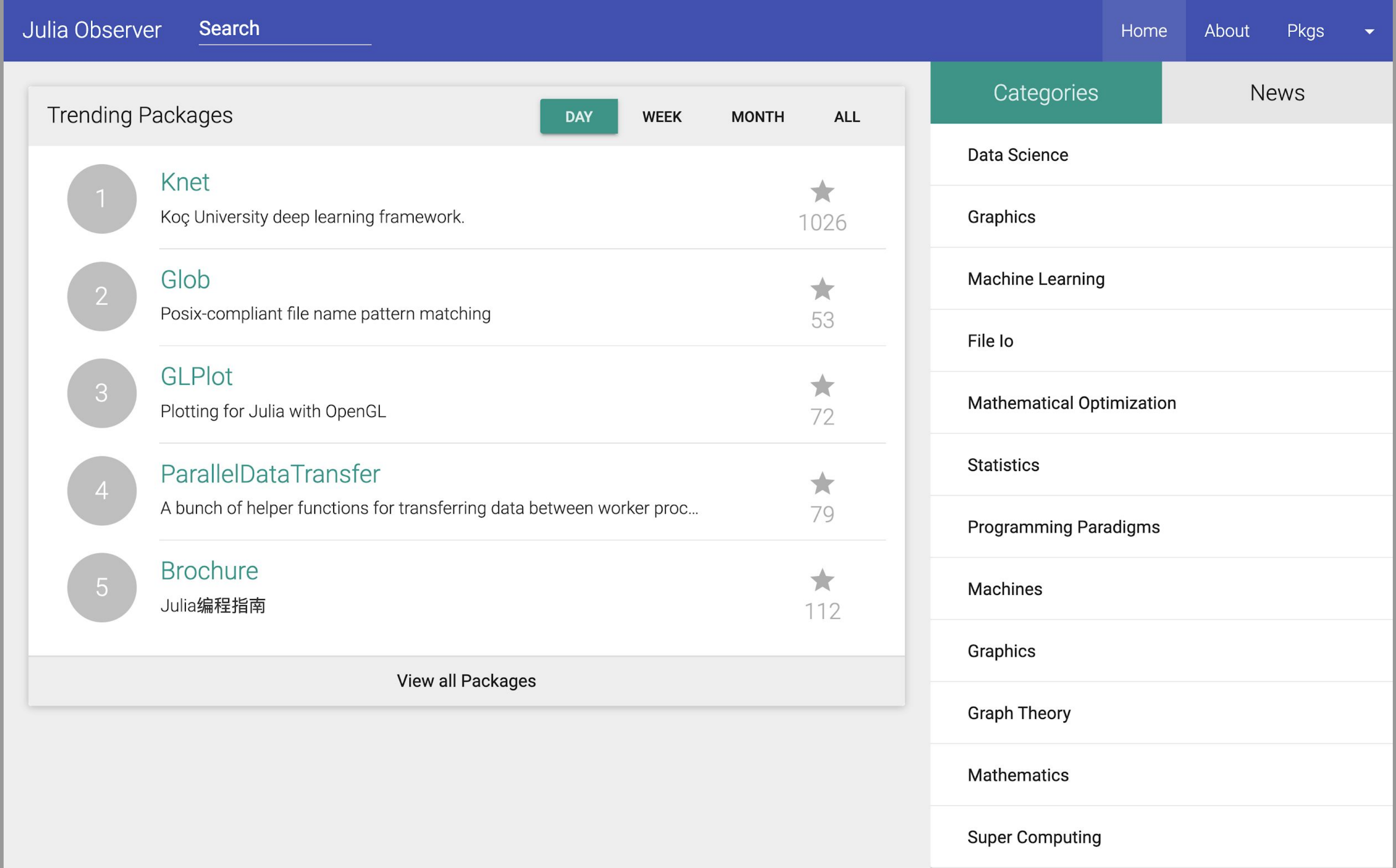
Categorizing Julia Packages using Semi-Supervised Learning

Dan Segal
@djsegal

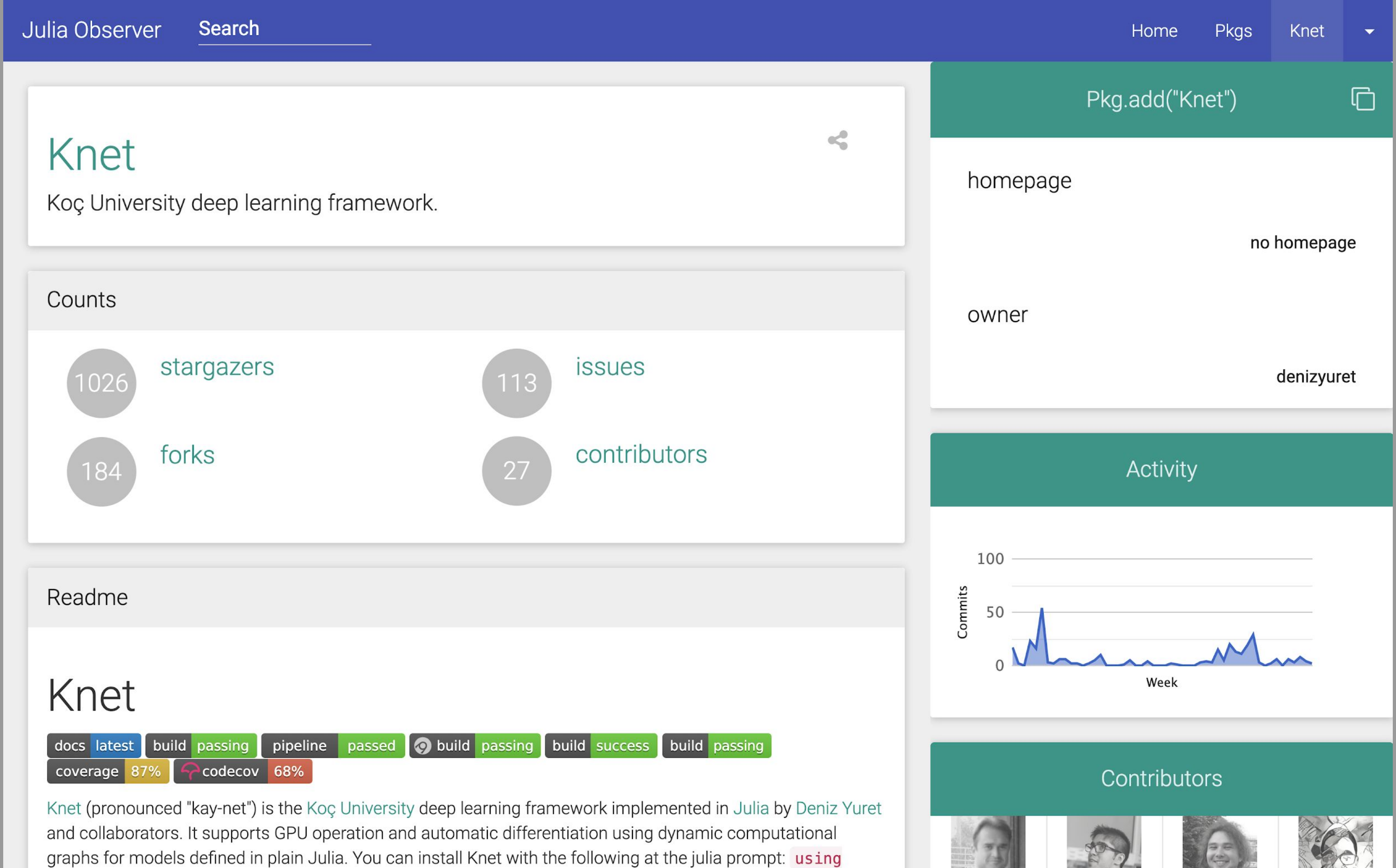
Julia Observer is a website for finding julia packages



Trending Packages

	Package	Description	★
1	Knet	Koç University deep learning framework.	1026
2	Glob	Posix-compliant file name pattern matching	53
3	GLPlot	Plotting for Julia with OpenGL	72
4	ParallelDataTransfer	A bunch of helper functions for transferring data between worker proc...	79
5	Brochure	Julia编程指南	112

- Data Science
- Graphics
- Machine Learning
- File Io
- Mathematical Optimization
- Statistics
- Programming Paradigms
- Machines
- Graphics
- Graph Theory
- Mathematics
- Super Computing



Knet

Koç University deep learning framework.

Counts

1026	stargazers	113	issues
184	forks	27	contributors

Readme

Knet

Knet (pronounced "kay-net") is the Koç University deep learning framework implemented in Julia by Deniz Yuret and collaborators. It supports GPU operation and automatic differentiation using dynamic computational graphs for models defined in plain Julia. You can install Knet with the following at the julia prompt: `using`

```
Pkg.add("Knet")
```

homepage

no homepage

owner

denizyuret

Activity

Contributors

P: Currently, the categories come from svaksha's Julia.jl database

S: Our goal is to build a set of workers that add more packages

This is a 4 step semi-supervised learning problem

I. Collect two databases

- **svaksha/Julia.jl**

Name	Readme	Category
JuMP.jl	...	Optimization
Plots.jl	...	Graphics
PyCall.jl	...	API
ScikitLearn.jl	...	ML

- **JuliaRegistries/General**

Name	Readme	Category
Reddit.jl	...	?
Seaborn.jl	...	?
TimeSeries.jl	...	?
Tokenizers.jl	...	?

II. Combine data sets and cluster them

Name	Readme (TF-IDF)	Category	Cluster
JuMP.jl	...	Optimization	7
Plots.jl	...	Graphics	**– 3 –**
...	...	...	...
Reddit.jl	...	?	24
ScikitLearn.jl	...	ML	1
Seaborn.jl	...	?	**– 3 –**
TimeSeries.jl	...	?	11

III. Train classifier on categorized rows

Name	Readme (TF-IDF)	Category	Cluster
JuMP.jl	...	Optimization	7
Plots.jl	...	Graphics	**– 3 –**
...	...	...	...
Reddit.jl	...	?	24
ScikitLearn.jl	...	ML	1
Seaborn.jl	...	?	**– 3 –**
TimeSeries.jl	...	?	11

IV. Label only most certain uncategorized packages

Name	Readme (TF-IDF)	Category	Cluster
JuMP.jl	...	Optimization	7
Plots.jl	...	Graphics	– 3 –
...	...	...	...
Reddit.jl	...	API	24
ScikitLearn.jl	...	ML	1
Seaborn.jl	...	Graphics	– 3 –
TimeSeries.jl	...	**– ? –**	11

Using classification on labeled set we got a R^2 = 55% with 22 unbalanced labels

Package	Real Label	Our Label
DecisionTree	Machine Learning	Machine Learning
TimeModels	Statistics	Statistics
PlotlyJS	API	API
JuMP	Optimization	Optimization
Redis	Database	Database
Measurements	Mathematics	Physics
SparseVectors	Mathematics	Mathematics

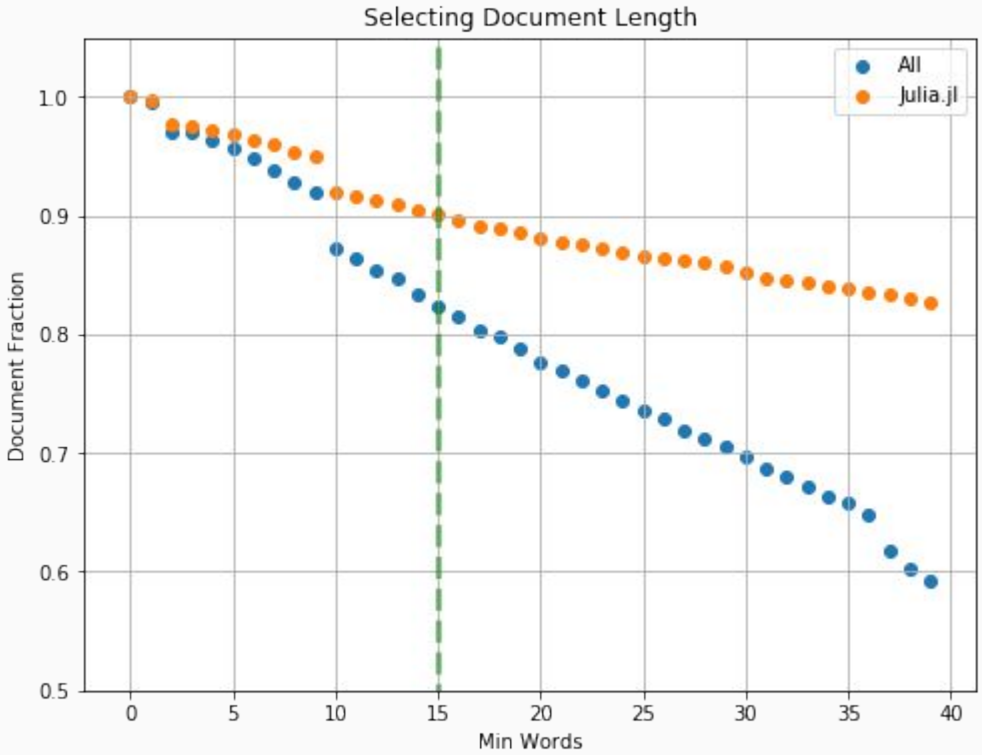
Next steps to get project working on JuliaObserver:

1 Connect clusterer and classifier

2 Setup workers on server

3 Add docs and description

Balanced document length between corpus size and usefulness


Selecting Document Length
All
Julia.jl
Document Fraction
Min Words