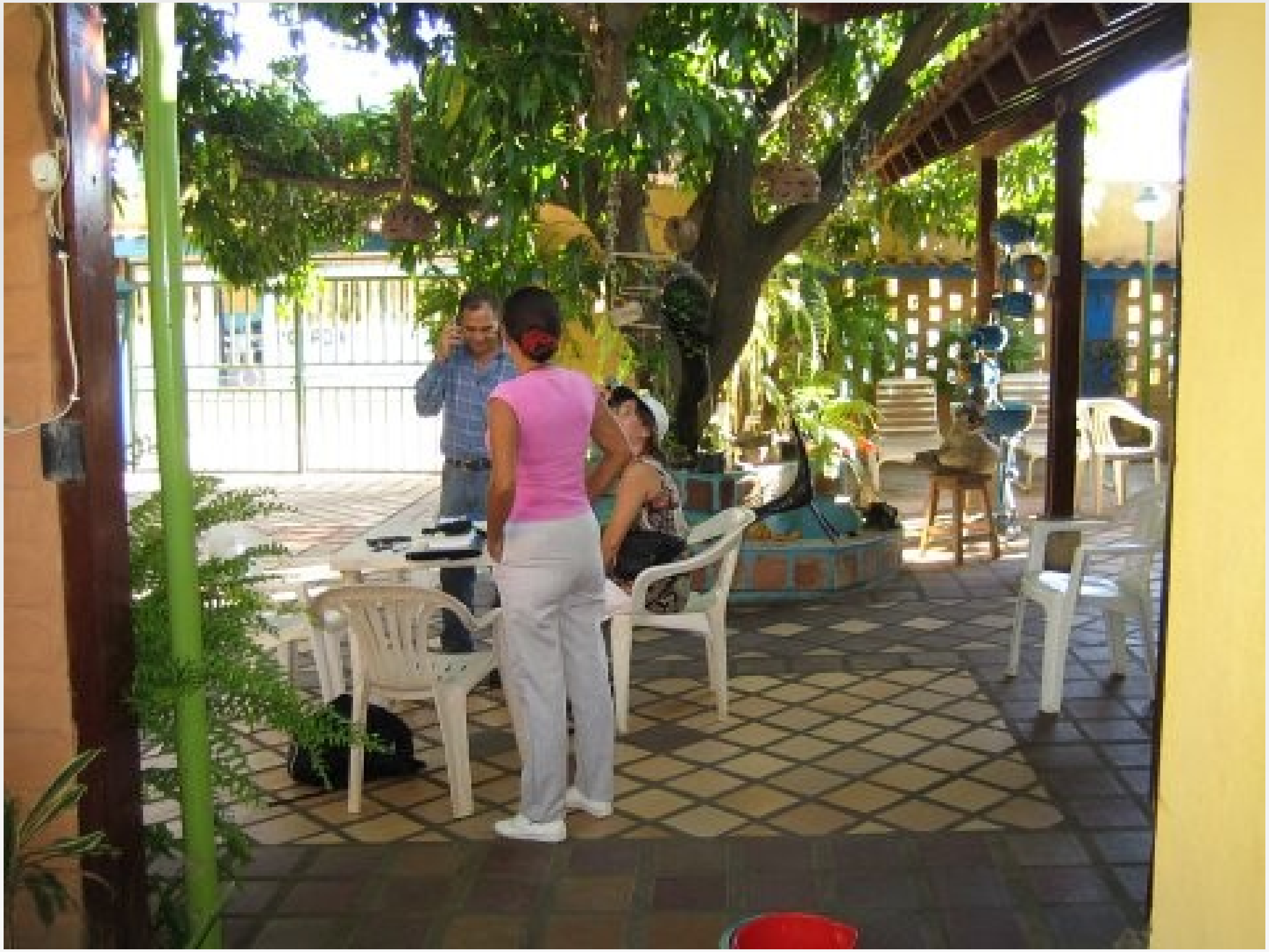
Automatic Learning of Descriptive Factors

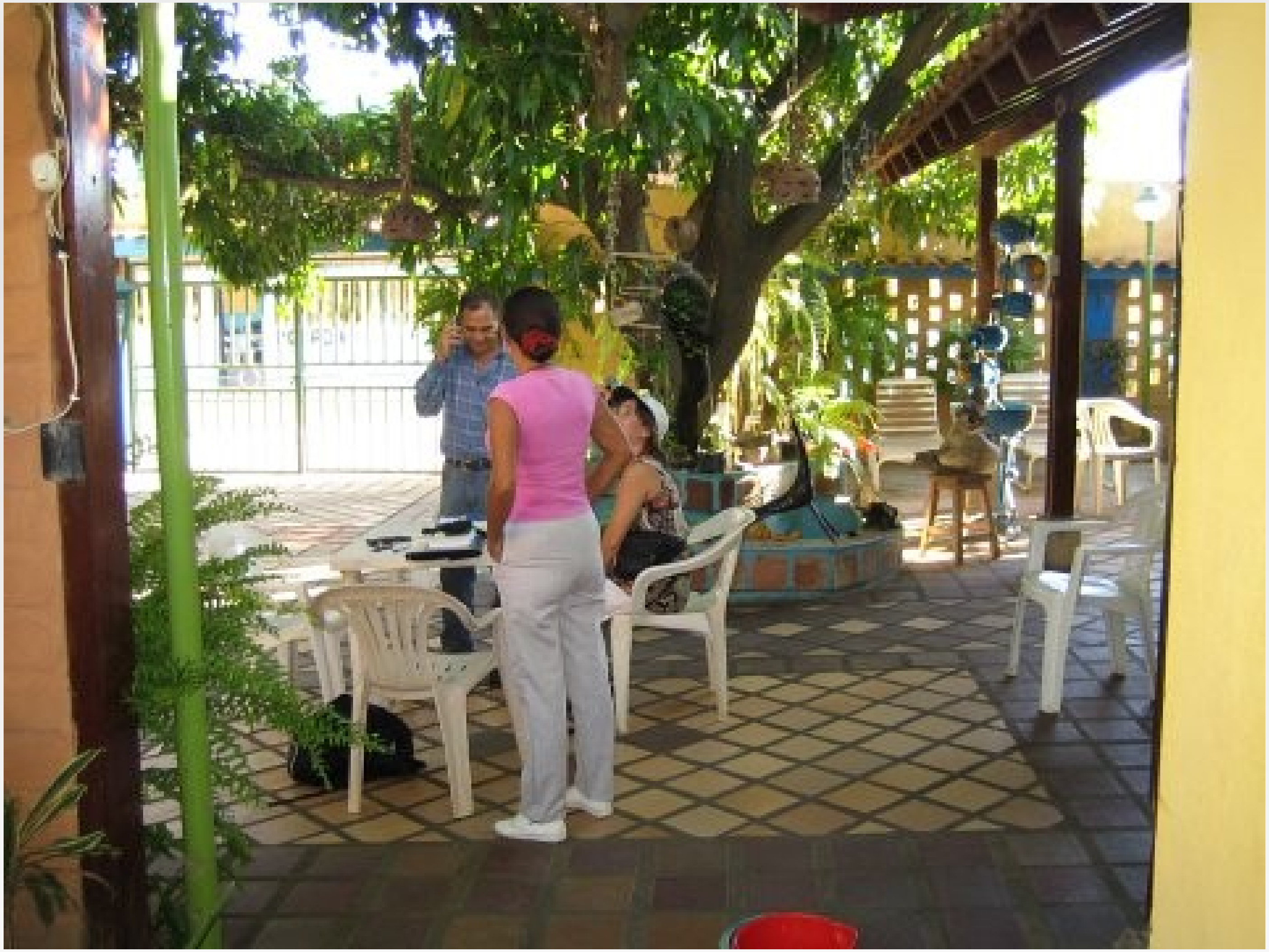
Karl Stratos

University of Rochester
Columbia University

Describe this Image!

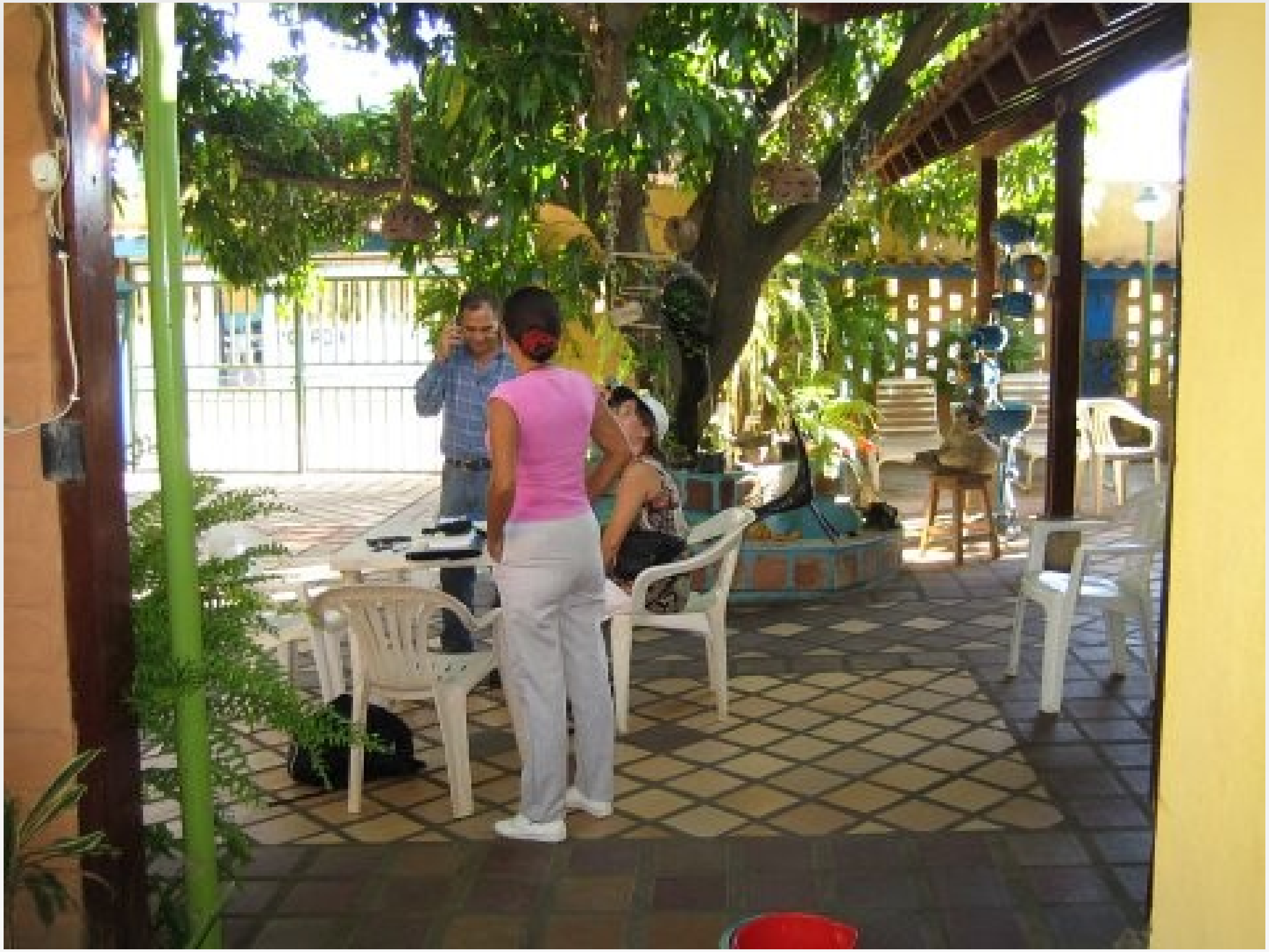


Describe this Image!



A bag, four chairs, one tree, three people, two walls, a floor...

Describe this Image!

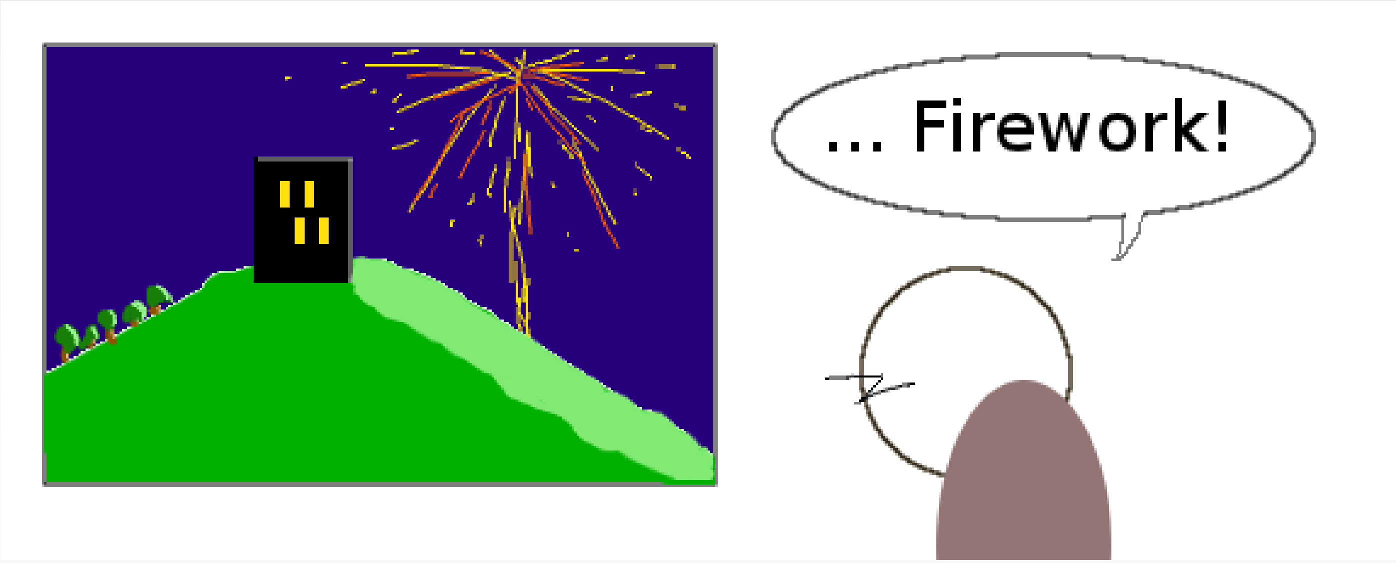


A bag, four chairs, one tree, three people, two walls, a floor...

People between walls

Task

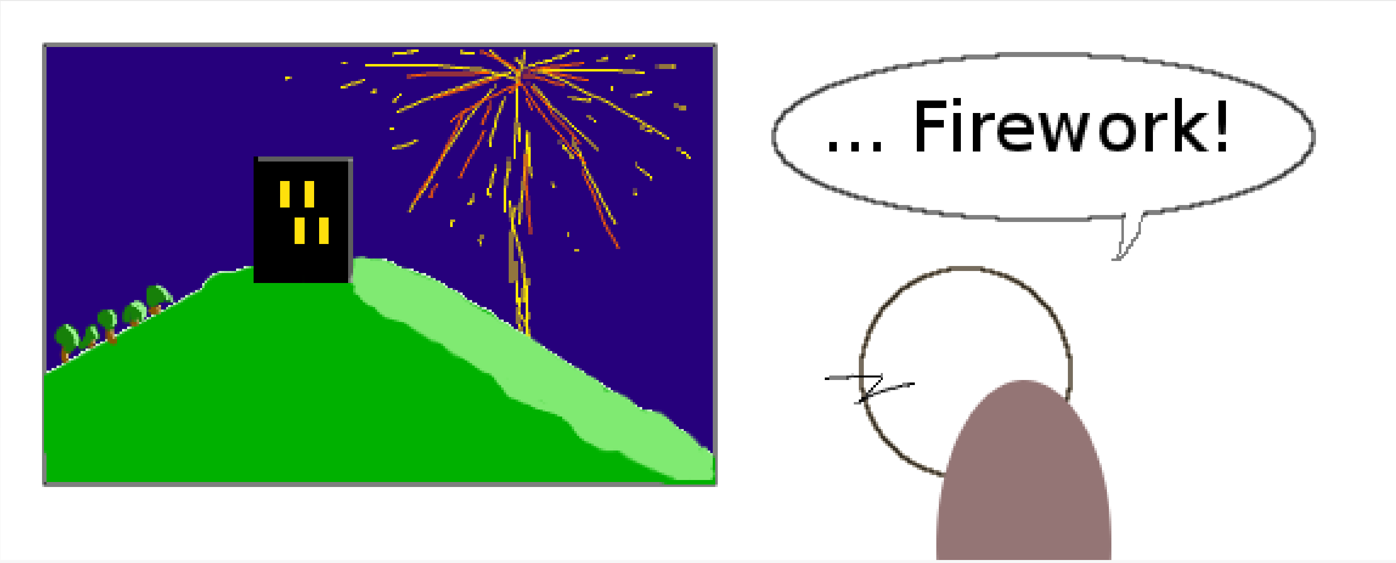
- Study what people choose to describe



Task

- Study what people choose to describe
- Learn what influence their choice ("Descriptive Factors”)

  - Type, size, location, etc.

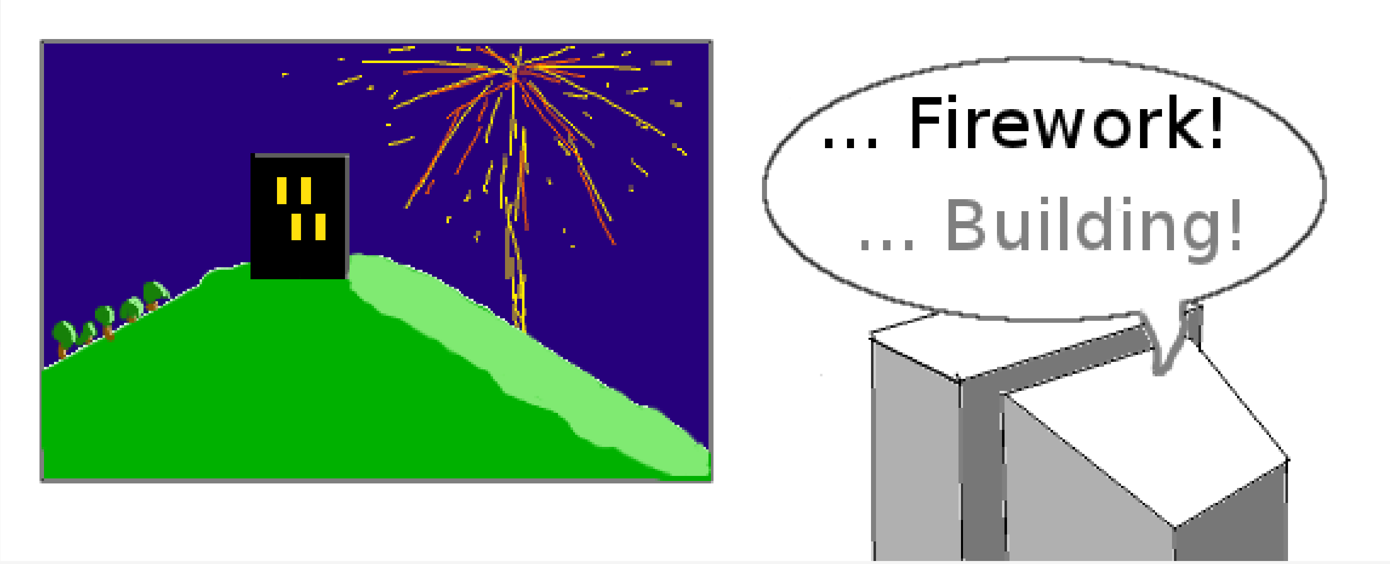


Task

- Study what people choose to describe
- Learn what influence their choice ("Descriptive Factors")

  - Type, size, location, etc.

- Model the process



What Makes an Object Salient?

- Spain and Perona (2008):
  - What object is likely to be mentioned first?
  - $P(O_1 = O \mid [O_1,...,O_n])$
- Us:
  - What objects should be mentioned at all?
  - $P(O \text{ is mentioned} \mid O \text{ is an object in } I)$

What Makes an Object Salient?

- Spain and Perona (2008):
  - What object is likely to be mentioned first?
  - P(O_1 = O | [O_1,...,O_n])
- Us:
  - What objects should be mentioned at all?
  - P(O is mentioned | O is an object in I)

Information we need:

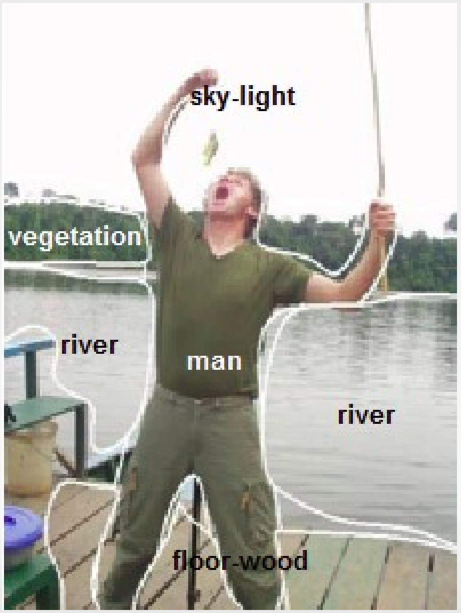
1. All objects Os "noticeable" in *I*
2. If *O* is described

ImageCLEF

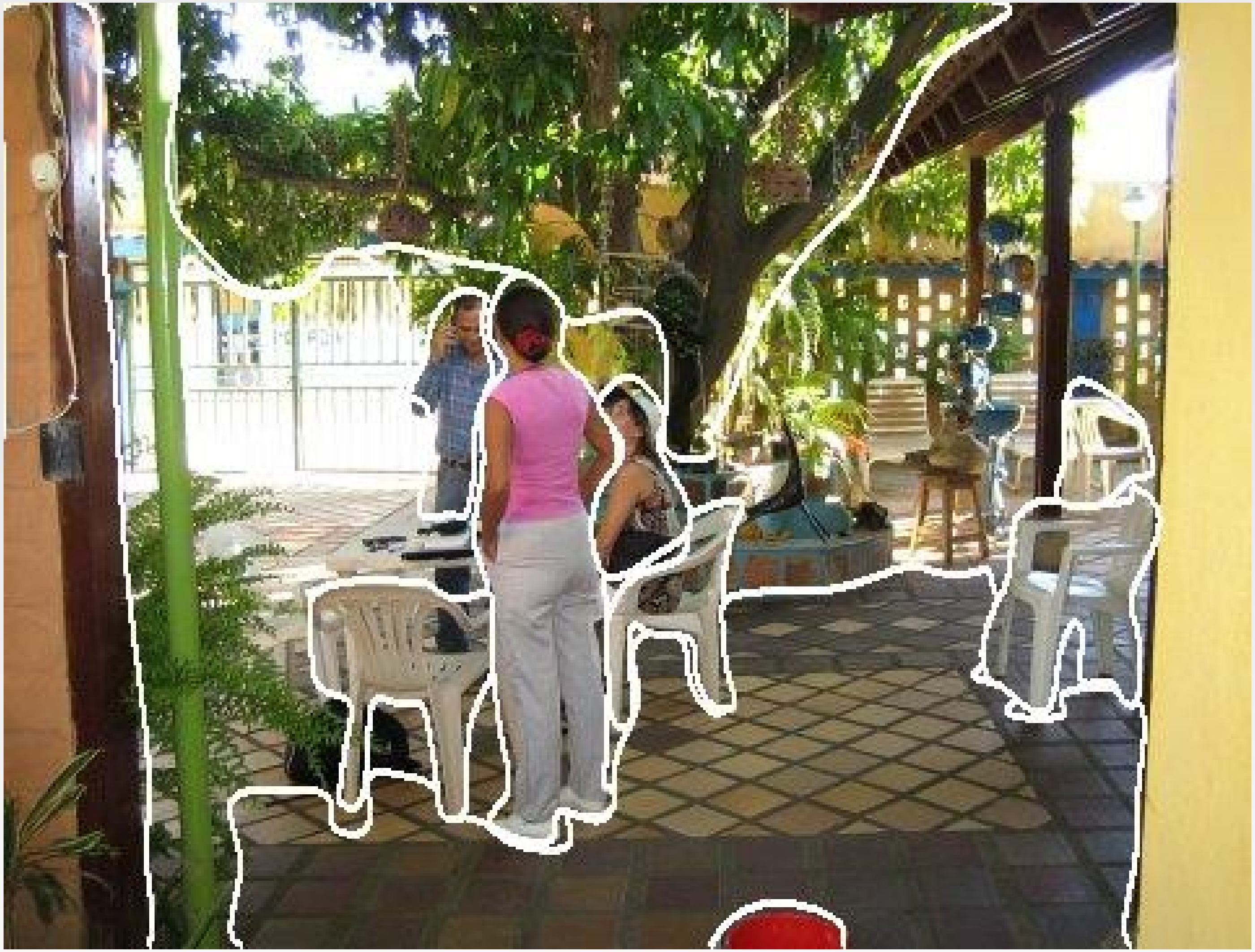
- 20K images of various aspects of contemporary life
  - Sports, cities, animals, people, landscapes
- Annotated with **free-text descriptions**

ImageCLEF

- 20K images of various aspects of contemporary life
  - Sports, cities, animals, people, landscapes
- Annotated with **free-text descriptions**
- And also **segmented** according to a small set of labels

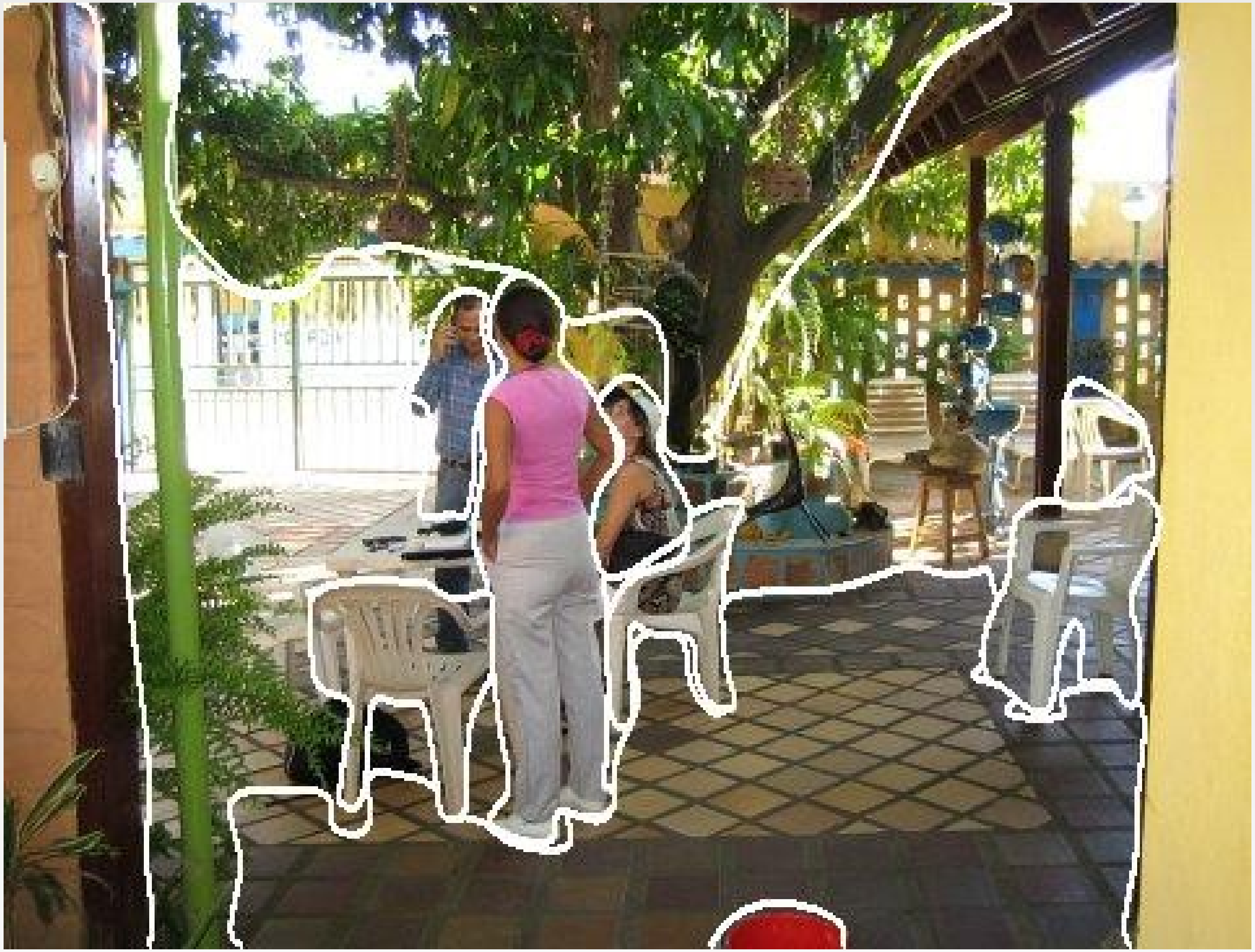


ImageCLEF



ImageID: 1236

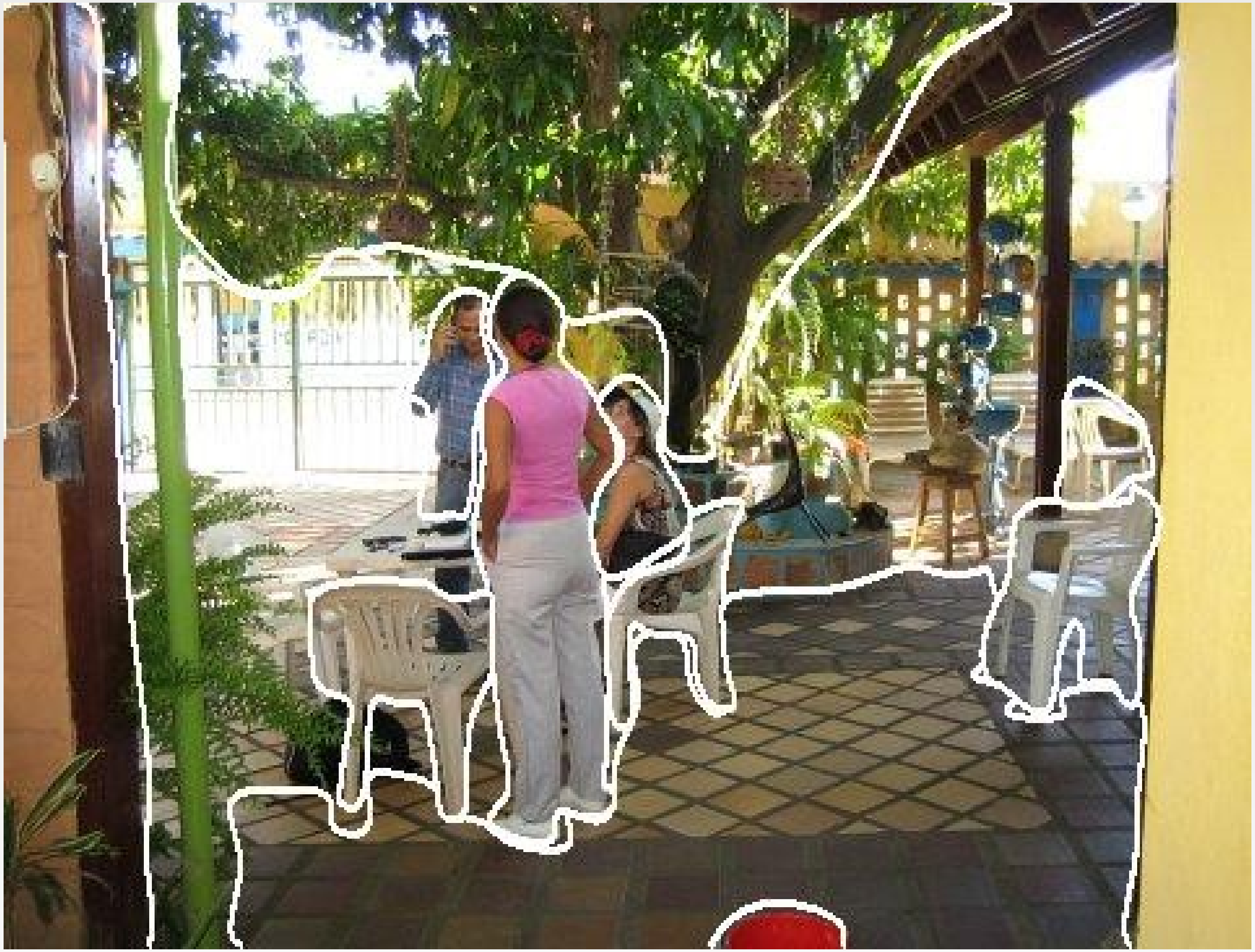
ImageCLEF



ImageID: 1236

Labels: ‘tree’, ‘floor’, ‘chair’, ‘chair’, ‘chair’, ‘chair’, ‘woman’, ‘man’, ‘woman’, ‘door’, ‘wall’

ImageCLEF



ImageID: 1236

Labels: ‘tree’, ‘floor’, ‘chair’, ‘chair’, ‘chair’, ‘chair’, ‘woman’, ‘man’, ‘woman’, ‘door’, ‘wall’

Descriptions:

"Two women and a man are standing and sitting in a yard on white chairs around a white table in the foreground"

ImageCLEF

- Idea

  - Use labels as proxies for all things in the image

  - P(*O* is mentioned | *O* is an object in *I*)

    ≈ P(*L* is referred to | *L* is a label of *I*)

    = C(L is referred to) / C(L)

ImageCLEF

- Idea
  - Use labels as proxies for all things in the image
  - $P(O \text{ is mentioned} \mid O \text{ is an object in } I)$

    $\approx P(L \text{ is referred to} \mid L \text{ is a label of } I)$

    $= C(\underbrace{\text{L is referred to}}_{\text{How do we know?}}) / C(\text{L})$

WordNet-Based Mapping

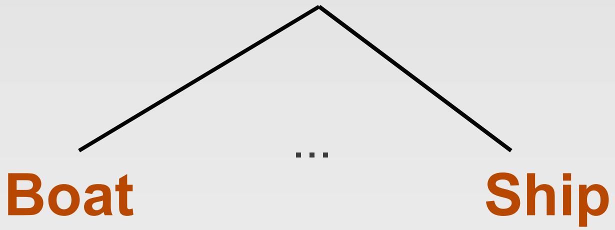
- WordNet Hierarchy

  1. If two words are "close”, they mean the same thing

WordNet-Based Mapping

- WordNet Hierarchy

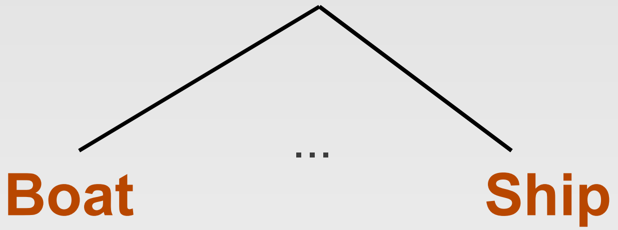
1. If two words are "close”, they mean the same thing



WordNet-Based Mapping

- WordNet Hierarchy

1. If two words are "close”, they mean the same thing



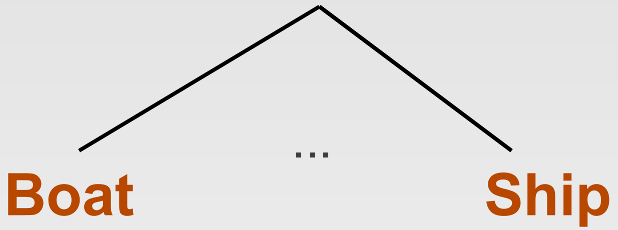
Boat ... Ship

2. If one word is an ancestor of the other, they mean the same thing

WordNet-Based Mapping

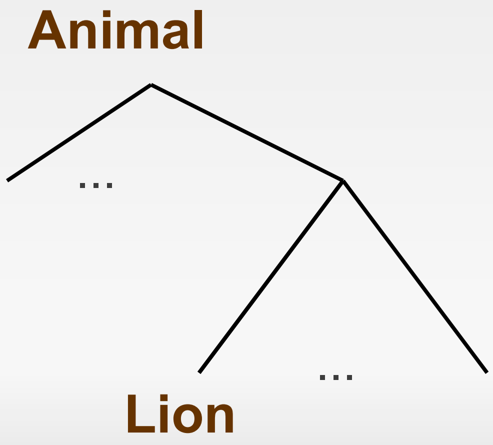
- WordNet Hierarchy

1. If two words are "close”, they mean the same thing



Boat ... Ship

2. If one word is an ancestor of the other, they mean the same thing



Animal

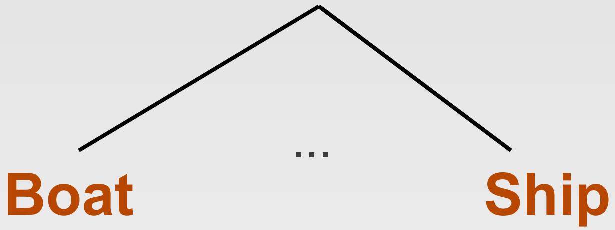
...

Lion ...

WordNet-Based Mapping

- WordNet Hierarchy

1. If two words are "close”, they mean the same thing

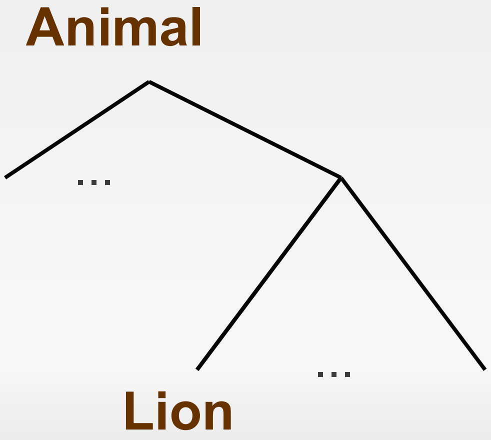


Boat

...

Ship

2. If one word is an ancestor of the other, they mean the same thing



Animal

...

Lion

...

F1 score of 94

ImageCLEF

‘tree’, ‘floor’, ‘chair’, ‘chair’, ‘chair’, ‘chair’, ‘woman’, ‘man’, ‘woman’, ‘door’, ‘wall’

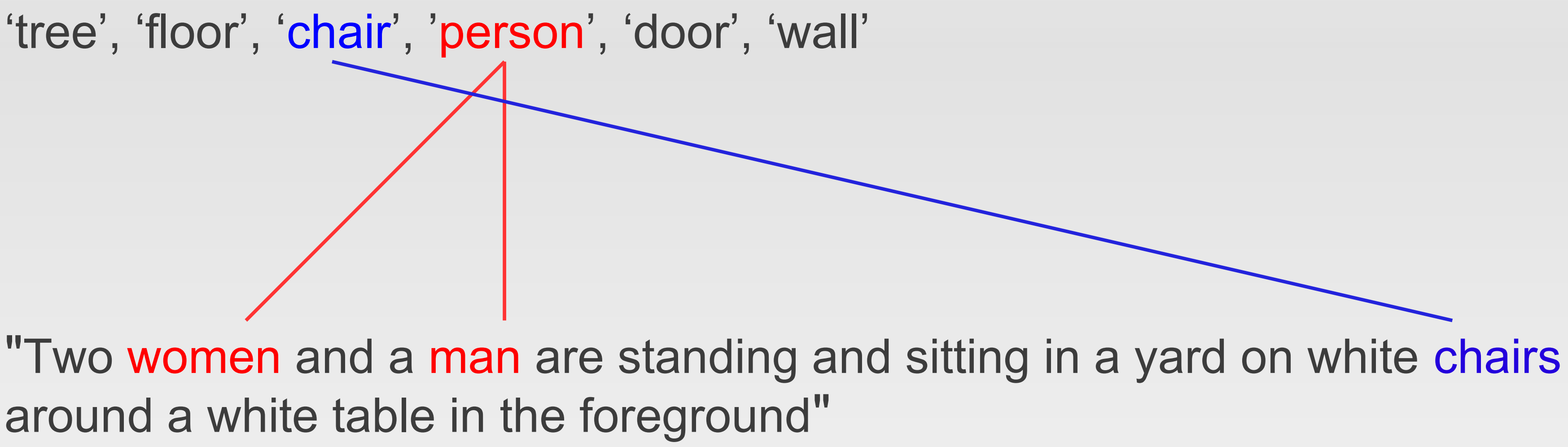
"Two women and a man are standing and sitting in a yard on white chairs around a white table in the foreground"

ImageCLEF

‘tree’, ‘floor’, ‘chair’, ’person’, ‘door’, ‘wall’

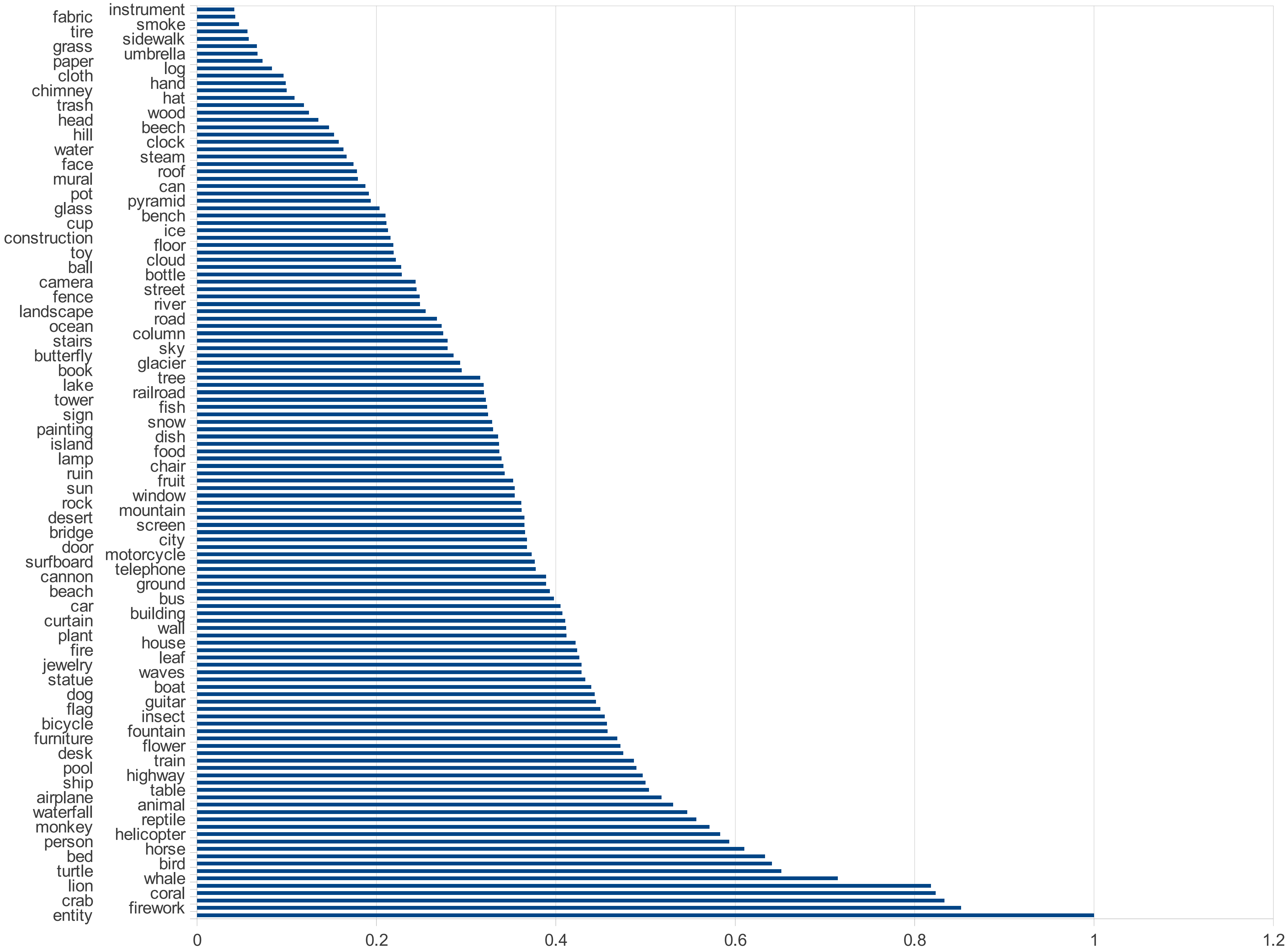
"Two women and a man are standing and sitting in a yard on white chairs around a white table in the foreground"

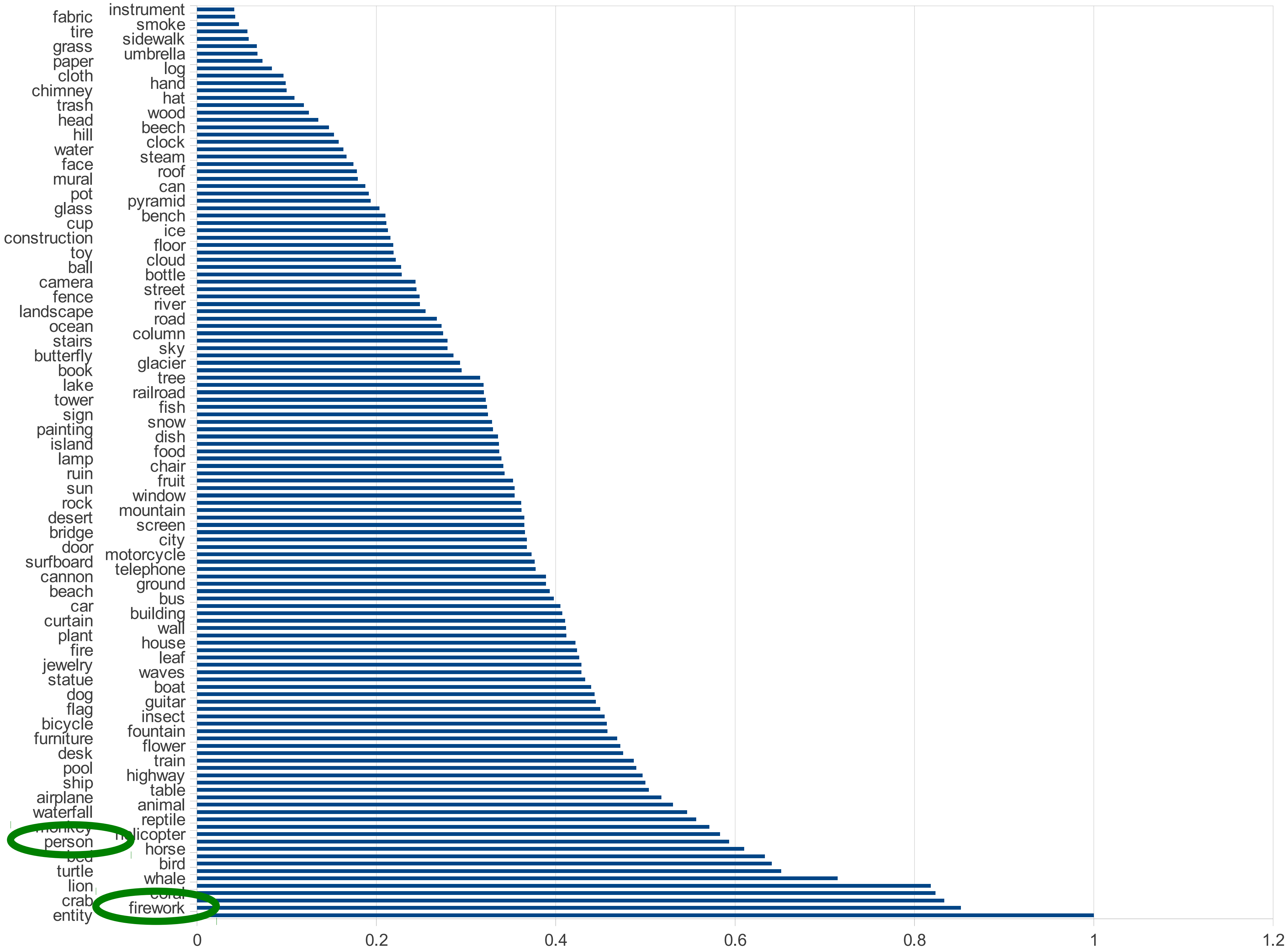
ImageCLEF

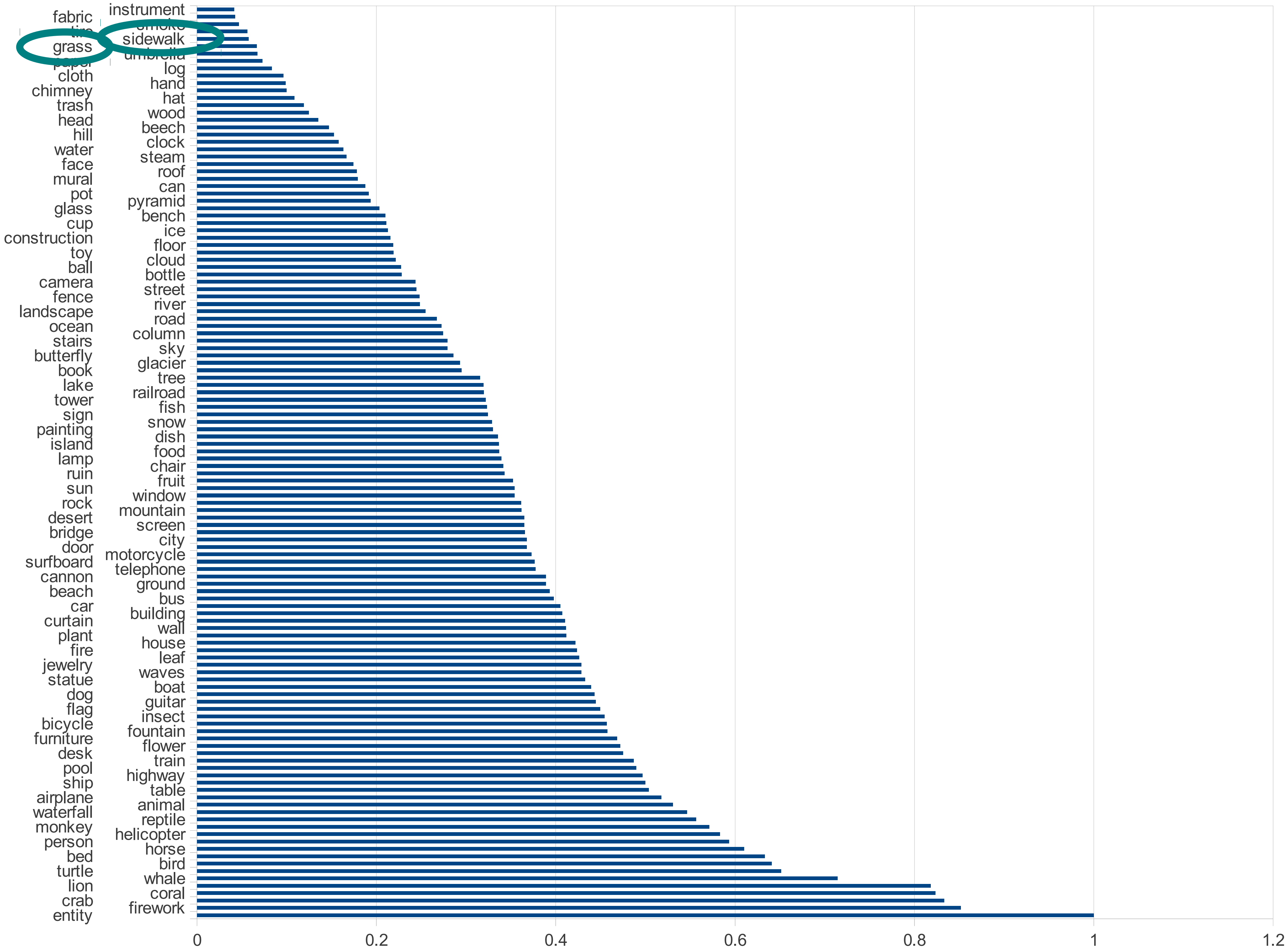


‘tree’, ‘floor’, ‘chair’, ’person’, ‘door’, ‘wall’

"Two women and a man are standing and sitting in a yard on white chairs around a white table in the foreground"


instrument
fabric
smoke
tire
sidewalk
grass
umbrella
paper
log
cloth
hand
chimney
hat
trash
wood
head
beech
hill
clock
water
steam
face
roof
mural
can
pot
pyramid
glass
bench
cup
ice
construction
floor
toy
cloud
ball
bottle
camera
street
fence
river
landscape
road
ocean
column
stairs
sky
butterfly
glacier
book
tree
lake
railroad
tower
fish
sign
snow
painting
dish
island
food
lamp
chair
ruin
fruit
sun
window
rock
mountain
desert
screen
bridge
city
door
motorcycle
surfboard
telephone
cannon
ground
beach
bus
car
building
curtain
wall
plant
house
fire
leaf
jewelry
waves
statue
boat
dog
guitar
flag
insect
bicycle
fountain
furniture
flower
desk
train
pool
highway
ship
table
airplane
animal
waterfall
reptile
monkey
helicopter
person
horse
bed
bird
turtle
whale
lion
coral
crab
firework
entity
0
0.2
0.4
0.6
0.8
1
1.2


instrument
fabric
smoke
tire
sidewalk
grass
umbrella
paper
log
cloth
hand
chimney
hat
trash
wood
head
beech
hill
clock
water
steam
face
roof
mural
can
pot
pyramid
glass
bench
cup
ice
construction
floor
toy
cloud
ball
bottle
camera
street
fence
river
landscape
road
ocean
column
stairs
sky
butterfly
glacier
book
tree
lake
railroad
tower
fish
sign
snow
painting
dish
island
food
lamp
chair
ruin
fruit
sun
window
rock
mountain
desert
screen
bridge
city
door
motorcycle
surfboard
telephone
cannon
ground
beach
bus
car
building
curtain
wall
plant
house
fire
leaf
jewelry
waves
statue
boat
dog
guitar
flag
insect
bicycle
fountain
furniture
flower
desk
train
pool
highway
ship
table
airplane
animal
waterfall
reptile
monkey
helicopter
person
horse
bed
bird
turtle
whale
lion
coral
crab
firework
entity
0
0.2
0.4
0.6
0.8
1
1.2


instrument
fabric
smoke
tire
sidewalk
grass
umbrella
paper
log
cloth
hand
chimney
hat
trash
wood
head
beech
hill
clock
water
steam
face
roof
mural
can
pot
pyramid
glass
bench
cup
ice
construction
floor
toy
cloud
ball
bottle
camera
street
fence
river
landscape
road
ocean
column
stairs
sky
butterfly
glacier
book
tree
lake
railroad
tower
fish
sign
snow
painting
dish
island
food
lamp
chair
ruin
fruit
sun
window
rock
mountain
desert
screen
bridge
city
door
motorcycle
surfboard
telephone
cannon
ground
beach
bus
car
building
curtain
wall
plant
house
fire
leaf
jewelry
waves
statue
boat
dog
guitar
flag
insect
bicycle
fountain
furniture
flower
desk
train
pool
highway
ship
table
airplane
animal
waterfall
reptile
monkey
helicopter
person
horse
bed
bird
turtle
whale
lion
coral
crab
firework
entity
0
0.2
0.4
0.6
0.8
1
1.2

Other Descriptive Factors

- Size?

  size(O) = relative number of pixels of O

- Location?

  loc(O) = relative distance from the center

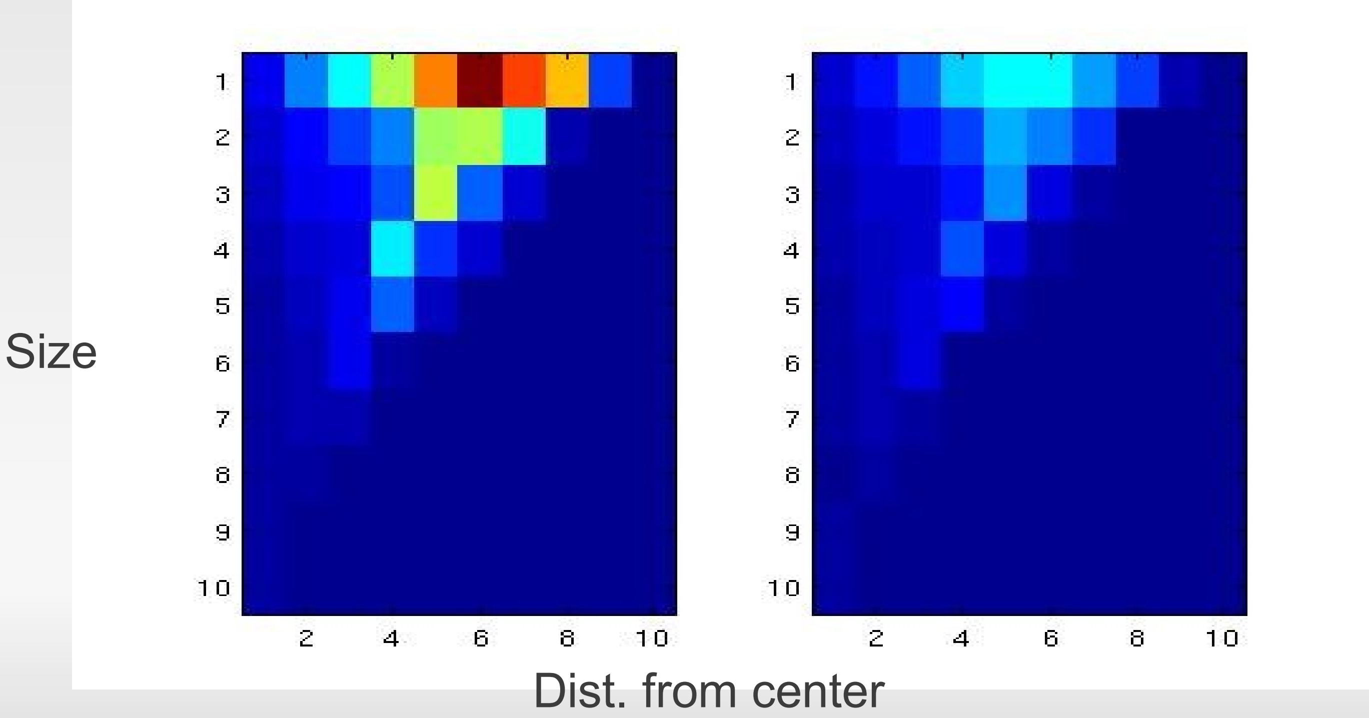
Other Descriptive Factors

- Size?

  size(O) = relative number of pixels of O

- Location?

  loc(O) = relative distance from the center



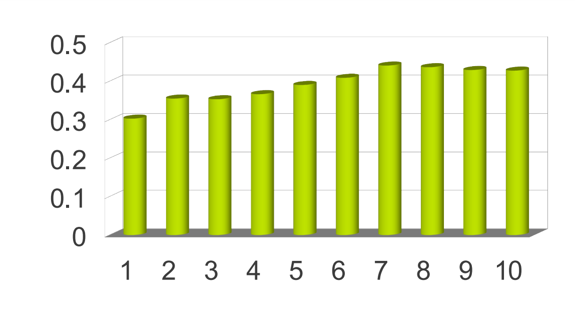
Other Descriptive Factors

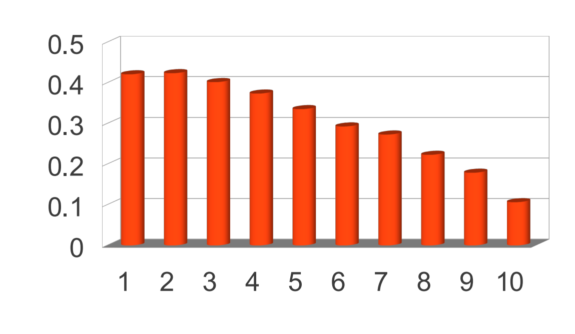
- Size?

  size(O) = relative number of pixels of O

- Location?

  loc(O) = relative distance from the center

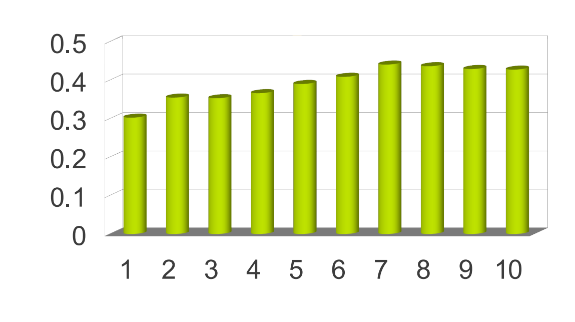




Other Descriptive Factors

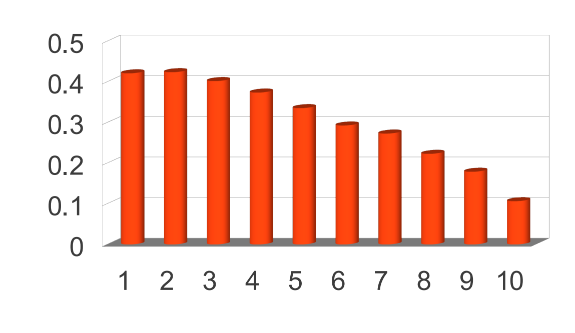
- Size?

  size(O) = relative number of pixels of O



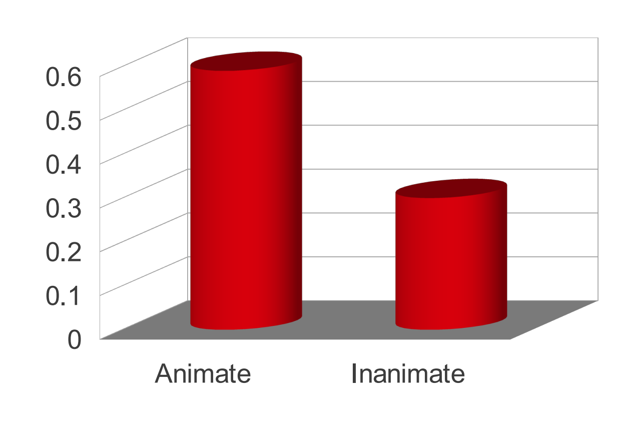
- Location?

  loc(O) = relative distance from the center



- Animacy?

  ani(O) = O is animate

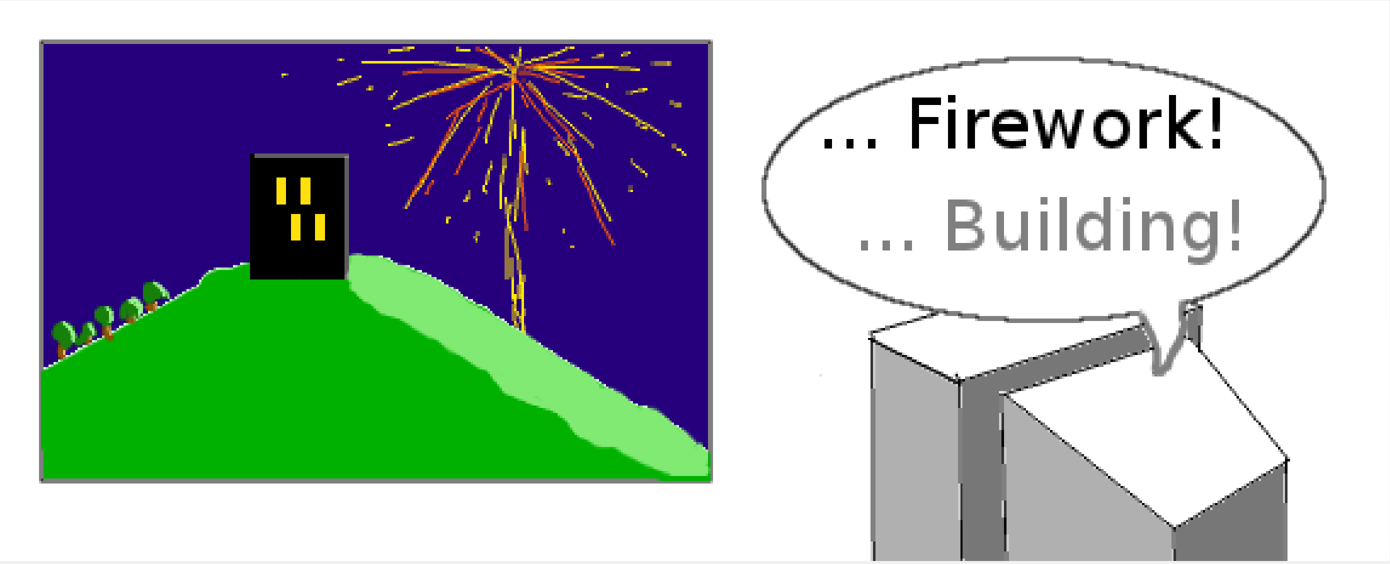


Modeling

- Use descriptive factors as features to train a classifier
  $\Phi(O)$ = (type, size, location, …)

Modeling

- Use descriptive factors as features to train a classifier
  $\Phi(O)$ = (type, size, location, …)
- Learn h: X → Y that minimizes the error on the data
  - X = {$\Phi(O)$: O an object in the image}
  - Y = {Mention, Ignore}



Modeling

	Accuracy (%)
Baseline: "Ignore" to all	63.1

T: Type
S: Size
L: Location
A: Animacy

Modeling

	Accuracy (%)
Baseline: "Ignore" to all	63.1
"Homemade" perceptron T+S+L+A	59.2

T: Type
S: Size
L: Location
A: Animacy

Modeling

T: Type
S: Size
L: Location
A: Animacy

	Accuracy (%)
Baseline: "Ignore" to all	63.1
"Homemade" perceptron T+S+L+A	59.2
Linear SVM S+L	60.3
Linear SVM T	68.2
Linear SVM **T+S+L**	**69.7**
Linear SVM **T+S+L+A**	**69.7**

Modeling

T: Type
S: Size
L: Location
A: Animacy

	Accuracy (%)
Baseline: "Ignore" to all	63.1
"Homemade" perceptron T+S+L+A	59.2
Linear SVM S+L	60.3
Linear SVM T	68.2
Linear SVM **T+S+L**	**69.7**
Linear SVM **T+S+L+A**	**69.7**

"Baseline": Ignore when
P(O mentioned | O present) < 0.5

83.1

Takeaway Message

- We can automatically learn what people choose to describe by exploiting an existing dataset

- Semantic features in making that choice tell us about human behavior, and can be helpful in modeling the process

Thanks!

tree floor chair person door wall