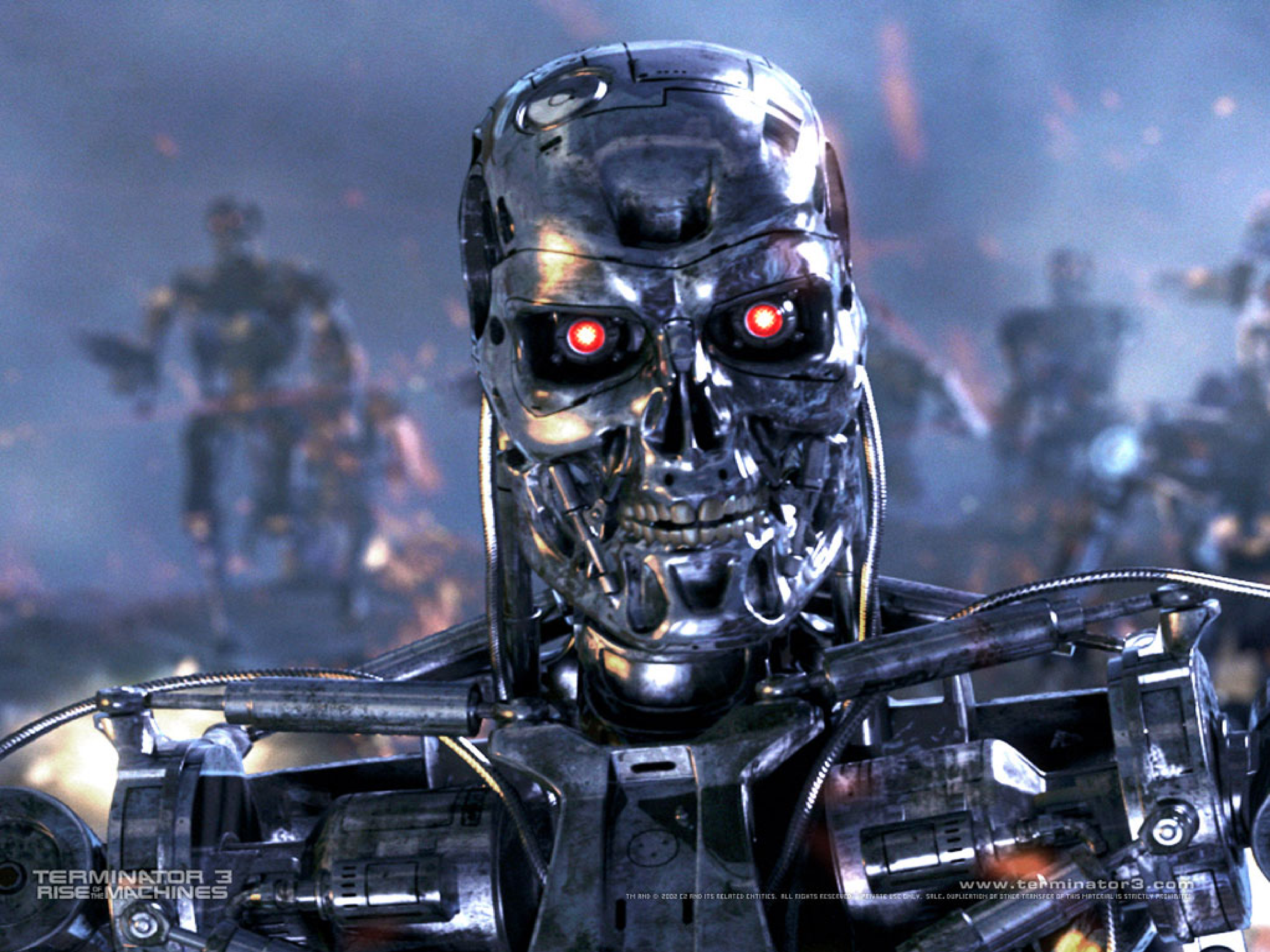
A Snake Learns

Machine Learning and Python

Igor Guerrero
@igorgue

What's Machine Learning?


TERMINATOR 3
RISE OF THE MACHINES
www.terminator3.com
TM AND © 2002 C2 AND ITS RELATED ENTITIES. ALL RIGHTS RESERVED. PRIVATE USE ONLY. SALE, DUPLICATION OR OTHER TRANSFER OF THIS MATERIAL IS STRICTLY PROHIBITED.

*"A branch of **artificial intelligence**, is a scientific discipline concerned with the design and development of **algorithms** that allow **computers** to evolve behaviors based on empirical **data**, such as from **sensor** data or **databases"**.*

*- **Wikipedia** (http://en.wikipedia.org/wiki/Machine_Learning)*

Cool Story, Bro!

*Machine Learning is more than just **algorithms!***

Machine Learning in real life

Data Input

Algorithms

Data Output

Runtime

Big Data is Big

mongoDB
{name: “mongo”, type:“DB”}

hadoop

CouchDB
relax

redis

riak

*I'm **not** telling you to switch database...*

*If your current **relational database** doesn't cut it for **ML** there are alternatives!*

*And **really** good ones!*

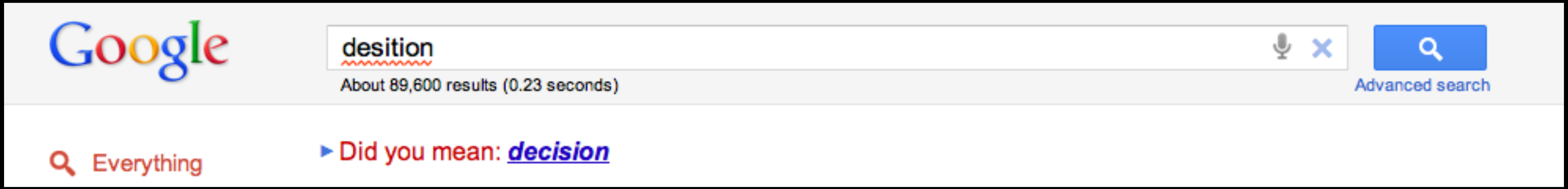
amazon
webservices™

http://aws.amazon.com/elasticmapreduce/
(let them run your stuff, based on Hadoop)

Brute-force "learning"

Data is the ***algorithm***

Silly **Google** practices this!

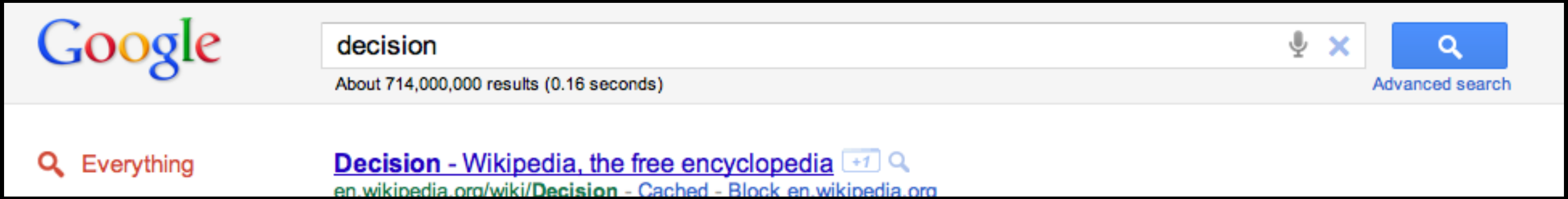


desition

About 89,600 results (0.23 seconds)

Did you mean: ***decision***

89,600 < 714,000,000



decision

About 714,000,000 results (0.16 seconds)

Decision - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/**Decision** - Cached - Block en.wikipedia.org

Brute-forcing their spell checker...

Not so genius now right?

http://code.google.com/apis/predict/

The Netflix Challenge winner was a collection of results generated by multiple algorithms:

http://www.netflixprize.com/leaderboard

NLP

Natural Language Processing, I knew grammar was useful.

A field of ***computer science*** *and* ***linguistics*** *concerned with the interactions between computers and human (natural) languages*

Guess the first word!

dataisbig

Word?(d) + ataisbig
Word?(da) + taisbig
Word?(dat) + aisbig
Word?(data) + isbig
(repeat procedure with the rest)

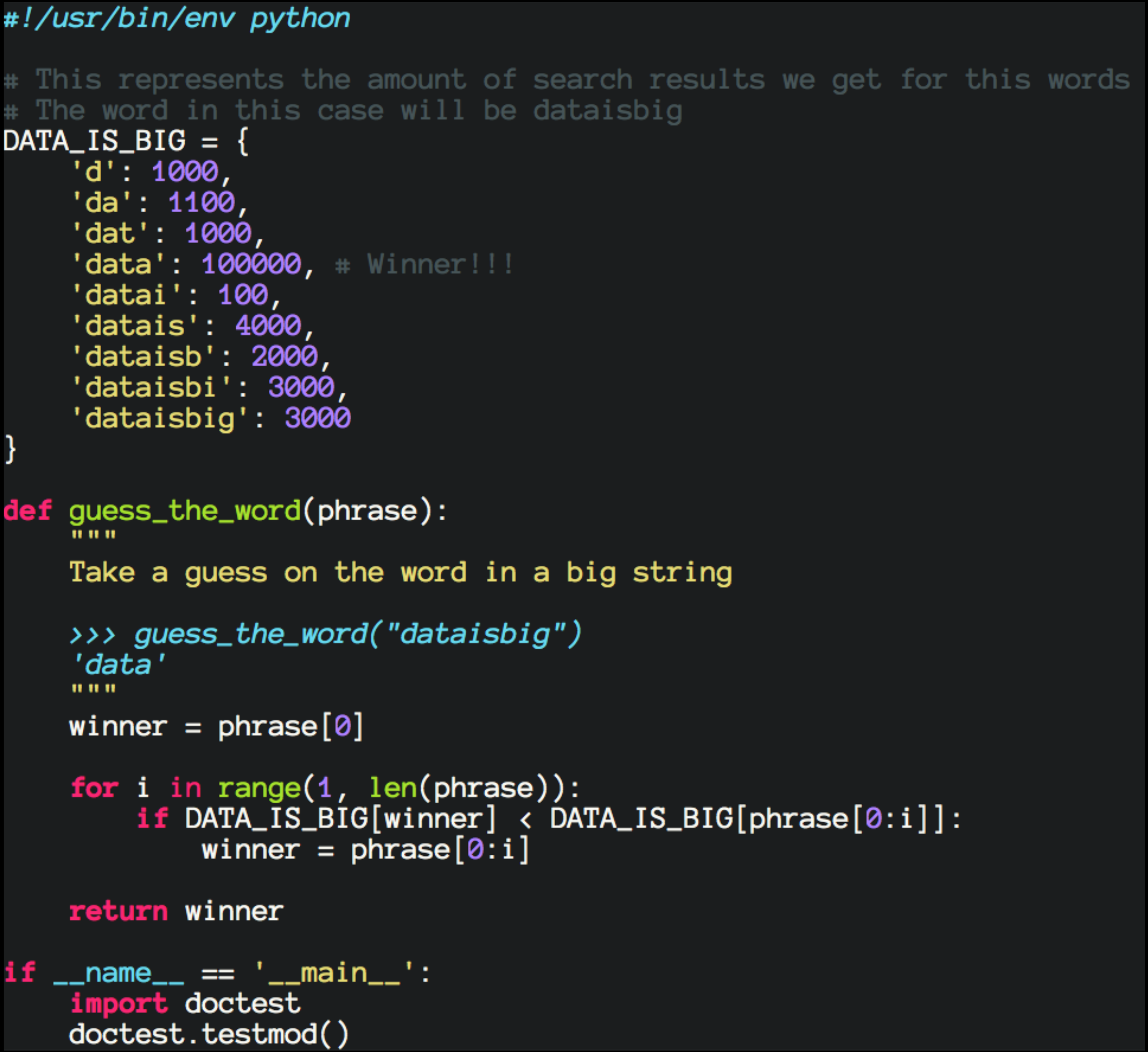
This is known as **word segmentation** very useful in foreign languages search!

Word?(word) = #Google hits / ~#pages of the web

It works, I promise!

http://ngrams.googlelabs.com/datasets

Google ngram database from scans from Google Books.

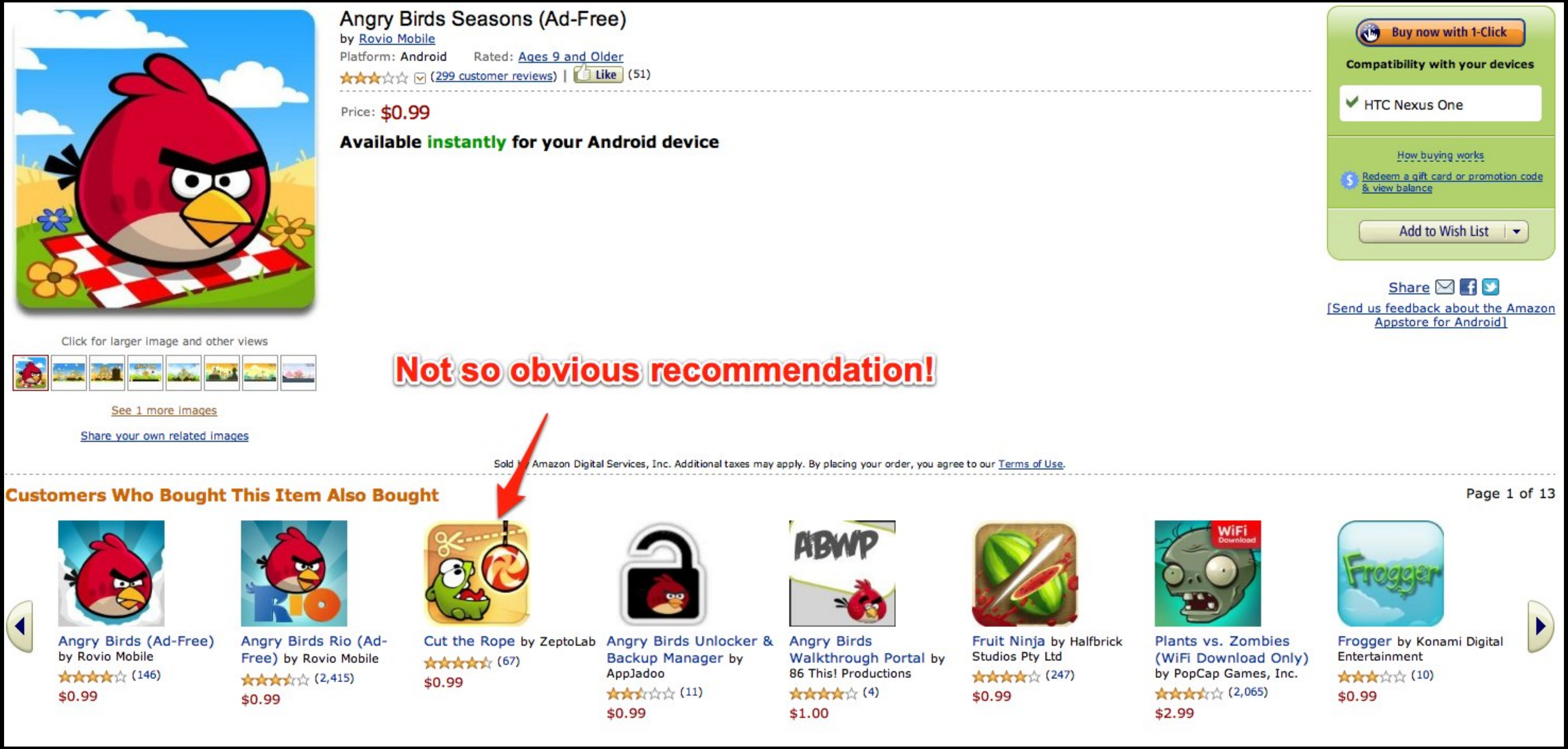


```python
#!/usr/bin/env python

# This represents the amount of search results we get for this words
# The word in this case will be dataisbig
DATA_IS_BIG = {
    'd': 1000,
    'da': 1100,
    'dat': 1000,
    'data': 100000, # Winner!!!
    'datai': 100,
    'datais': 4000,
    'dataisb': 2000,
    'dataisbi': 3000,
    'dataisbig': 3000
}

def guess_the_word(phrase):
    """
    Take a guess on the word in a big string

    >>> guess_the_word("dataisbig")
    'data'
    """
    winner = phrase[0]

    for i in range(1, len(phrase)):
        if DATA_IS_BIG[winner] < DATA_IS_BIG[phrase[0:i]]:
            winner = phrase[0:i]

    return winner

if __name__ == '__main__':
    import doctest
    doctest.testmod()
```

Recommendations

Based on your viewing history you might like "Snakes on a Plane"...

Amazon loves these



Angry Birds Seasons (Ad-Free)

by Rovio Mobile

Platform: Android   Rated: Ages 9 and Older

(299 customer reviews) | Like (51)

Price: $0.99

Available instantly for your Android device

Click for larger image and other views

See 1 more images

Share your own related images

Not so obvious recommendation!

Sold by Amazon Digital Services, Inc. Additional taxes may apply. By placing your order, you agree to our Terms of Use.

Buy now with 1-Click
Compatibility with your devices
✓ HTC Nexus One
How buying works
Redeem a gift card or promotion code & view balance
Add to Wish List
Share
[Send us feedback about the Amazon Appstore for Android]

Customers Who Bought This Item Also Bought   Page 1 of 13

Angry Birds (Ad-Free) by Rovio Mobile (146) $0.99

Angry Birds Rio (Ad-Free) by Rovio Mobile (2,415) $0.99

Cut the Rope by ZeptoLab (67) $0.99

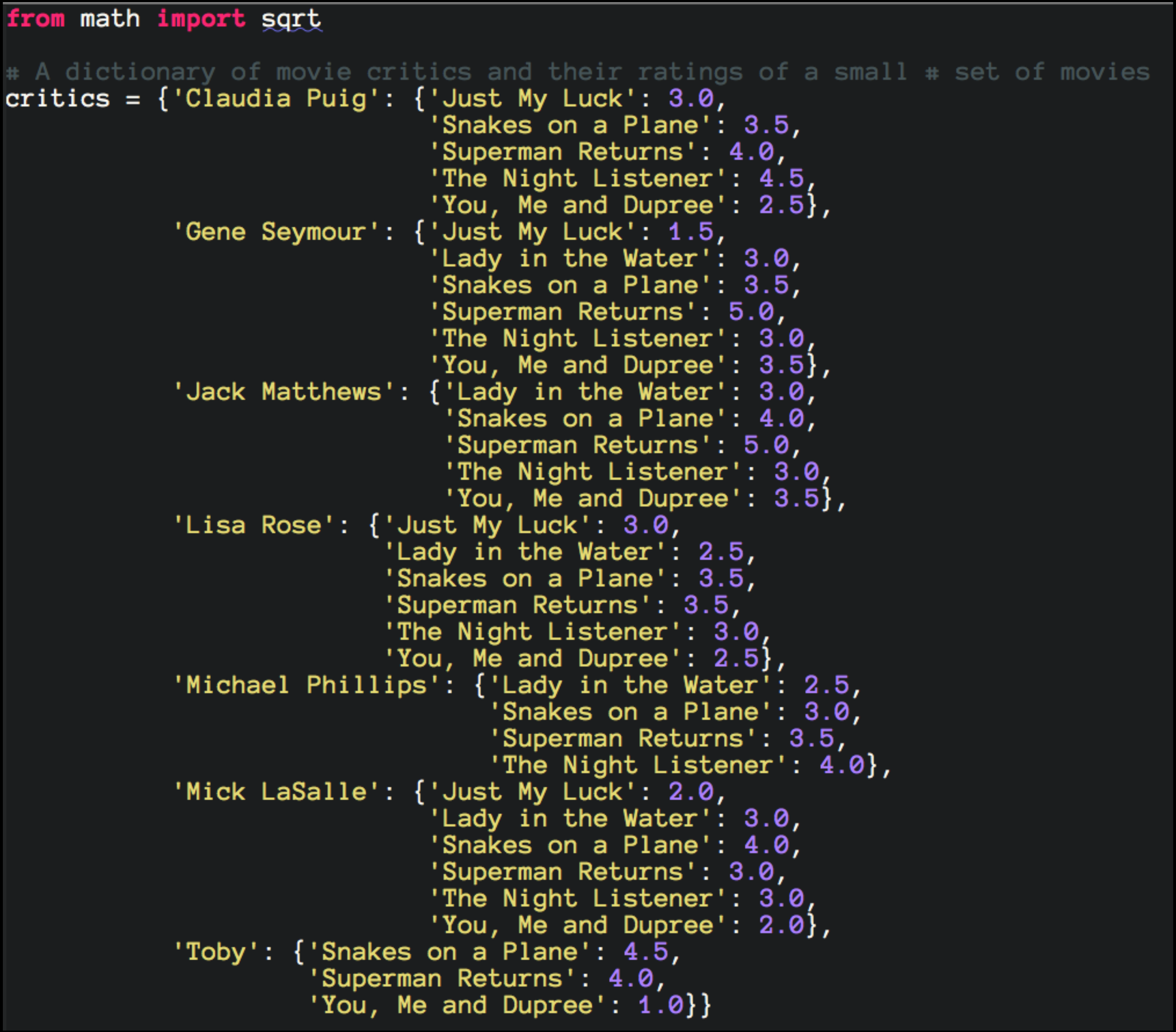
Angry Birds Unlocker & Backup Manager by AppJadoo (11) $0.99

Angry Birds Walkthrough Portal by 86 This! Productions (4) $1.00

Fruit Ninja by Halfbrick Studios Pty Ltd (247) $0.99

Plants vs. Zombies (WiFi Download Only) by PopCap Games, Inc. (2,065) $2.99

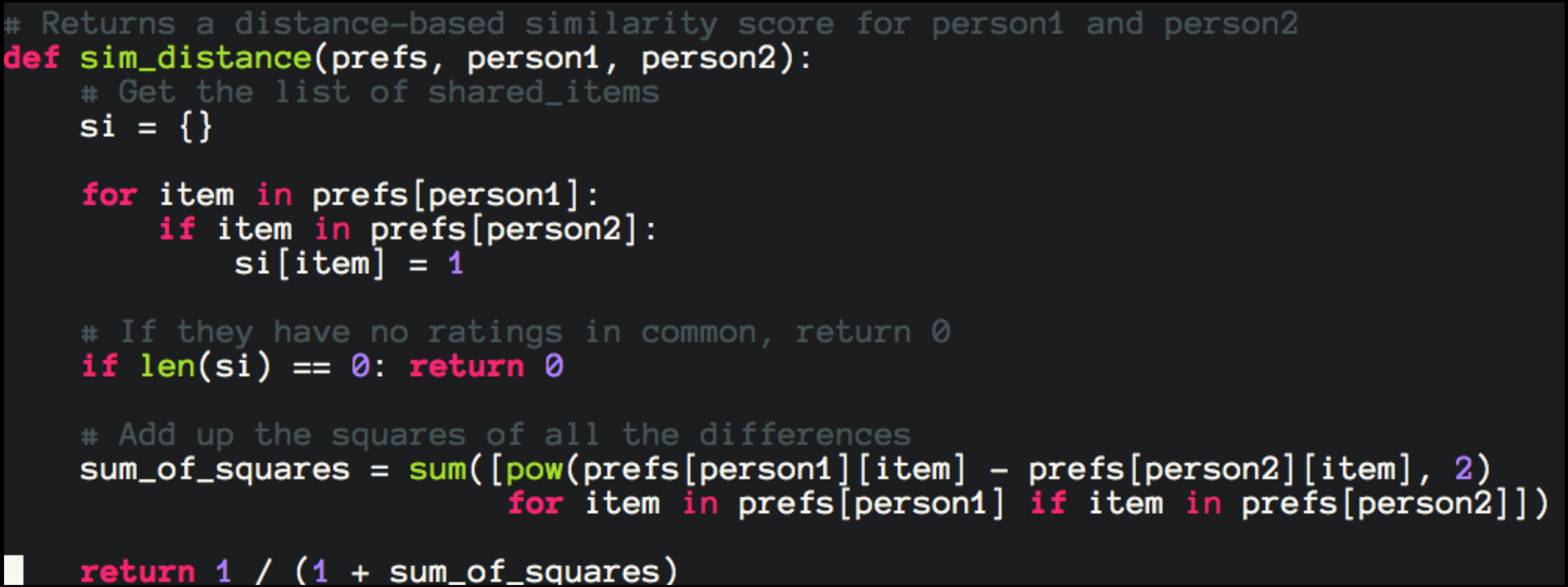
Frogger by Konami Digital Entertainment (10) $0.99



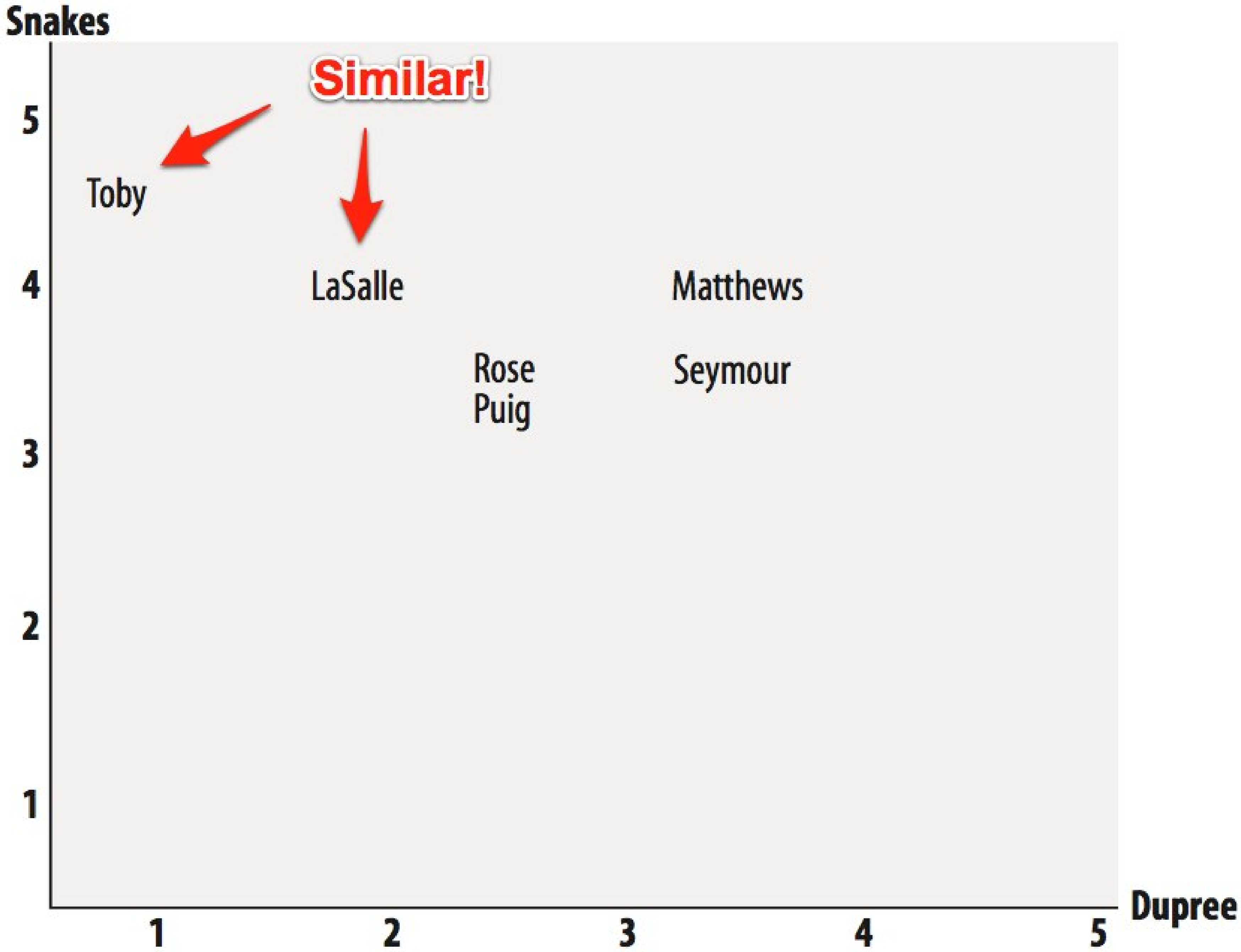
```python
from math import sqrt

# A dictionary of movie critics and their ratings of a small # set of movies
critics = {'Claudia Puig': {'Just My Luck': 3.0,
                            'Snakes on a Plane': 3.5,
                            'Superman Returns': 4.0,
                            'The Night Listener': 4.5,
                            'You, Me and Dupree': 2.5},
           'Gene Seymour': {'Just My Luck': 1.5,
                            'Lady in the Water': 3.0,
                            'Snakes on a Plane': 3.5,
                            'Superman Returns': 5.0,
                            'The Night Listener': 3.0,
                            'You, Me and Dupree': 3.5},
           'Jack Matthews': {'Lady in the Water': 3.0,
                             'Snakes on a Plane': 4.0,
                             'Superman Returns': 5.0,
                             'The Night Listener': 3.0,
                             'You, Me and Dupree': 3.5},
           'Lisa Rose': {'Just My Luck': 3.0,
                         'Lady in the Water': 2.5,
                         'Snakes on a Plane': 3.5,
                         'Superman Returns': 3.5,
                         'The Night Listener': 3.0,
                         'You, Me and Dupree': 2.5},
           'Michael Phillips': {'Lady in the Water': 2.5,
                                'Snakes on a Plane': 3.0,
                                'Superman Returns': 3.5,
                                'The Night Listener': 4.0},
           'Mick LaSalle': {'Just My Luck': 2.0,
                            'Lady in the Water': 3.0,
                            'Snakes on a Plane': 4.0,
                            'Superman Returns': 3.0,
                            'The Night Listener': 3.0,
                            'You, Me and Dupree': 2.0},
           'Toby': {'Snakes on a Plane': 4.5,
                    'Superman Returns': 4.0,
                    'You, Me and Dupree': 1.0}}
```

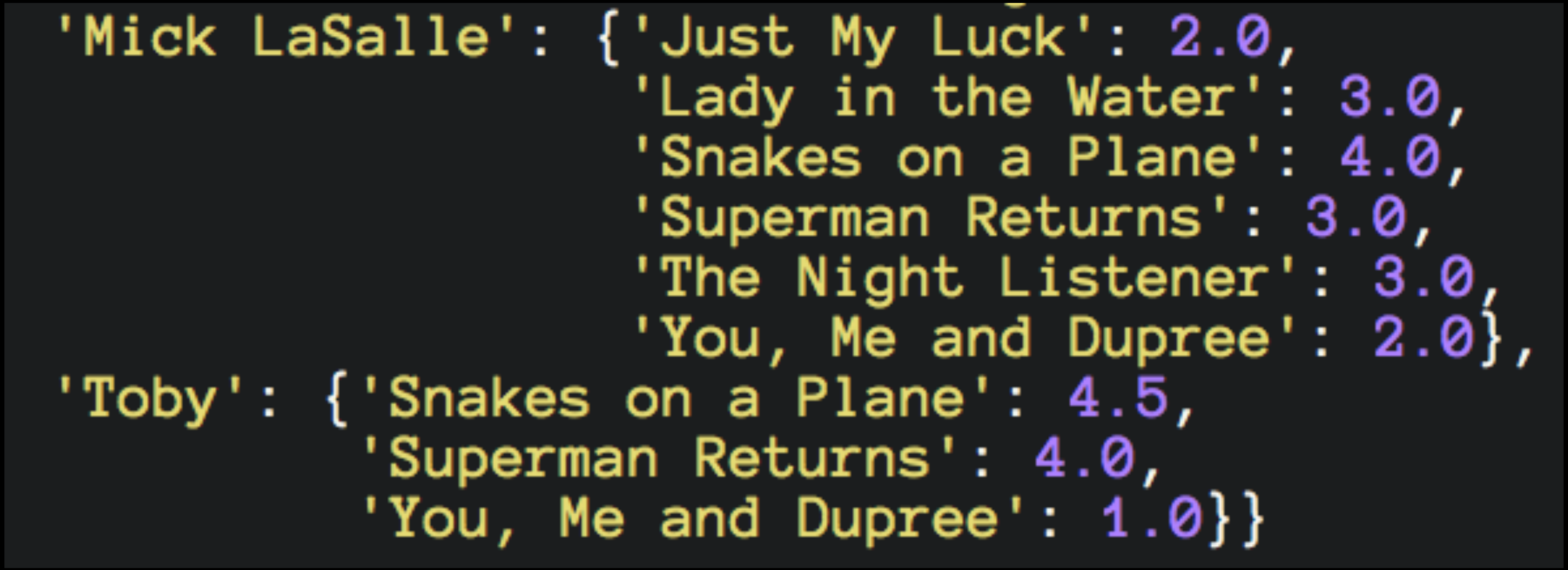
Euclidean Distance Algorithm

$$d(p,q) = (p_1 - q_1)^2 + (p_2 - q_2)^2$$



```
# Returns a distance-based similarity score for person1 and person2
def sim_distance(prefs, person1, person2):
    # Get the list of shared_items
    si = {}

    for item in prefs[person1]:
        if item in prefs[person2]:
            si[item] = 1

    # If they have no ratings in common, return 0
    if len(si) == 0: return 0

    # Add up the squares of all the differences
    sum_of_squares = sum([pow(prefs[person1][item] - prefs[person2][item], 2)
                         for item in prefs[person1] if item in prefs[person2]])

    return 1 / (1 + sum_of_squares)
```


Snakes
5
4
3
2
1
Similar!
Toby
LaSalle
Matthews
Rose
Puig
Seymour
1
2
3
4
5
Dupree



```
'Mick LaSalle': {'Just My Luck': 2.0,
                 'Lady in the Water': 3.0,
                 'Snakes on a Plane': 4.0,
                 'Superman Returns': 3.0,
                 'The Night Listener': 3.0,
                 'You, Me and Dupree': 2.0},
'Toby': {'Snakes on a Plane': 4.5,
         'Superman Returns': 4.0,
         'You, Me and Dupree': 1.0}}
```

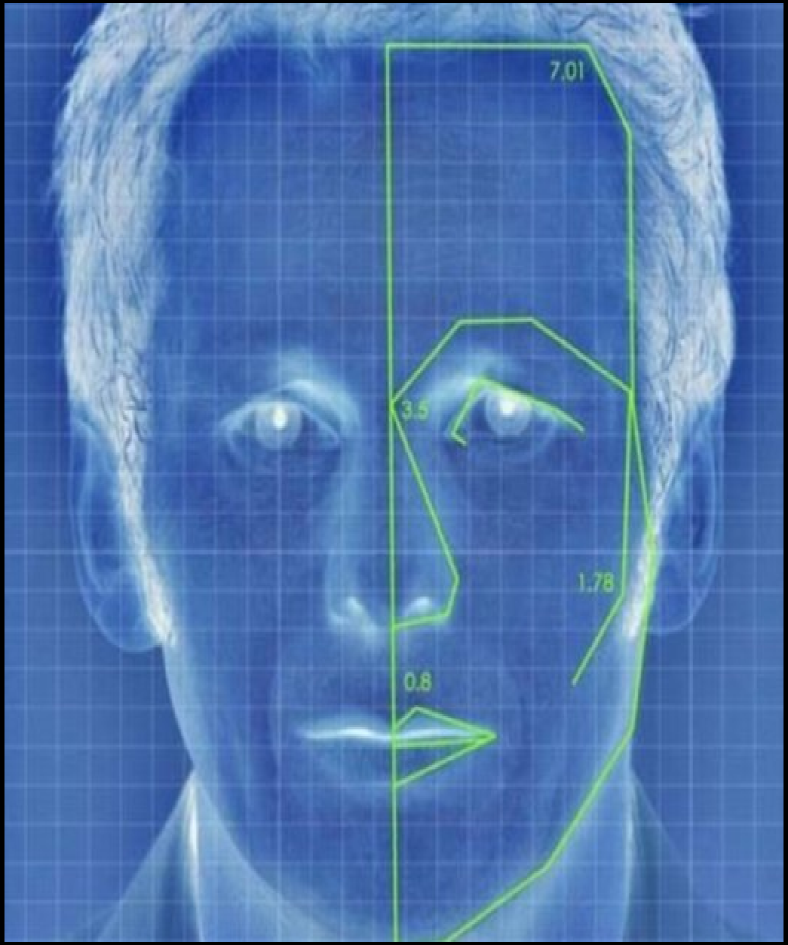
Toby might enjoy "Lady in the Water" and "The Night Listener".

And he'd hate "Just My Luck"...

Classification

"Dividing" data sets

Great for face recognition!



Facebook implemented it!

http://face.com offers a Free API!

Support Vector Machines

The calculation the line that divide objects is done via **SVM.**

http://www.csie.ntu.edu.tw/~cjlin/libsvm/

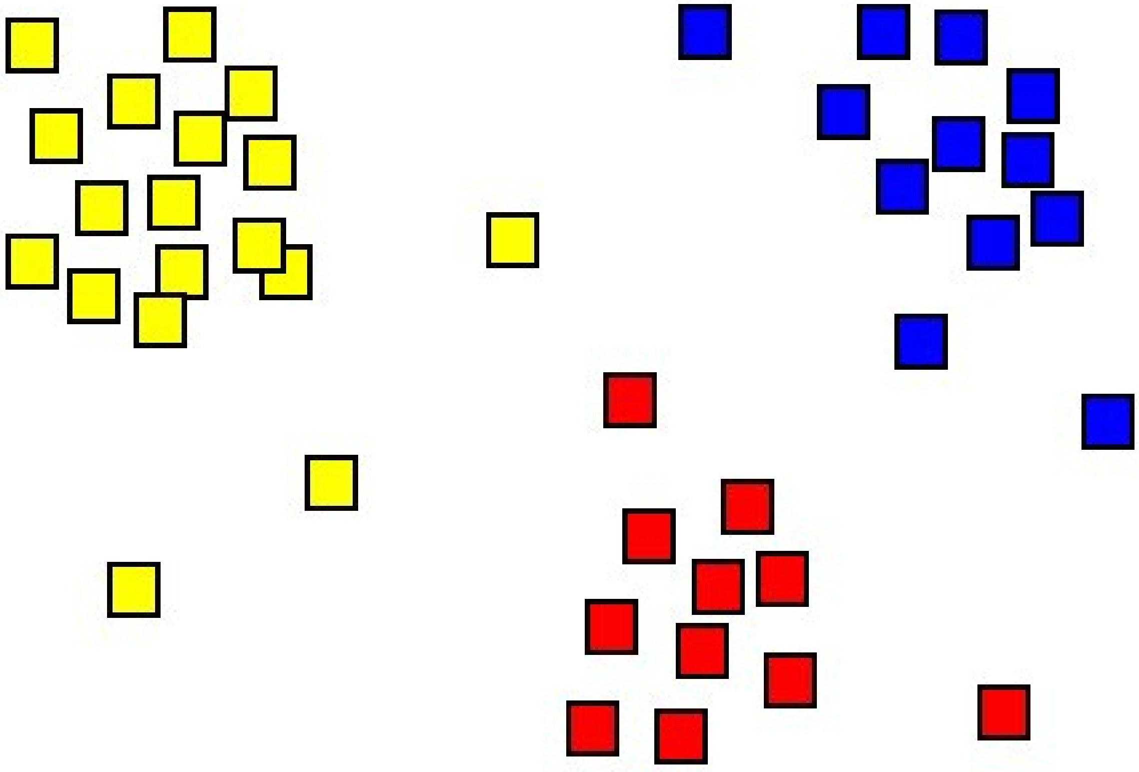
Clustering

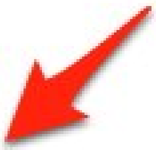
"Similarities" between different sets

This is how compression algorithms work

1. **AAAA AAA AA AAAAAA**
2. **BB BBBBB BBB BBBBBB**
3. **CCC CCCC CCCC CCC**

Use Euclidean Distance to know what elements are similar!





Similar

Resources

- Programming Collective Intelligence: http://oreilly.com/catalog/9780596529321
- Hadoop tutorial: http://developer.yahoo.com/hadoop/tutorial/
- R Programming language: http://www.r-project.org/
- My favorite Machine Learning community members:
    - Ilya Grigorik (Google): http://www.igvita.com/
    - Jonathan Harris (We Feel Fine): http://www.wefeelfine.org/
- Contact me: http://igorgue.com