Harmonisation methods

Eve Cheng

The goal

A bunch of survey questions:

1. How old are you?
2. What is your occupation?
3. How old is your dog?
4. Who would you vote for?
5. How old is your father?
6. How old is your mother?

?

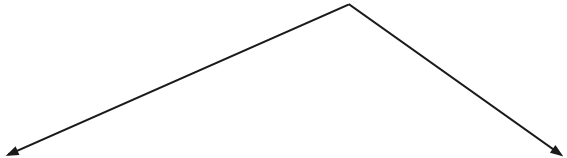
1. How old are you?
2. How old is your dog?
3. How old is your father?
4. How old is your mother?

1. What is your occupation?

1. Who would you vote for?

Current methods

Machine learning: requires training your model or downloading existing models



word2vec (word mover)

SentenceTransformer

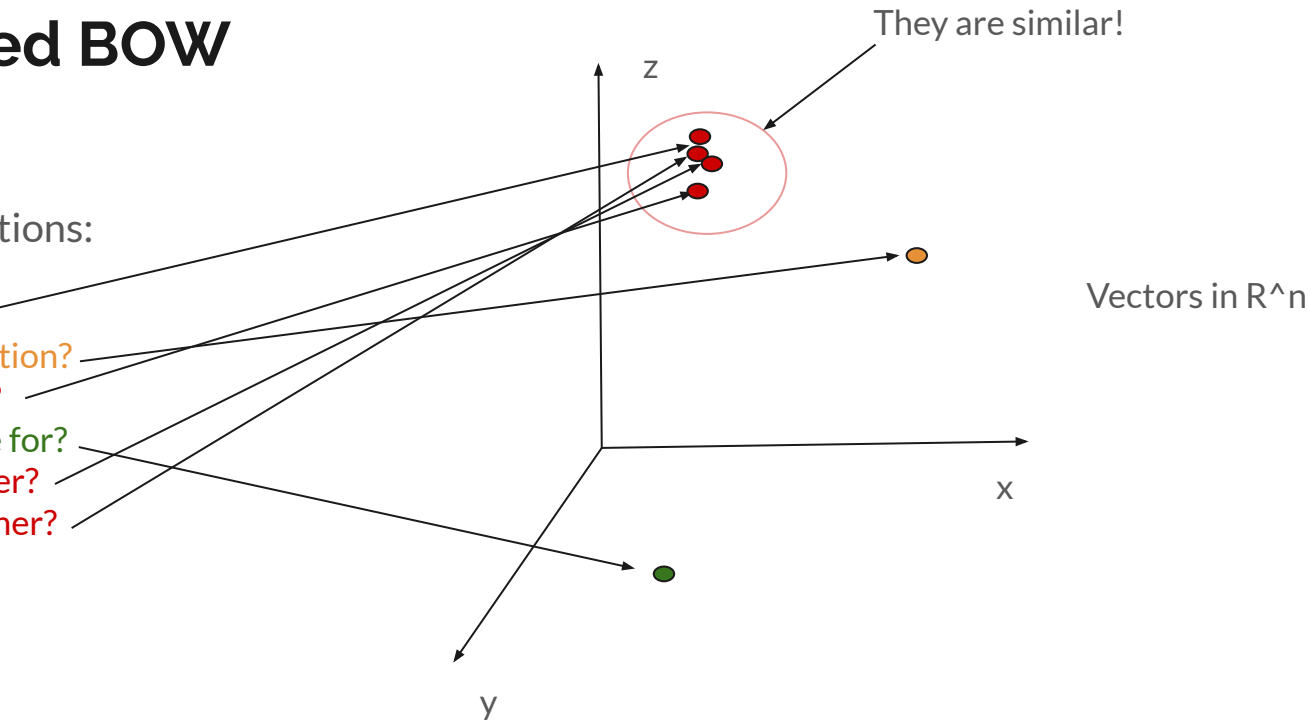
No training required

Bag of words

Normalised BOW

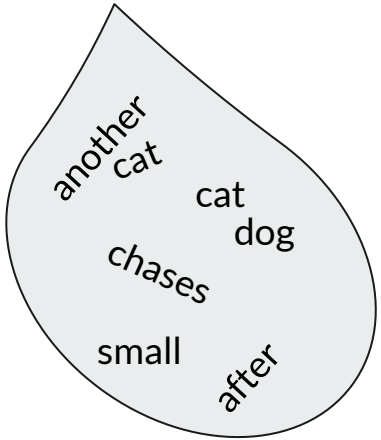
A bunch of survey questions:

1. How old are you?
2. What is your occupation?
3. How old is your dog?
4. Who would you vote for?
5. How old is your father?
6. How old is your mother?



Step 1: calculate the vectors

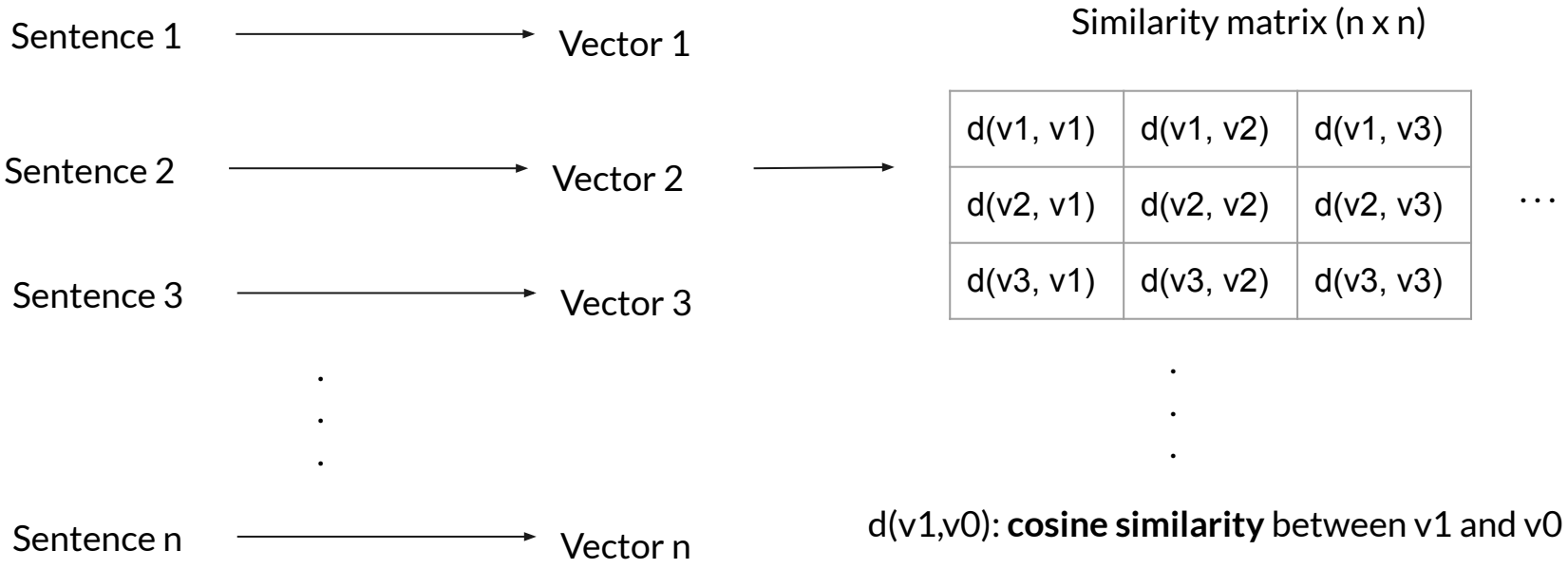
The cat chases after the dog and another small cat



word	freq
cat	2
dog	1
chases	1
small	1
after	1
another	1

Vector in R^6
(2,1,1,1,1,1,1)

Step 2: calculate pair-wise distances



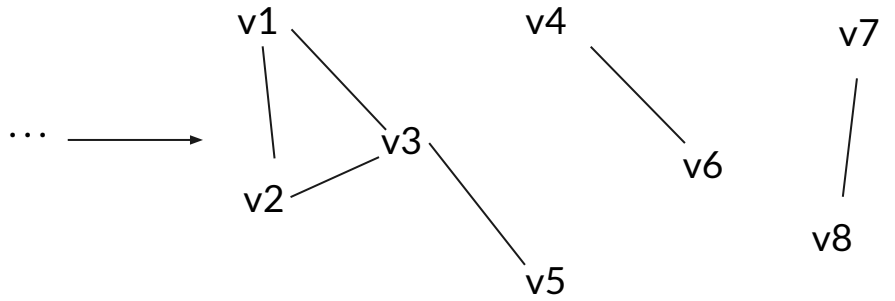
Step 3: create a network of sentences

Similarity matrix (n x n)

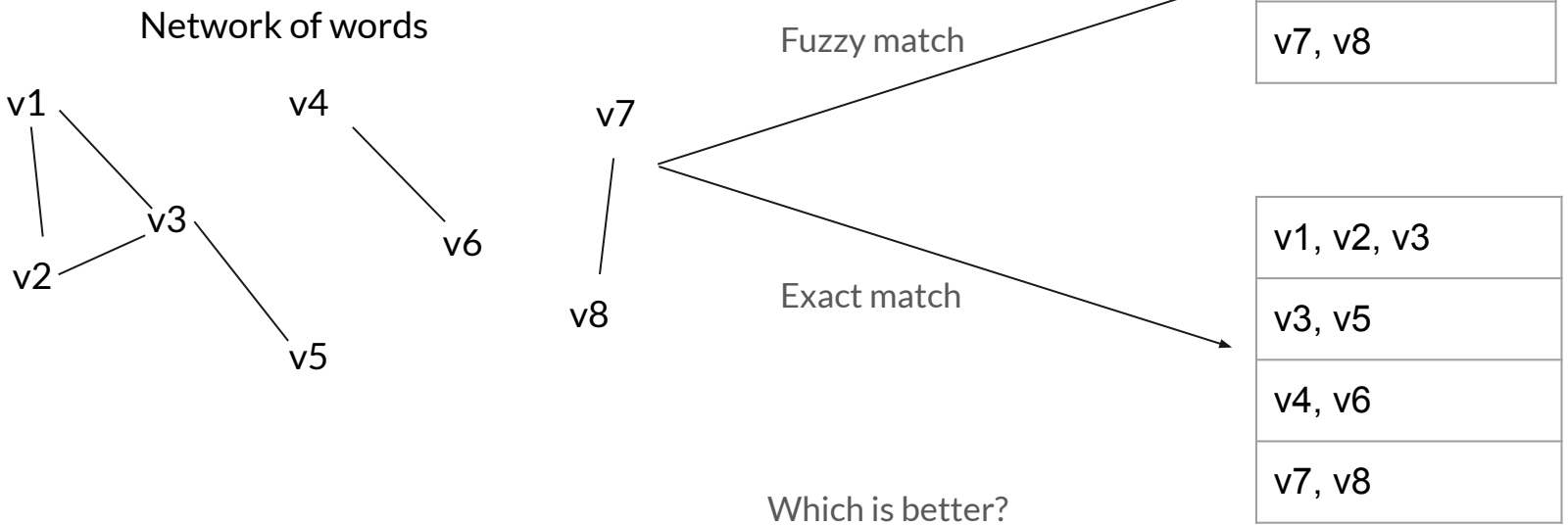
d(v1, v1)	d(v1, v2)	d(v1, v3)
d(v2, v1)	d(v2, v2)	d(v2, v3)
d(v3, v1)	d(v3, v2)	d(v3, v3)

.
.
.

Network of words

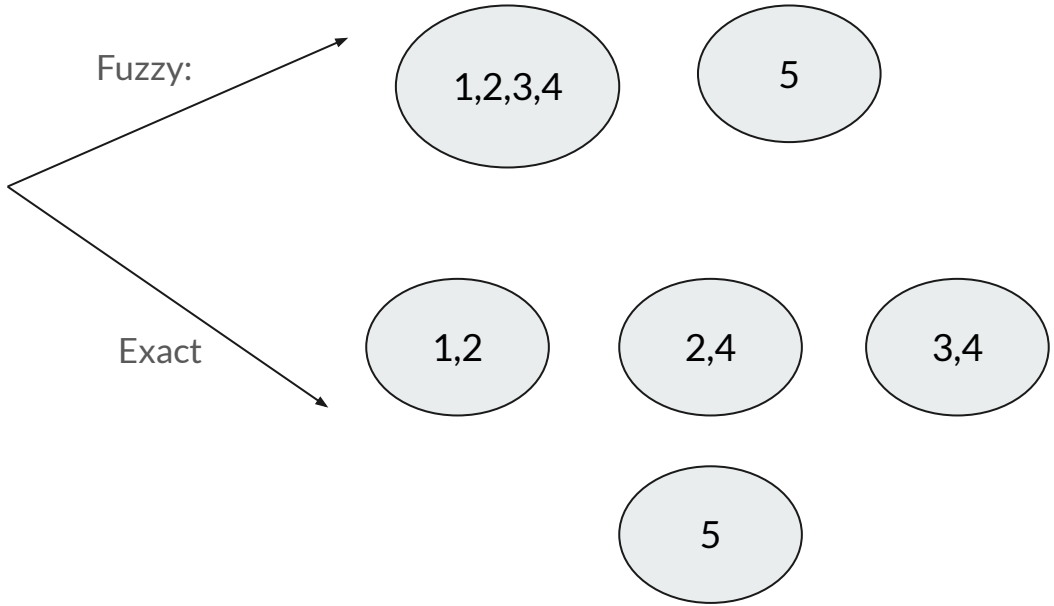


Step 4: detect clusters



Example. Input 2: threshold (set to 0.7 here)

1	Country of Birth - mother
2	Country of Birth - mother - extended
3	Country of Birth - father
4	Country of Birth - father - extended
5	Age father left school



Inputs for BOW so far:

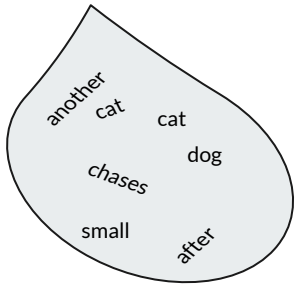
1. Threshold: above what similarity do you consider similar. The program requires a number from 0 to 1.

2. Cluster option: fuzzy or exact

One last input: the N in the Ngram method

Vanilla BOW

The cat chases after the dog and another small cat



Ngram: for example, N=3

The cat chases after the dog and another small cat.

The cat chases

Another small cat

cat chases after

. . .

Summary

Threshold

Cluster option

N in Ngram

Sentences that are deemed similar

What about all those machine learning stuff?