

01.06.2012 – Defesa de Dissertação

# Etiquetagem de Micromensagens no Twitter: Uma Abordagem Linguística



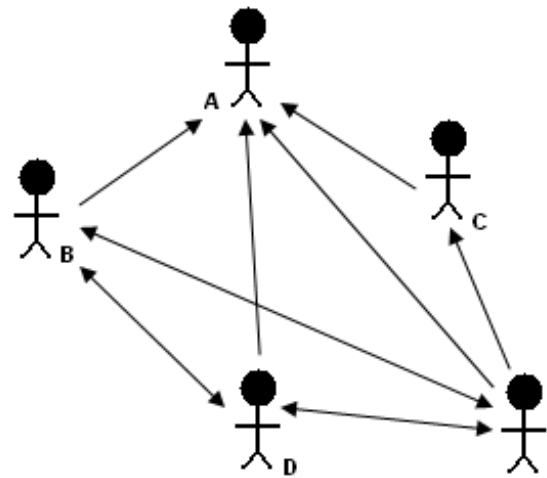
Evandro Landulfo Teixeira Paradela Cunha  
Mestrado – Programa de Pós-Graduação em Ciência da Computação  
Universidade Federal de Minas Gerais  
Orientador: Prof. Dr. Virgílio Augusto Fernandes Almeida

# #Estrutura da apresentação

- Introdução
  - Twitter e etiquetagem
  - Variação linguística
- O processo de etiquetagem textual
  - Folksonomias
  - Motivações para etiquetagem
- Apresentação e análise dos dados
  - Caracterização geral
  - Fatores linguísticos e sociais
- Conclusões

# #Introdução: **twitter**

- Rede social online



- Servidor de microblogging



**Série MPB & Jazz** @seriempbejazz

19 h

Gil @gilbertogil recebe convidados e se emociona após o concerto no Municipal| Série MPB & Jazz 2012

[seriempbejazz.com.br/2012/gilberto-...](http://seriempbejazz.com.br/2012/gilberto-...) via @seriempbejazz

 Retweetado por Gilberto Gil

# #Introdução: twitter



Buscar

Você possui uma conta? [Entrar](#)



**Gilberto Gil**   
**@gilbertogil**  
*Twitter atualizado pela equipe de produção do Gilberto Gil*  
Rio de Janeiro · <http://www.gilbertogil.com.br/>



**2.886** TWEETS  
**7** SEGUINDO  
**438.317** SEGUIDORES

**Siga Gilberto Gil**

Inscriva-se

**Tweets**

Seguindo

Seguidores

Favoritos

Listas

Imagens recentes

**Tweets**



**Gilberto Gil** @gilbertogil 3 h  
"Uma pessoa linda e formidável. Além d músico estelar, autor d composições lindas, estava sempre transmitindo alegria."  
[g1.globo.com/pop-arte/music...](http://g1.globo.com/pop-arte/music...)  
[Expandir](#)



**Gilberto Gil** @gilbertogil 5 h  
Nelson Jacobina in memoriam: "Herói das Estrelas" (Mautner/Jacobina) – [youtube.com/watch?v=-lyllo...](http://youtube.com/watch?v=-lyllo...)  
 Ver vídeo













**Série MPB & Jazz** @seriempbejazz 19 h  
Gil @gilbertogil recebe convidados e se emociona após o concerto no Municipal| Série MPB & Jazz 2012  
[seriempbejazz.com.br/2012/gilberto-...](http://seriempbejazz.com.br/2012/gilberto-...) via @seriempbejazz  
 Retweetado por Gilberto Gil  
[Expandir](#)



**Gilberto Gil** @gilbertogil 30 maio  
"Dia 120: Chegou o Momento do Inverto e Moré por Moço Cassinelli"

# #Introdução: twitter

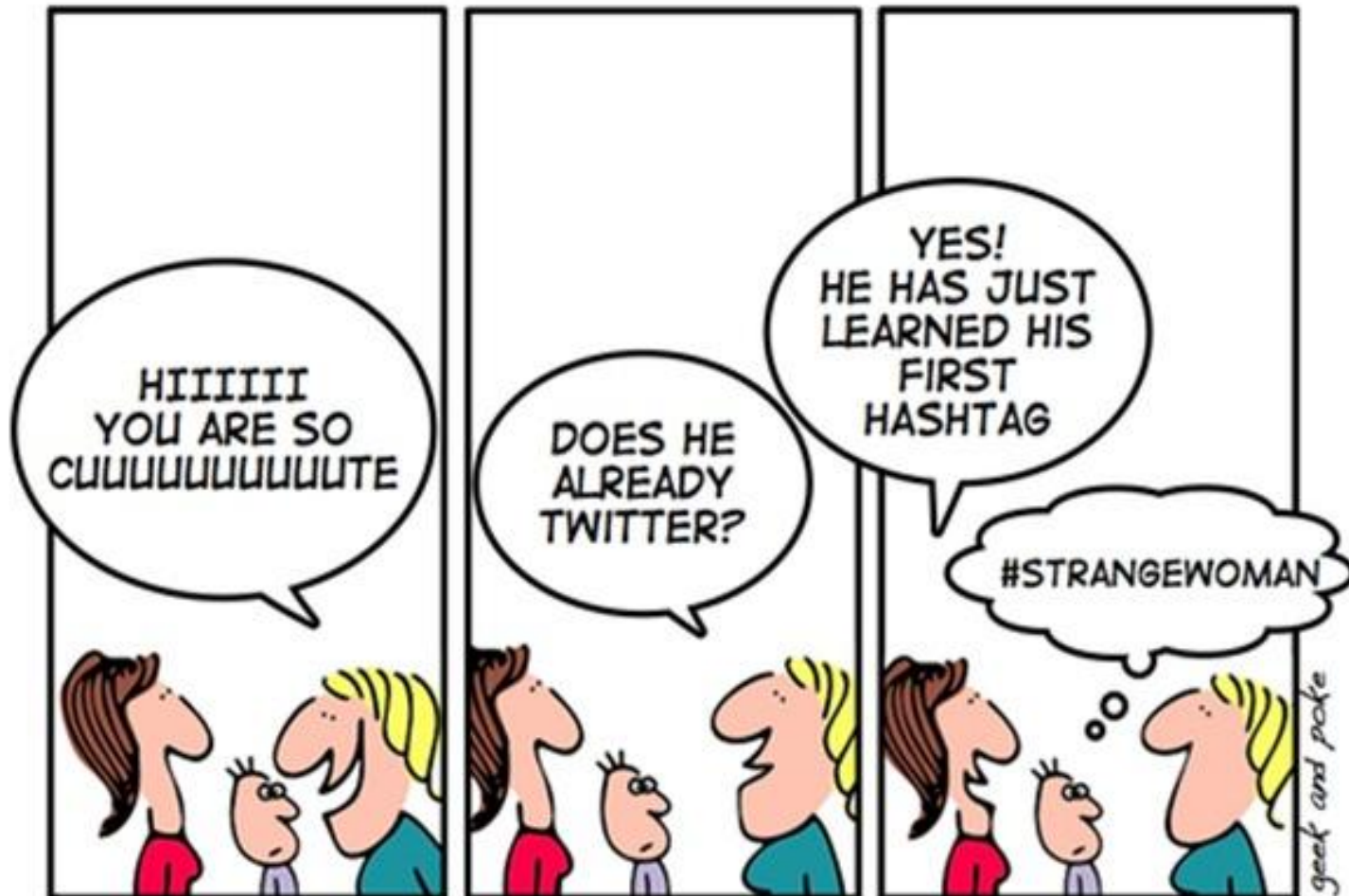
The following list of the **Most Popular Websites** was updated on **Thursday, May 31st 2012...**

1.	 Google.com	www.google.com
2.	 Facebook.com	www.facebook.com
3.	 Youtube.com	www.youtube.com
4.	 Yahoo.com	www.yahoo.com
5.	 Baidu.com	www.baidu.com
6.	 Wikipedia.org	www.wikipedia.org
7.	 Live.com	www.live.com
8.	 Twitter.com	www.twitter.com
9.	 Qq.com	www.qq.com
10.	 Amazon.com	www.amazon.com

<http://mostpopularwebsites.net/>

- Membros ativos = 140 milhões
- Tweets por dia = 340 milhões

# #Introdução: as hashtags



# #Introdução: as hashtags



**Nara Gil** @NaraGil

25 maio

A linda poetisa Belina autografando. #lançamento #livro @ Biblioteca Universitária de Saúde (BUS) UFBA [instagr.am/p/LENGd-qU2I/](https://www.instagram.com/p/LENGd-qU2I/)

 Retweetado por Gilberto Gil

# #Introdução: as hashtags

**Resultados para #lançamento**



**Tweets** Top / Todos



→ **show man** ← @Carekinhaa\_ 30 min  
@NayllaPop sim sim Uiiiiiiiiii adooooooo essa e #Lançamento só @deus sabe quem é aUAuahHUAhuaHUA  
Ver conversa



**Cia dos Descontos** @Ciadosdescontos 1 h  
Fifa Street 4 Edição Limitada Por R\$179,90 no Wal-Mart - [tinyurl.com/cen5u7r](http://tinyurl.com/cen5u7r) #Lançamento #Wal-Mart #FifaStreet  
Expandir



**MÓ CHAVÃO** @\_FrasesDeFuuunk 1 h  
Mc Luciano SP - Mestre do Bonde ( #Lançamento 2012) Clique e Tweet: [clicktotweet.com/an071](http://clicktotweet.com/an071)  
Expandir



**Nara Gil** @NaraGil 1 h  
A linda poetisa Belina autografando. #lançamento #livro @ Biblioteca Universitária de Saúde (BUS) UFBA [instagr.am/p/LENGd-qU2I/](http://instagr.am/p/LENGd-qU2I/)  
Ver foto

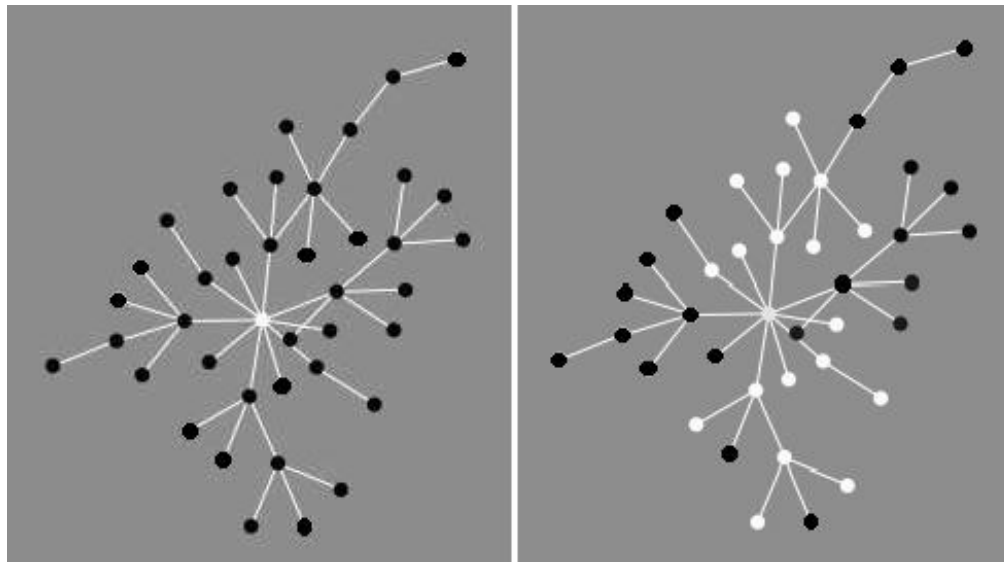


# #Introdução: as hashtags



# #Introdução: variação linguística

- Teoria da Variação e Mudança Linguística



- Fatores linguísticos e sociais

- *Ocê/Cê me viu?*
- *Eu vi ocê/\*cê.*

# #Introdução: o problema

- Há variação no uso de etiquetas no Twitter?
  - Quais são os fatores que condicionam essa variação?
- Por que os usuários etiquetam suas mensagens?

# #O processo de etiquetagem textual

---

- Taxonomia x Folksonomia
- Ambientes de livre etiquetagem

# #O processo de etiquetagem textual

---

- Motivações
  - Informações contextuais
  - Recuperação posterior
  - Organização
  - Compartilhamento

# #O processo de etiquetagem textual

- 400 questionários aplicados entre janeiro e abril de 2012
  - Informações demográficas (gênero e idade)
  - Frequência de postagem no Twitter
  - Frequência de uso de hashtags
  - Principal motivo para o uso de hashtags

# #O processo de etiquetagem textual

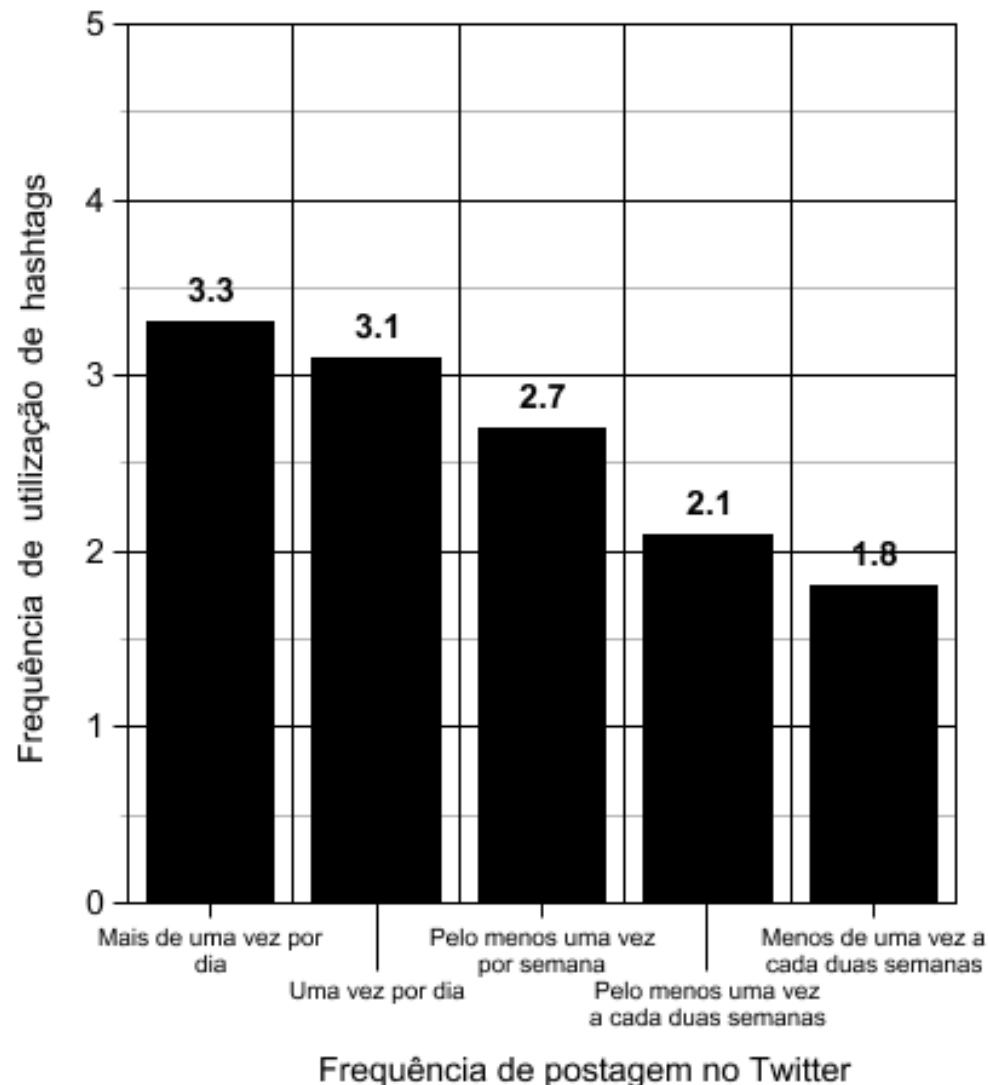
5 = em praticamente todos os tweets

4 = na maioria dos tweets

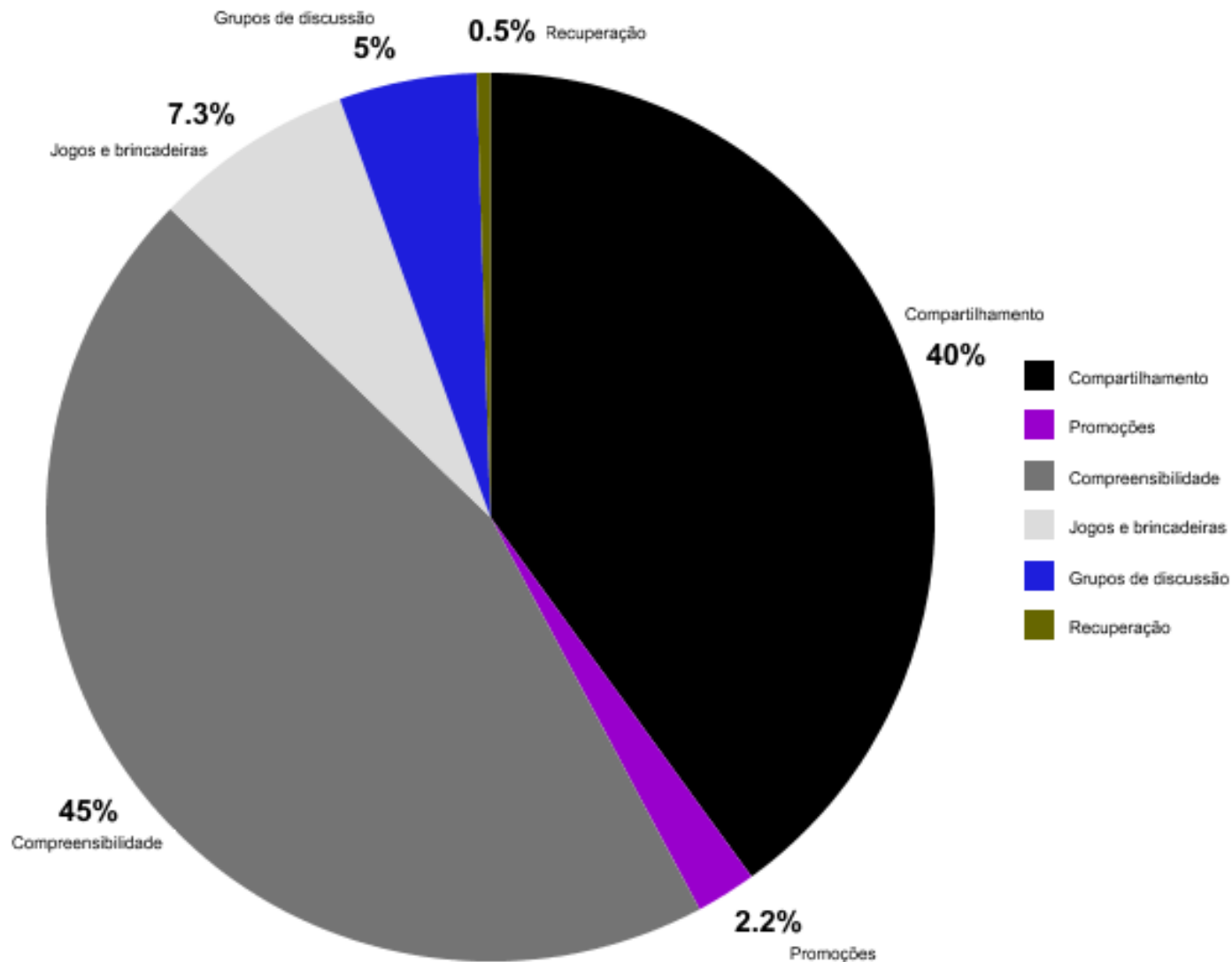
3 = em alguns tweets

2 = em poucos tweets

1 = nunca utilizou

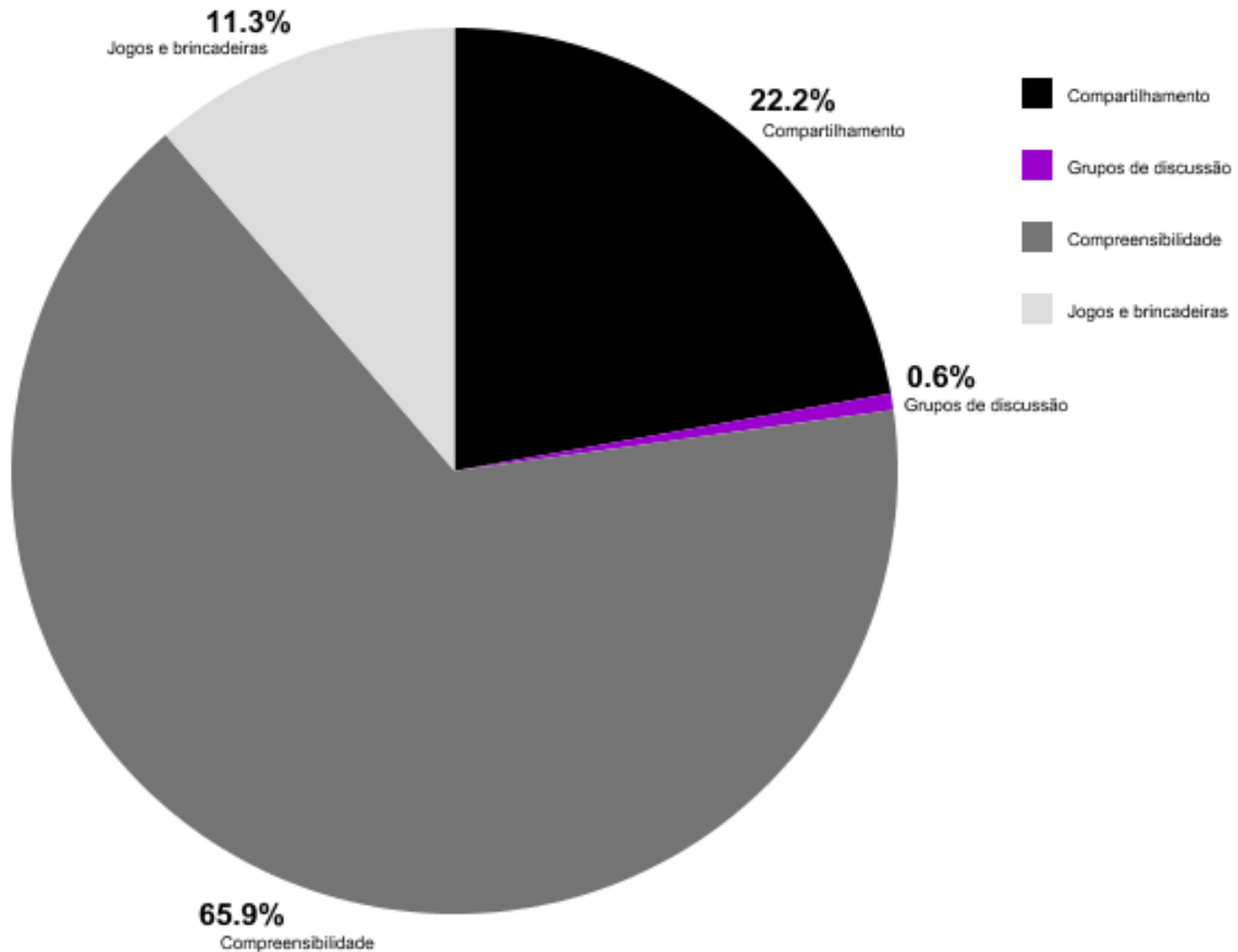


# #O processo de etiquetagem textual





# #O processo de etiquetagem textual



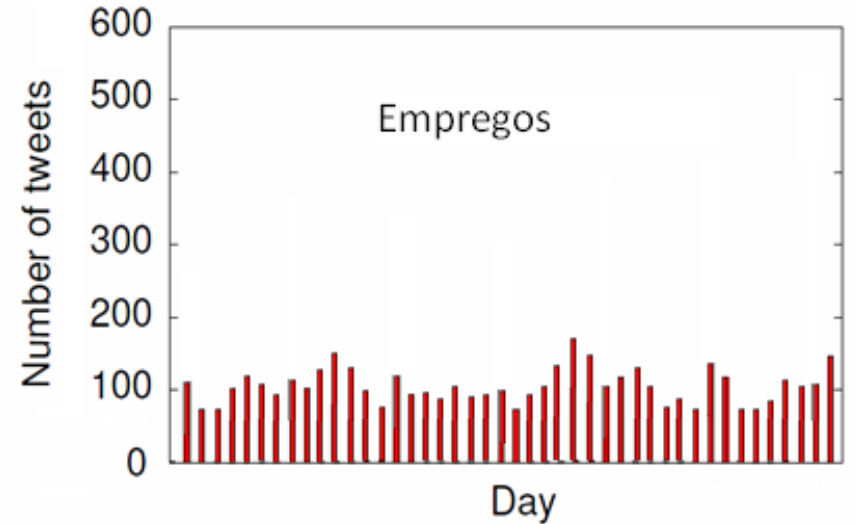
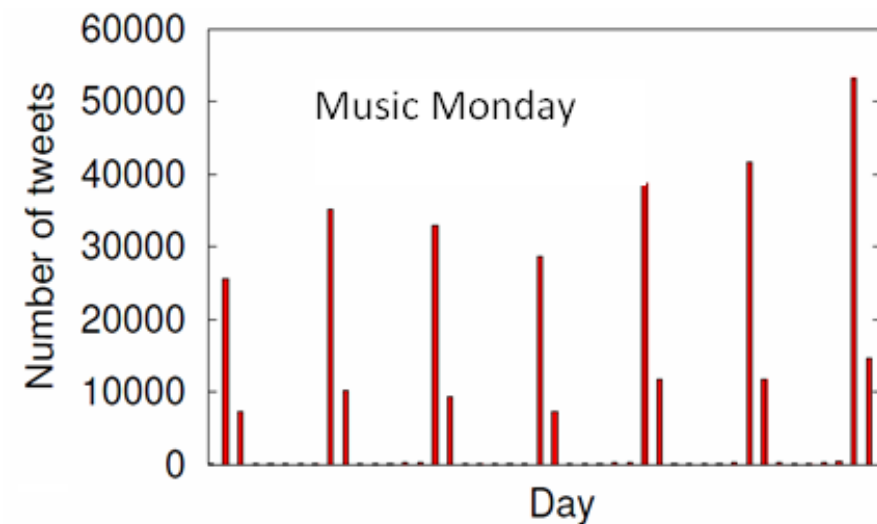
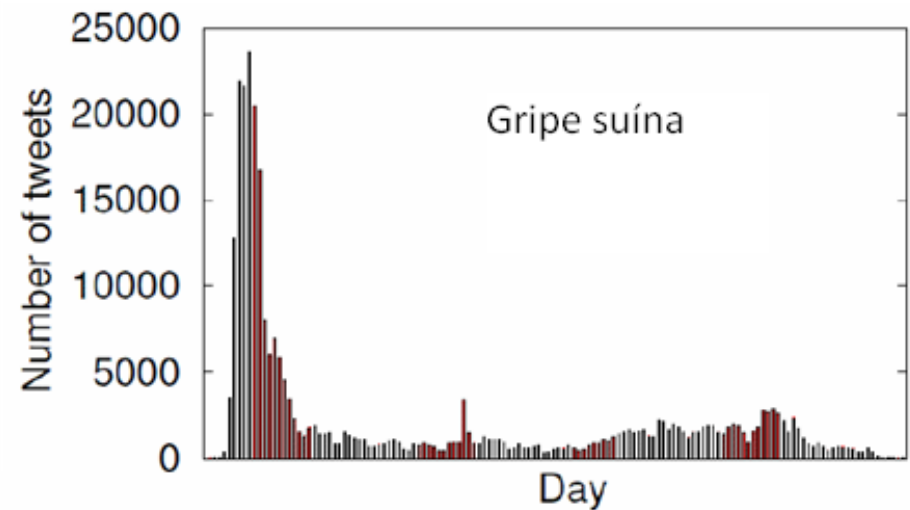
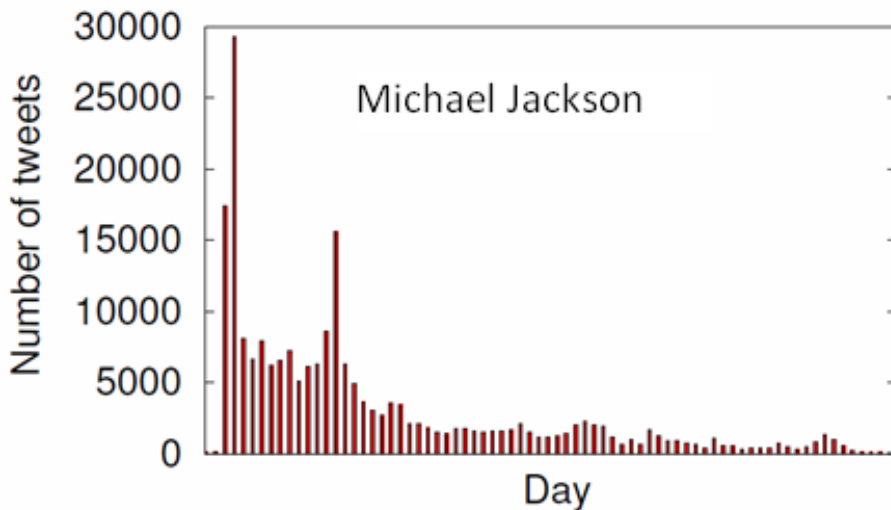
# #Análise dos dados: datasets

- 1.8 bilhão de tweets
- 55 milhões de usuários
- Entre julho/2006 e agosto/2009
- Tópicos: Michael Jackson, gripe suína e music Monday

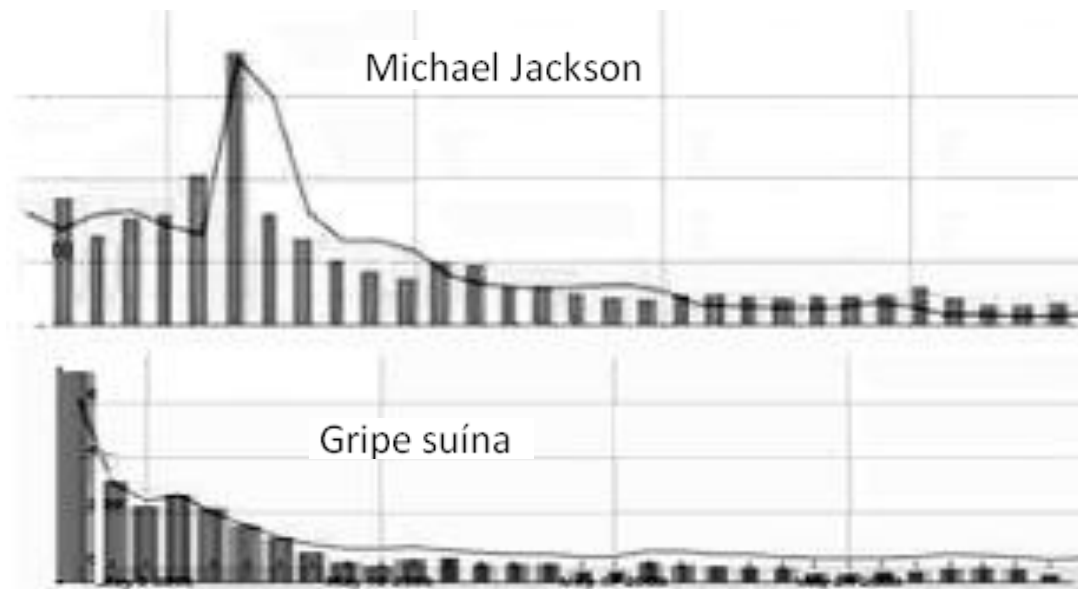
--

- 465 mil tweets
- 460 mil usuários
- Entre março e dezembro/2010
- Tópico: eleições 2010

# #Análise dos dados: caracterização

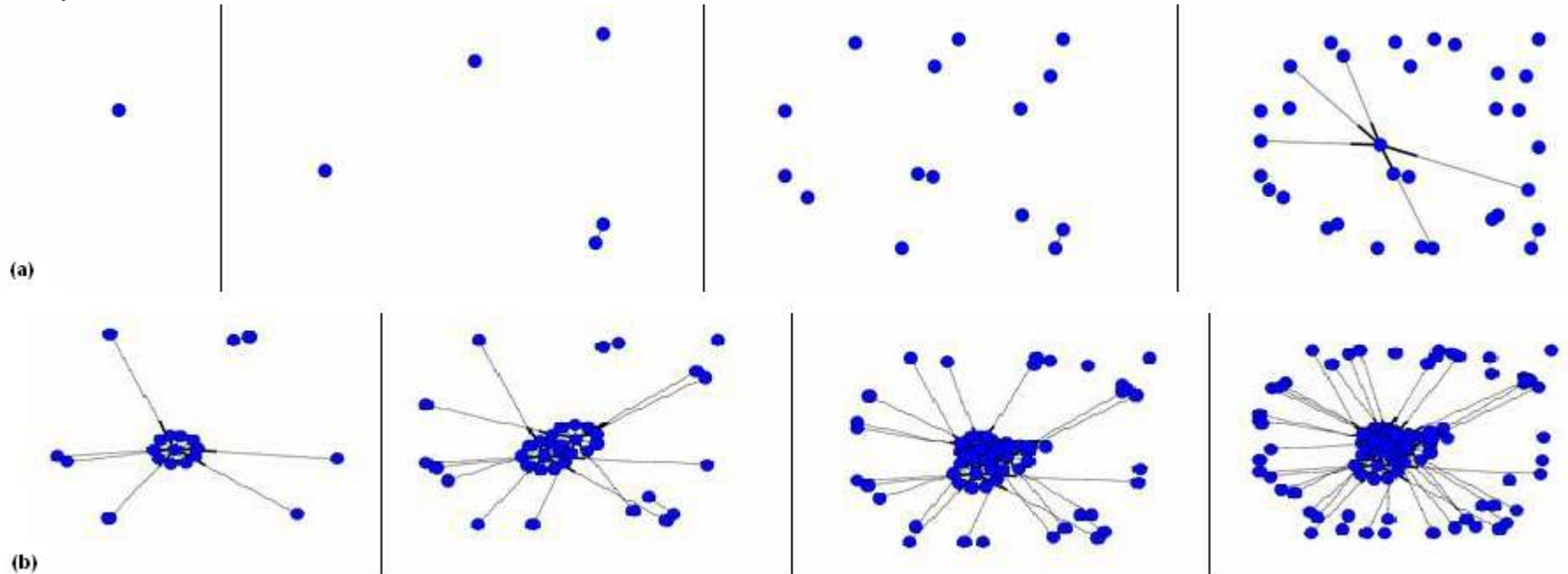


# #Análise dos dados: caracterização



# #Análise dos dados: caracterização

Gripe suína



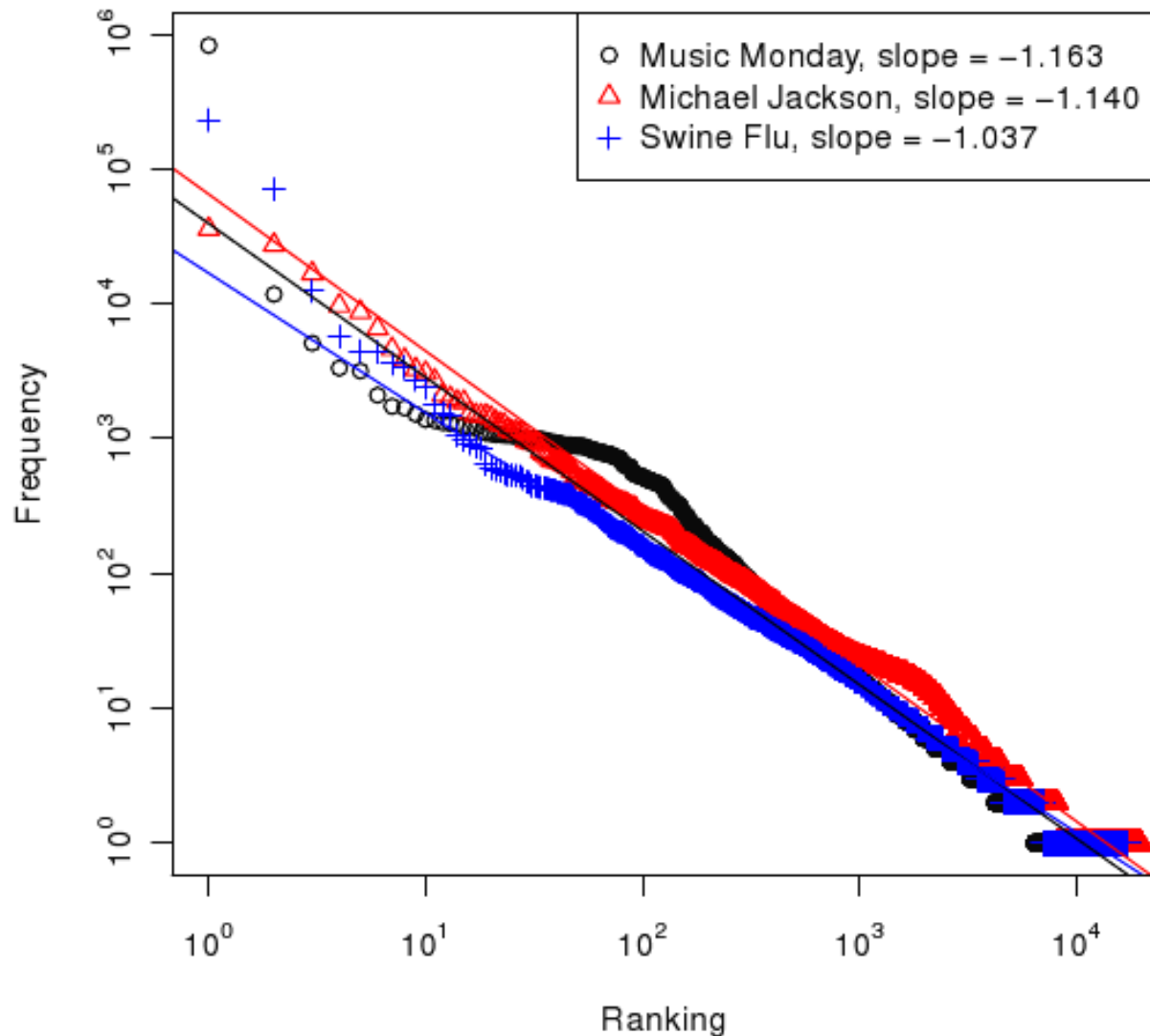
Music Monday

# #Análise dos dados: caracterização

Tópico	% de hashtags com $i$ utilizações		
	$i=1$	$i=2$	$i=10$
Michael Jackson	59%	72%	88%
Gripe Suína	59%	73%	92%
Music Monday	60%	74%	91%

Tópico	Número de hashtags com $j$ utilizações		
	$j=10.000$	$j=5.000$	$j=1.000$
Michael Jackson	3	6	28
Gripe Suína	3	4	14
Music Monday	2	3	28

# #Análise dos dados: caracterização



# #Análise dos dados

---

- Há variação no uso de etiquetas no Twitter?
  - Quais são os fatores que condicionam essa variação?



# #Análise dos dados: fatores linguísticos

---

- Comprimento
- Presença do sinal \_

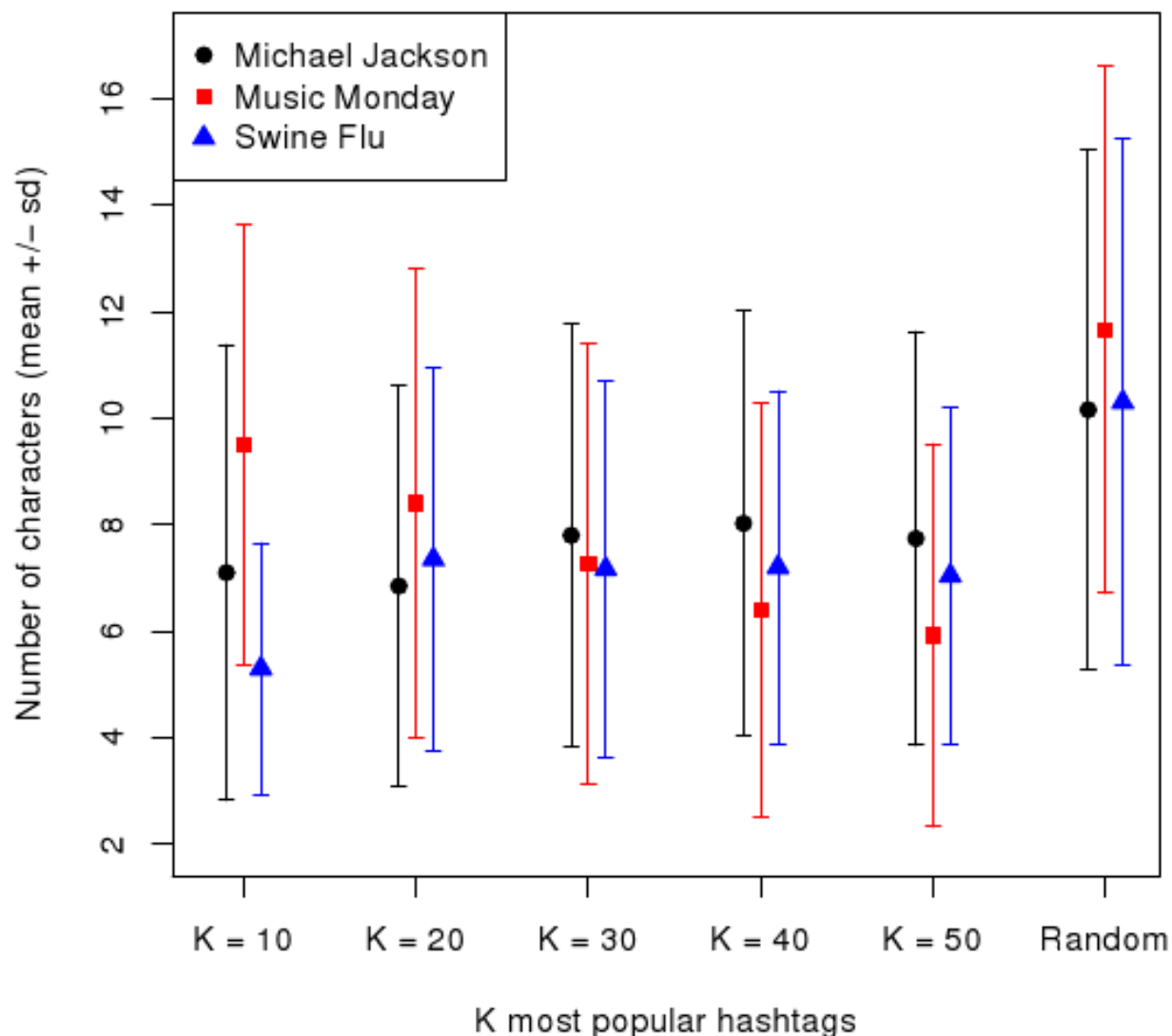
# #Análise dos dados: fatores linguísticos - comprimento

Hashtags mais comuns (número de tweets)	Hashtags mais comuns com 15 ou mais caracteres (número de tweets)
#michaeljackson (35.861) #michael (27.298) #mj (16.758)	#nothingpersonal (962) #iwillneverforget (912) #thankyoumichael (690)
#swineflu (230.457) #h1n1 (70.693) #swine (12.444)	#swinefluhatesyou (1.056) #crapnamesforpubs (145) #superhappyfunflu (124)
#musicmonday (824.778) #musicmondays (11.770) #music (5.106)	#musicmondayhttp (540) #fatpeoplearesexier (471) #crapurbanlegends (23)

# #Análise dos dados: fatores linguísticos - comprimento

Base	Comprimento médio das...					
	... $k$ hashtags mais populares					...hashtags menos populares
	$k=10$	$k=20$	$k=30$	$k=40$	$k=50$	
MJ	7,10	6,85	7,80	8,02	7,74	10,16
SF	5,30	7,35	7,17	7,20	7,04	10,30
MM	9,50	8,40	7,27	6,40	5,92	11,66

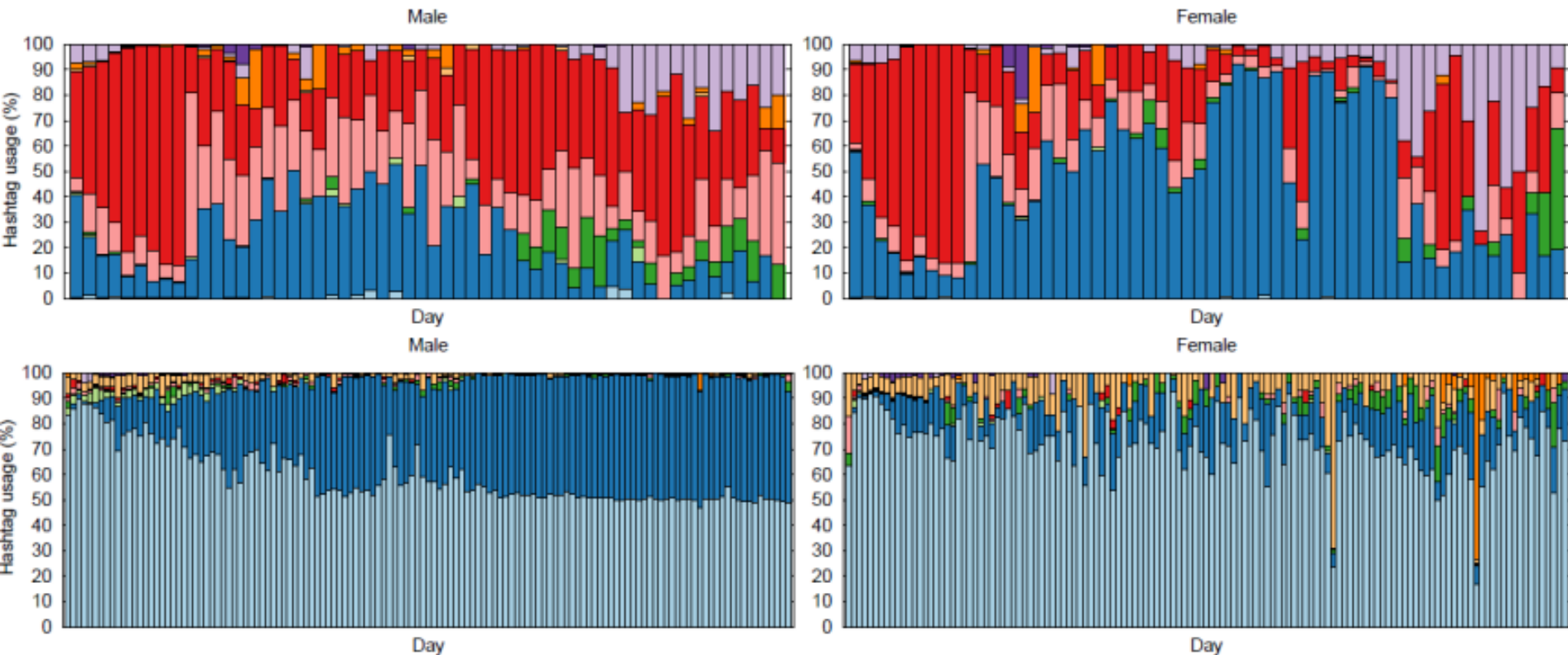
# #Análise dos dados: fatores linguísticos - comprimento



# #Análise dos dados: fatores linguísticos – sinal \_

Tópico	Número de hashtags contendo _	% de hashtags contendo o sinal _ entre as hashtags usadas $i$ vezes	
		$i=2$	$i=10$
Michael Jackson	251 (1,2%)	89%	97%
Gripe Suína	155 (0,9%)	87%	97%
Music Monday	143 (0,9%)	89%	98%

# #Análise dos dados: fator social – gênero



# #Análise dos dados: fator social – gênero

Hashtag fortemente feminina (HFF)	Hashtag feminina (HF)	Hashtag neutra (HN)	Hashtag masculina (HM)	Hashtag fortemente masculina (HFM)
$z > 1,96$	$1,96 \geq z > 1$	$1 \geq z \geq -1$	$-1 > z \geq -1,96$	$z < -1,96$

Dataset	HFF	HF	HN	HM	HFM
Total	5,4%	10,8%	68,8%	14,0%	1,1%
EB-1	0,0%	20,0%	70,0%	10,0%	0,0%
EB-2	0,0%	22,2%	55,6%	22,2%	0,0%
EB-3	10,0%	0,0%	80,0%	10,0%	0,0%
EB-4	7,4%	7,4%	70,4%	14,8%	0,0%
MJ	6,7%	6,7%	66,7%	20,0%	0,0%
SF	4,5%	13,6%	68,2%	9,1%	4,5%

# #Análise dos dados: fator social – gênero

- Formas padrão x Formas não-padrão

Tópico	Escores z	
	Formas padrão	Formas não-padrão
Eleições - apoiadores da Dilma	0,974	-0,145
Eleições - apoiadores do Serra	0,450	-0,215
Eleições - opositores da Dilma	1,024	-1,512
Eleições - opositores do Serra	0,885	0,031
Michael Jackson	1,467	-0,024
Gripe Suína	0,002	0,079



# #Análise dos dados: fator social – gênero

- Envolvimento pessoal x Persuasão

Tópico	Escores z	
	Envolvimento pessoal	Persuasão clara
Eleições - apoiadores da Dilma	0,601	-1,894
Eleições - apoiadores do Serra	1,477	-0,957

# #Conclusões

- Motivações para a etiquetagem no Twitter
- Caracterização das dinâmicas de utilização
- Processo de conexão preferencial
- Semelhanças qualitativas e quantitativas entre comunidades de fala online e offline
- Abordagem funcionalista da linguagem na Web