Word Vectors - I

Pawan Goyal

CSE, IIT Kharagpur

January 20th, 2017

Semantics

What is semantics?

Semantics

What is semantics?

Why worry about semantics?

Semantics

What is semantics?

Why worry about semantics?

How to capture semantics?

Let's start with the words ...

Word Representation

In general, words are treated as atomic symbols.

Word Representation

In general, words are treated as atomic symbols.

One-hot representation

Distributional Similarity Based Representations

You know a word by the company it keeps

Distributional Similarity Based Representations

You know a word by the company it keeps

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

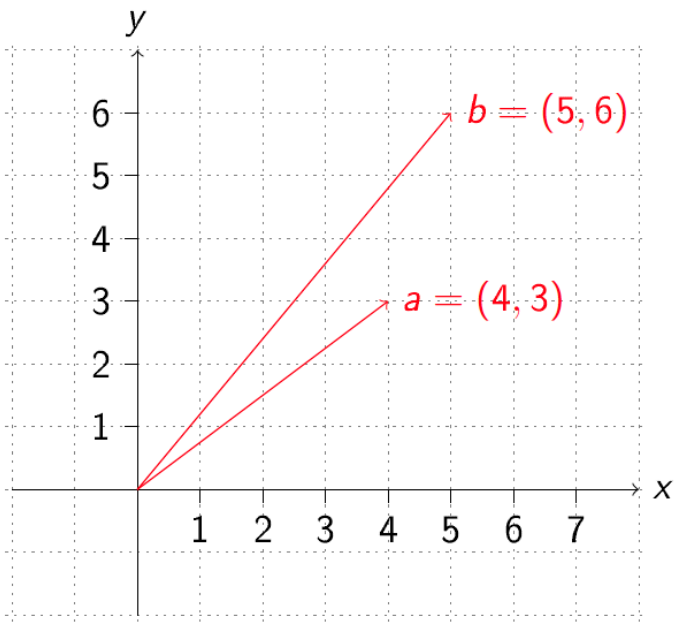
Distributional Similarity Based Representations

You know a word by the company it keeps

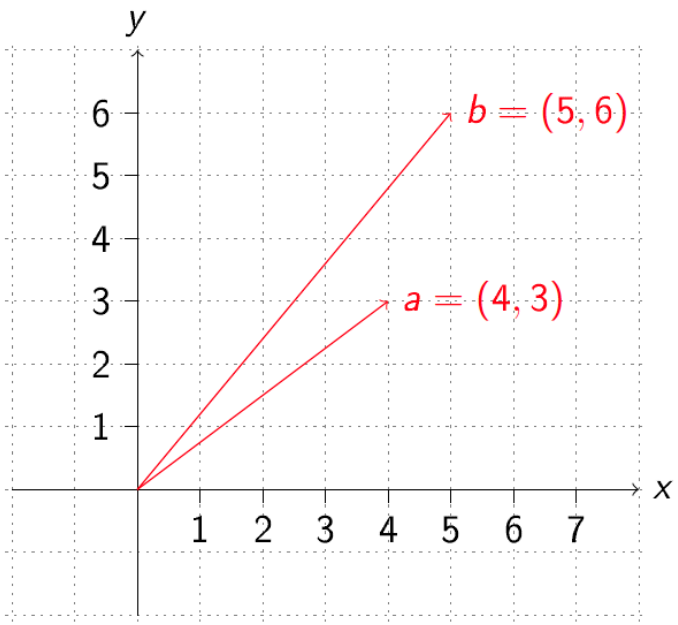
government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

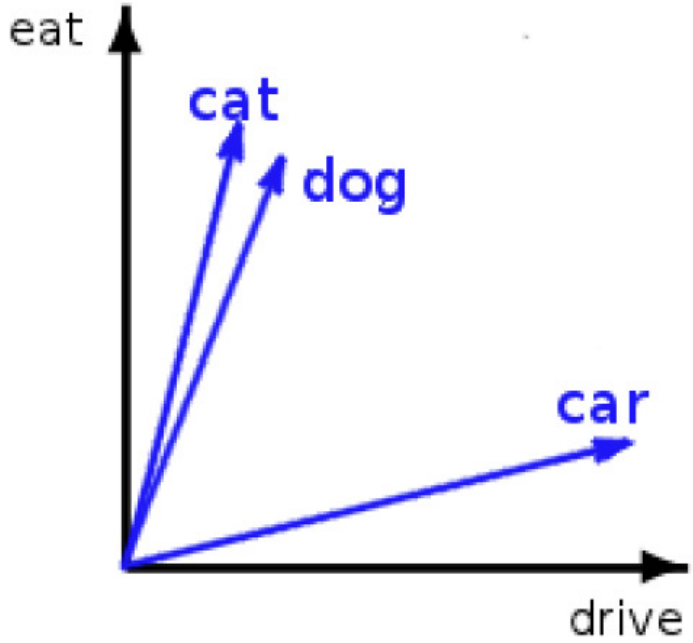
These words will represent banking

Vector Space Model

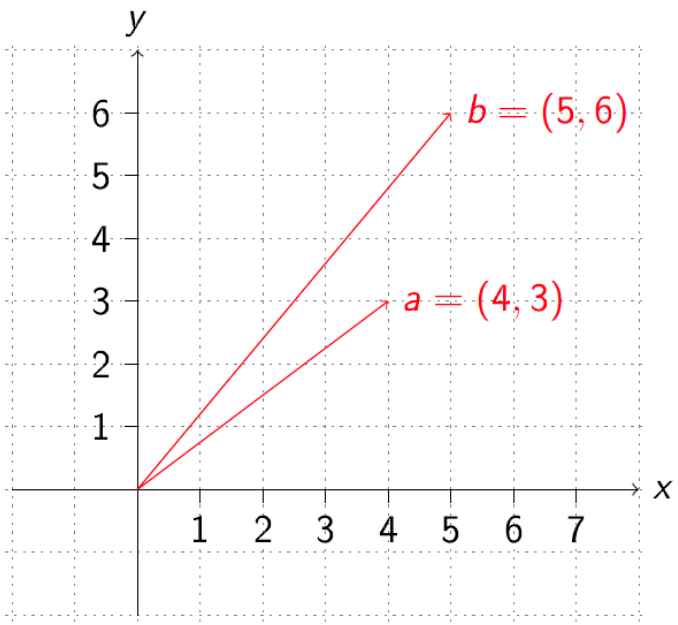


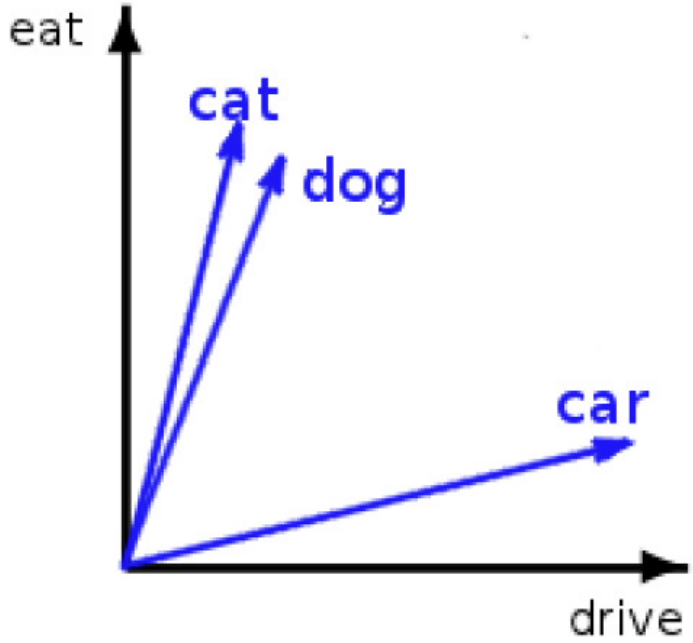
Vector Space Model





Vector Space Model





In practice, many more dimensions are used.

$cat =$ [...dog 0.8, eat 0.7, joke 0.01, mansion 0.2,...]

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

$\Downarrow$

Select the targets and the contexts

Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)

⇓

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

⇓

Weight the contexts (optional)

⇓

Build the distributional matrix

⇓

Reduce the matrix dimensions (optional)

⇓

Compute the vector distances on the (reduced) matrix