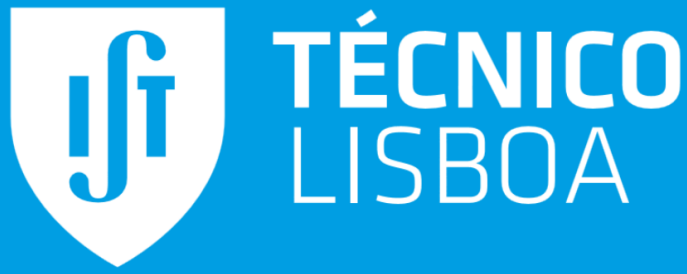


Homophilic Self Organizing Feature Maps

Bernardo Simões
Instituto Superior Técnico

October 27, 2014

Outline

Outline

Introduction

What is Twitter:

- Twitter is a Microbloging Platform.
- Users share what is happening in less than 140 characters.

Why do Topic Detection on Twitter?

- 19% of all tweets are about brands and 78% of Internet users put their trust on other users.
- Before events hit the news reports, they are being commented on Twitter.

Topic Detection on Twitter

Hard to Detect Topics on Individual Tweets

- Slang words.
- Typos.
- Small body of text.
- General *topic detection* mechanisms rely heavily on *TF-IDF*.

More Information Than Just Words

- Hashtags; Replies; Timestamps; Geo Coordinates.
- Social network behind the author of the tweet.

Clustering for Topic Detection and Tracking

Document Clustering

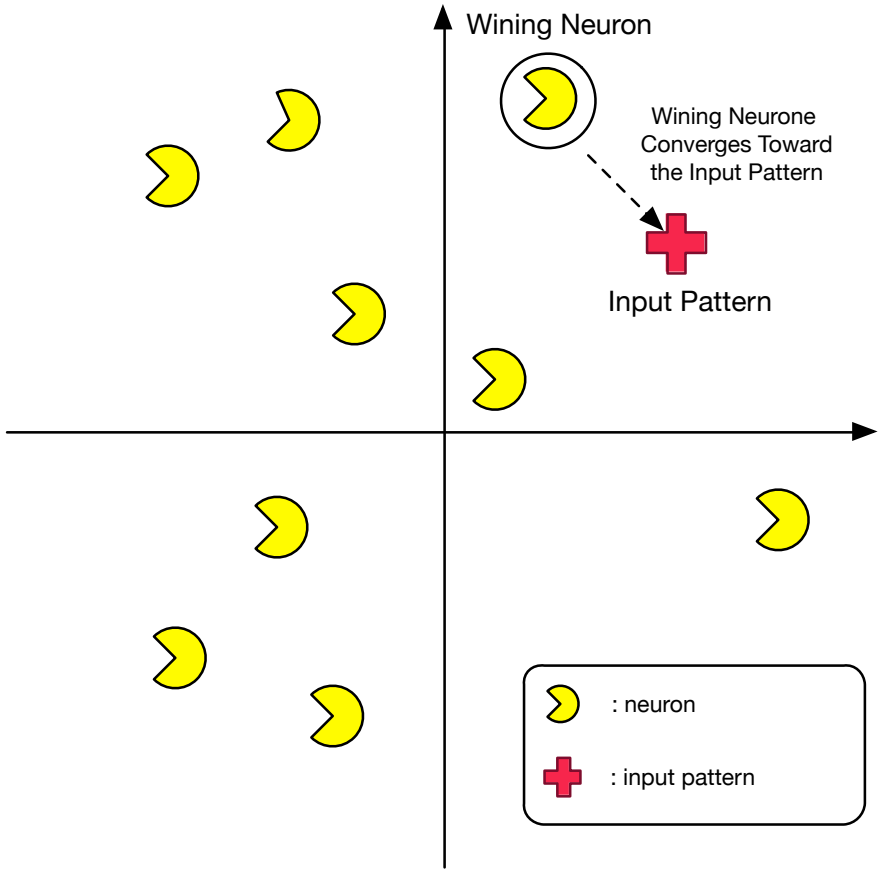
Cluster analysis or *clustering* is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

The Self Organizing Map

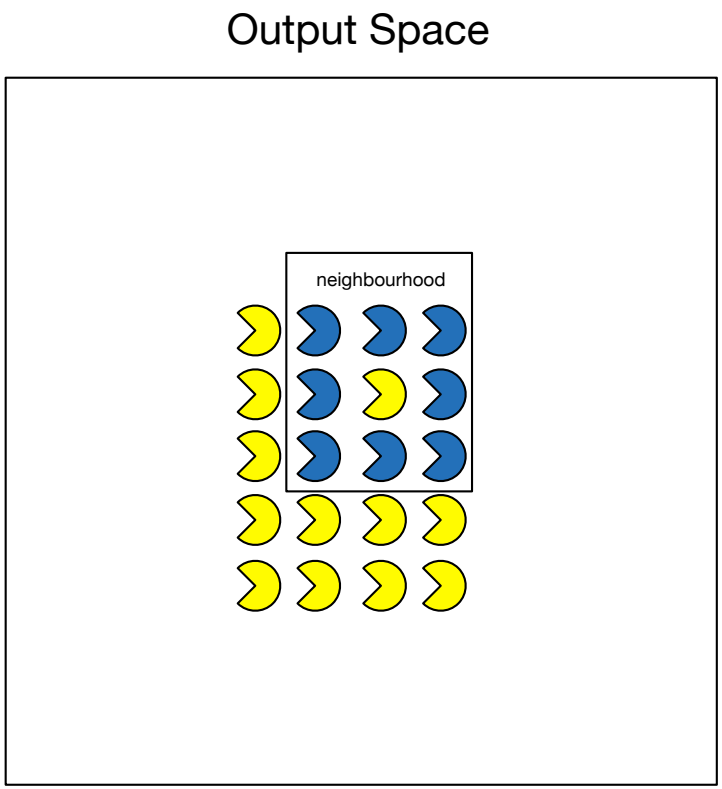
The Self Organizing Map

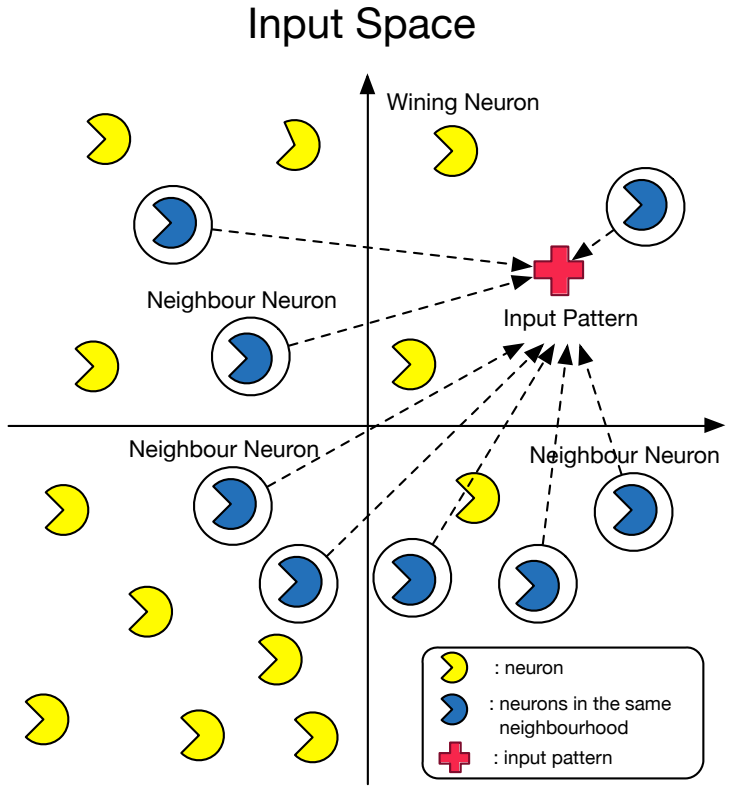
- A self-organizing map is a type of artificial neural network that is trained using *unsupervised* learning to produce a low-dimensional representation of the input space of the training samples, called a map.
- Self-organizing maps use a neighborhood function to preserve the topological properties of the input space.
- Mimics the way the cortex of highly developed animals brain (are supposed to) work.

SOM Input Space



SOM Output Space





Outline

The GEO SOM

GEO SOM

- Applies the first law of geography "Everything is related to everything else, but near things are more related than distant things".
- Defining a variable k which is used as a "geographical tolerance" that forces the winning neuron to be geographically near the input pattern.
- selection of the winning neuron is done in two steps. First, geographic neurons inside the tolerance k with the input data as a center are selected. Only after that, comparisons are made with the rest of data present in the input data.

WEBSOM

The WEBSOM self organizing maps

- First SOM is called *word category map* used to find words that have similar meaning.
- Second SOM, called *text*, used to cluster the documents.

Twitter Natural Language Processing

ARK Tweet NLP:

- Built using a maximum entropy Markov model.
- Tags words, such as nouns, verbs, etc..
- Can tag words, such as abbreviations, emojis and spelling errors.

Example:

ikr	smh	he	asked	fir	yo	last
!	G	O	V	P	D	A
name	so	he	can	add	u	on
N	P	O	V	V	O	P
fb	lolololol					
^	!					

Homophily in Social Networks

Miller et al, 2001

- Similarity breeds connection.
- Homophily means *"people like us*.
- In diverse societies, race, and race-like ethnicity create the most stark divides.
- Sex, age, religion, and education strongly structure our relations with others.

Outline

VSM Conversion

Problems with direct convertion:

- Words without meaningful content where used.
- Similar words where categorized differently.
- Symbols near words would render different word.

VSM Reduction with String Reducers 1

Tweets:

1. OMG these 2 cats are so adorable!! http://bit.ly/edThyy http://bit.ly/edThyy
2. The adorability of those caaaaats is to much
3. OOOOMG this CAT is adorable

Total number of unique words: 21

1) Remove URL

Tweets:

1. OMG these 2 cat are so adorable!! http://bit.ly/edThyy http://bit.ly/edThyy
2. The adorability of those caaaaats is to much
3. OOOOMG this CAT is adorable

Total number of unique words: 19

2) Remove non letters

VSM Reduction with String Reducers 2

Tweets:

1. OMG these **2** cat are so adorable!!
2. The adorability of those caaaaats is to much
3. OOOOMG this CAT is adorable

Total number of unique words: 17

3) Downcase

Tweets:

1. omg these cat are so adorable
2. The adorability of those caaaaats is to much
3. oooomg this cat is adorable

Total number of unique words: 16

4) Squeeze

VSM Reduction with String Reducers 3

Tweets:

1. omg these cat are so adorable
2. The adorability of those cats is to much
3. omg this cat is adorable

Total number of unique words: 15

5) Remove Small Words

Tweets:

1. omg these cat are so adorable
2. The adorability of those cats is to much
3. omg this cat is adorable

Total number of unique words: 11

6) Remove Stop Words

VSM Reduction with String Reducers 4

Tweets:

1. omg these cat are adorable
2. The adorability those cats much
3. omg this cat adorable

Total number of unique words: 5

7) Stem

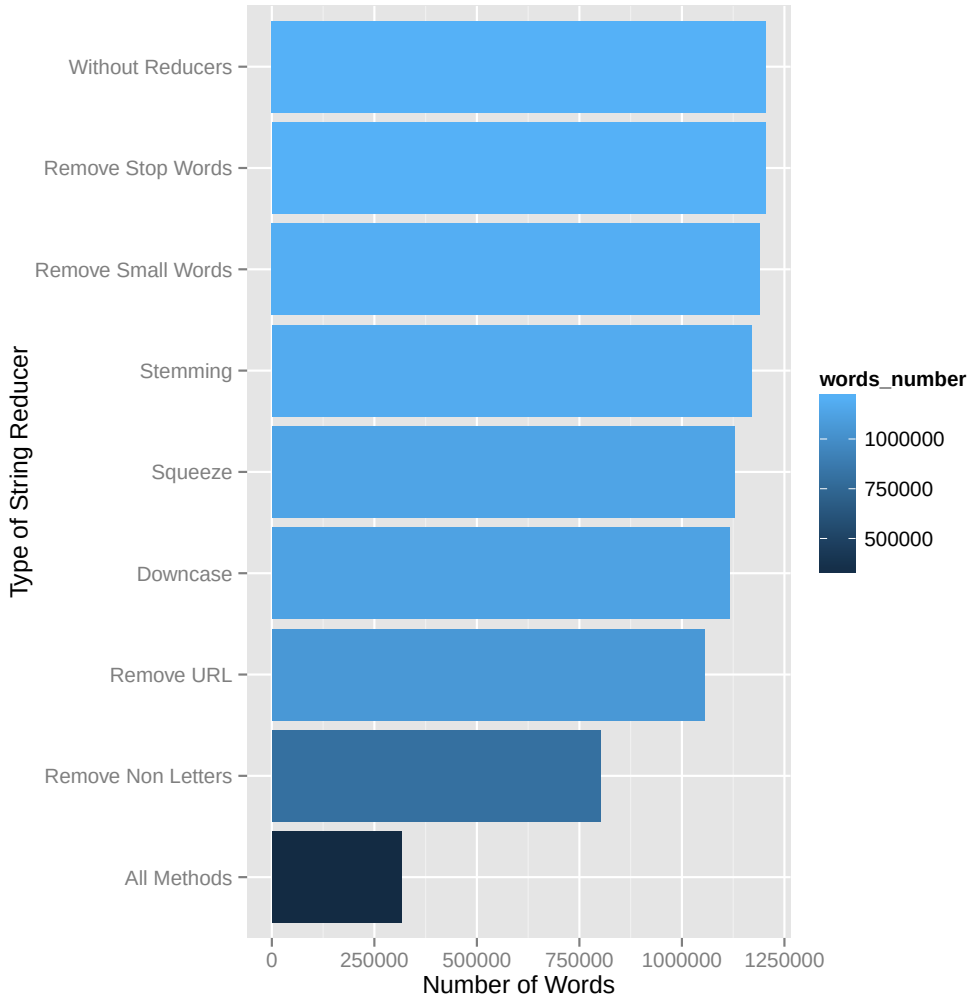
Tweets:

1. omg cat adorable
2. adorability cats
3. omg cat adorable

Total number of unique words: 3

VSM Reduction Results

75% reduction of total unique words



VSM Reduction with NLP 1

Tweets:

1. OMG these 2 cats are so adorable!! http://bit.ly/edThyy http://bit.ly/edThyy
2. The adorability of those caaaaats is to much #cat
3. OOOOMG this CAT is adorable

Total number of unique words: 22

1) NLP find :
- proper nouns
- common nouns
- hashtags

Tweets:

1. OMG these 2 cats are so adorable!! http://bit.ly/edThyy http://bit.ly/edThyy
2. The adorability of those caaaaats is to much #cat
3. OOOOMG this CAT is adorable

Total number of unique words: 4

2) Running all string reduction techniques on the tagged words

VSM Reduction with NLP 2

Tweets:

1. cat
2. cat cat
3. cat

Total number of unique words: 1

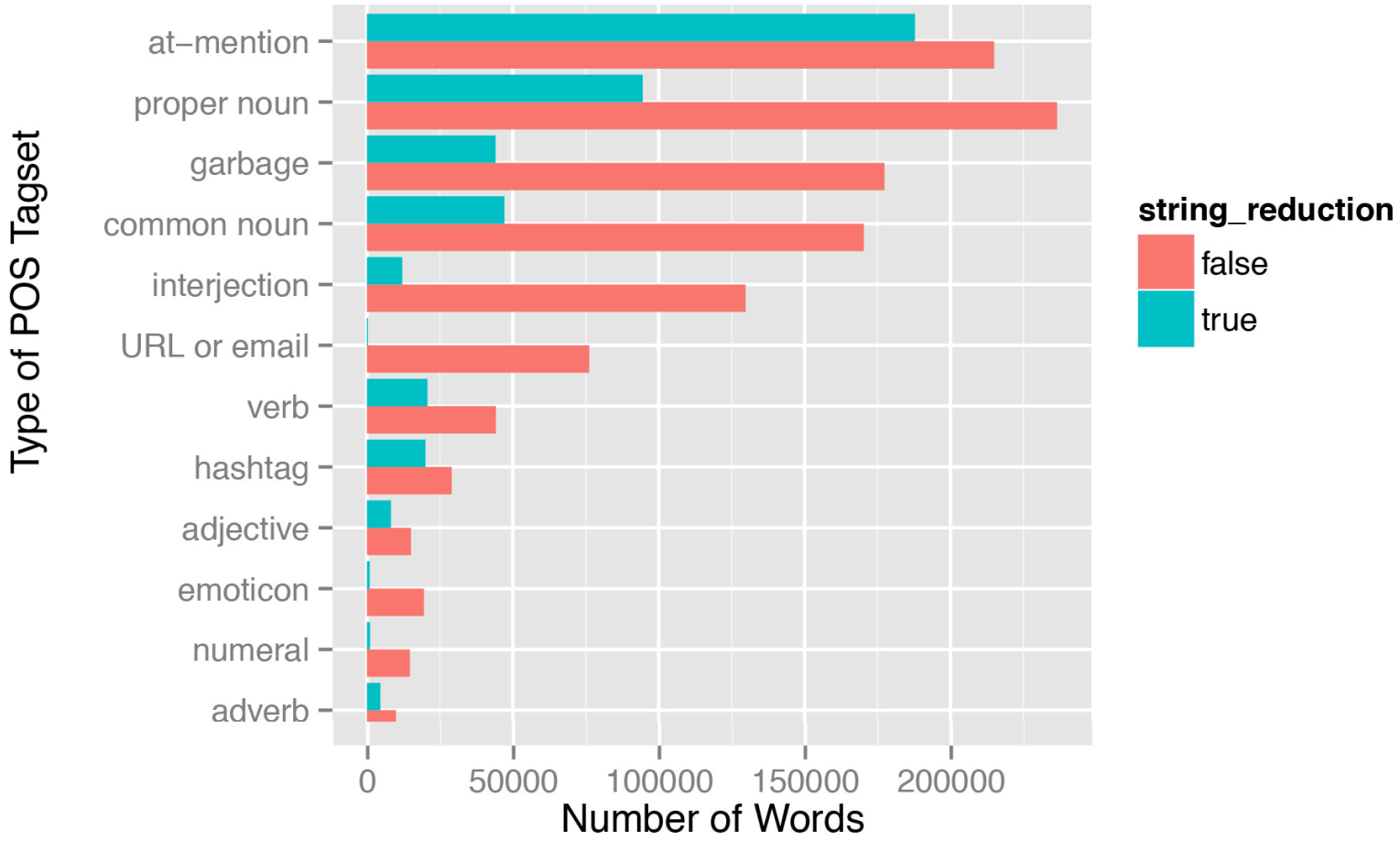
3) Conversion to VSM

VSM:

	cat
1.	1
2.	1
3.	1

VSM Reduction Results

95% reduction of total unique words



Outline

1 Introduction

2 Related Work
SOMs
Twitter

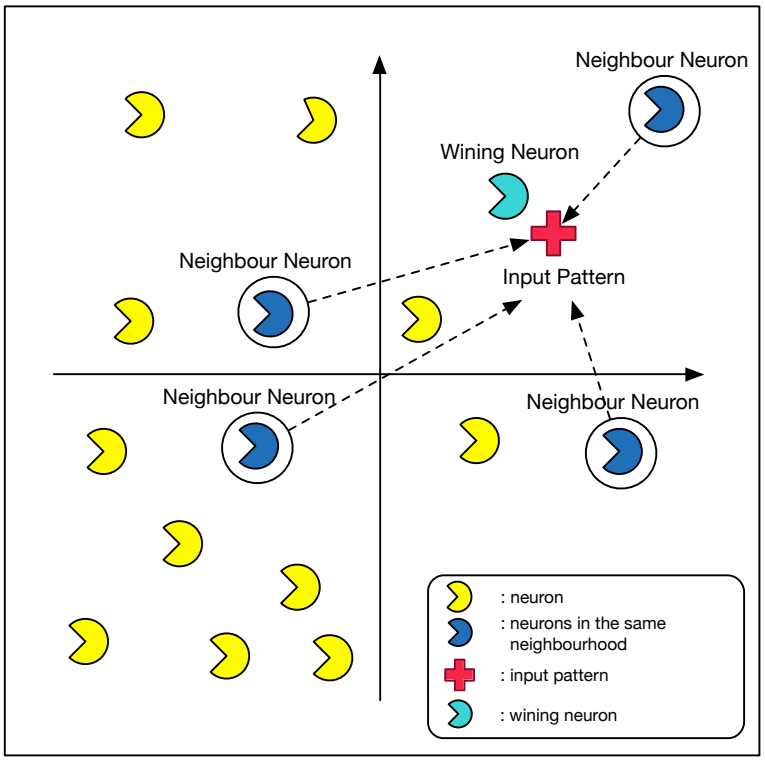
3 Clustering Tweets with SOMs

4 Enhancing SOM for socially connected data

5 Conclusions and Future Work

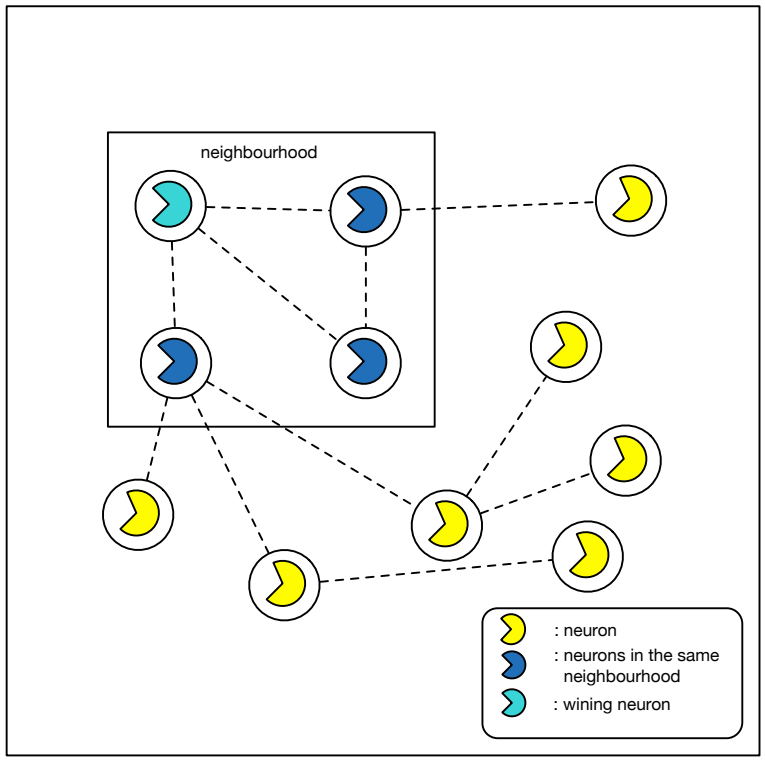
Homophilic SOM

Homophilic SOM Input Space



Homophilic SOM

Homophilic SOM Output Space



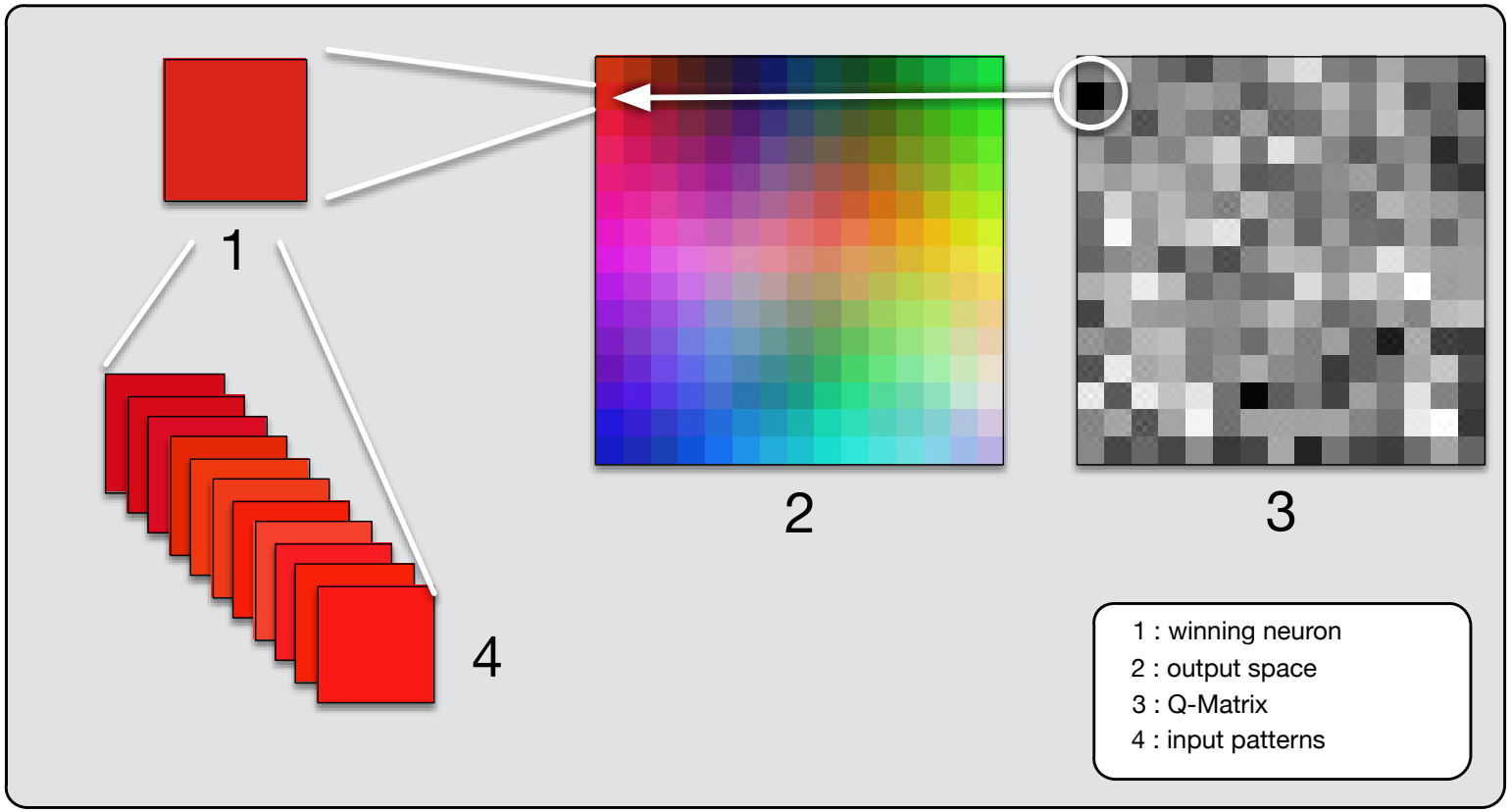
SOM Framework

Why develop another SOM library?

- Current SOM libraries don't allow the neighborhood function to be defined before training.
- A lot of customizations to the SOM algorithm where made, and each has its own code.
- SOM implementations on higher level languages rely on C.
- Easy to test new SOM approaches.

Testing SOM Framework

Using the SOM framework to train a SOM to identify different colors:



Homophilic SOM Implementation

- On top of the SOM framework.
- Changed the Output Space class, which was no longer squared.
- Everything keeps working as it was supposed to.

Homophilic SOM Results

- I think I haven't had a segmentation fault in years
- Just bought a banana phone at #bananamarket
- Real Software Engineering by @glv via @confreaks. @daviddias you're going to enjoy this (it is not about ruby)
- R vrs SAS, interesting debate:
- I'm finding @duckduckgo to be pretty more reliable than google when searching for code. Gonna try it as my default se.
- Mind blowing results! Taking Commercial 3DP into the Nano Dimension - #3DPrinting — @scoopit

Outline

Conclusions

- Homophilic SOM yielded some interesting results.
- Able to greatly reduce VSM size for tweets.
- Developed a SOM framework, to create custom SOMs.

Future Work

- Improve results display on the SOM framework.
- Improve SOM framework performance.
- Improve Homophilic SOM performance and result analysis.

Thank you!

Questions?