Crab
A Python Framework for Building Recommendation Engines
Scipy 2011, Austin TX

Marcel Caraciolo  @marcelcaraciolo
Ricardo Caspirro  @ricardocaspirro
Bruno Melo  @brunomelo
What is **Crab**?

A **python** framework for building recommendation engines

A **Scikit** module for collaborative, content and hybrid filtering

Mahout Alternative for Python Developers :D

Open-Source under the BSD license

https://github.com/muricoca/crab
When **started**?

It began **one year ago**

Community-driven, **4 members**

Since **April, 2011** the open-source labs Muriçoca incorporated it

Since **April, 2011** rewritting it as **Scikit**

https://github.com/muricoca/
Knowing **Scikits**

Scikits are **Scipy** Toolkits - independent and projects hosted under a common namespace.

Scikits Image
Scikits MlabWrap
Scikits AudioLab
**Scikit Learn**

....

http://scikits.appspot.com/scikits
Knowing **Scikits**

**Scikit-Learn**

Machine Learning Algorithms + scientific Python packages (Numpy, Scipy and Matplotlib)

http://scikit-learn.sourceforge.net/

**Our goal:** Incorporate the Crab as Scikit and incorporate some parts of them at Scikit-learn
Why **Recommendations**?

The world is an over-crowded place
Why Recommendations?

We are overloaded!

Thousands of news articles and blog posts each day

Millions of movies, books and music tracks online

Several Places, Offers and Events

Even Friends sometimes we are overloaded!
Why **Recommendations**?

We really need and consume only a few of them!

“A lot of times, people don’t **know** what they want until you **show** it to them.”

Steve Jobs

“We are leaving the **Information** age, and entering into the **Recommendation** age.”

Chris Anderson, from book Long Tail
Why **Recommendations**?

Can **Google** help?

Yes, but only when we really know what we are looking for.

But, what’s does it mean by “interesting”?

Can **Facebook** help?

Yes, I tend to find my friends’ stuffs interesting.

What if i had only few friends and what they like do not always attract me?

Can **experts** help?

Yes, but it won’t scale well.

But it is what they like, not me! Exactly same advice!
Why Recommendations?

Recommendation Systems

Systems designed to recommend to me something I may like.
Why Recommendations?

Recommendation Systems

Graph Representation
The current Crab

- Collaborative Filtering algorithms
  - User-Based, Item-Based and Slope One

- Evaluation of the Recommender Algorithms
  - Precision, Recall, F1-Score, RMSE

Precision-Recall Charts
The current **Crab**

Precision-Recall Charts
The current Crab

```python
from crab.models.datamodel import *
from crab.recommender.topmatches import *
from crab.similarities.similarity import UserSimilarity, ItemSimilarity
from crab.similarities.similarity_distance import *
from crab.neighborhood.itemstrategies import PreferredItemsNeighborhoodStrategy
from crab.recommender.recommender import ItemRecommender

movies={'Marcel Caraciolo': {'Lady in the Water': 2.5, 'Snakes on a Plane': 3.5,
                       'Just My Luck': 3.0, 'Superman Returns': 3.5, 'You, Me and Dupree': 2.5,
                       'The Night Listener': 3.0},
       'Luciana Nunes': {'Lady in the Water': 3.0, 'Snakes on a Plane': 3.5,
                         'Just My Luck': 1.5, 'Superman Returns': 5.0, 'The Night Listener': 3.0,
                         'You, Me and Dupree': 3.5},
       'Leopoldo Pires': {'Lady in the Water': 2.5, 'Snakes on a Plane': 3.0,
                          'Superman Returns': 3.5, 'The Night Listener': 4.0},
       'Lorena Abreu': {'Snakes on a Plane': 3.5, 'Just My Luck': 3.0,
                        'The Night Listener': 4.5, 'Superman Returns': 4.0,
                        'You, Me and Dupree': 2.5},
       'Steve Gates': {'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0,
                        'Just My Luck': 2.0, 'Superman Returns': 3.0, 'The Night Listener': 3.0,
                        'You, Me and Dupree': 2.0},
       'Sheldon': {'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0,
                   'The Night Listener': 3.0, 'Superman Returns': 5.0, 'You, Me and Dupree': 3.5},
       'Penny Frewman': {'Snakes on a Plane': 4.5, 'You, Me and Dupree': 1.0, 'Superman Returns': 4.0},
       'Maria Gabriela': {}}

model = DictDataModel(movies)
similarity = ItemSimilarity(model, sim_euclidian)
strategy = PreferredItemsNeighborhoodStrategy()

UserID = 'Leopoldo Pires'
recSys = ItemRecommender(model, similarity, strategy, False)
print recSys.recommend(UserID, 4)
```
The current **Crab**

**Using REST APIs to deploy the recommender**

django-piston, django-rest, django-tastypie
Crab is already in production

Brazilian Social Network called Atepassar.com

Educational network with more than **60,000 students** and **3000 video-classes**

Running on Python + Numpy + Scipy and Django

Backend for Recommendations

**MongoDB** - mongoengine

**Daily** Recommendations with **Explanations**
Evaluating your recommender

Crab implements the most used recommender metrics. Precision, Recall, F1-Score, RMSE

Using `matplotlib` for a plotter utility

**Implement** new metrics

**Simulations** support maybe (??)
Evaluating your recommender

All you have to do is implement your **Evaluator**

```python
from evaluators.statistics
    import IRStatsRecommenderEvaluator

# initialize the recommender
# initialize the IR Evaluator
evaluator = IRStatsRecommenderEvaluator()

# call evaluate considering the top 4 items recommended.
print evaluator.evaluate(recommender, model, 2, 1.0)
# {'precision': 0.75, 'recall': 1.0,
#   'f1Score': 0.6777}
```
Distributing the recommendation computations

Use Hadoop and Map-Reduce intensively

Investigating the Yelp mrjob framework  https://github.com/pfig/mrjob

Develop the Netflix and novel standard-of-the-art used

Matrix Factorization, Singular Value Decomposition (SVD), Boltzmann machines

The most commonly used is Slope One technique.
Why migrate?

Old Crab running only using Pure Python

Recommendations demand heavy maths calculations and lots of processing

Compatible with Numpy and Scipy libraries

High Standard and popular scientific libraries optimized for scientific calculations in Python

Scikits projects are amazing!

Active Communities, Scientific Conferences and updated projects (e.g. scikit-learn)

Turn the Crab framework visible for the community

Join the scientific researchers and machine learning developers around the Globe coding with Python to help us in this project

Be Fast and Furious
Why migrate?

Numpy optimized with PyPy

2.x - 48.x Faster

How are we **working**?

Sprints, Online Discussions and Issues

---

**Sprint planning: 21 April 2011 (02:00PM GMT -3:00)**

---

**Places**

- In Recife/PE - Brazil: Send an email to caraciolo@gmail.com for further information about the place.
- On IRC (#muricoca on irc.freenode.net)

---

**People present**

- Marcel Caraciolo
- Bruno Melo

---

**Tasks**

- **Study the Architecture of Scikit-learn**: Understand the architecture, run examples and discuss about its architecture and how crab can be implemented in the framework. Decisions about the next steps will be also discussed in this sprint. Issues will be also created for the next sprint.
- **Improve the Wiki and README**: Explain better the README about the project and How-To-Start page for developers.
- **New WebSite**: Talk with Rodrigo to design and port the new github homepage for muricoca-crab project.

---

**Past sprints**

Future Releases

✓ Planned Release 0.1
   Collaborative Filtering Algorithms working, sample datasets to load and test

✓ Planned Release 0.11
   Evaluation of Recommendation Algorithms and Database Models support

✓ Planned Release 0.12
   Recommendation as Services with REST APIs

....
Join **us**!

1. **Read our Wiki Page**
   
   https://github.com/muricoca/crab/wiki/Developer-Resources

2. **Check out our current sprints and open issues**
   
   https://github.com/muricoca/crab/issues

3. **Forks, Pull Requests mandatory**

4. **Join us at irc.freenode.net #muricoca or at our discussion list**
   
   http://groups.google.com/group/scikit-crab
Crab
A Python Framework for Building Recommendation Engines

https://github.com/muricoca/crab

Marcel Caraciolo @marcelcaraciolo
Ricardo Caspirro @ricardocaspirro
Bruno Melo @brunomelo

{marcel, ricardo, bruno}@muricoca.com