NYU
DATA
SCIENCE

word2vec and friends

https://github.com/bmtgoncalves/word2vec-and-friends/

Bruno Gonçalves
www.bgoncalves.com

Teaching machines to read!

- Computers are really good at crunching numbers but not so much when it comes to words.

- Perhaps can we represent words numerically?

a	1
about	2
above	3
after	4
again	5
against	6
all	7
am	8
an	9
and	10
any	11
are	12
aren't	13
as	14
…	…

Teaching machines to read!

- Computers are really good at crunching numbers but not so much when it comes to words.

- Perhaps can we represent words numerically?

$$v_{after} = (0, 0, 0, 1, 0, 0, \cdots)^T$$
$$v_{above} = (0, 0, 1, 0, 0, 0, \cdots)^T$$

One-hot encoding

- Can we do it in a way that preserves semantic information?

"You shall know a word by the company it keeps"
(J. R. Firth)

- Words that have similar meanings are used in similar contexts and the context in which a word is used helps us understand it's meaning.

The red house is beautiful.
The blue house is old.
The red car is beautiful.
The blue car is old.

Teaching machines to read!

"You shall know a word by the company it keeps"
(J. R. Firth)

➡Words with similar meanings should have similar representations.

➡From a word we can get some idea about the context where it might appear

___ ___ house __ ____.
___ ___ car __ ______.

$$\max p\left(C|w\right)$$

➡And from the context we have some idea about possible words

The red _____ is beautiful.
The blue _____ is old.

$$\max p\left(w|C\right)$$

word2vec

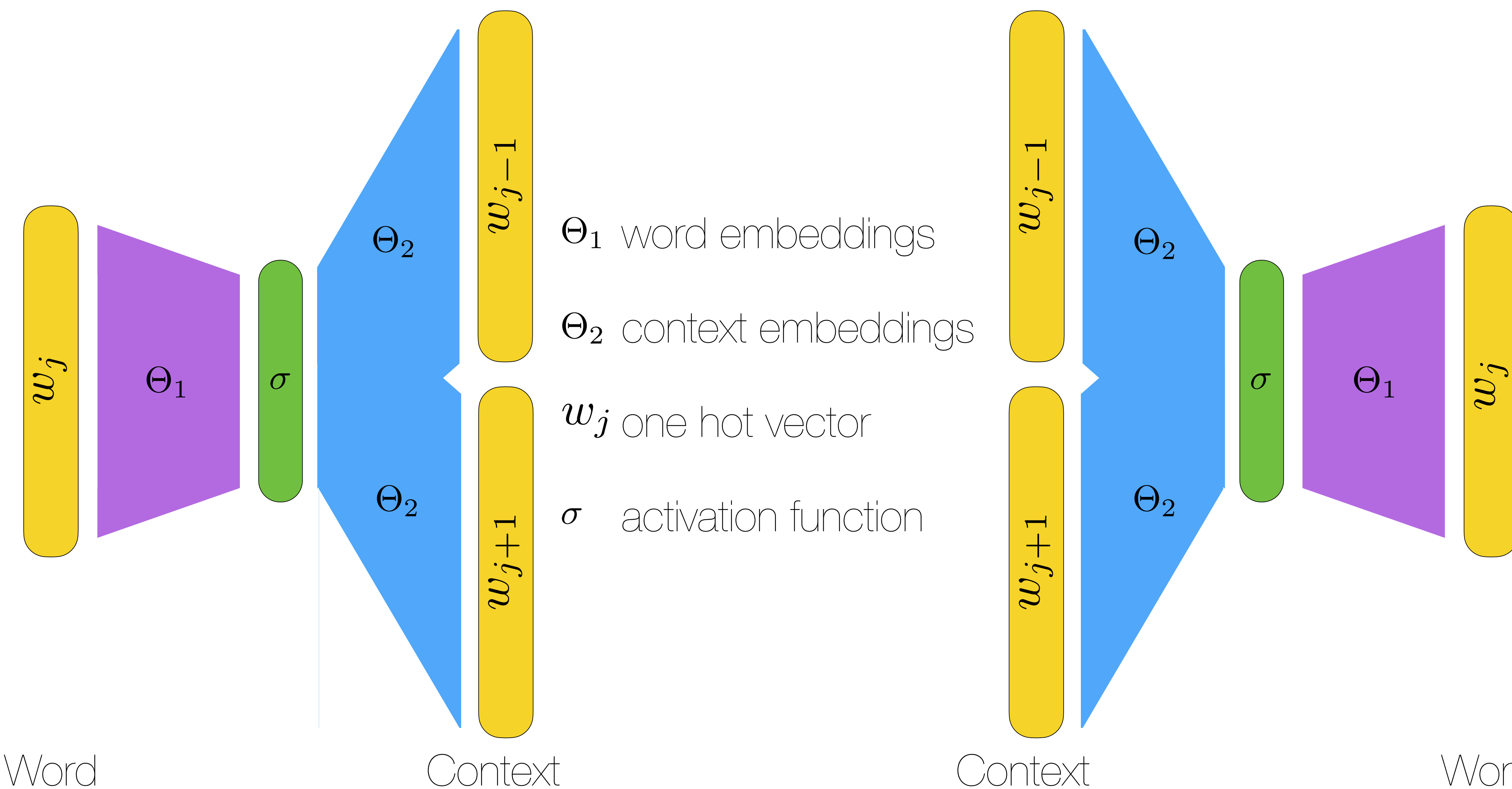
Mikolov 2013

Skipgram

$$\max p\left(C|w\right)$$

Continuous Bag of Words

$$\max p\left(w|C\right)$$



Skipgram

- Let us take a better look at a simplified case with a single context word
- Words are one-hot encoded vectors $w_j = (0,0,1,0,0,0,\cdots)^T$ of length V
- Θ_1 is an $(M \times V)$ matrix so that when we take the product:

$$\Theta_1 \cdot w_j$$

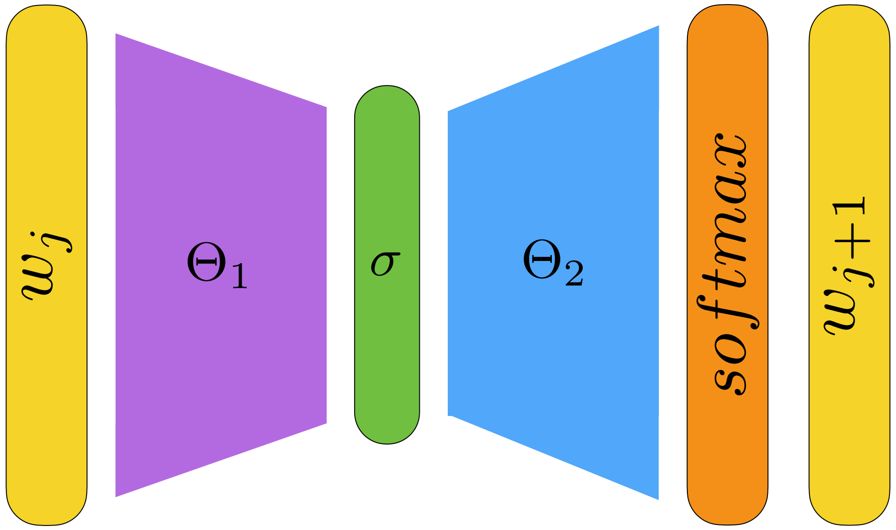
- We are effectively selecting the j'th column of Θ_1:

$$v_j = \Theta_1 \cdot w_j$$

- The linear activation function simply passes this value along which is then multiplied by Θ_2, a $(V \times M)$ matrix.
- Each element k of the output layer its then given by:

$$u_k^T \cdot v_j$$

- We convert these values to a normalized probability distribution by using the softmax



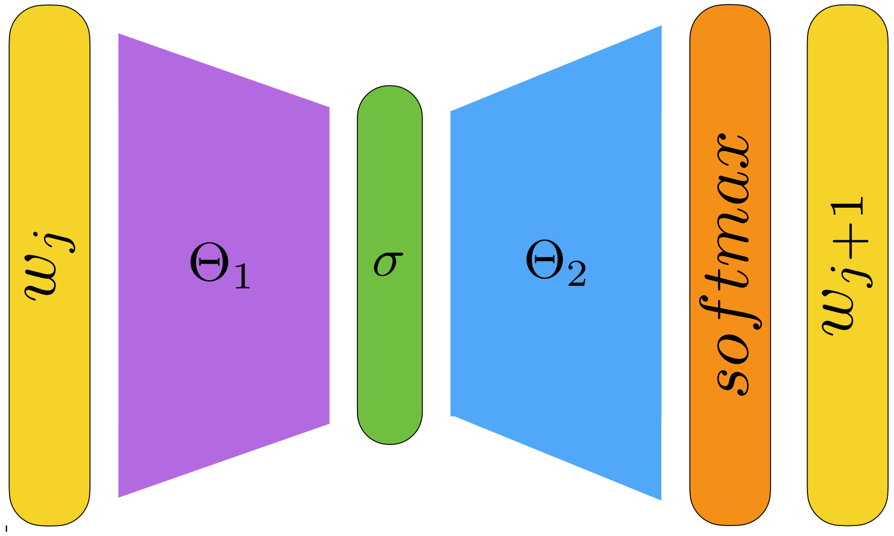
Softmax

- A standard way of converting a set of number to a normalized probability distribution:

$$softmax(x) = \frac{\exp(x_j)}{\sum_l \exp(x_l)}$$

- With this final ingredient we obtain:

$$p(w_k|w_j) \equiv softmax\left(u_k^T \cdot v_j\right) = \frac{\exp\left(u_k^T \cdot v_j\right)}{\sum_l \exp\left(u_l^T \cdot v_j\right)}$$



- Our goal is then to learn:

$$\Theta_1 \qquad \Theta_2$$

- so that we can predict what the next word is likely to be using:

$$p(w_{j+1}|w_j)$$

- But how can we quantify how far we are from the correct answer? Our error measure shouldn't be just binary (right or wrong)…

Cross-Entropy

- First we have to recall that what we are, in effect, comparing two probability distributions:

$$p\left(w_k|w_j\right)$$

- and the one-hot encoding of the context:

$$w_{j+1} = \left(0, 0, 0, 1, 0, 0, \cdots\right)^T$$

- The Cross Entropy measures the distance, in number of bits, between two probability distributions p and q:

$$H\left(p, q\right) = -\sum_k p_k \log q_k$$

- In our case, this becomes:

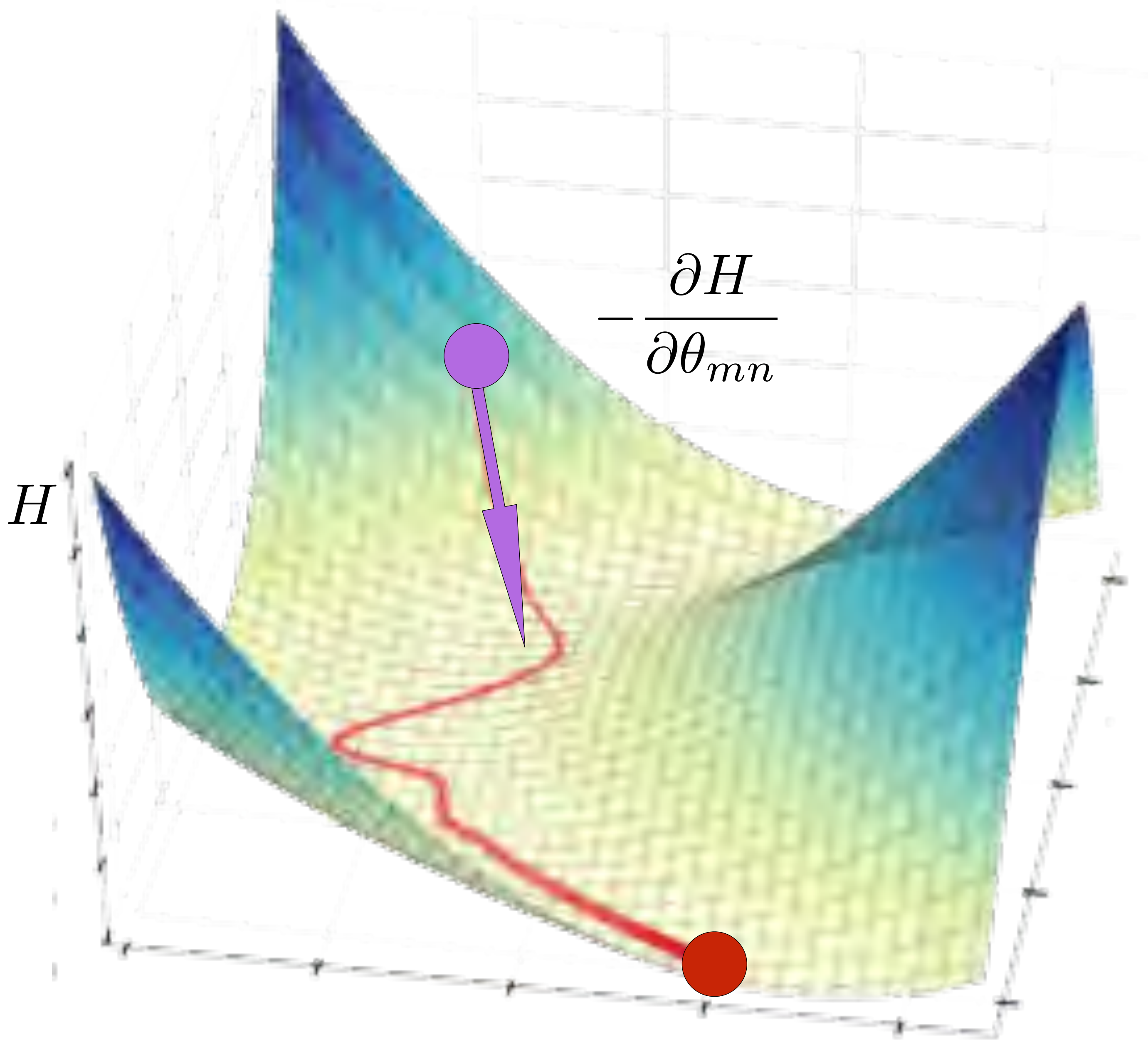
$$H\left[w_{j+1}, p\left(w_k|w_j\right)\right] = -\sum_k w_{j+1}^k \log p\left(w_k|w_j\right)$$

- So it's clear that the only non zero term is the one that corresponds to the "hot" element of w_{j+1}

$$H = -\log p\left(w_{j+1}|w_j\right)$$

- This is our Error function. But how can we use this to update the values of Θ_1 and Θ_2?

Gradient Descent



- Find the gradient for each training batch
- Take a step **downhill** along the direction of the gradient

$$\theta_{mn} \leftarrow \theta_{mn} - \alpha \frac{\partial H}{\partial \theta_{mn}}$$

- where α is the step size.
- Repeat until "convergence".

Chain-rule

- How can we calculate

$$\frac{\partial H}{\partial \theta_{mn}} = \frac{\partial}{\partial \theta_{mn}} \log p\left(w_{j+1} | w_j\right) \qquad \theta_{mn} = \left\{\theta_{mn}^{(1)}, \theta_{mn}^{(2)}\right\}$$

- we rewrite:

$$\frac{\partial H}{\partial \theta_{mn}} = \frac{\partial}{\partial \theta_{mn}} \log \frac{\exp\left(u_k^T \cdot v_j\right)}{\sum_l \exp\left(u_l^T \cdot v_j\right)}$$

- and expand:

$$\frac{\partial H}{\partial \theta_{mn}} = \frac{\partial}{\partial \theta_{mn}} \left[u_k^T \cdot v_j - \log \sum_l \exp\left(u_l^T \cdot v_j\right)\right]$$

- Then we can rewrite:

$$u_k^T \cdot v_j = \sum_q \theta_{kq}^{(2)} \theta_{qj}^{(1)}$$

- and apply the chain rule:

$$\frac{\partial f\left(g\left(x\right)\right)}{\partial x} = \frac{\partial f\left(g\left(x\right)\right)}{\partial g\left(x\right)} \frac{\partial g\left(x\right)}{\partial x}$$

Training procedures

- online learning - update weights after each case
  - might be useful to update model as new data is obtained
  - subject to fluctuations

- mini-batch - update weights after a "small" number of cases
  - batches should be balanced
  - if dataset is redundant, the gradient estimated using only a fraction of the data is a good approximation to the full gradient.

- momentum - let gradient change the velocity of weight change instead of the value directly

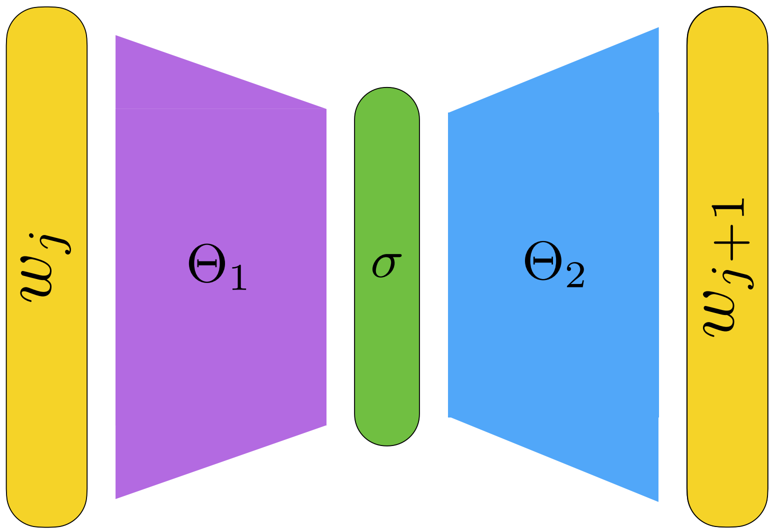
- rmsprop - divide learning rate for each weight by a running average of "recent" gradients

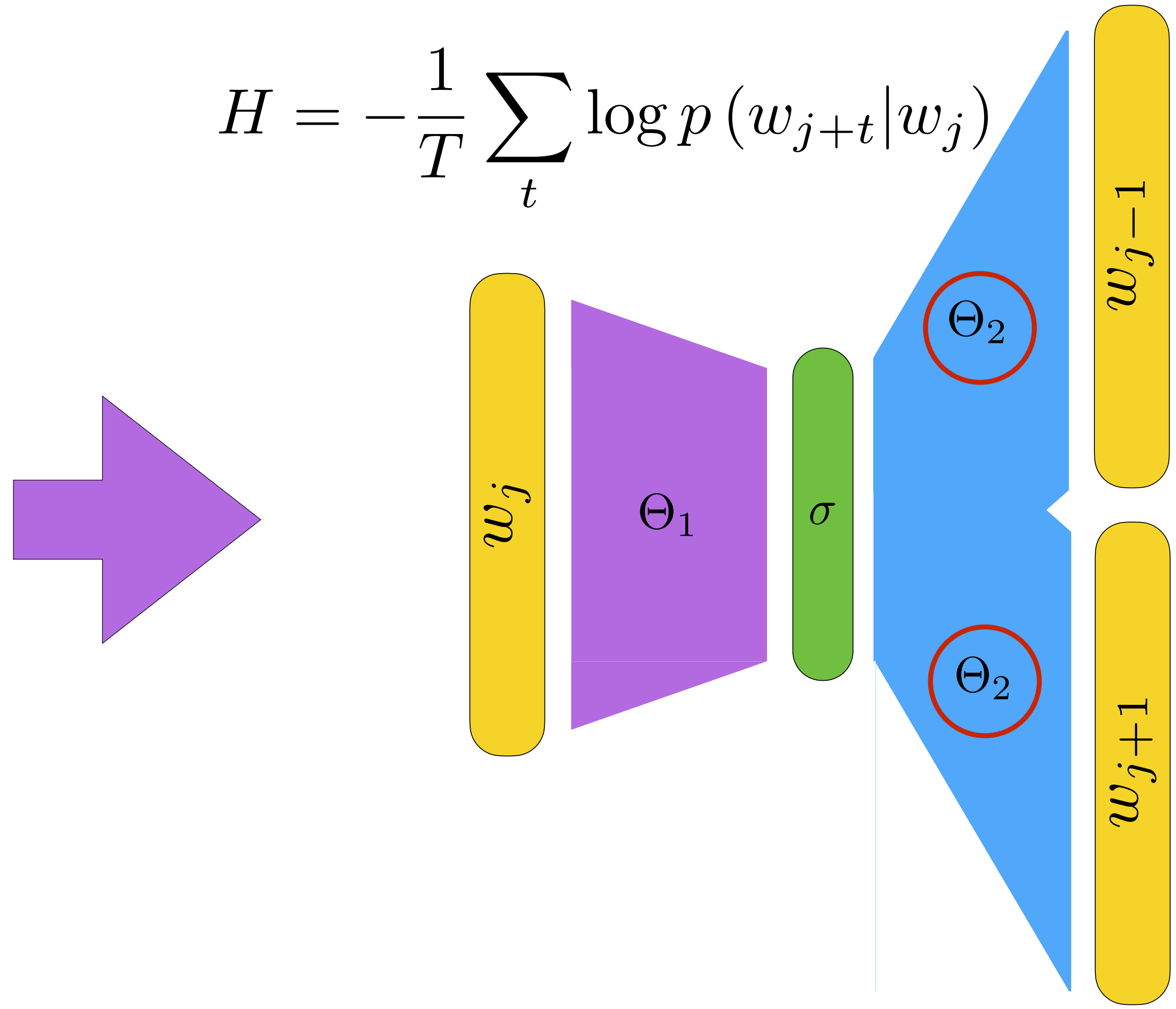
- learning rate - vary over the course of the training procedure and use different learning rates for each weight

SkipGram with Larger Contexts

$$H = -\log p\left(w_{j+1}|w_j\right)$$

$$H = -\frac{1}{T}\sum_t \log p\left(w_{j+t}|w_j\right)$$

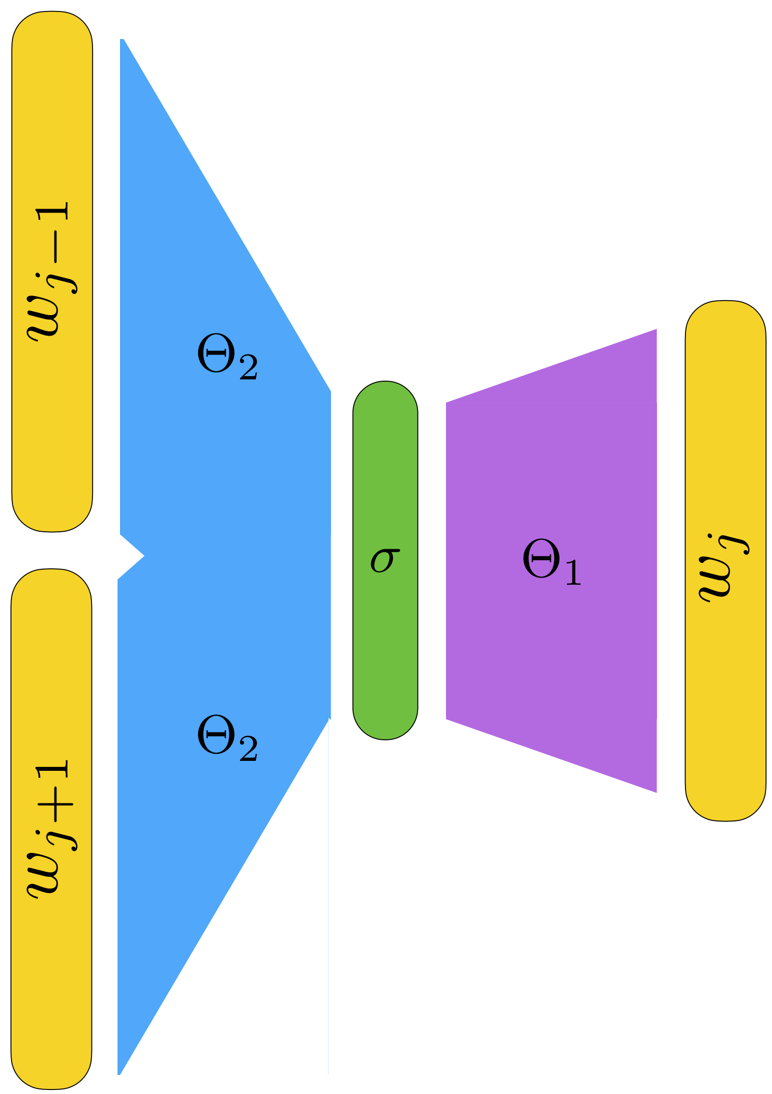




- Use the same Θ_2 for all context words.
- Use the average of cross entropy.
- word order is not important (the average does not change)
- Can essentially be trained one context word at at time..

Continuous Bag of Words

- The process is essentially the same

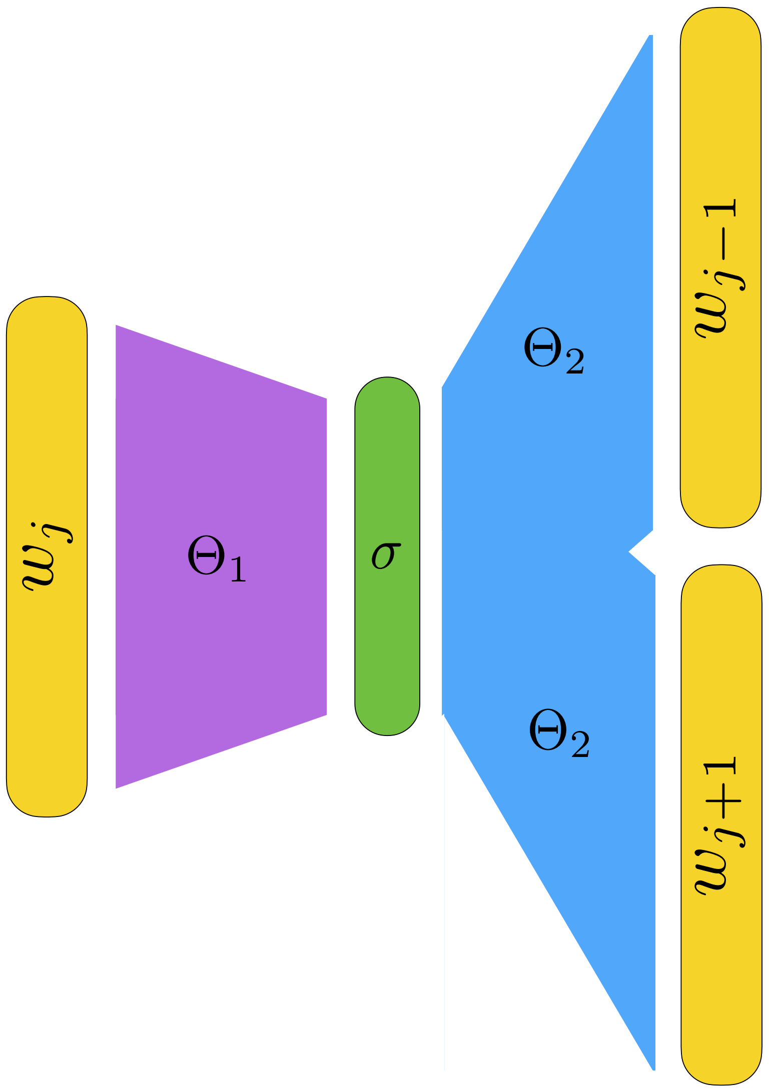


Variations

- Hierarchical Softmax:
    - Approximate the softmax using a binary tree
    - Reduce the number of calculations per training example from V to $\log_2 V$ and increase performance by orders of magnitude.
- Negative Sampling:
    - Under sample the most frequent words by removing them from the text before generating the contexts
    - Similar idea to removing stop-words — very frequent words are less informative.
    - Effectively makes the window larger, increasing the amount of information available for context

Comments

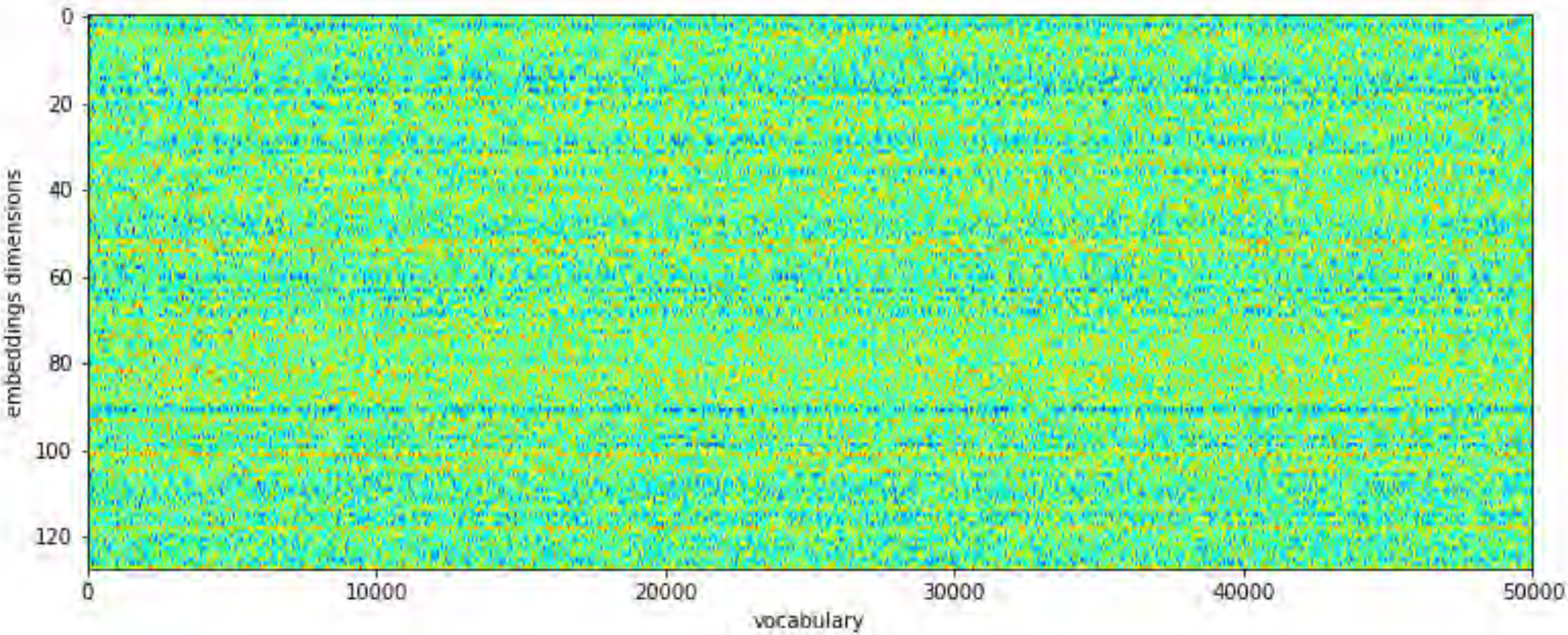
- word2vec, even in its original formulation is actually a family of algorithms using various combinations of:
  - Skip-gram, CBOW
  - Hierarchical Softmax, Negative Sampling
- The output of this neural network is deterministic:
  - If two words appear in the same context ("blue" vs "red", for e.g.), they will have similar internal representations in Θ_1 and Θ_2
  - Θ_1 and Θ_2 are vector embeddings of the input words and the context words respectively
- Words that are too rare are also removed.
- The original implementation had a dynamic window size:
  - for each word in the corpus a window size k' is sampled uniformly between 1 and k



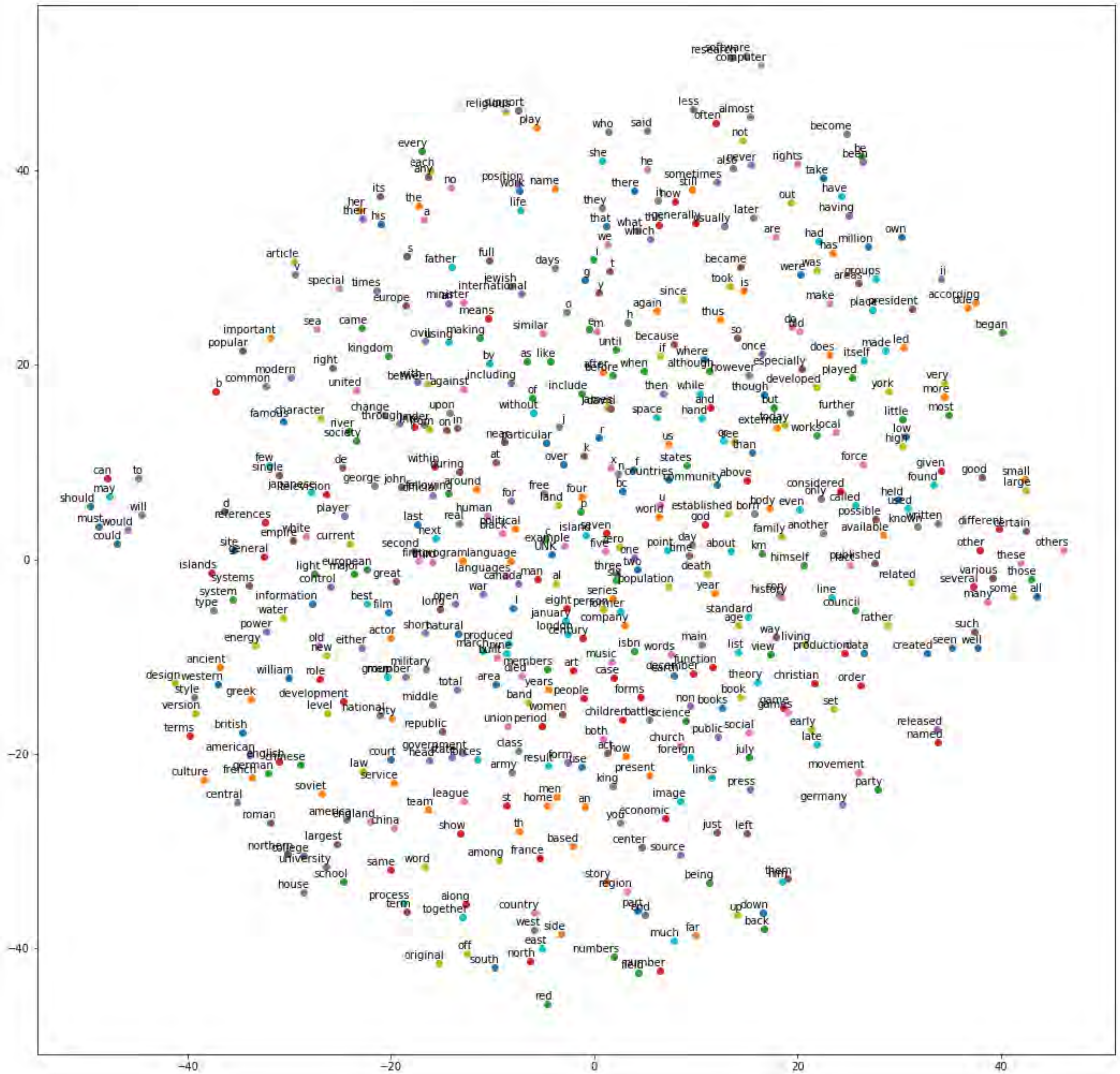
Online resources

- C - https://code.google.com/archive/p/word2vec/ (the original one)
- Python/tensorflow - https://www.tensorflow.org/tutorials/word2vec
    - Both a minimalist and an efficient versions are available in the tutorial
- Python/gensim - https://radimrehurek.com/gensim/models/word2vec.html
- Pretrained embeddings:
    - 30+ languages, https://github.com/Kyubyong/wordvectors
    - 100+ languages trained using wikipedia: https://sites.google.com/site/rmyeid/projects/polyglot

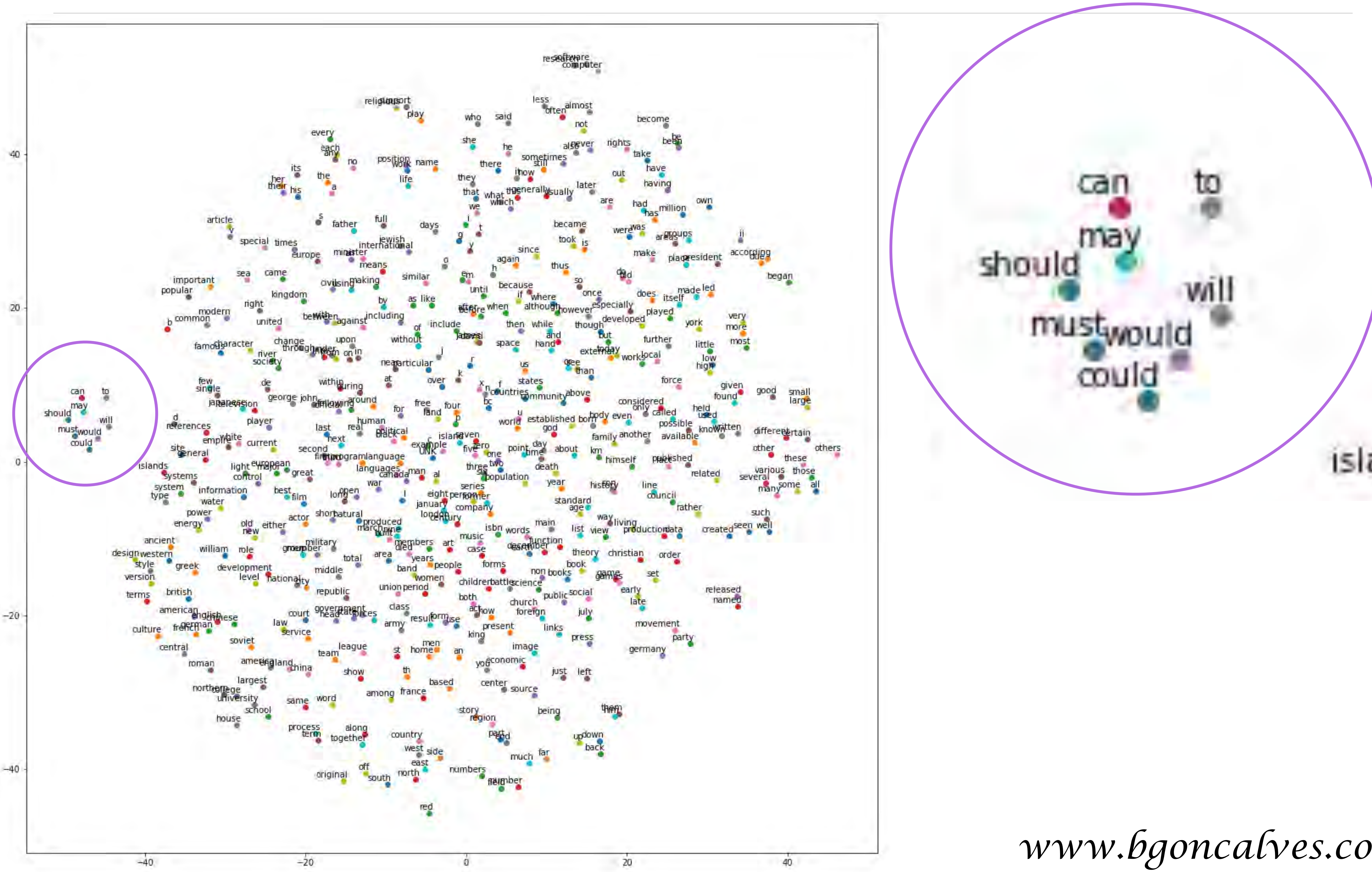
Visualization



Visualization

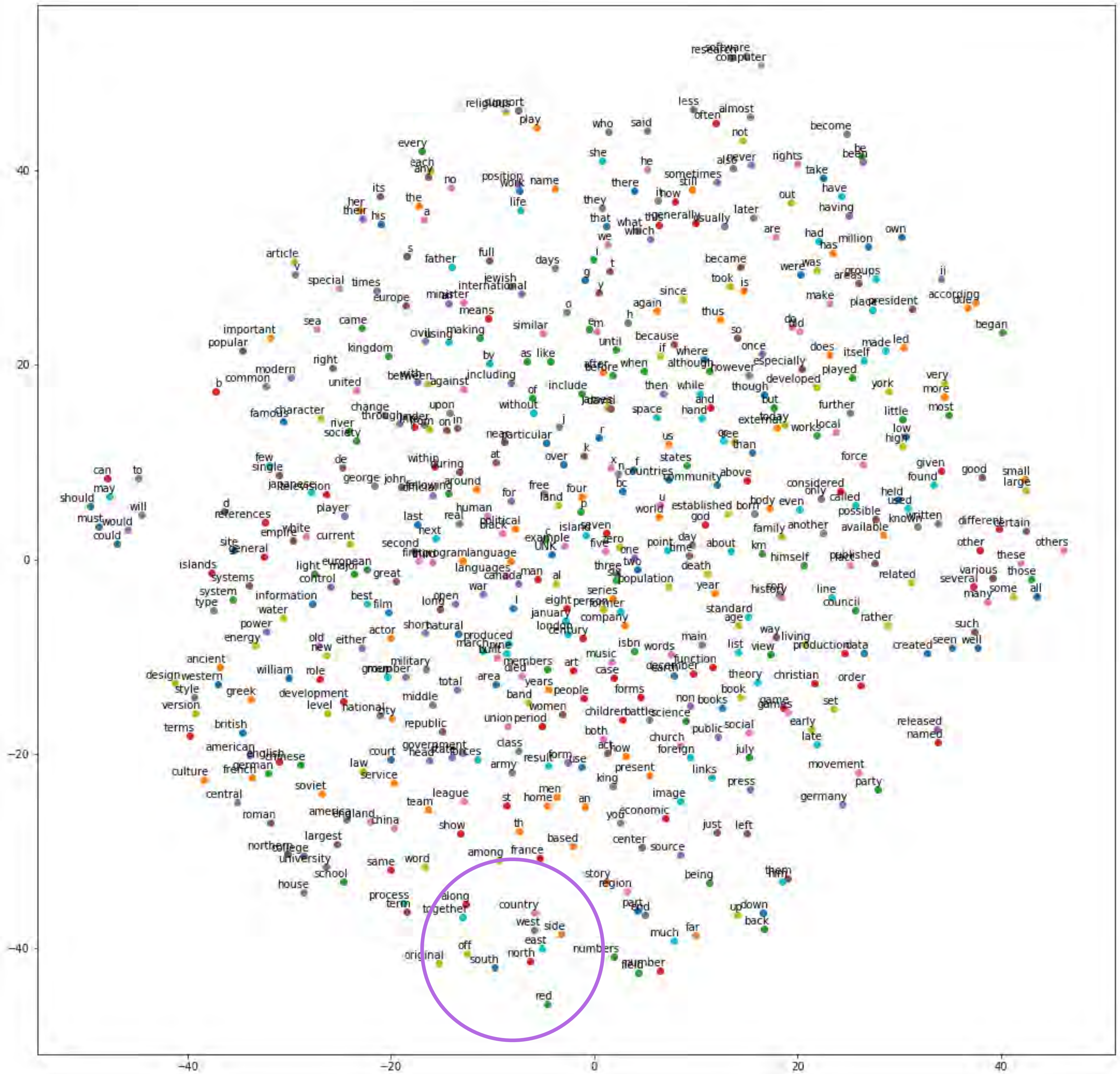


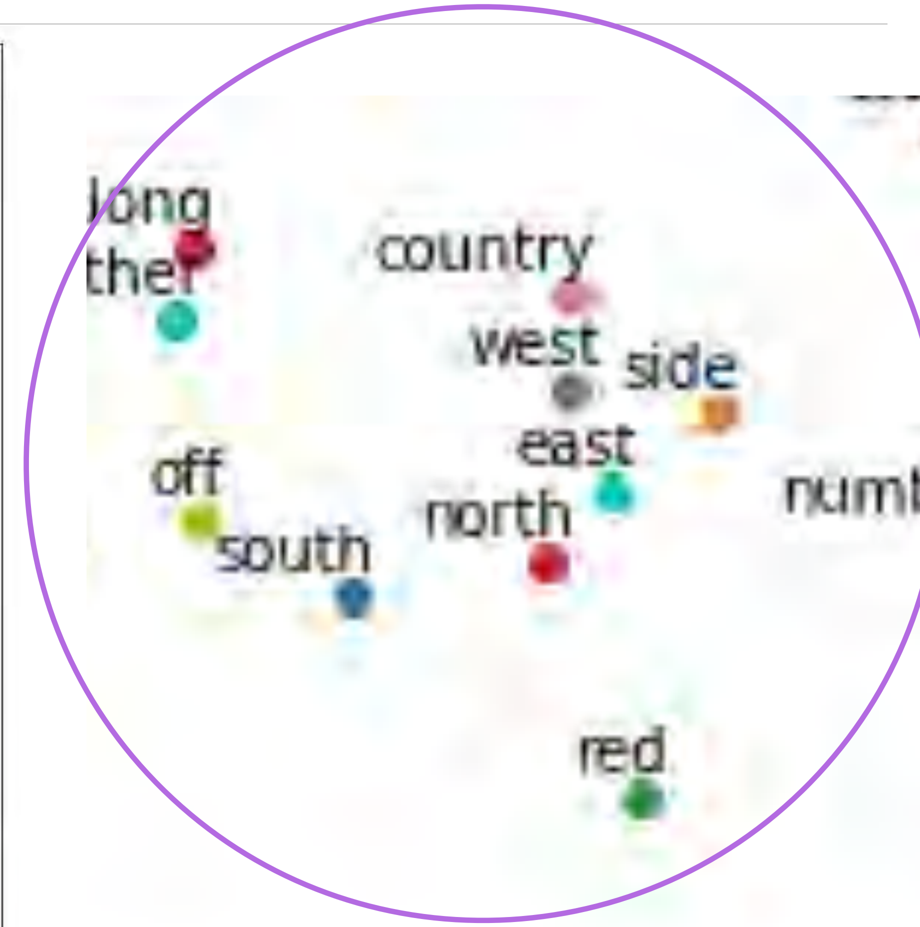
Visualization



www.bgoncalves.com

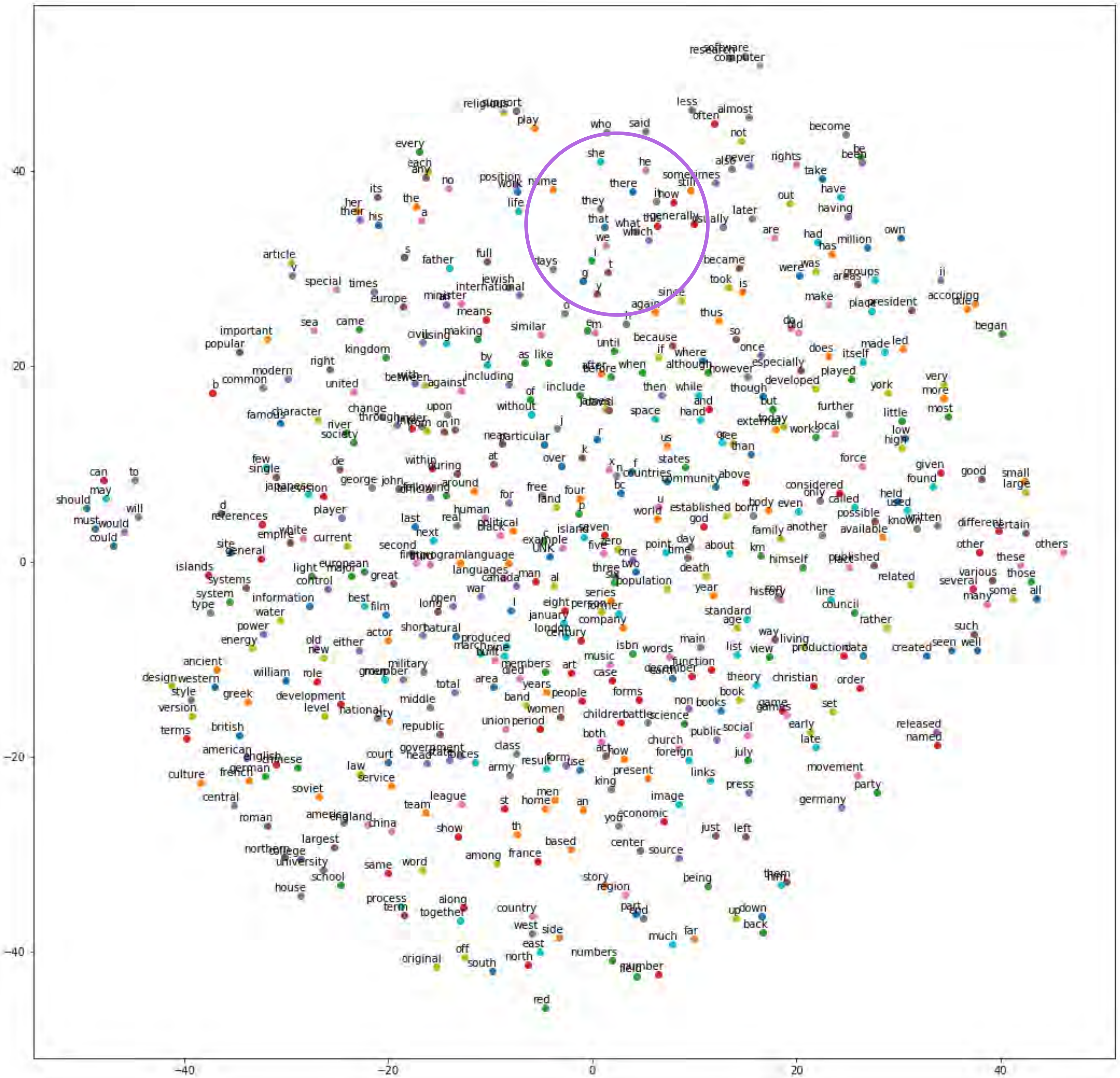
Visualization

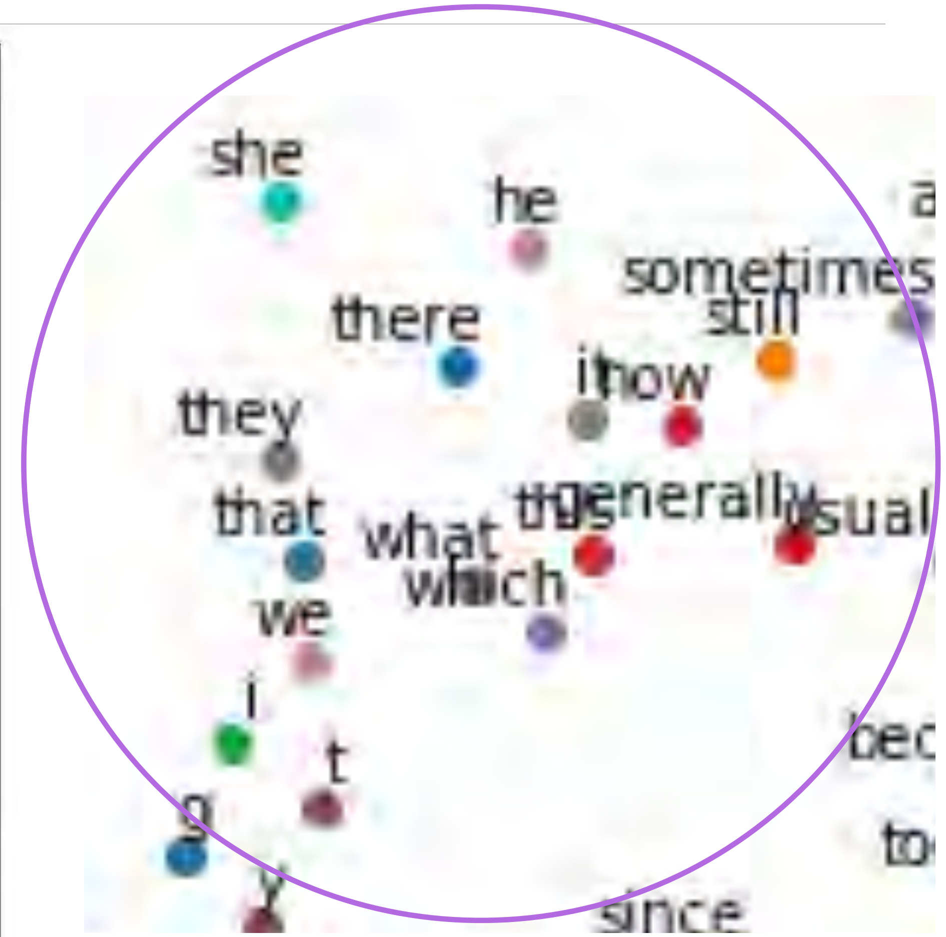




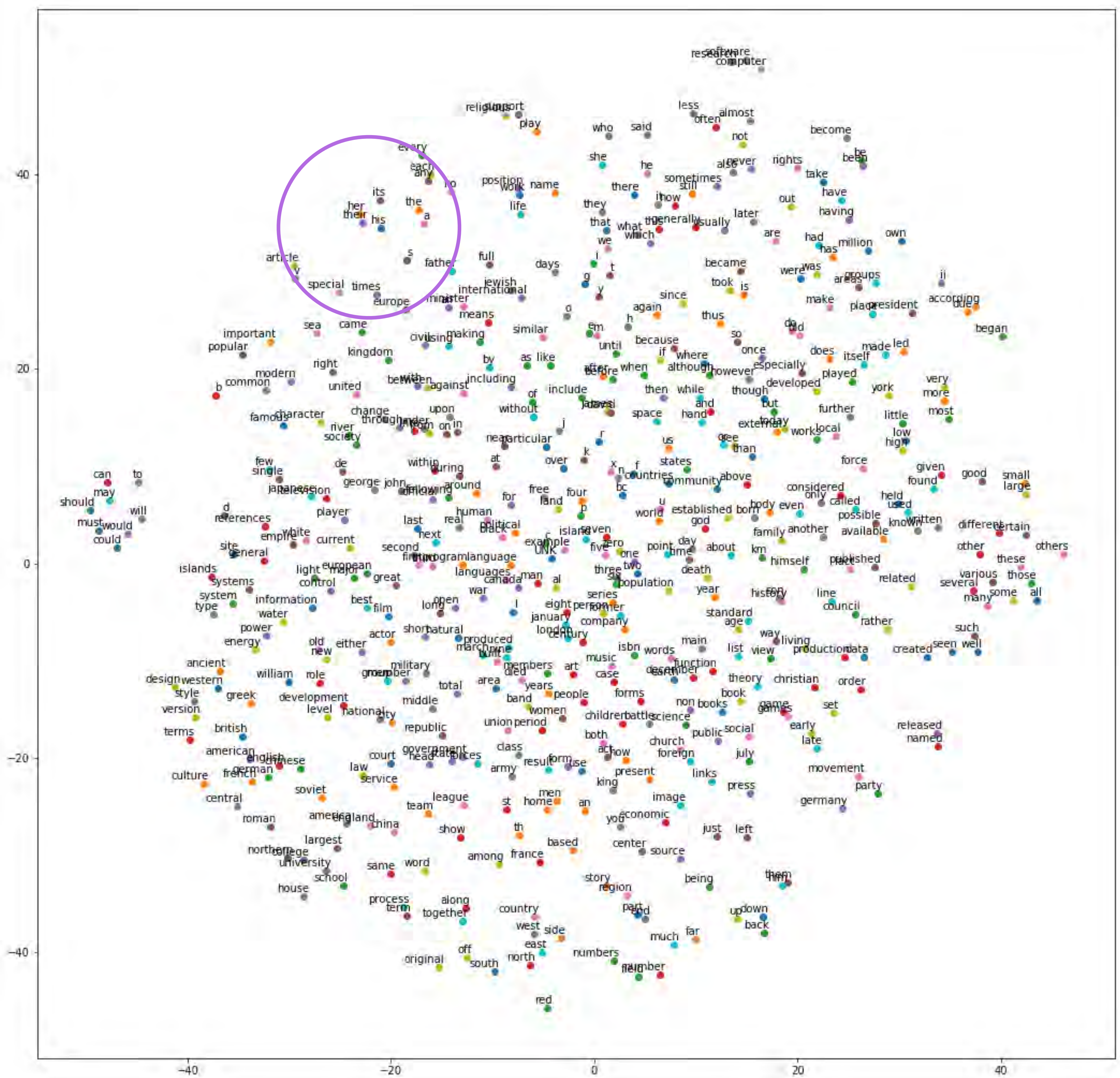
www.bgoncalves.com

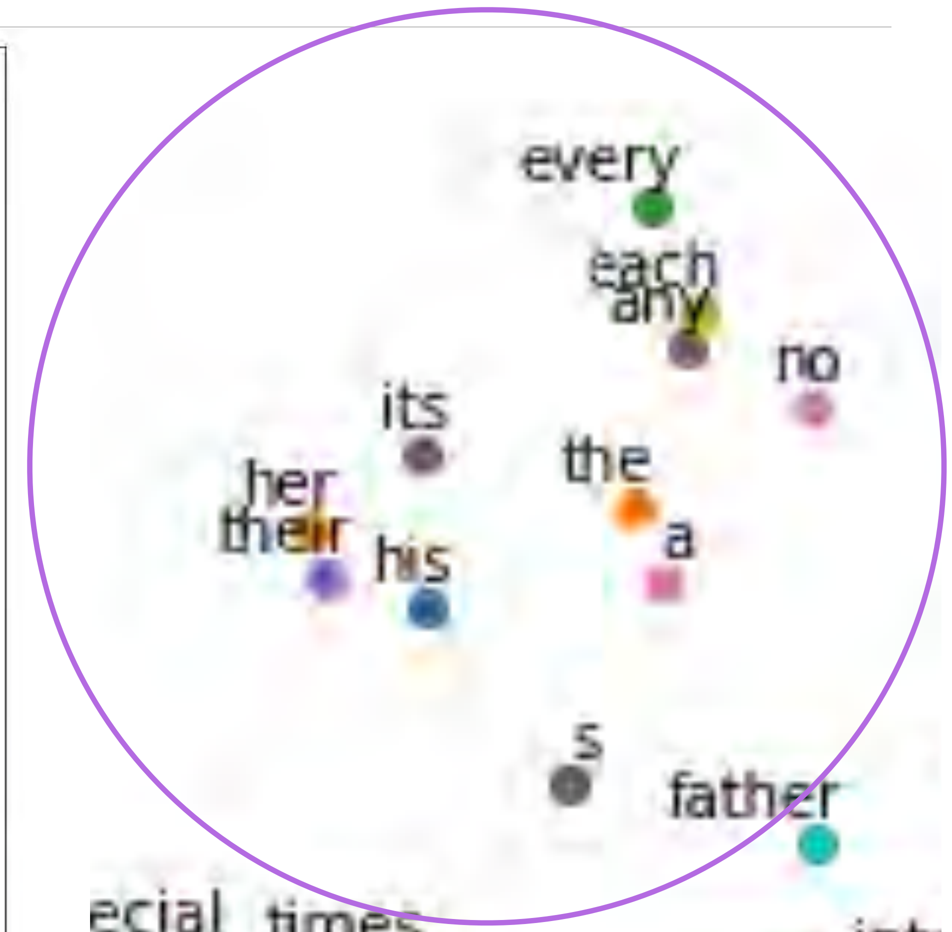
Visualization





Visualization

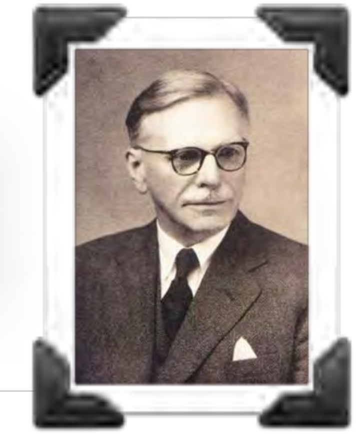




www.bgoncalves.com

Analogies

"You shall know a word by the company it keeps"
(J. R. Firth)



- The embedding of each word is a function of the context it appears in:

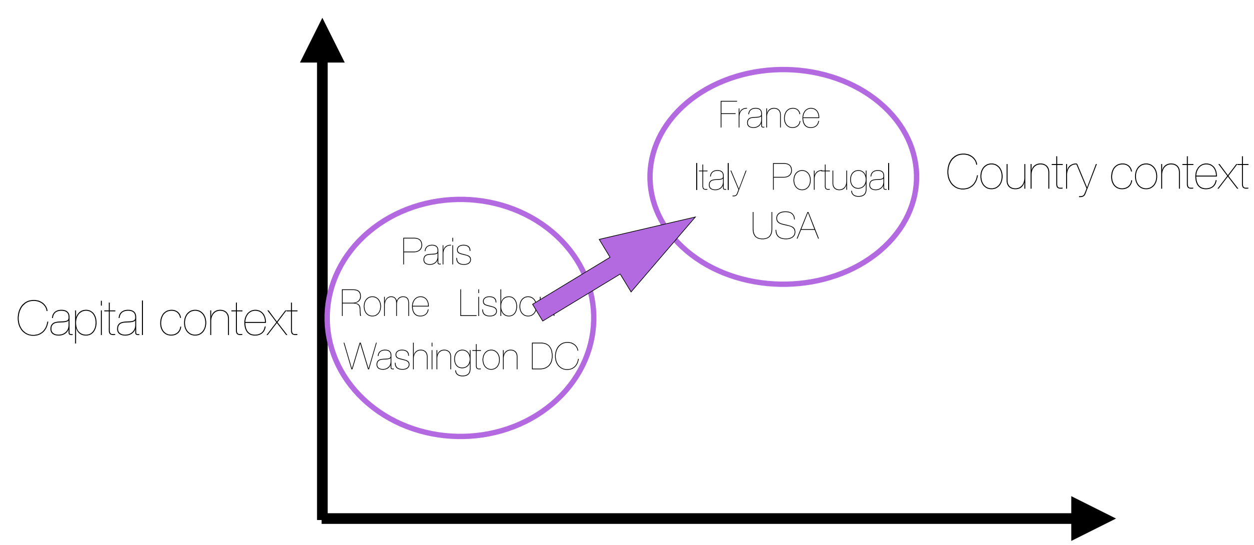
$$\sigma\left(red\right)=f\left(context\left(red\right)\right)$$

- words that appear in similar contexts will have similar embeddings:

$$context\left(red\right)\approx context\left(blue\right)\implies\sigma\left(red\right)\approx\sigma\left(blue\right)$$

- "Distributional hypotesis" in linguistics

Geometrical relations between contexts imply semantic relations between words!



$$\sigma\left(France\right)-\sigma\left(Paris\right)+\sigma\left(Rome\right)=\sigma\left(Italy\right)$$

$$\vec{b}-\vec{a}+\vec{c}=\vec{d}$$

Analogies

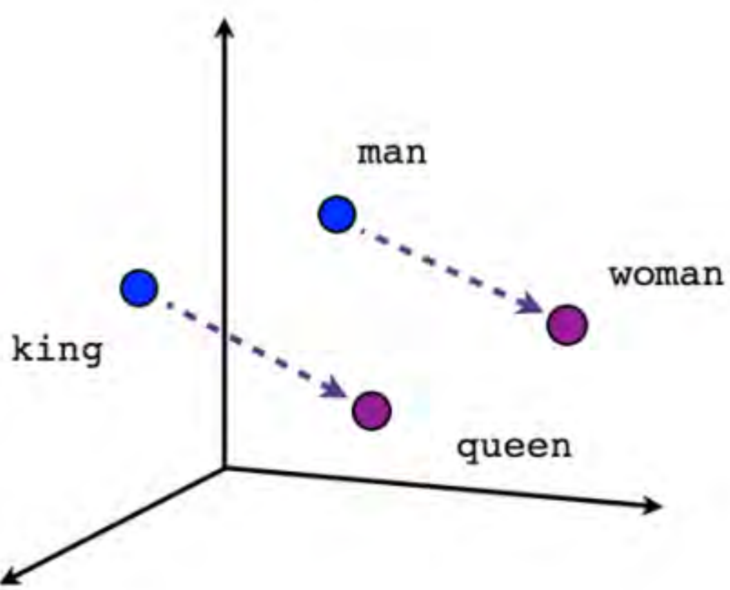
https://www.tensorflow.org/tutorials/word2vec

$$\vec{b} - \vec{a} + \vec{c} = \vec{d}$$

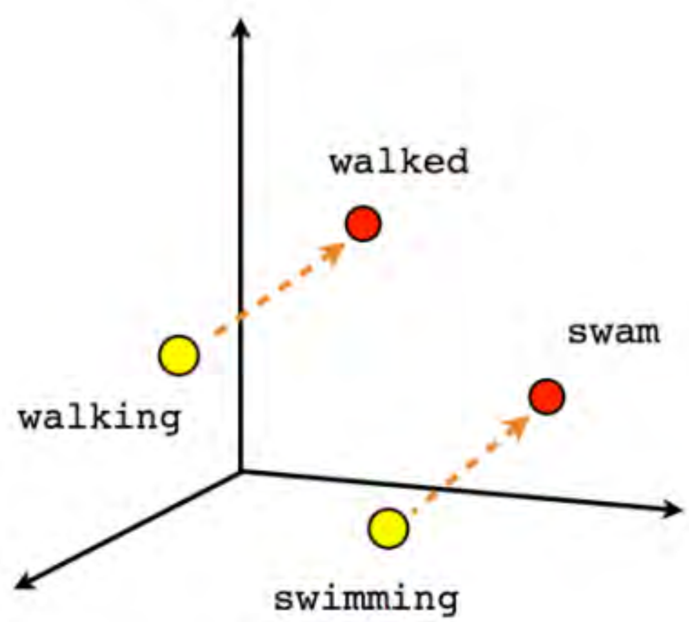
$$d^{\dagger} = \underset{x}{\operatorname{argmax}} \frac{\left(\vec{b} - \vec{a} + \vec{c}\right)^{T}}{\left|\left|\vec{b} - \vec{a} + \vec{c}\right|\right|} \vec{x}$$

$$d^{\dagger} \sim \underset{x}{\operatorname{argmax}} \left(\vec{b}^{T} \vec{x} - \vec{a}^{T} \vec{x} + \vec{c}^{T} \vec{x}\right)$$

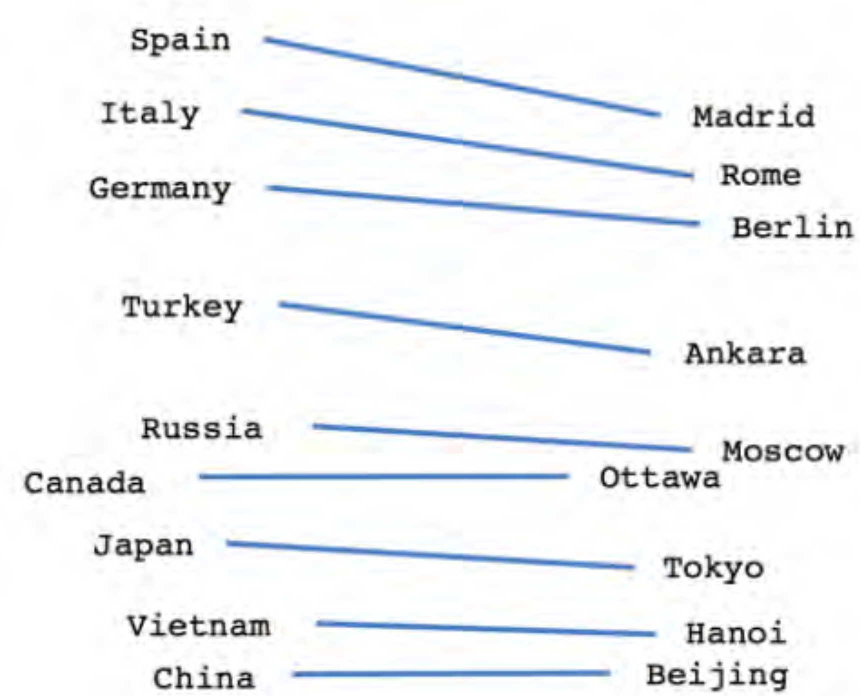
What is the word d that is most similar to b and c and most dissimilar to a?



Male-Female



Verb tense

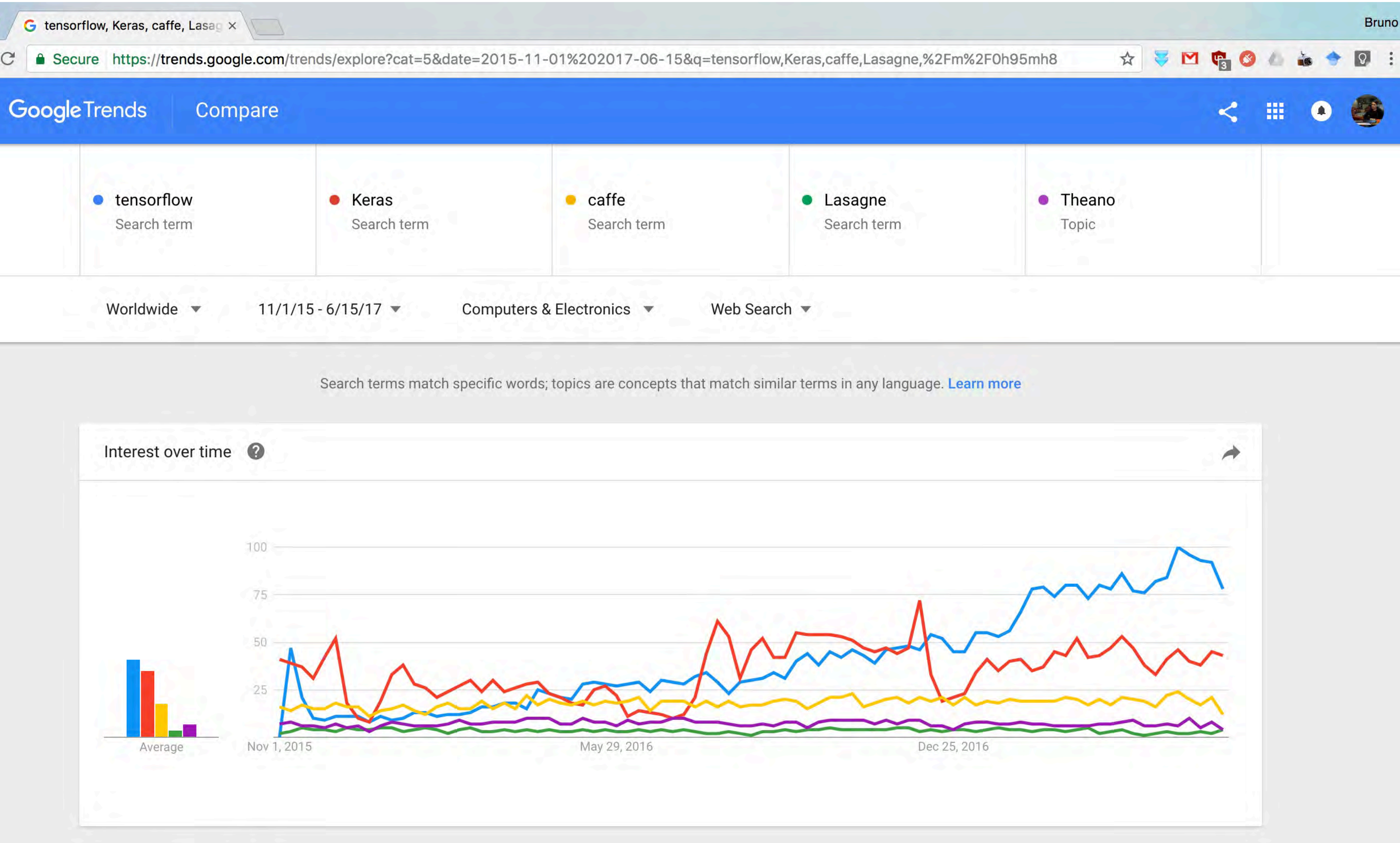


Country-Capital

TensorFlow

https://github.com/bmtgoncalves/word2vec-and-friends/

Tensorflow


tensorflow, Keras, caffe, Lasag
Bruno
Secure https://trends.google.com/trends/explore?cat=5&date=2015-11-01%202017-06-15&q=tensorflow,Keras,caffe,Lasagne,%2Fm%2F0h95mh8
Google Trends
Compare
tensorflow
Search term
Keras
Search term
caffe
Search term
Lasagne
Search term
Theano
Topic
Worldwide
11/1/15 - 6/15/17
Computers & Electronics
Web Search
Search terms match specific words; topics are concepts that match similar terms in any language. Learn more
Interest over time
100
75
50
25
Average
Nov 1, 2015
May 29, 2016
Dec 25, 2016

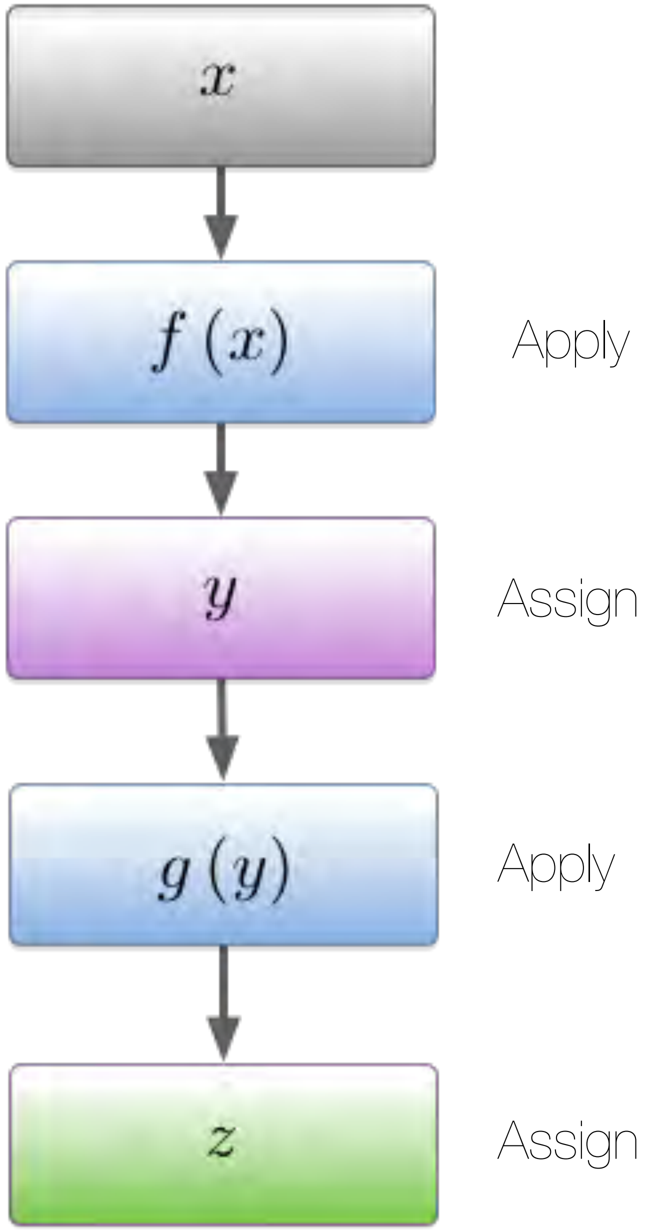
A diversion...

- Let's imagine I want to perform these calculations:

$$y = f(x)$$

$$z = g(y)$$

- for some given x .
- To calculate z we must follow a certain sequence of operations.



A diversion…

TensorFlow

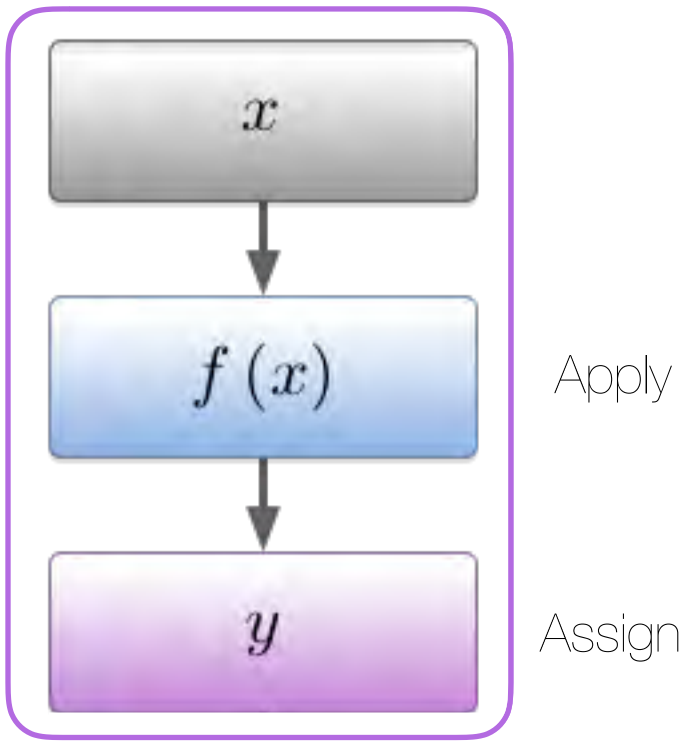
https://www.tensorflow.org/

- Let's imagine I want to perform these calculations:

$$y = f(x)$$

$$z = g(y)$$

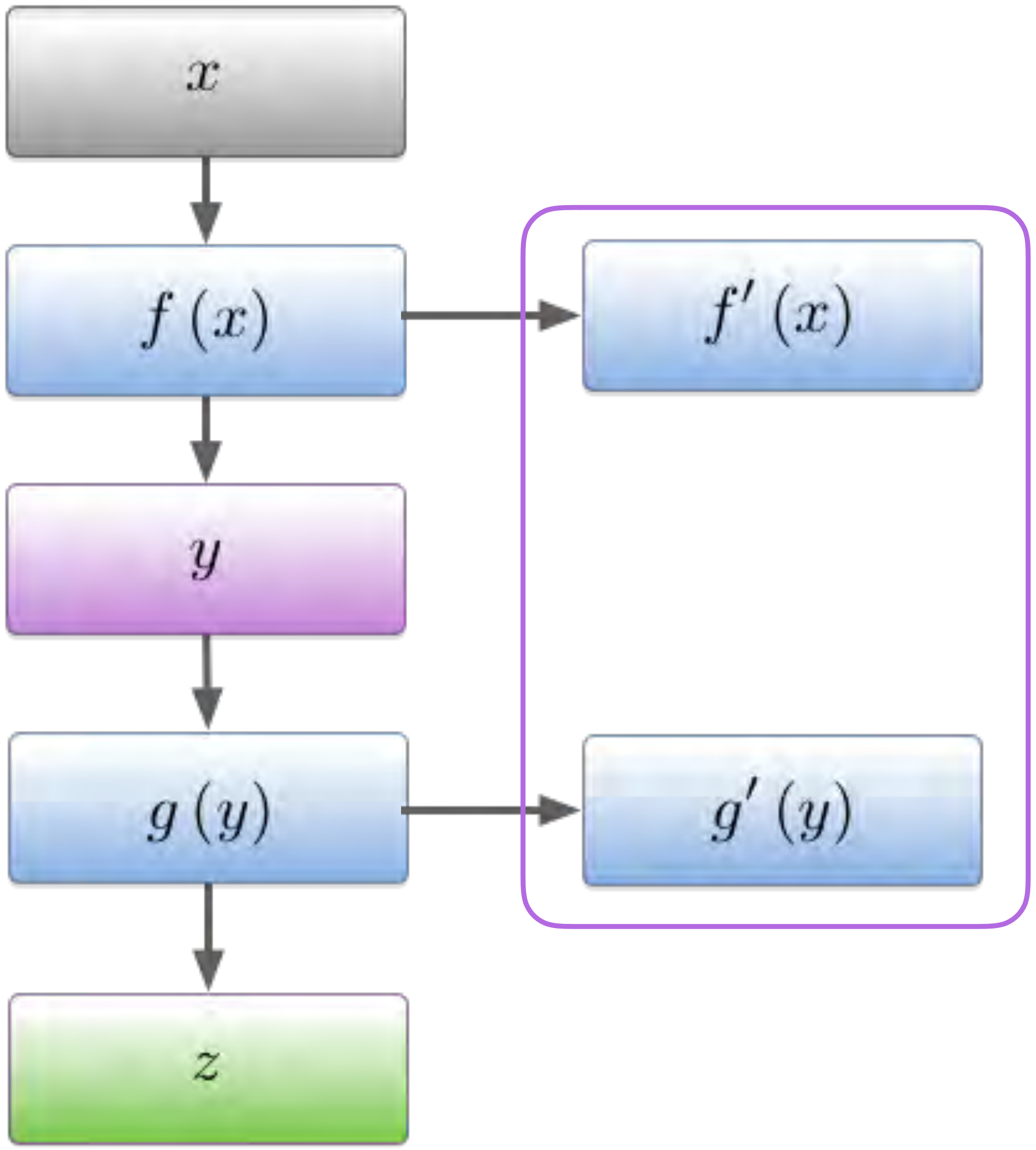
- for some given x .
- To calculate z we must follow a certain sequence of operations.
- Which can be shortened if we are interested in just the value of y
- In Tensorflow, this is called a Computational Graph and it's the most fundamental concept to understand
- Data, in the form of tensors, flows through the graph from inputs to outputs
- Tensorflow, is, essentially, a way of defining arbitrary computational graphs in a way that can be automatically distributed and optimized.



Computational Graphs

https://www.tensorflow.org/

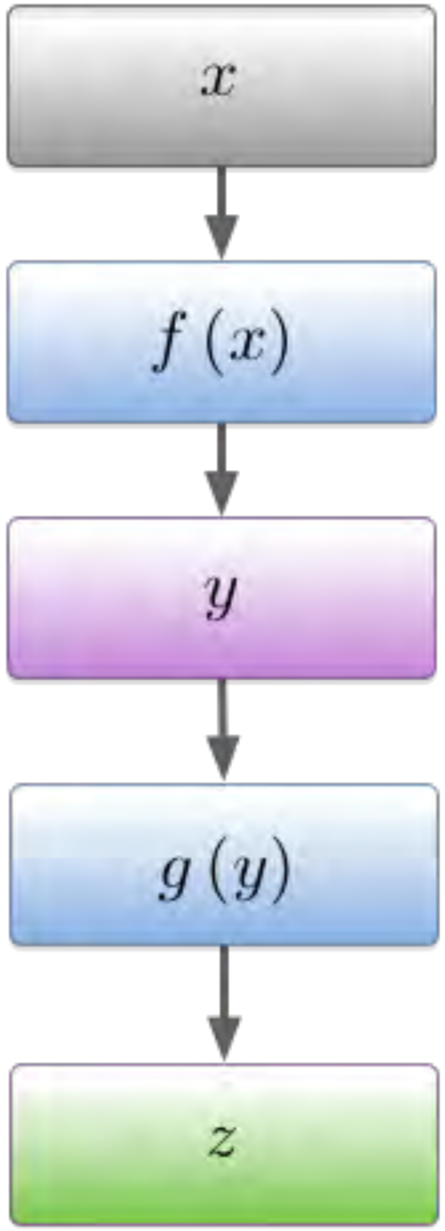
- If we use base functions, tensorflow knows how to automatically calculate the respective gradients
- Automatic BackProp
- Graphs can have multiple outputs
  - Predictions
  - Cost functions
  - etc…



Sessions

https://www.tensorflow.org/

- After we have defined the computational graph, we can start using it to make calculations
- All computations must take place within a "session" that defines the values of all required input values
- Which values are required for a specific computation depend on what part of the graph is actually being executed.
- When you request the value of a specific output, tensorflow determines what is the specific subgraph that must be executed and what are the required input values.
- For optimization purposes, it can also execute independent parts of the graph in different devices (CPUs, GPUs, TPUs, etc) at the same time.



Installation

https://www.tensorflow.org/install/

Install TensorFlow

Assuming the prerequisite software is installed on your Mac, take the following steps:

1. Install TensorFlow by invoking **one** of the following commands:

```
$ pip install tensorflow      # Python 2.7; CPU support
$ pip3 install tensorflow     # Python 3.n; CPU support
```

If the preceding command runs to completion, you should now validate your installation.

A basic Tensorflow program

TensorFlow
https://www.tensorflow.org/

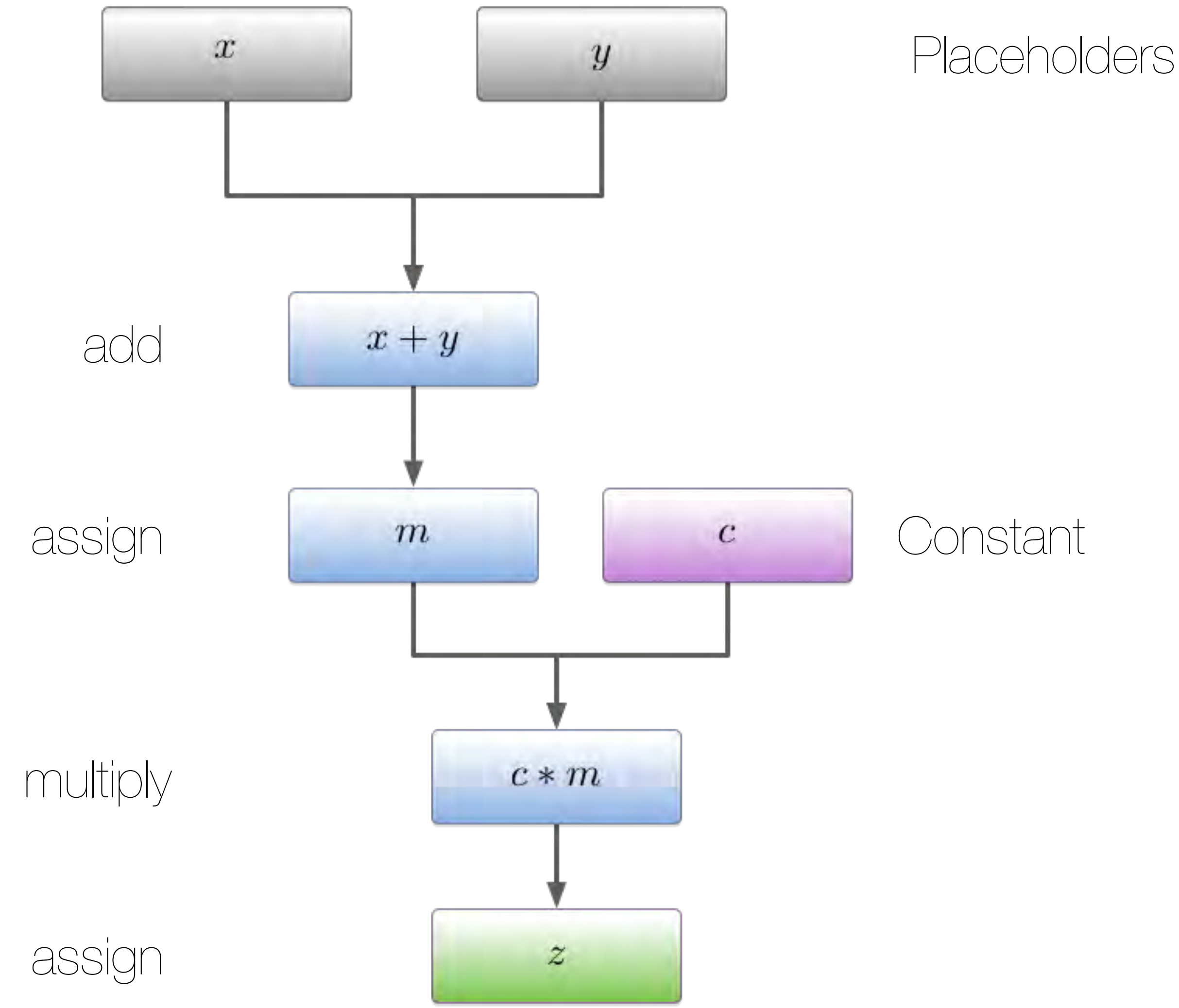
$$z = c * (x + y)$$

```
import tensorflow as tf

x = tf.placeholder(tf.float32)
y = tf.placeholder(tf.float32)
c = tf.constant(3.)

m = tf.add(x, y)
z = tf.multiply(m, c)

with tf.Session() as sess:
    output = sess.run(z, feed_dict={x: 1., y: 2.})
    print("Output value is:", output)
```



@bgoncalves

basic.py

A basic Tensorflow program

https://www.tensorflow.org/

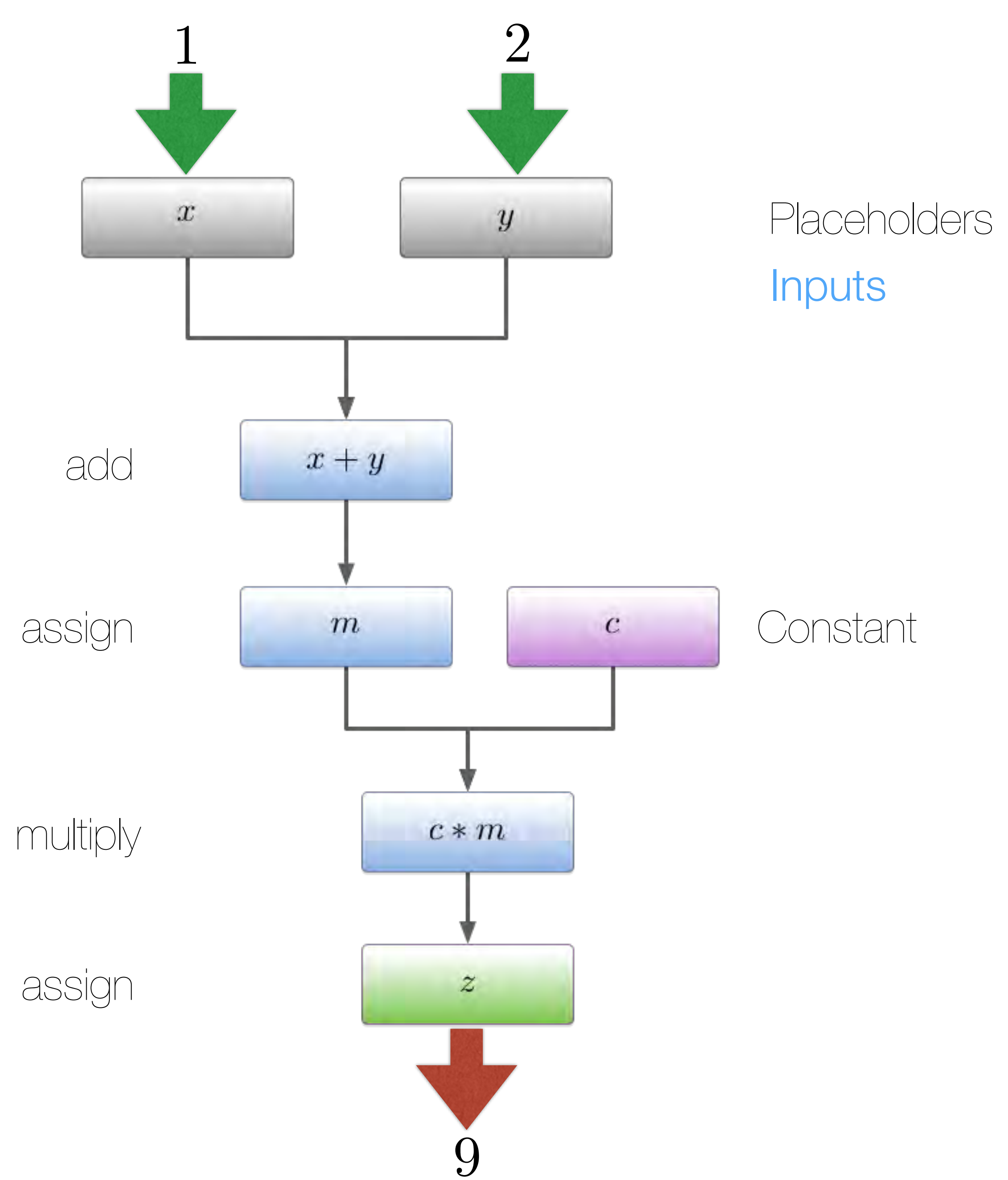
$$z = c * (x + y)$$

```
import tensorflow as tf

x = tf.placeholder(tf.float32)
y = tf.placeholder(tf.float32)
c = tf.constant(3.)

m = tf.add(x, y)
z = tf.multiply(m, c)

with tf.Session() as sess:
    output = sess.run(z, feed_dict={x: 1., y: 2.})
    print("Output value is:", output)
```

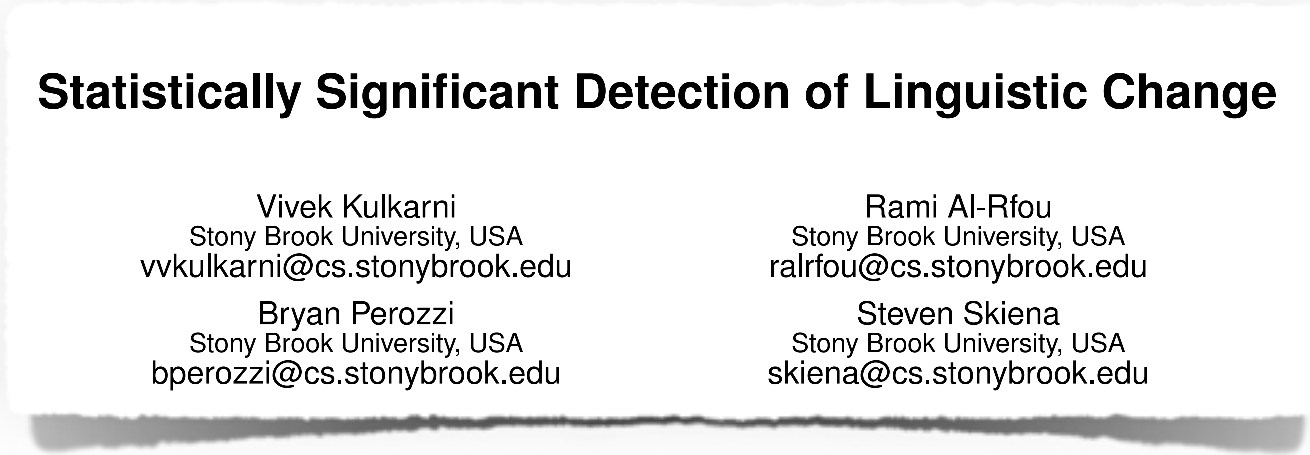


@bgoncalves

Jupyter Notebook

Linguistic Change

WWW'15, 625 (2015)



Statistically Significant Detection of Linguistic Change

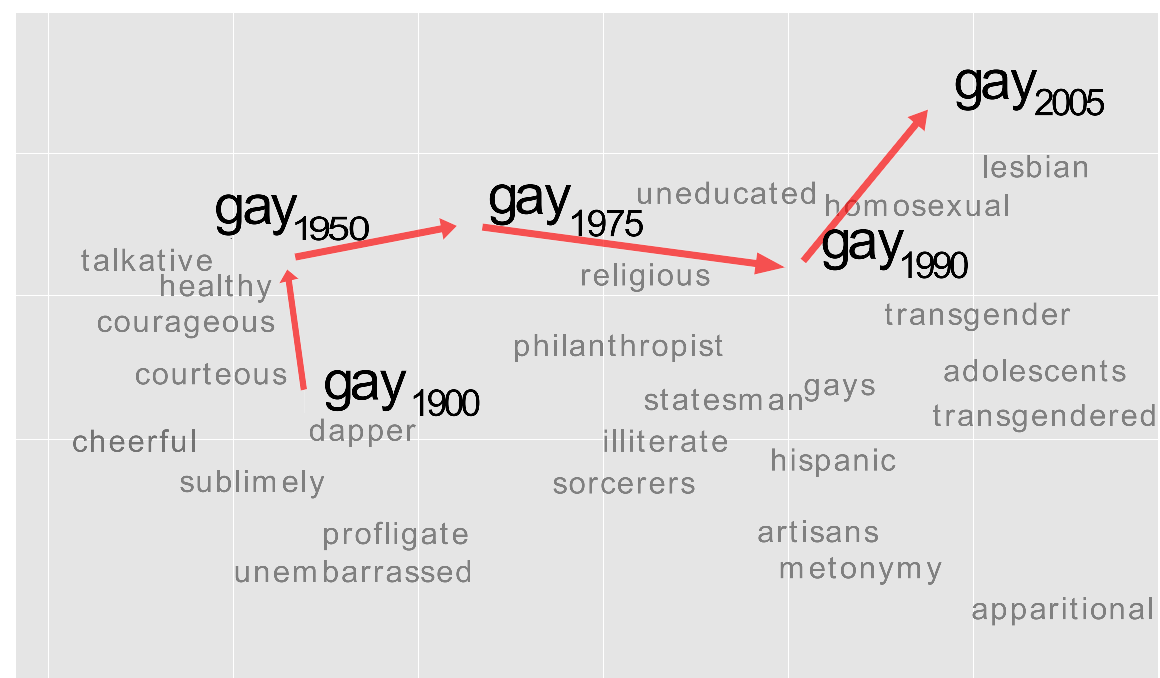
Vivek Kulkarni
Stony Brook University, USA
vvkulkarni@cs.stonybrook.edu

Rami Al-Rfou
Stony Brook University, USA
ralrfou@cs.stonybrook.edu

Bryan Perozzi
Stony Brook University, USA
bperozzi@cs.stonybrook.edu

Steven Skiena
Stony Brook University, USA
skiena@cs.stonybrook.edu

- Train word embeddings for different years using Google Books
- Independently trained embeddings differ by an arbitrary rotation
- Align the different embeddings for different years
- Track the way in which the meaning of words shifted over time!



node2vec

KDD'16, 855 (2016)

node2vec: Scalable Feature Learning for Networks

Aditya Grover
Stanford University
adityag@cs.stanford.edu

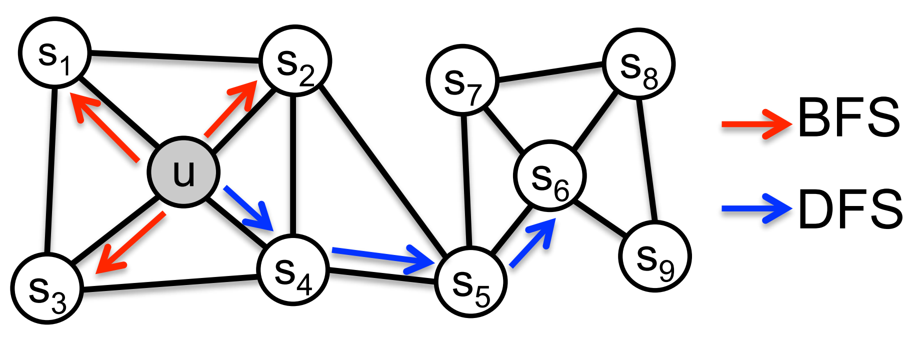
Jure Leskovec
Stanford University
jure@cs.stanford.edu

- You can generate a graph out of a sequence of words by assigning a node to each word and connecting the words within their neighbors through edges.
- With this representation, a piece of text is a walk through the network. Then perhaps we can invert the process? Use walks through a network to generate a sequence of nodes that can be used to train node embeddings?
- node embeddings should capture features of the network structure and allow for detection of similarities between nodes.

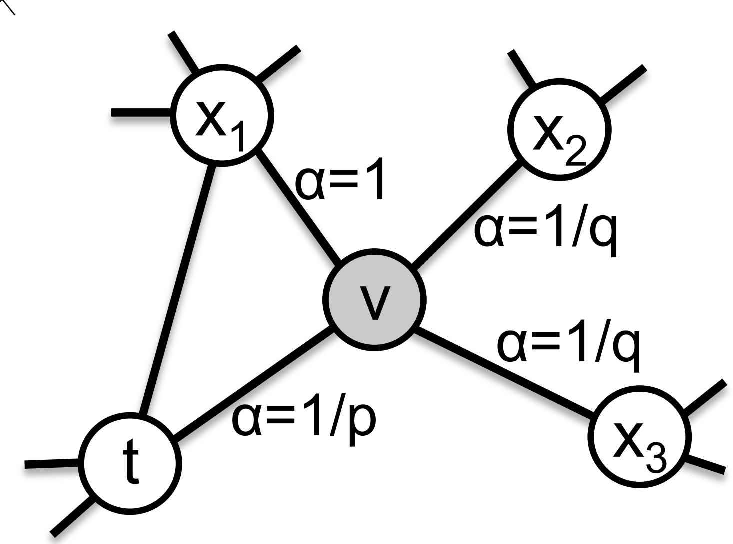
node2vec

KDD'16, 855 (2016)

- The features depends strongly on the way in which the network is traversed
- Generate the contexts for each node using Breath First Search and Depth First Search



- Perform a biased Random Walk



- BFS - Explores only limited neighborhoods. Suitable for structural equivalences
- DFS - Freely explores neighborhoods and covers homophiles communities
- By modifying the parameter of the model it can interpolate between the BFS and DFS extremes

dna2vec

arXiv: 1701.06279 (2017)

dna2vec: Consistent vector representations of variable-length k-mers

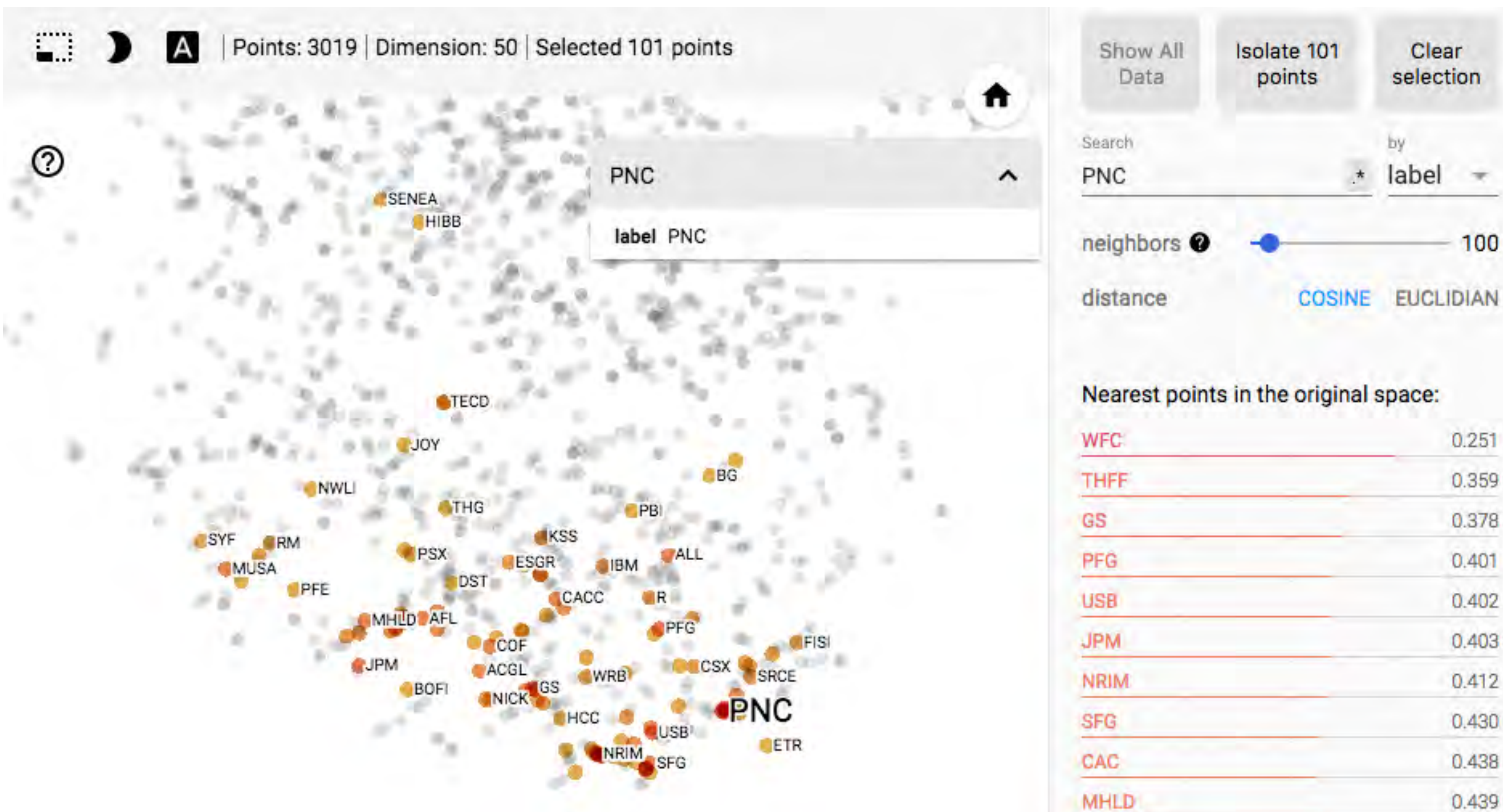
Patrick Ng
ppn3@cs.cornell.edu

- Separate the genome into long non-overlapping DNS fragments.
- Convert long DNA fragments into overlapping **variable** length k-mers
- Train embeddings of each k-mer using **Gensim** implementation of SkipGram.
    - Summing embeddings is related to concatenating **k-mers**
    - Cosign similarity of k-mer embeddings reproduces a biologically motivated similarity score (Needleman-Wunsch) that is used to align nucleoti

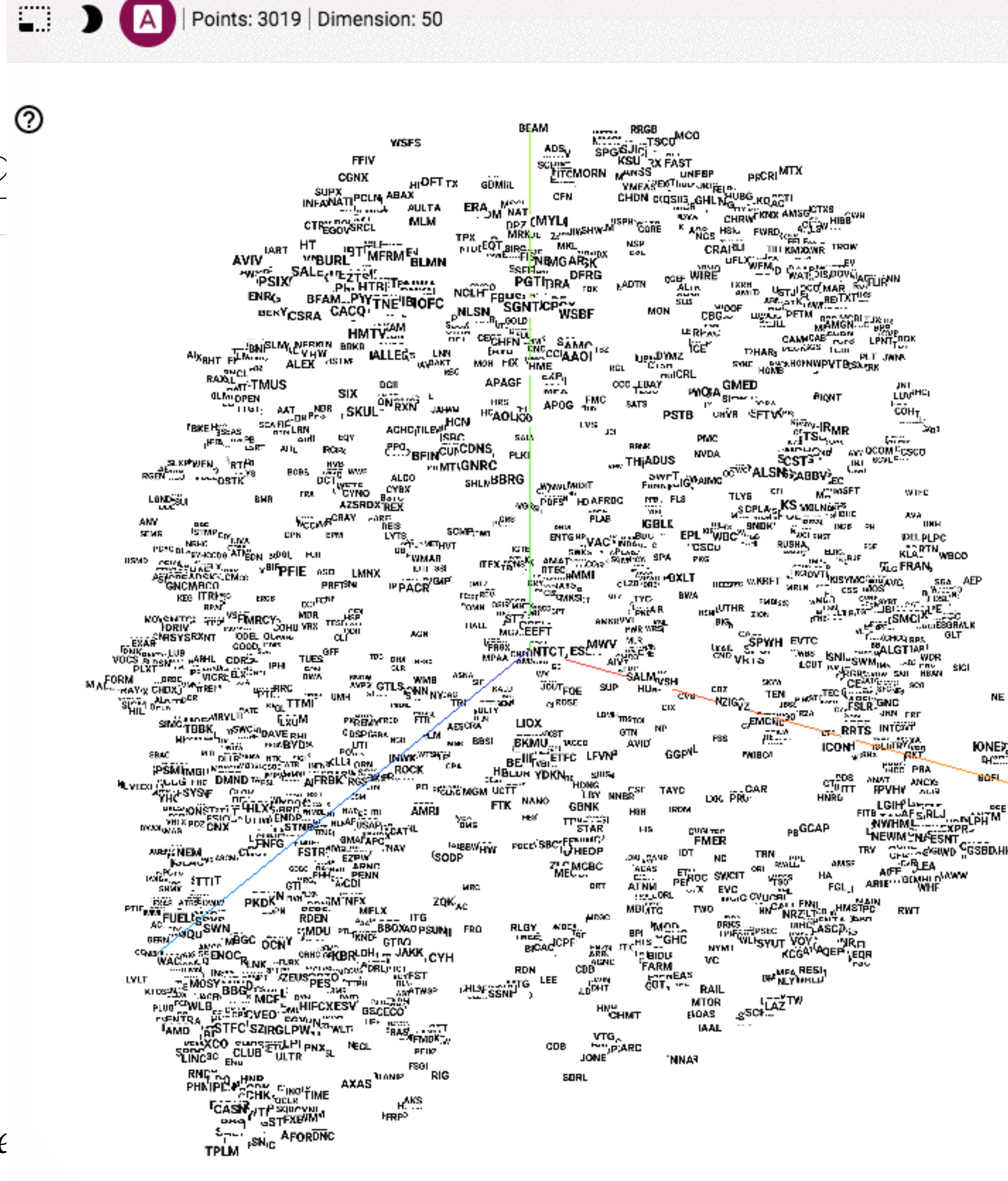
stock2vec

https://medium.com/towards-data-science/stock2vec-from-ml-to-p-e-2e6ba407c24

- Apply word2vec to the 40 years of stock market data
- Identify significant semantic similarities between companies working in the same area



stock2



ml-to-p-e-2e6ba407c24

Thank you!

You can hear me speak more about word2vec in this weeks podcast!