

Trabalho de Conclusão do Curso **Ciência da Computação**



MINERAÇÃO DE DADOS DA REDE SOCIAL *LINKEDIN* UTILIZANDO A LINGUAGEM DE PROGRAMAÇÃO PYTHON E BIBLIOTECAS PARA ANÁLISE E MINERAÇÃO DE DADOS

Acadêmico: **Thiago Medeiros de Souza**
Orientador: Valmei Abreu Júnior
Co-Orientador: João Paulo de Lima Barbosa

Foz do Iguaçu - PR, Dezembro de 2015

Agenda

- **Introdução**

Justificativa

Objetivos

- **Revisão Bibliográfica**

- **Fundamentação Teórica**

KDD

Data Mining

Python

Machine Learning

- **Materiais e Métodos**

Bibliotecas do Python

API

OAuth

LinkedIn

Etapas Mineração de Dados

- **Considerações Finais**

Introdução

Mineração e análise da rede social *LinkedIn*, para predizer um perfil de um profissional de TI, a verificação de fortes conexões e construção de um perfil mais influente.

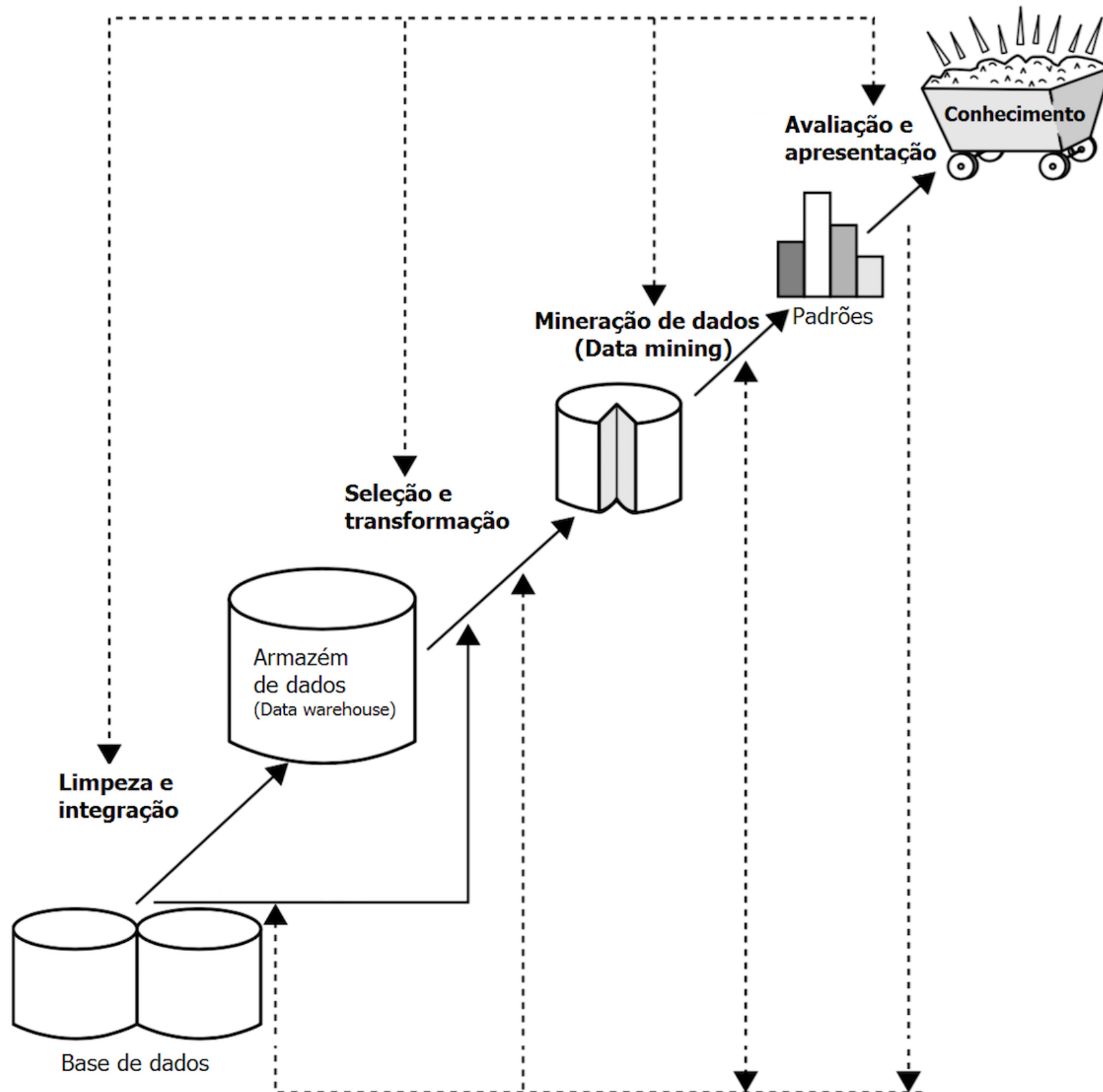
Introdução

- API do *LinkedIn*
- Python e bibliotecas para análise de dados
- Estudo e manipulação de APIs da rede social e métodos de autenticação
- Algoritmos para a mineração de dados
- Gráficos e planilhas
- Apresentação dos testes e resultados

Revisão Bibliográfica

- Análise de crédito bancário
- Classificação de pacientes em prováveis ictéricos com câncer ou cálculo
- Mineração de *logs* em páginas *web*

Fundamentação Teórica - KDD



Fundamentação Teórica - *Data Mining*

Data mining é o processo de extração de informações de algum conjunto de dados para tomada de decisões

- Classificação
- Modelos de Relacionamento entre Variáveis
- Análise de Agrupamento (*Cluster*)
- Sumarização
- Modelo de Dependência
- Regras de Associação
- Análise de Séries Temporais

Fundamentação Teórica - *Machine Learning*

- Aprendizado Supervisionado
 - Classificação
 - Regressão
- Aprendizado Não Supervisionado
 - Clustering*

Fundamentação Teórica - Python

- Sintaxe simples e clara
- Bibliotecas para a mineração de dados
- Solução e apenas uma única linguagem

Materiais e Métodos

Bibliotecas do Python - *pandas*

```
[In [5]: obj = Series([4, 7, -5, 3])
```

```
[In [6]: obj
```

```
Out[6]:
```

```
0      4
```

```
1      7
```

```
2     -5
```

```
3      3
```

```
dtype: int64
```

```
[In [9]: frame
```

```
Out[9]:
```

	pop	state	year
0	1.5	Ohio	2000
1	1.7	Ohio	2001
2	3.6	Ohio	2002
3	2.4	Nevada	2001
4	2.9	Nevada	2002

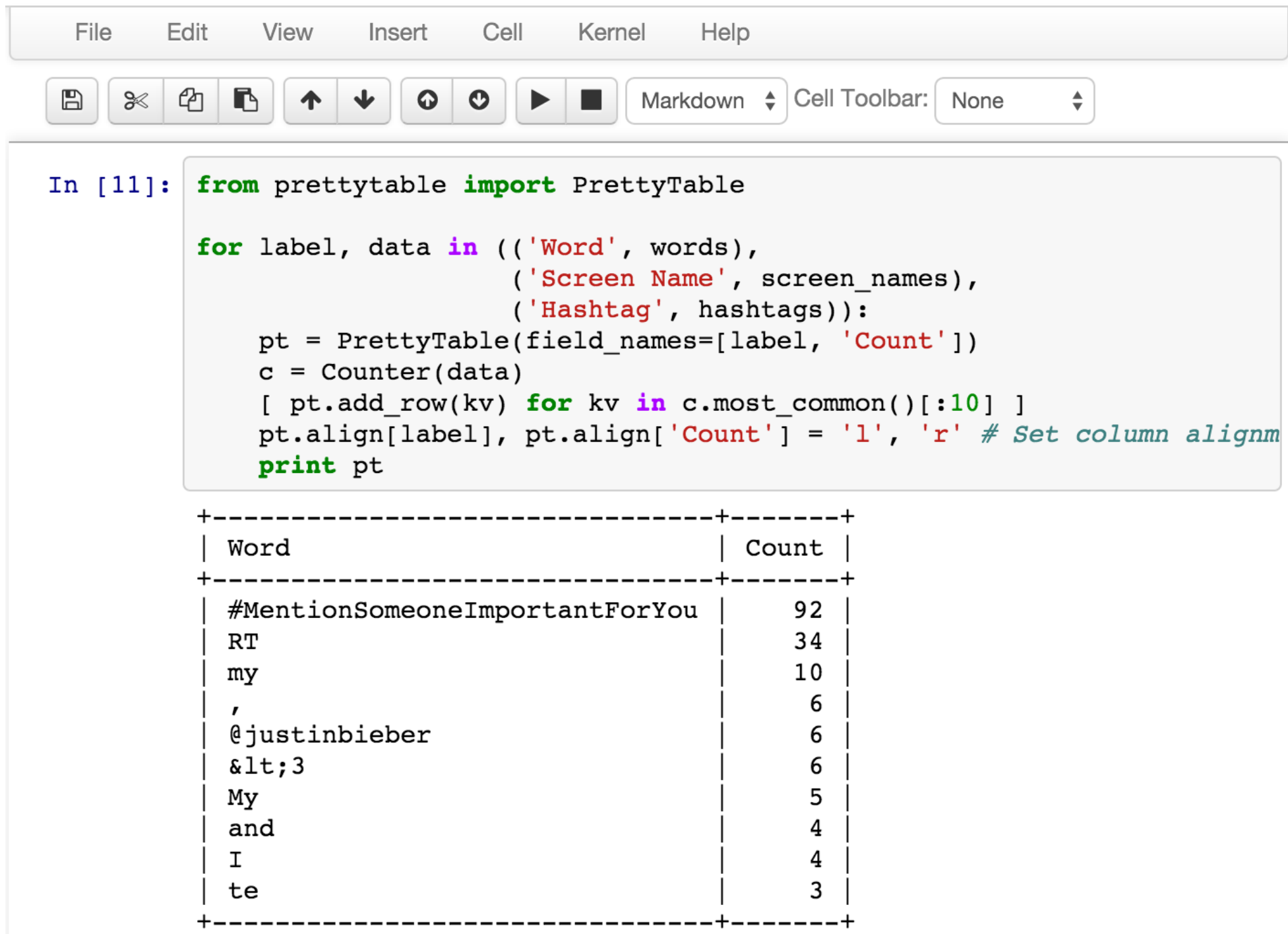
```
data = {'state': ['Ohio', 'Ohio', 'Ohio', 'Nevada', 'Nevada'],  
        'year': [2000, 2001, 2002, 2001, 2002],  
        'pop': [1.5, 1.7, 3.6, 2.4, 2.9]}
```

```
frame = DataFrame(data)
```

Materiais e Métodos

Bibliotecas do Python - *IPython*

IP[y]: Notebook Chapter 1 - Mining Twitter



The screenshot shows an IPython Notebook interface. At the top is a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', and 'Help'. Below the menu bar is a toolbar with icons for saving, cutting, copying, pasting, undo, redo, running, and a cell toolbar dropdown set to 'None'. The main area contains a code cell labeled 'In [11]:' with the following Python code:

```
from prettytable import PrettyTable

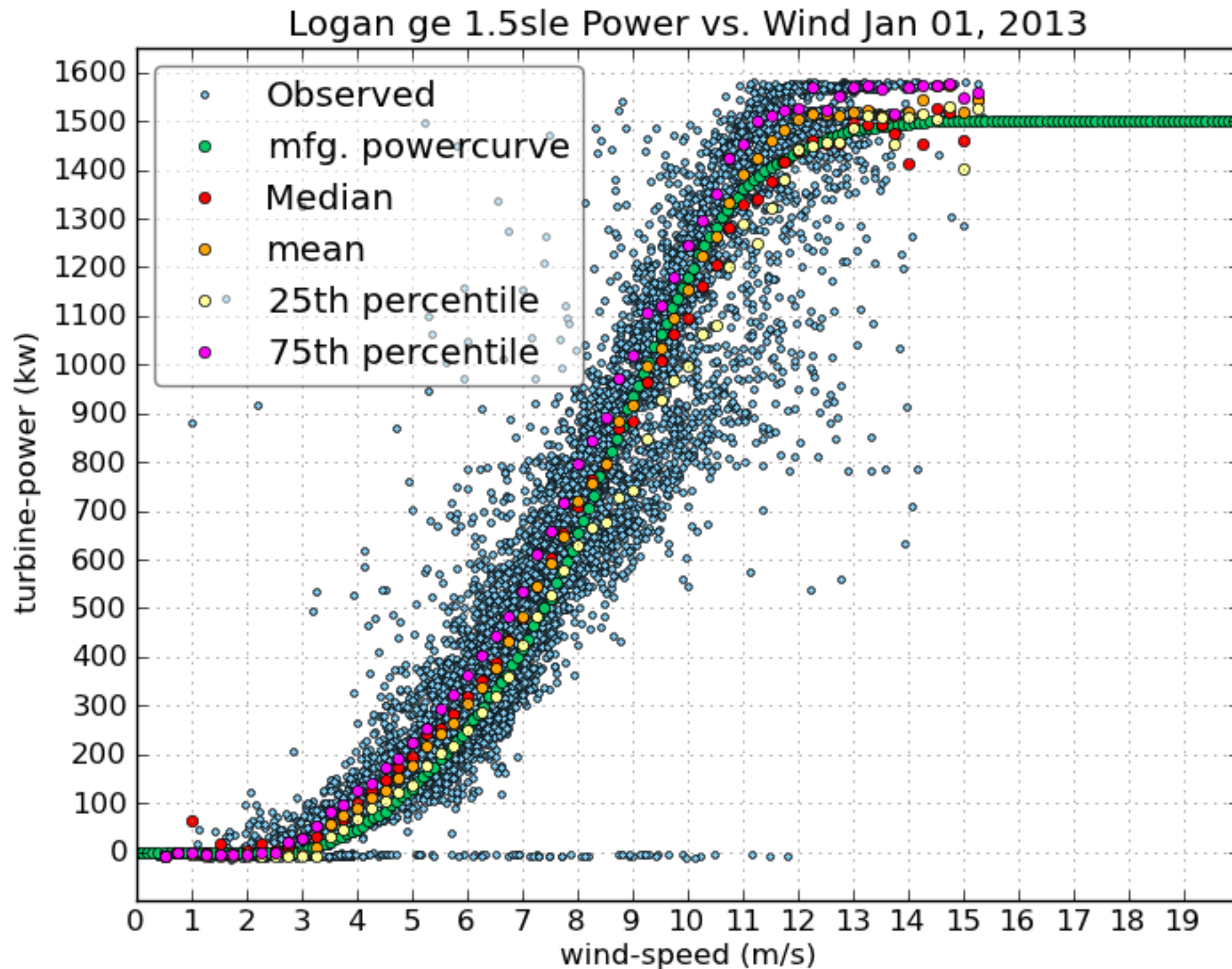
for label, data in (('Word', words),
                   ('Screen Name', screen_names),
                   ('Hashtag', hashtags)):
    pt = PrettyTable(field_names=[label, 'Count'])
    c = Counter(data)
    [ pt.add_row(kv) for kv in c.most_common()[:10] ]
    pt.align[label], pt.align['Count'] = 'l', 'r' # Set column alignment
    print pt
```

The output of the code is a PrettyTable with two columns: 'Word' and 'Count'. The table is formatted with dashed lines and the 'Count' column is right-aligned.

Word	Count
#MentionSomeoneImportantForYou	92
RT	34
my	10
,	6
@justinbieber	6
<3	6
My	5
and	4
I	4
te	3

Materiais e Métodos

Bibliotecas do Python - *matplotlib*



Materiais e Métodos

Bibliotecas do Python

- *NumPy* - criação e manipulação de matrizes
- *SciPy* - coleção de pacotes para computação científica
- *python-linkedin* - provê a API do *LinkedIn*

Materiais e Métodos

- API - Interface de Programação para Aplicações
 - REST - Transferência de Estado Representacional
- OAuth
 - OAuth 1.0a
 - OAuth 2.0
- *LinkedIn*

Materiais e Métodos

Etapas para a mineração de dados do *LinkedIn*

- *Clustering*
- Normalização dos dados
- Computação de similaridade

Considerações Finais

- Bibliotecas do Python
- API do *LinkedIn*
- Processos de *data mining*

Referências

HAN, J. *et al.* **Data Mining: Concepts and Techniques.** Elsevier, 2012.

HARRINGTON, P. **Machine Learning in Action.** Manning Publications, 2012. ISBN 9781617290183

KALDERO, N. **Why is Python a language of choice for data scientists?** 2015. Acesso em 27 de outubro de 2015. Disponível em: <<http://qr.ae/RkleiB>>.

MCKINNEY, W. **Python for Data Analysis.** O'Reilly, 2013

RUSSEL, M. A. **Mining the Social Web.** O'Reilly Media, 2013. ISBN 978-1-449-36761-9.