CSCI E-29 Week 2 Section

Agenda

- Concepts
  - Word vector representation
  - Word2Vec
  - Visualization / t-SNE
  - Distance metrics
- Pset 1
  - overview
  - q&a

Mapping words to vectors

“cat”	→	[0.13 0.62 0.98 … 0.71]
“dog”	→	[0.53 0.61 0.53 … 0.90]
“octopus”	→	[0.81 0.62 0.98 … 0.04]

Word vector representations

Key aspects:

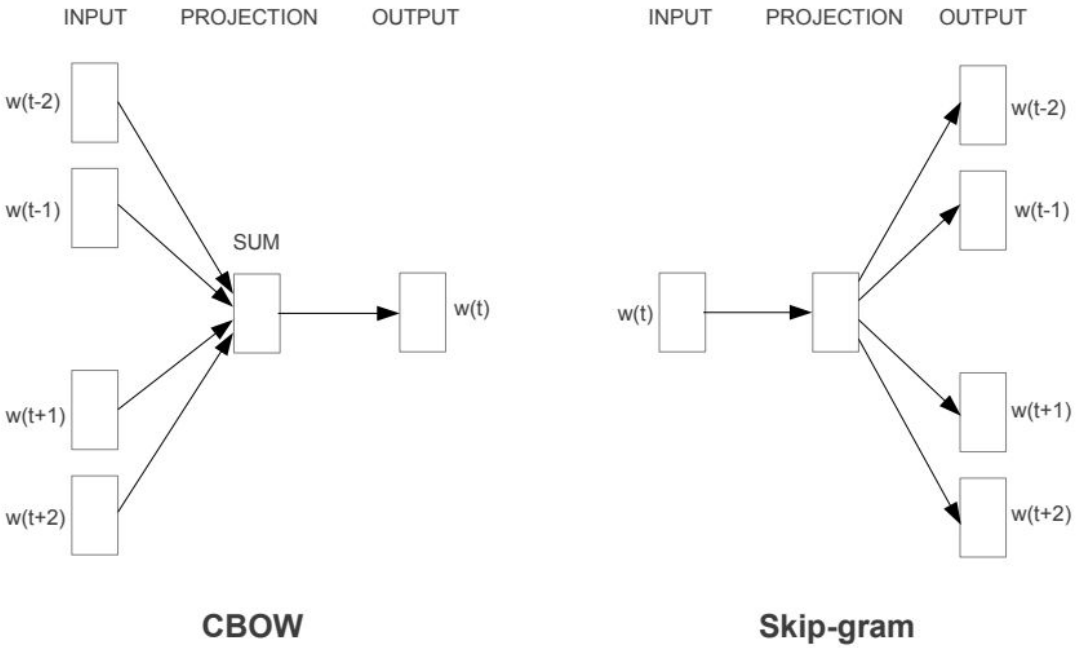
- Number of dimensions, D
- Vocabulary size (number of words), V

50〜300 *dim*

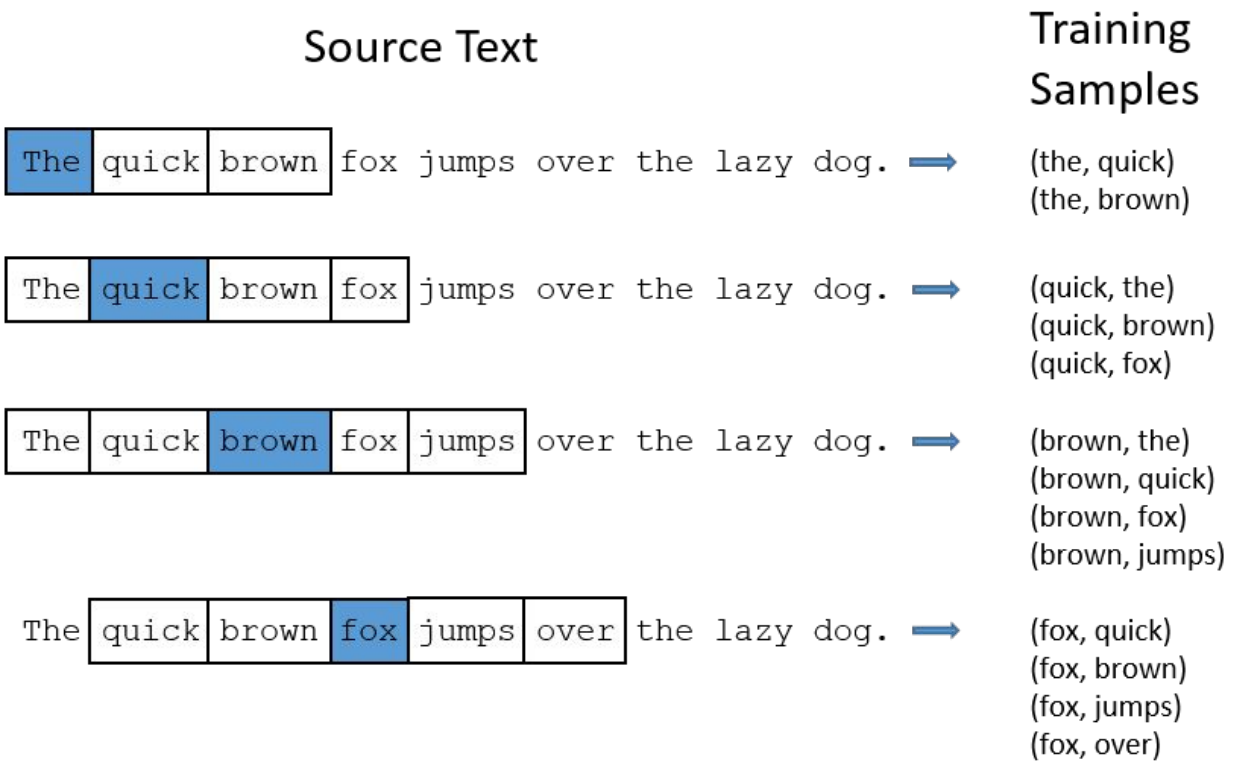
python	0.52	0.21	0.37	…
ruby	0.48	0.21	0.33	…
word	0.05	0.23	0.06	…

Word2vec

Efficient Estimation of Word Representations in Vector Space (Mikolov et al, 2013)

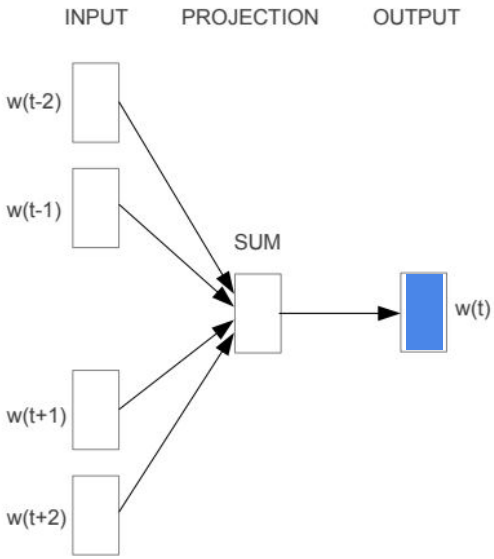


Context “windows”: Target word vs context words

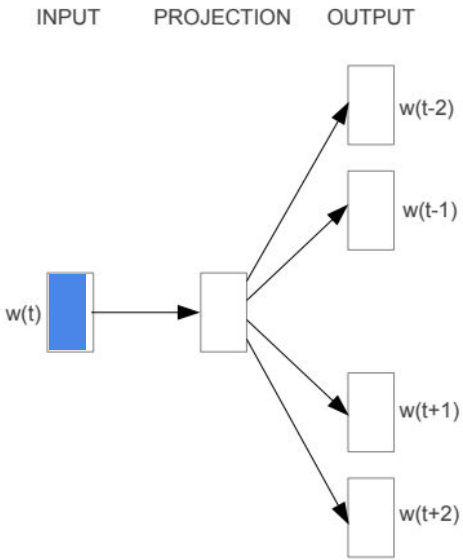


CBOW (Continuous bag of words) vs Skip-gram

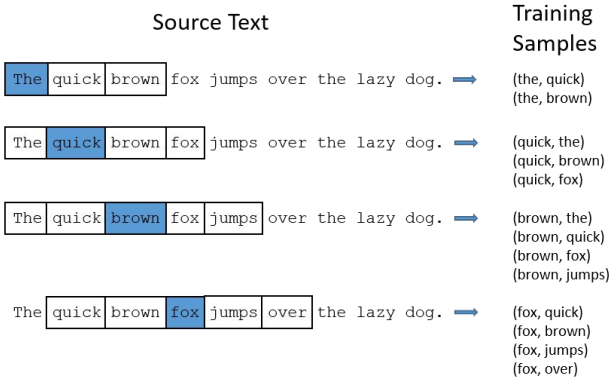
Predict target from context vs. predict context from target



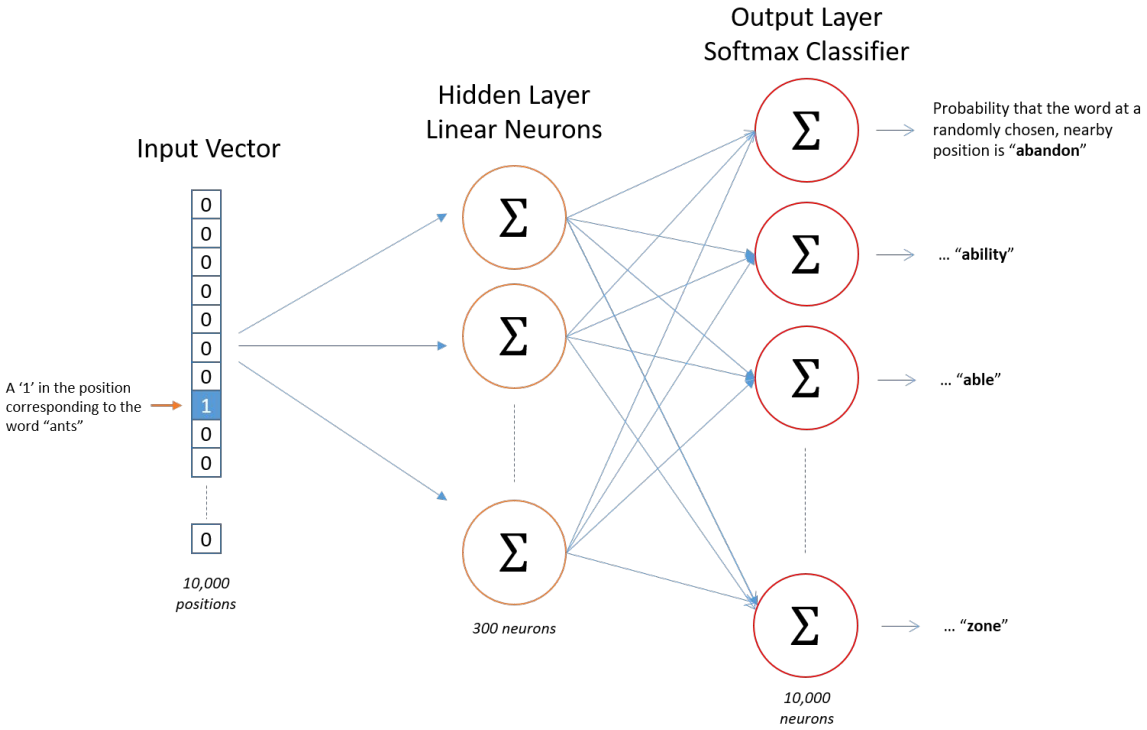
CBOW



Skip-gram

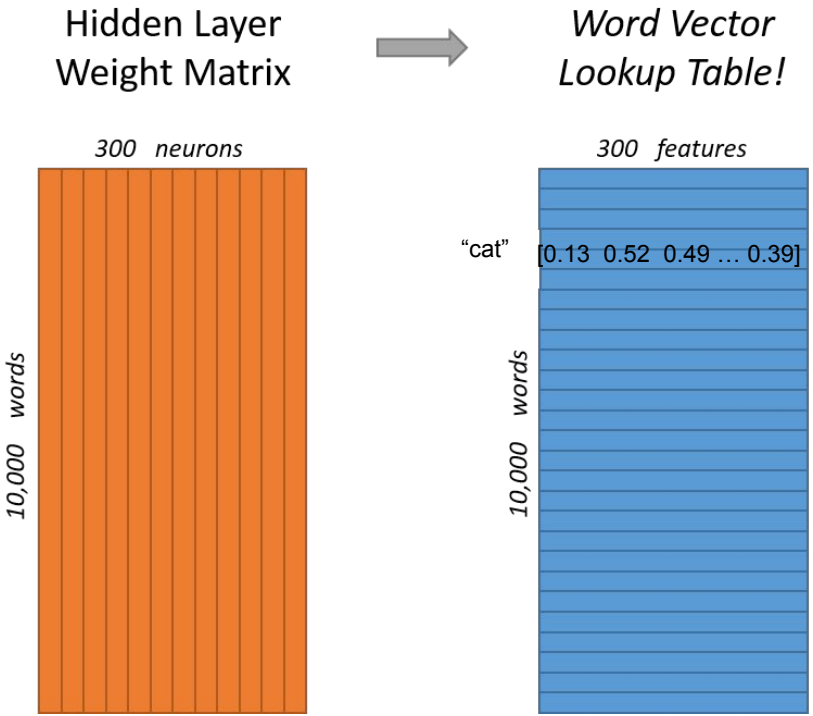


Word2vec network architecture

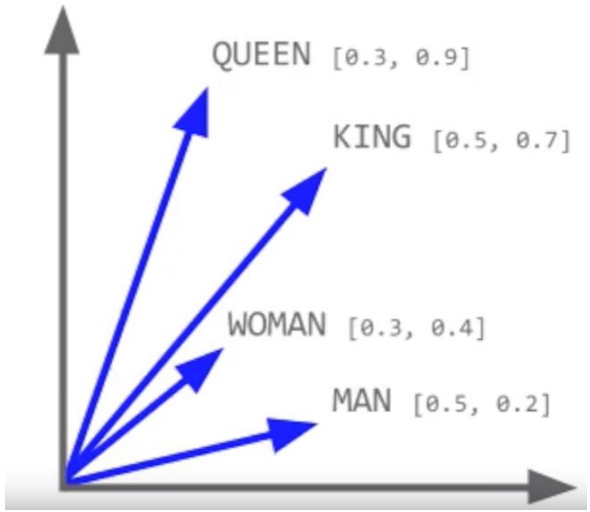


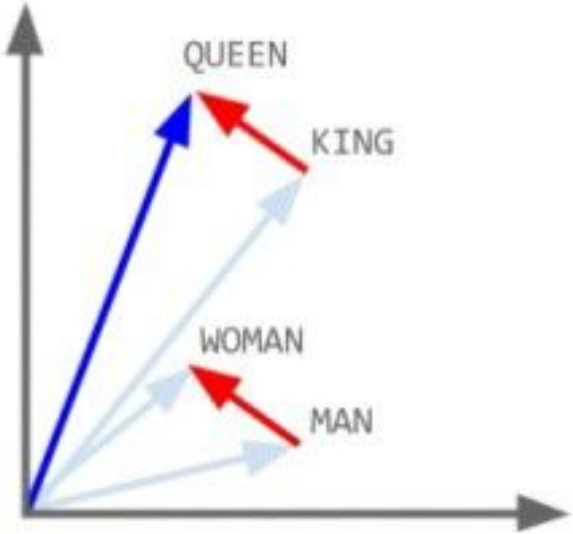
Middle layer has size D, each neuron has V weights = weight matrix W for layer is [D x W]

Network weights from the trained network are vector representations

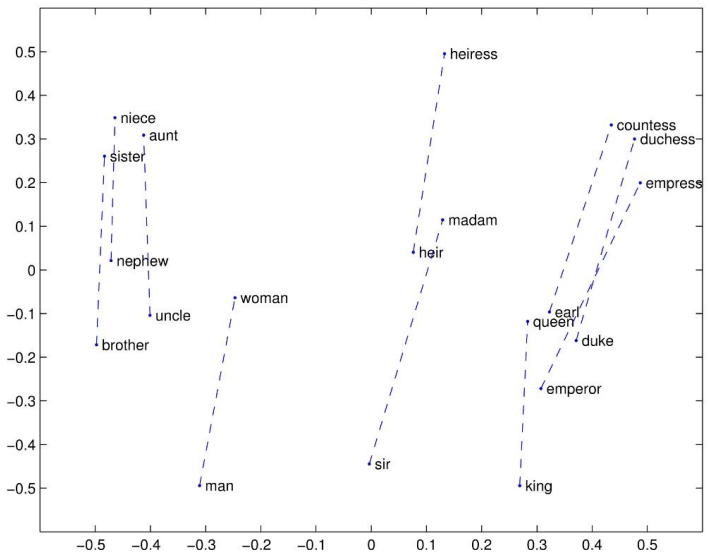


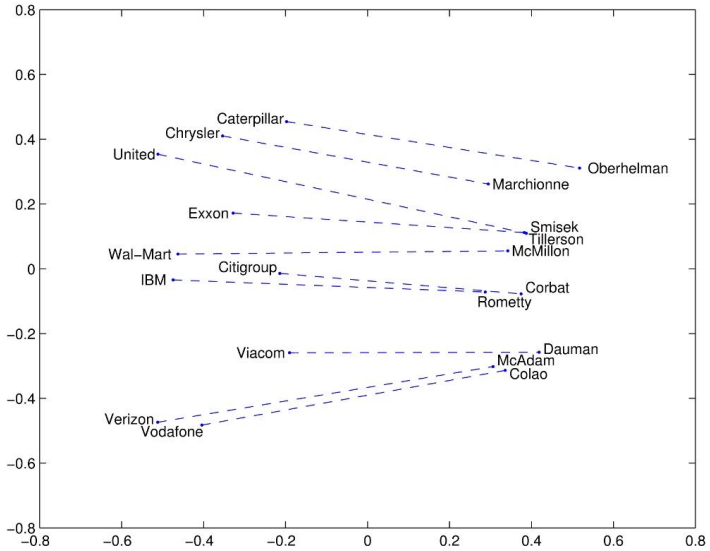
Concepts represented by vector difference: “man is to woman as king is to queen”





Concepts represented by vector difference





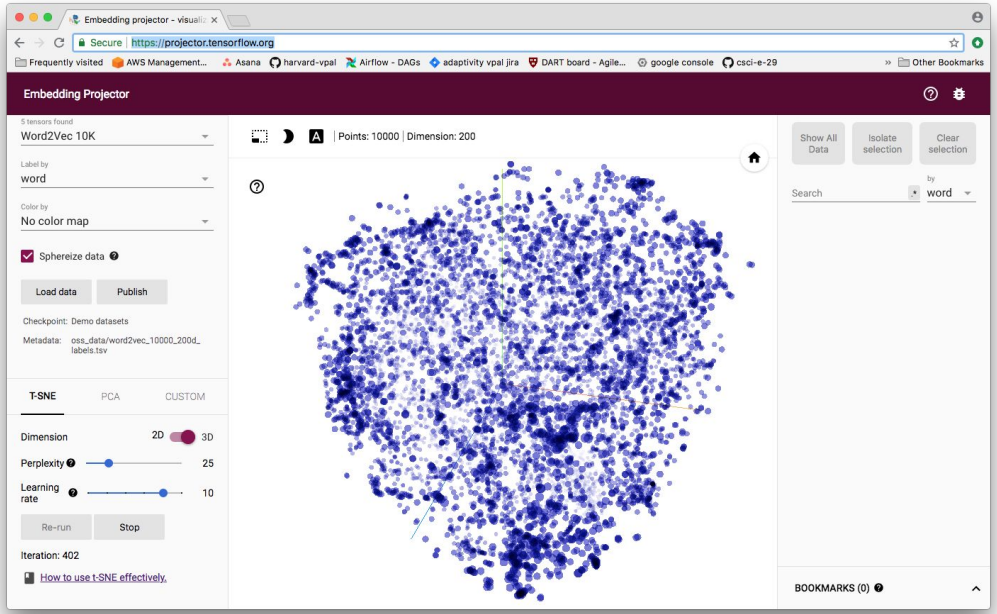
https://nlp.stanford.edu/projects/glove/

Dimensionality reduction methods

- Visualization requires reducing a 300-dimensional vector into 2-3 dimensions
- PCA - principal component analysis
  - deterministic
  - linear transform
  - optimize for variance explained
- t-SNE - t-distributed stochastic neighbor embedding
  - probabilistic
  - non-linear transform
  - prioritizes keeping neighboring points close

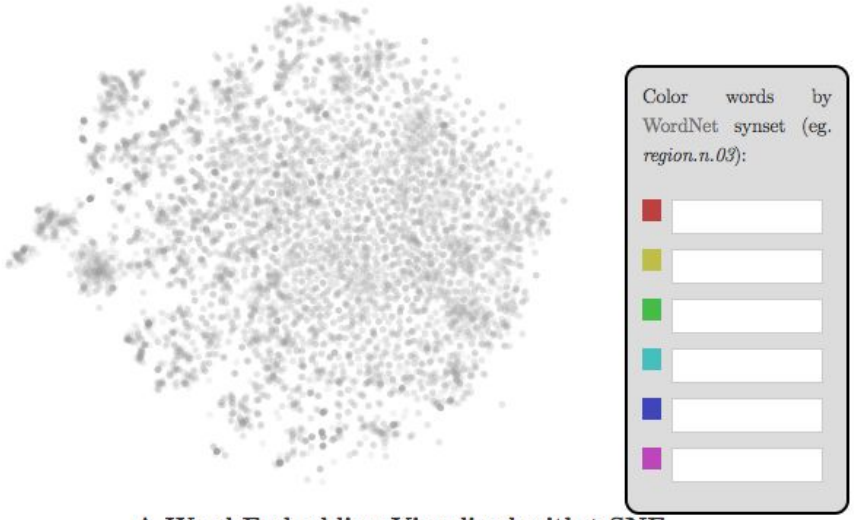
Dimensionality reduction - visualization

https://projector.tensorflow.org/



Visualizing word vector representations with t-SNE

https://colah.github.io/posts/2015-01-Visualizing-Representations/



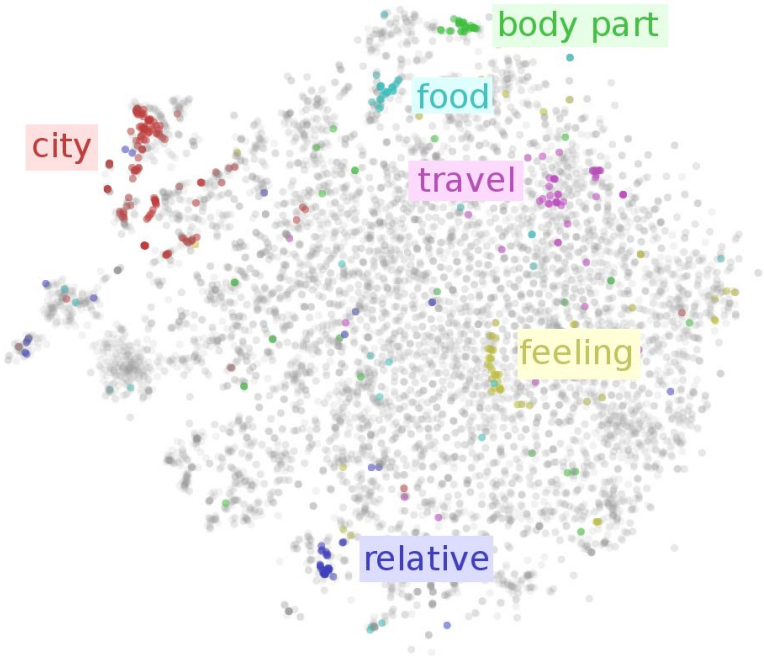
A Word Embedding Visualized with t-SNE

(Hover over a point to see the word.)

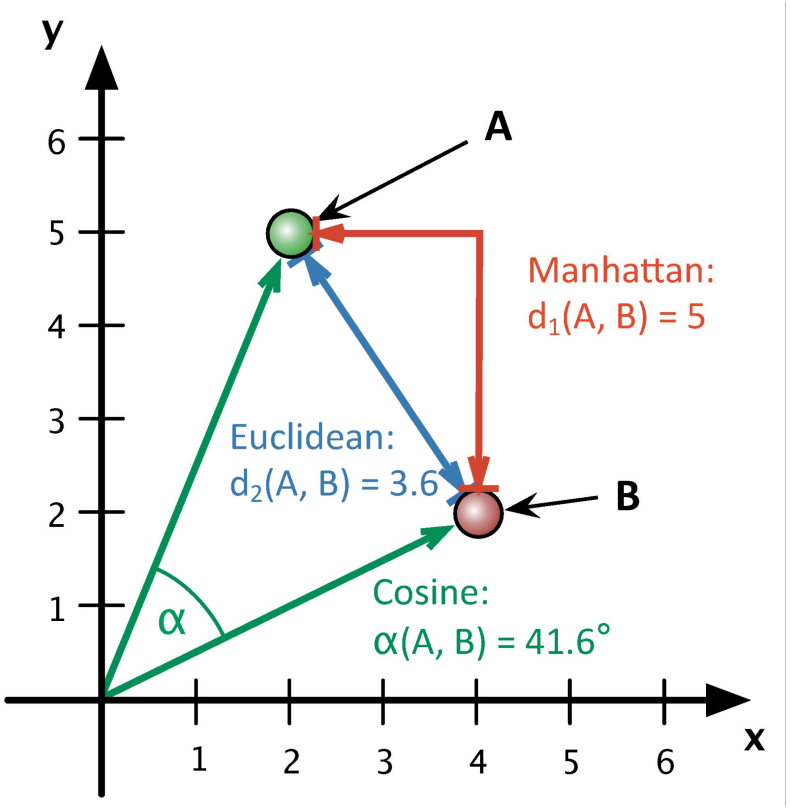
(See this with 50,000 points!)

Visualizing word vector representations with t-SNE

https://colah.github.io/posts/2015-01-Visualizing-Representations/



“Distance” as a metric for word similarity



Applications of word embeddings

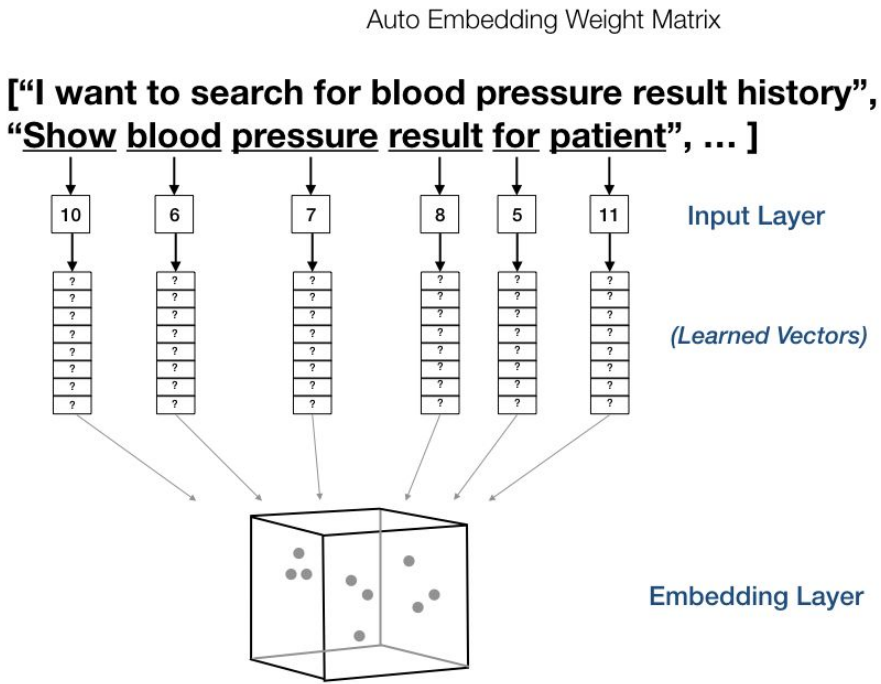
San Francisco

Word	Cosine distance
los_angeles	0.666175
golden_gate	0.571522
oakland	0.557521
california	0.554623
san_diego	0.534939
pasadena	0.519115
seattle	0.512098
taiko	0.507570
houston	0.499762
chicago_illinois	0.491598

France

Word	Cosine distance
spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033
portugal	0.577154
russia	0.571507
germany	0.563291
catalonia	0.534176

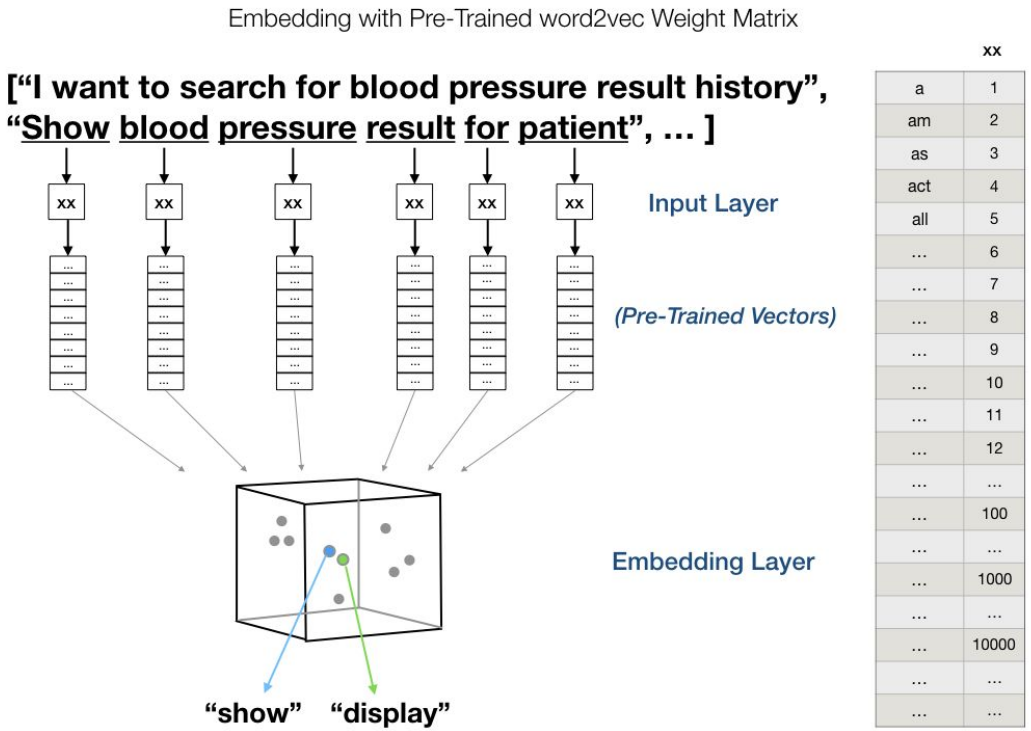
Applications of word embeddings



i	1
want	2
to	3
search	4
for	5
blood	6
pressure	7
result	8
history	9
show	10
patient	11
...	
LAST	20

https://medium.com/@JMangia/coreml-with-glove-word-embedding-and-recursive-neural-network-part-2-ab238ca90970

Applications of word embeddings



https://medium.com/@JMangia/coreml-with-glove-word-embedding-and-recursive-neural-network-part-2-ab238ca90970

pset1 overview

- Templating:
  - Customize template config for
    - project structure/naming
    - environment - pipenv and docker
    - testing - pytest and travisCI
- Render the template
  - provide overrides for template values or use default
- Implement packages
  - Atomic writes
  - Hashing strings

Templating

- Templating with cookiecutter
  - Adjust the template
  - Defaults are defined
  - Render the template, all instances of {{cookiecutter.VARIABLE}} get replaced with variable defined at runtime
- Design pattern: separating configuration from code.
  - Key for reusability - not just a one-off script

Package structure

Rendered package:

```
pset_utils_kunanit/
    pset_utils/
        io/
            __init__.py
            [io.py]
        hashing/
            __init__.py
            [hashing.py]
```

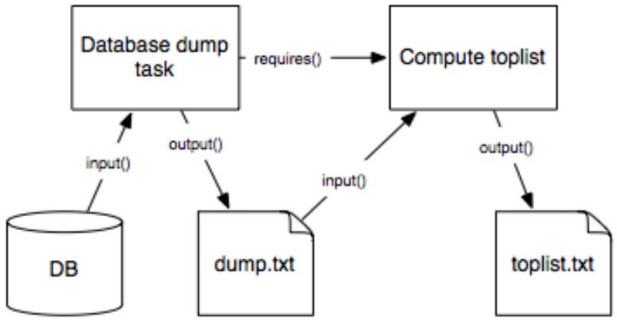
Package structure affects import API:

```
from pset_utils.io import atomic_write

from pset_utils.io.io import atomic_write
```

Atomic writes

- Thanksgiving bug
- Multi-stage data pipelines - need to know whether a step completed successfully or not



Hashing

Some ways hashing can be useful

- File contents checksum - only download and reprocess file if MD5 checksum changes
- Converting URLs to a filename friendly name
- Representing user ids to remote systems
- Authentication without sending sensitive key - both parties know the key and compute hash (with other user info added as salt)