com capac ... **http://t.co/DUDYIuUL6A** #infomoney #vale5

PROPN, VERB, SCONJ, NOUN, ADP, PROPN, ADP, DET, PROPN, AUX, ADJ, PUNCT, CCONJ, ADV, ADJ, PUNCT, DET, NOUN, PUNCT, ADP, NOUN, PUNCT, **SYM**, PROPN, PROPN

Algumas diretrizes de anotação de PoS

- Símbolo que indica variação da bolsa → **SYM**

A LIGHT S.A. fechou o dia de hoje ao preço de R$ 18,65 (-1,37%) (original)

A LIGHT S.A. fechou o dia de hoje a o preço de R$ 18,65 (**-** 1,37 %) (tokenizado)

DET, PROPN, PROPN, VERB, DET, NOUN, ADP, NOUN, ADP, DET, NOUN, ADP, SYM, NUM, PUNCT, **SYM**, NUM, SYM, PUNCT

Algumas diretrizes de anotação de PoS

- Emoticons e emojis → **SYM** ou PoS *tag* correspondente à função

Algo me diz que vou ver a #PETR4 em a casa de os 12 hoje ainda . vamos aguardar **:)** PRON, PRON, VERB, SCONJ, AUX, VERB, DET, PROPN, ADP, DET, NOUN, ADP, DET, NUM, ADV, ADV, PUNCT, AUX, VERB, **SYM**
Eu ♥ café ☕ PRON, **VERB**, NOUN, **SYM**

Diferentemente das URLs, entendemos que esses elementos já foram incorporados ao léxico dos *tweets*

Em andamento...

- Aperfeiçoamento da *tokenização* para o DANTE
- Treinamento do UD-Pipe para o DANTE
  - Conjunto dos 200 *tweets* anotados no treinamento dos anotadores

Próxima fase...

- Revisão efetiva da anotação PoS do UD-Pipe

  - Mesma metodologia de revisão do *corpus* Folha-Kaggle
  - Início em **maio** (término previsto para julho*)
  - Pacote semanal de **380** *tweets* para os 3 anotados
    - Todos revisam tudo
    - Ariani → adjudicação