Word Representation

From Counts to Dense Vectors

Ramaseshan Ramachandran

Word Semantic Space

WHAT IS CONTEXT?

The main challenge in understanding how words relate to real-world situations has increasingly become a matter of context.

- All words within a window or ideally within a sentence
- All content words within a window or sentence that fall in a certain frequency range
- All content words which stand in closest proximity to the word in question in the grammatical schema of each window or sentence

CONTEXT - THE INFLUENCER

Phrase	Meaning
Spill the tea	Share the gossip or details
On the same page	Everyone understanding or agreeing on something.
I'm so over it	Tired of something
That's so cool	Approval or exclamation
It's raining cats and dogs	It's raining heavily

- Context influences the word meaning
- Small boy, small car, small house, small island
- Words that occur in similar contexts will tend to have similar meanings
- Semantic similarity between two words (w_x, w_y) is a function of how frequently they appear in similar linguistic contexts
  - $\vec{w_x} \approx \vec{w_y}$ when the frequency of the context $(f_{C_{xy}(k)})$ with a window of size k in which both words w_x and w_y appeared is higher
- If $f_{C_{xy}}(k)$ is higher, then the semantic relationship of (w_x, w_y) is stronger
- Extending to multiple similar words for w_x:

  $\vec{w_x} \approx \vec{w_{y_i}}$ when the frequency of $C_{xy_i}(k)$ is higher, where $i = 1 \ldots n$

  Note: Here $\approx$ represents similarity

SEMANTIC SPACE OR CONCEPT SPACE

Hyponyms and hypernyms are linguistic terms that describe the relationship between words based on their meaning.

- A space where similar words (synonyms, hyponyms, hypernyms) are classified and arranged along various axes
  - Colour (hypernym) - $\underbrace{\text{red}, \text{Green}, \text{Orange}}_{\text{co-hyponyms}}$ (hyponyms) - Attributional Similarity
- Models that automatically find similar words are known as Distributional Semantic Models (DSM)
- Semantically similar words are found automatically using co-occurrences/co-locations/context

  OR
- Words connected by similar patterns are probably semantically similar

Side-effect: Lexical co-occurrences associate semantically dis-similar words

LAWS OF ASSOCIATION

- **The law of similarity**: If two things are similar, the thought of one will tend to trigger the thought of the other
- **The law of frequency**: The more often two things or events are linked, the more powerful will be that association.
- **The law of contiguity**: Things or events that occur close to each other in space or time tend to get linked together in the mind.
- **The law of contrast**: Seeing or recalling something may also trigger the recollection of something completely opposite.

DISTRIBUTED SEMANTIC MODELS

- Extract the meaning of the words using distributed linguistic properties
- Compute lexical **co-occurrence** of every word (co-locates with certain distance) with every other word in the Vocabulary
  - Linear proximity of words within a window is considered
  - They need not represent any relations

  Example He drove the car through a red bridge.
  The verb drove relates to red and bridge only through the proximity,
  but carries no relations with red and bridge in terms of semantics
- Build a co-occurrence matrix using co-occurrence statistics
- Rows/columns in the matrix represent distributed semantic information of words

Distributed Semantic Models

DISTRIBUTED SEMANTIC MODELS

Context is essential in the mapping the relationship between words and concepts.

I cook dinner every Sunday
...
I cooked dinner last Sunday
...
I am cooking dinner today
...
My son cooks dinner every Sunday
...

- The words in this corpus are related by association
- The verb forms cook, cooked, cooks, and cooking are related through their co-occurrence statistics - a semantic relationship
- The words dinner and Sunday are linked by an associative relationship arising from co-occurrence

- In the COVID19 research corpus, one may not find the phrase *needle in a haystack*
- You will find the word needle will be lexically associated with $\underbrace{\texttt{pain}, \texttt{illness}, \texttt{blood}, \texttt{drugs}, \texttt{syringe}}_{\text{Associative relationship}}$, but not to thread, knitting, cloth, etc.

You shall know a word by the company it keeps

- Firth, 1957

WORDS AND TERMS

The atomic unit is a ***word***

- Words are built from the alphabet of a language; we treat the ***word*** (or ***term***, which may span co-occurring words) as the atomic unit
- For computation, every word needs a numerical representation
- The vocabulary $V = \{w_1, w_2, w_3, \ldots, w_n\}$ is the set of the n unique words of a corpus
- A word of V may appear in a document of the collection $D = \{D_1, D_2, D_3, \ldots, D_m\}$ once, several times, or not at all

TERM-FREQUENCY WEIGHTING SCHEMES

$$\text{Boolean} - 0, 1 \tag{1}$$

$$\text{Raw count} - \mathrm{tf}_{i,d} \tag{2}$$

$$\text{Adjusted to document length} - \frac{\mathrm{tf}_{i,d}}{M_d} \tag{3}$$

$$\text{Log weighting} - \begin{cases} 1 + \log \mathrm{tf}_{i,d} & \text{if } \mathrm{tf}_{i,d} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathrm{tf}_{i,d}$ is the count of term i in document d and M_d is the number of tokens in d

DISADVANTAGES OF RAW FREQUENCY

- All terms are treated as equally important
- A common term such as ***the*** carries no information about the document, yet receives a high score
- Classification suffers when the score is dominated by common words shared across documents

INVERSE DOCUMENT FREQUENCY

We want to scale down the weight of frequently occurring terms and, at the same time, boost the weight of rarely occurring terms.
Inverse document frequency (IDF) is defined as

$$\mathrm{IDF}_t = \log\left(\frac{N}{df_t}\right) \tag{5}$$

where N is the total number of documents in the collection, and df_t is the number of documents containing the term t

- Rare terms get a significantly higher weight
- Commonly occurring terms are attenuated
- It is a measure of informativeness
- The tf weight of a term is reduced by a factor that grows with the number of documents it appears in
- If a term appears in every document, its IDF is zero. The term does not discriminate between documents

TF-IDF

Composition of TF and IDF produces a composite scaling for each term in the document

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_{t,d} \tag{6}$$

- The value is high when t occurs many times within a few documents
- The value is very low when a term appears in all documents

INVERSE DOCUMENT FREQUENCY-IDF

IDF measures how rare the term t is across the document collection. It is computed as

$$\text{IDF of a term t} = \log_{10}\left(\frac{\text{Total number of documents in a corpus}}{\text{Count of documents with term t}}\right)$$

IDF is the measure of ***informativeness***

Example:

Consider a corpus with 100K documents. The word ***moon*** occurs in some documents (say, 100) with the following frequency:

$\text{TF}_{d_1} = \frac{20}{427}, \text{TF}_{d_2} = \frac{30}{250}, \text{TF}_{d_3} = \frac{20}{250}, \text{TF}_{d_9} = \frac{5}{125}$ and $\text{TF}_{d_{1000}} = \frac{20}{1000}$

The total number of documents in the corpus $= 100{,}000$

$$\therefore \text{IDF}_{d_1} = \log_{10}\left(\frac{100000}{100}\right)$$

$$\text{TF}_{d_1} * \text{IDF} = 0.141$$

If the word ***Andromeda*** appears only once d_1, then $\text{TF}_{d_1} * \text{IDF} = 0.0117$. If the word ***the*** appeared in every document and 45 times in d1, then $\text{TF} * \text{IDF} = 0.210$

DOCUMENT RANKING USING TF-IDF

Using TF-IDF, the documents can be rank-ordered for the query term *moon*.

Document Name	tf-idf	Rank
d1	0.14	3
d2	0.36	1
d3	0.24	2
d9	0.12	4
d1000	0.06	5

BAG OF WORDS

Representing a document by its term weights alone, $[\mathrm{tf}_1, \mathrm{tf}_2, \mathrm{tf}_3, \ldots, \mathrm{tf}_n]$, is known as the *bag-of-words* model

- The order of the terms is ignored. Only their counts matter
- Two documents with similar bags of words are assumed to be similar in content
- It is a purely quantitative representation of the document
- Contrast: modern contextual models (transformers) explicitly model word order through position information

Empirical Laws

ZIPF'S LAW

This empirical law models the frequency distribution of words across various languages. Zipf's law [1] states that, in a given corpus, the frequency of any word is inversely proportional to its rank in the term-frequency table.

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{7}$$

where $\alpha \approx 1$, r is the *frequency rank* of a word and $f(r)$ is the frequency in the corpus.

Distribution of terms/words

This empirical law models the frequency distribution of words in languages. This distribution is observed across several languages when a large corpus is used.

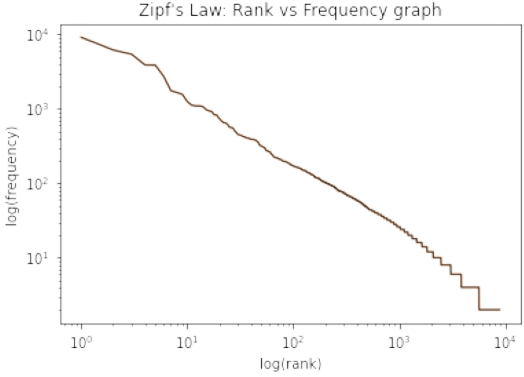
ZIPF'S LAW - FREQUENCY VS RANK

Output generated from Indian budget speeches using the bash script [26] with minimal preprocessing

Word	Frequency	Rank
the	59042	1
of	46110	2
to	35873	3
and	24661	4
in	22916	5
for	14426	6
a	14366	7
is	11155	8
be	9760	9
I	9070	10
...	...	...

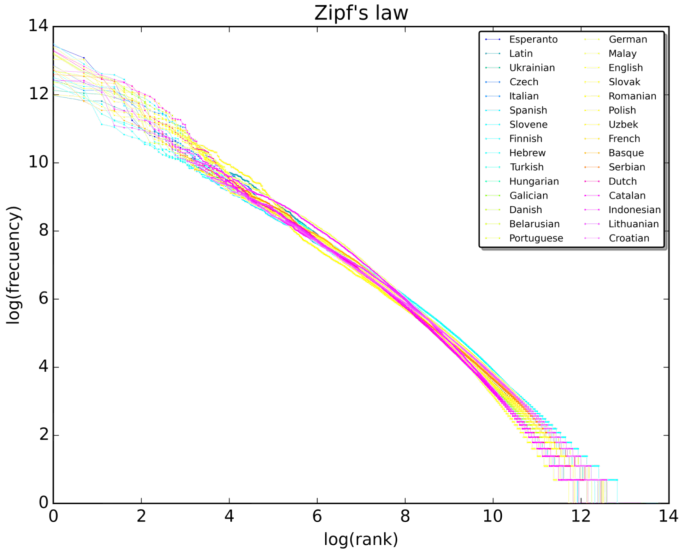
Word	Frequency	Rank
agricultural	479	233
important	478	234
policy	477	235
reduced	476	236
balance	476	237
...	...	...
(SFB)	1	92363
(SFAC),1	1	92364
(SFAC)	1	92365
(SEZs).,1	1	92366
...	...	...

ZIPF'S LAW I



$$f \propto \frac{1}{r^{\alpha}}$$

A plot of the rank versus frequency in a log-log scale.



$$f \propto \frac{1}{r^{\alpha}}$$

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedia (dumps from October 2015) in a log-log scale.

Source: Zipf's law -Wikipedia

Web Traffic Patterns

- Google.com dominates with billions of daily visits
- The top 20% of websites account for roughly 80% of total traffic
- Follows the Pareto principle

Caching Strategies

- CDNs and single-page apps exploit Zipf's law to optimize content delivery
- Small number of web pages receive the majority of requests
- Caching popular pages to improve response times

Search Algorithms

- Most searches focus on a limited set of popular topics

Text Analytics

- Helps in identifying stopwords
- Helps in identifying the important words to focus on
- Reduces the sample space
- Motivates frequency-based subsampling and vocabulary truncation in word-embedding and language models (e.g., word2vec subsampling, BPE vocabularies)

BASH SCRIPT FOR COMPUTING RANK AND FREQUENCY

```
# 10. Concatenate all files in TXT directory
# 11. Remove form feed character (^L)
# 12. Split into words (replace spaces with newlines)
# 13. Sort in reverse
# 14. Count unique words
# 15. Sort by frequency, descending
# 16. Add word, frequency rank
# 17. output as word_freq_rank.csv

cat xx.txt | \
tr -d '\f' | \
tr -s ' ' '\n' | \
sort -r | \
uniq -c | \
sort -nr | \
awk '{print $2 "," $1 "," NR }' \
> word_freq_rank.csv
```

MANDELBROT APPROXIMATION

Mandelbrot derived a more general law that fits the observed frequency distribution more closely by adding an offset to the rank

$$f(r) \propto \frac{1}{(r+\beta)^{\alpha}} \tag{8}$$

The offset β mainly improves the fit for the most frequent words (small r), where plain Zipf overestimates the frequency.

TYPE–TOKEN RATIO (TTR)

- $\text{TTR} = \dfrac{\text{number of distinct words (types)}}{\text{total number of words (tokens)}}$
- A measure of lexical diversity: higher TTR $\Rightarrow$ richer vocabulary
- TTR falls as text length grows (a consequence of Heaps' law), so compare documents only over equal-sized samples (standardized TTR)

APPLICATIONS OF TTR

- Monitor vocabulary usage in a text or by a writer
- Track child vocabulary development
- Estimate vocabulary variation across genres or authors
- Detect templated or repetitive machine-generated text

EXERCISES

- Write a program to find out whether Mandelbrot's approximation provides a better fit than Zipf's law. Use the same corpus for Zipf and Mandelbrot approximation.
- Find the percentage of tokens that are stop words.
- Find the top ten words by frequency.
- Fit Heaps' law to a corpus and predict its vocabulary size. Compare the prediction with the actual vocabulary size of the same corpus.

OPEN-CLASS AND STOP WORDS

Open-class words: Carry the main semantic content of a sentence

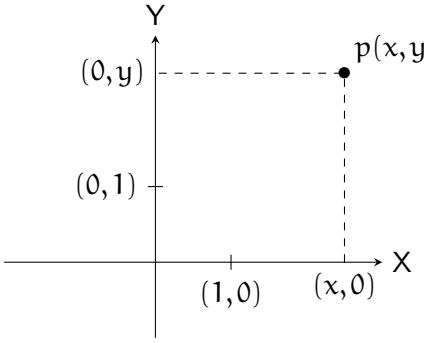
- Contribute to the overall meaning of a sentence
- Include nouns, verbs, adjectives, and adverbs
- **Expansive**: New words can be added to this category over time (e.g., "vlog," "emoji", “Mulligatawny”)
- **Adaptability**: New terms to express emerging concepts, technologies, and innovation

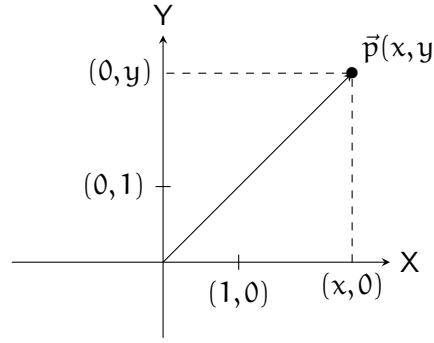
Closed-class (function) words, the usual stop words, serve grammatical purposes

- Include
  - prepositions (in, on, with, ...)
  - conjunctions (and, but, or..)
  - articles (a, an, the)
  - pronouns(he, she, they...)
- Essential for the grammatical integrity of sentences
- Contribute less to the overall meaning of the sentence

2-D VECTOR SPACE

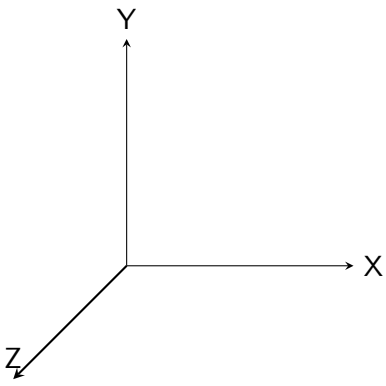
A 2-D vector-space is defined as a set of linearly independent basis vectors with 2 axes. Each axis corresponds to a dimension in the vector-space

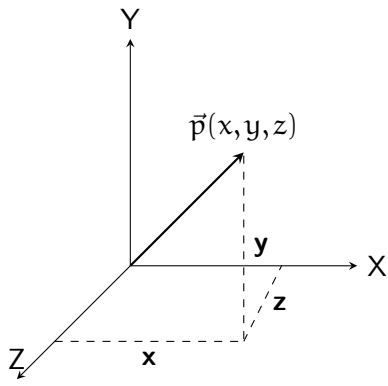




3-D VECTOR SPACE

A 3-D vector-space is defined as a set of linearly independent basis vectors with 3 axes. Each axis corresponds to a dimension in the vector-space





Linearly independent vectors of size $\mathcal{N}$ will result in $\mathcal{N}$-dimensional axes which are mutually orthogonal to each other

VECTOR SPACE MODEL FOR WORDS

Let us assume that the words in a corpus are considered as linearly independent basis vectors.

VECTOR SPACE MODEL FOR WORDS

Let us assume that the words in a corpus are considered as linearly independent basis vectors.
If a corpus contains $|\mathbb{V}|$ words which are linearly independent, then every word represents an axis in the continuous vector space $\mathbb{R}$.

VECTOR SPACE MODEL FOR WORDS

Let us assume that the words in a corpus are considered as linearly independent basis vectors.
If a corpus contains $|\mathbb{V}|$ words which are linearly independent, then every word represents an axis in the continuous vector space $\mathbb{R}$.
Each word takes an independent axis which is orthogonal to other words/axes.

VECTOR SPACE MODEL FOR WORDS

Let us assume that the words in a corpus are considered as linearly independent basis vectors.
If a corpus contains $|\mathbb{V}|$ words which are linearly independent, then every word represents an axis in the continuous vector space $\mathbb{R}$.
Each word takes an independent axis which is orthogonal to other words/axes.
Then $\mathbb{R}$ will contain $|\mathbb{V}|$ axes.

Let us assume that the words in a corpus are considered as linearly independent basis vectors.
If a corpus contains $|\mathbb{V}|$ words which are linearly independent, then every word represents an axis in the continuous vector space $\mathbb{R}$.
Each word takes an independent axis which is orthogonal to other words/axes.
Then $\mathbb{R}$ will contain $|\mathbb{V}|$ axes.

Examples

1. The vocabulary size of *emma corpus* is 7079. If we plot all the words in the real space $\mathscr{R}$, we get 7079 axes
2. The vocabulary size of *Google News Corpus corpus* is 3 million. If we plot all the words in the real space $\mathscr{R}$, we get 3 million axes

DOCUMENT VECTOR SPACE MODEL

- Vector space models are used to represent words in a continuous vector space $\mathscr{R}$

DOCUMENT VECTOR SPACE MODEL

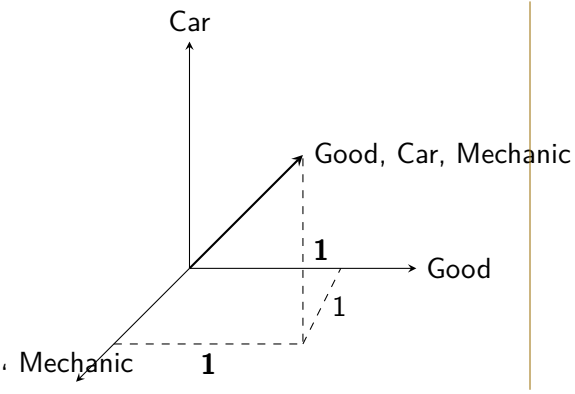
- Vector space models are used to represent words in a continuous vector space $\mathscr{R}$
- Combination of Terms represent a document vector in the word vector space

DOCUMENT VECTOR SPACE MODEL

- Vector space models are used to represent words in a continuous vector space $\mathscr{R}$
- Combination of Terms represent a document vector in the word vector space
- Very high dimensional space - several million axes, representing terms and several million documents containing several terms

EXAMPLE

Let us consider three words - *good, car, mechanic* and we will represent these words in a 3-D vector space



	Good	Car	Mechanic
D1	1	1	1
D2	1	0	1
D3	0	1	1

Term-term representation of words

DOCUMENT-TERM MATRIX

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12
t1	0.1	0.0	0.4	0.1	0.2	0.0	0.1	0.9	0.9	0.3	0.0	0.8
t2	0.1	0.0	0.4	0.1	0.2	0.0	0.1	0.9	0.9	0.3	0.0	0.8
t3	0.0	0.9	0.0	0.2	0.3	0.1	0.7	0.0	0.2	0.7	0.5	0.5
t4	0.0	0.9	0.3	0.9	0.5	0.1	0.9	0.3	0.8	0.4	0.1	0.4
t5	0.4	0.0	0.3	0.2	0.5	0.9	0.3	0.7	0.4	0.6	0.0	0.3
t6	0.6	0.0	0.4	0.7	0.3	0.3	0.9	0.1	0.9	0.0	0.0	0.3
t7	0.0	0.8	0.5	0.6	0.6	0.6	0.0	0.1	0.4	0.9	0.3	0.1
t8	0.4	0.0	0.6	0.5	0.5	0.1	0.7	0.1	0.5	0.3	0.8	0.1
t9	0.3	0.0	0.7	0.9	0.8	0.7	0.7	0.8	0.6	0.6	0.8	0.0
t10	0.0	0.5	0.5	0.0	0.2	0.0	0.0	0.1	0.3	0.4	0.5	0.3

The columns of the matrix represent the document as vectors. A document vector is represented by the terms present in the document

TERM-TERM MATRIX

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
t1	0.1	0.0	0.4	0.1	0.2	0.0	0.1	0.9	0.9	0.3	0.0	0.8
t2	0.1	0.0	0.4	0.1	0.2	0.0	0.1	0.9	0.9	0.3	0.0	0.8
t3	0.0	0.9	0.0	0.2	0.3	0.1	0.7	0.0	0.2	0.7	0.5	0.5
t4	0.0	0.9	0.3	0.9	0.5	0.1	0.9	0.3	0.8	0.4	0.1	0.4
t5	0.4	0.0	0.3	0.2	0.5	0.9	0.3	0.7	0.4	0.6	0.0	0.3
t6	0.6	0.0	0.4	0.7	0.3	0.3	0.9	0.1	0.9	0.0	0.0	0.3
t7	0.0	0.8	0.5	0.6	0.6	0.6	0.0	0.1	0.4	0.9	0.3	0.1
t8	0.4	0.0	0.6	0.5	0.5	0.1	0.7	0.1	0.5	0.3	0.8	0.1
t9	0.3	0.0	0.7	0.9	0.8	0.7	0.7	0.8	0.6	0.6	0.8	0.0
t10	0.0	0.5	0.5	0.0	0.2	0.0	0.0	0.1	0.3	0.4	0.5	0.3
t11	0.01	0.2	0.4	0.1	0.2	0.2	0.0	0.0	0.0	0.1	0.2	0.0
t12	0.1	0.12	0.54	0.01	0.02	0.0	0.0	0.0	0.0	0.6	0.7	0.0

The columns and rows of the matrix represent the words as vectors.

Do they represent the contextual relationship among words?

A similarity measure is a real-valued function that quantifies the similarity between two objects - in this case words Some of the similarity measures are given below.

$$\textbf{Euclidean Distance - } \mathscr{E}(\vec{w_1}, \vec{w_2}) = \sqrt{w_1^2 - w_2^2}$$

$$\textbf{Cosine Similarity} = \frac{\vec{w_1}.\vec{w_2}}{\|\vec{w_1}\| \|\vec{w_2}\|} = \frac{\vec{w_1}}{\|\vec{w_1}\|} \cdot \frac{\vec{w_2}}{\|\vec{w_2}\|}$$

$$\textbf{Cosine distance} = 1 - \frac{\vec{w_1}.\vec{w_2}}{\|\vec{w_1}\| \|\vec{w_2}\|} = \frac{\vec{w_1}}{\|\vec{w_1}\|} \cdot \frac{\vec{w_2}}{\|\vec{w_2}\|}$$

$$\textbf{Cluster similarity-} \mathscr{L}(\vec{w_1}, \vec{w_2}) = \frac{\vec{w_1}.\vec{w_2}}{\|\vec{w_1}\|}$$

Vector Semantics

VECTOR REPRESENTATION OF WORDS

Let V be the unique set of terms and $|V|$ be the size of the vocabulary. Then every vector representing the word $\mathscr{R}^{|V|x1}$ would point to a vector in the V-dimensional space

ONE-HOT VECTOR - 1

Consider all the ≈39000 words (estimated tokens in English is ≈ 13M) in the Oxford Learner's pocket dictionary. We can represent each word as an independent vector quantity as follows in the real space $\mathscr{R}^{|V|X1}$

$$t^{a} = \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} \quad t^{aback} = \begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} \quad \dots \; t^{zoom} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ \dots \\ 1 \\ 0 \end{pmatrix} \quad t^{zucchini} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ \dots \\ 0 \\ 1 \end{pmatrix}$$

This is a very simple codification scheme to represent words independently in the vector space. This is known as ***one-hot vector***.

ONE-HOT VECTOR - 2

In one-hot vector, every word is represented independently. The terms, *home, house, apartments, flats* are independently coded. With one-hot vector based model, the dot product

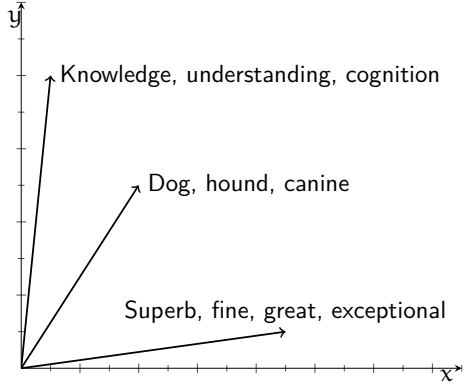
$$\left(t^{\text{House}}\right)^{T} \cdot t^{\text{Apartment}} = 0 \tag{9}$$

$$\left(t^{\text{Home}}\right)^{T} \cdot t^{\text{House}} = 0 \tag{10}$$

With one-Hot vector, there is no notion of similarity or synonyms.

RELATIONSHIP AMONG TERMS - SYNONYMS

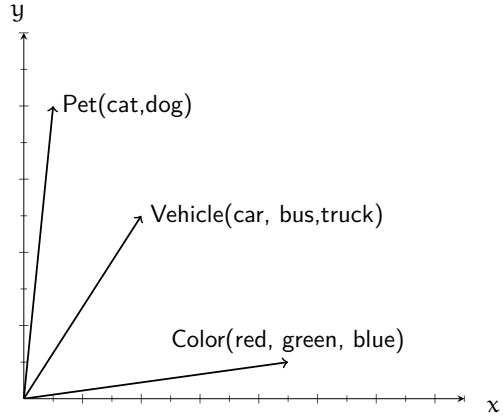
We could represent all the synonyms of a word in one axis



- Can we assume that similar words be situated/placed in the respective concept space?
- Can we assume that the properties of the word that inherit the properties of that space?

RELATIONSHIP AMONG TERMS - IS-A VECTOR

We could represent inheritance relationships of words as vectors.



SYNONYMS

small.a.01	['small', 'little']
minor.s.10	['minor', 'modest', 'small', 'small-scale', 'pocket-size', 'pocket-sized']
humble.s.01	['humble', 'low', 'lowly', 'modest', 'small']
little.s.07	['little', 'minuscule', 'small']
belittled.s.01	['belittled', 'diminished', 'small']
potent.a.03	['potent', 'strong', 'stiff']
impregnable.s.01	['impregnable', 'inviolable', 'secure', 'strong', 'unassailable', 'hard']
	He has such an impregnable defense (Cricket-Very hard to find the gap between the bat and the pad)
solid.s.07	['solid', 'strong', 'substantial']
strong.s.09	['strong', 'warm']
firm.s.03	['firm', 'strong'] - firm grasp of fundamentals

POLYSEMOUS WORD - BANK

Synset('bank.n.01')	sloping land (especially the slope beside a body of water)
Synset('depository-financial-institution.n.01')	a financial institution that accepts deposits and channels the money into lending activities
Synset('bank.n.03')	a long ridge or pile
Synset('bank.n.10')	a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)
Synset('trust.v.01')	have confidence or faith in

Bank appears in different word senses - or the meaning of the word is determined by the context in which appears

How do we place theses polysemous word in an axis? Using multiple vectors?

CONTEXTUAL UNDERSTANDING OF TEXT

You shall know a word by the company it keeps - (Firth, J. R. 1957)

- In order to understand the word and its meaning, it not enough if we consider only the individual word
- The *meaning* and *context* should be central in understanding word/text
- Exploit the context-dependent nature of words
- Language patterns cannot be accounted for in terms of a single entity
- The *collocation*, a particular word consistently co-occurs with the other words, gives enough clue to understand a word and its meaning

UNDERSTANDING A WORD FROM ITS CONTEXT

The view from the top of the mountain was
The view from the summit was
La vue du sommet de la montagne était
Mtazamo wa juu wa mlima huo ulikuwa

awesome/(*impressionnante*, *impressionnant*)
breathtaking
amazing, அற்புதமான/అద్భుతకరమాయ/
stunning/(*superbe*)
అద్భుతమైన/అద్భుతమైన
astounding अद्भुत/চমকপ্রদ
astonishing
awe-inspiring
extraordinary
incredible/(*incroyable*)
unbelievable
magnificent शानदार/ગંભીરમાય/ભવ્ય
wonderful/(*ajabu*)
spectacular
remarkable/(*yakuvutia*)

AJI AMARILLO

- Heard of Aji Amarillo?

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale
- It is not as hot as bhüt jolokia which measures 1 million on the same scale

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale
- It is not as hot as bhüt jolokia which measures 1 million on the same scale
- If you visit Peru, do not miss dishes made from this

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale
- It is not as hot as bhüt jolokia which measures 1 million on the same scale
- If you visit Peru, do not miss dishes made from this
- They are yellow in color and have a balanced, fruity flavor close to mango and even passion fruit

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale
- It is not as hot as bhüt jolokia which measures 1 million on the same scale
- If you visit Peru, do not miss dishes made from this
- They are yellow in color and have a balanced, fruity flavor close to mango and even passion fruit
- It is a member of the capsicum baccatum

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale
- It is not as hot as bhüt jolokia which measures 1 million on the same scale
- If you visit Peru, do not miss dishes made from this
- They are yellow in color and have a balanced, fruity flavor close to mango and even passion fruit
- It is a member of the capsicum baccatum
- Star ingredient in many Peru's dishes

AJI AMARILLO

- Heard of Aji Amarillo?
- It measures 30K/50K on the Scoville Heat scale
- It is not as hot as bhüt jolokia which measures 1 million on the same scale
- If you visit Peru, do not miss dishes made from this
- They are yellow in color and have a balanced, fruity flavor close to mango and even passion fruit
- It is a member of the capsicum baccatum
- Star ingredient in many Peru's dishes

Inference

Aji Amarillo is a hot food item
Aji Amarillo is a vegetable
Aji Amarillo is grown in Peru

Aji Amarillo is a member of capsicum baccatum family

IDEAL PROPERTIES OF WORD VECTORS

- Encode the relationship among words
- Identify similar words using
  - Law of similarity - "**car**" and "**automobile**" vectors are close due to shared context
  - Law of contiguity - "**coffee**" and "**cup**" vectors show proximity because they occur together
  - Law of contrast - "**hot**" and "**cold**" vectors are distinct yet related as antonyms
  - Reduce word-vector space into a smaller sub-space
- Extract semantic information - capture words that demonstrates context, represents relationship
  New Delhi, India, capital - semantically connected through geography and semantics
- Represent polysemous words - multiple senses of a word based on different contexts

SEMANTICALLY CONNECTED VECTORS

- Identify a model that enumerates the relationships between terms
- Identify a model that tries to place similar items closer to each other in some space or structure
- Build a model that discovers/uncovers the semantic similarity between words and documents in the latent semantic domain
- Develop a distributed word vectors or dense vectors that captures the linear combination of word vectors in the transformed domain
- Similar to the term-document space identify a semantic space for similar words

WORD SIMILARITY

- Sparse vectors are too long and not very convenient as features machine learning
- Abstracts more than just frequency counts
- It captures neighborhood words that are connected by synonyms

METHODS TO CREATE WORD VECTORS

- Brown clustering - statistical algorithms for assigning words to classes based on the frequency of their co-occurrence with other words. It creates a binary tree of word clusters where words in the same cluster tend to appear in similar contexts, helping capture semantic relatedness
- Hyperspace Analogue to Language - HAL
- Correlated Occurrence Analogue to Lexical Semantic - COALS
- Latent Semantic Analysis or Latent Semantic Indexing
- Global Vectors - GloVe
- Neural networks using skip grams and CBOW
  - CBOW - uses surrounding words to predict the center of words
  - Skip grams use center of words to predict the surrounding words

You shall know a word by the company it keeps

- Firth, 1957

ATTRIBUTIONAL SIMILARITY

Attributional Similarity

- Two words are similar if they shared similar attributes - cat and kitten, dog and puppy
- Refers to the degree of similarity between two words or phrases in terms of their shared attributes
- Words that share many collocates denote concepts that share many attributes

Relational Similarity

- Related by concepts/roles - King and queen are related by the roles in the monarchy
- Blood relationships - siblings, aunts, uncles, parents, etc.

Techniques to extract similarities - Distributional Semantic Models

VECTOR BASED MODELS

Assumption Context words within a certain distance from the target word are semantically relevant

- Ability to represent word meaning simply by using distributional statistics
- The context surrounding a given word provides important information about its meaning
  - Small number of words surrounding the target word is known as context
- Distributional patterns of co-occurrence with their neighboring words provide semantic properties of words

A SAMPLE CO-OCCURRENCE MATRIX WITH RAW FREQUENCY COUNT

	w_0	w_1	w_2	...	w_{n-3}	w_{n-2}	w_{n-1}	w_n
w_0	0	33	29	...	33	37	39	39
w_1	33	1	45	...	0	27	21	10
w_2	29	45	0	...	37	40	19	23
...	...	...	...	...	...	...	...	...
w_{n-3}	33	0	37	...	0	24	26	49
w_{n-2}	37	27	40	...	24	0	22	31
w_{n-1}	39	21	19	...	26	22	1	38
w_n	39	10	23	...	49	31	38	0

- This matrix captures the lexical space of the corpus
- Statistical knowledge about words and their relationship in terms of co-occurrence are static in nature

TECHNIQUES TO CAPTURE CO-OCCURRENCE INFORMATION

Let A denote the word-word co-occurrence matrix where each row corresponds to a unique/target word, and each column represents a context.
a_{ij} denotes every element of the A

a_i denotes the number of times the word i co-occurring with the word j, $a_i = \sum_j a_{ij}$

Now, we can fill the co-occurrence using:

1. **Raw Frequency Count:** Each element, a_{ij} denotes the raw frequency count of word i co-occurring with the word j
2. **Probability:** $p_{ij} = P(w_j \mid w_i) = \frac{a_{ij}}{a_i}$
3. **Positive Point-wise Mutual Information(PMI)**:
   $PPMI(w_i, w_j) = max\left(\log_2\left(\frac{p_{ij}}{p_i * p_j}\right), 0\right)$

The co-occurrence statistics are captured by scaning the entire corpus once using a windowing approach

IMPORTANCE OF PPMI IN WORD SEMANTIC RELATIONSHIPS

Quantifying True Semantic Association

Measures how much more often two words co-occur than expected by random chance, capturing strong contextual links (e.g., “ice” and “cold”).

Correcting Frequency Bias

Standard PMI overemphasizes rare words that appear together just once by chance. PPMI filters this uninformative, low-frequency background noise.

Eliminating Negative Noise

By replacing negative scores with $\mathbf{0}$ via $\max(0, \text{PMI})$, it yields highly accurate, zero-grounded semantic vector spaces ideal for similarity tasks.

EXAMPLE: COMPUTING PPMI WITH LAPLACE SMOOTHING

Corpus Setup: Total words $N = 1,000,000$. Smoothed Context Weight $\alpha = 0.75$.
Smoothed context counts: $c_\alpha(w) = N \times \left(\frac{c(w)}{N}\right)^\alpha$.

Frequent Pair: (data, science)

$c_{data} = 10,000$, $c_{science} = 1,000$
Raw Joint Count $= 100$

$$P_\alpha(\text{science}) \approx \frac{1,000^{0.75}}{1,000^{0.75} + \ldots} \approx 0.0031$$

$$P(\text{data}) = 0.01$$

$$P(\text{joint}) = 0.0001$$

$$PMI_\alpha = \log_2\left(\frac{0.0001}{0.01 \times 0.0031}\right) = 1.69$$

$$PPMI = \max(0, 1.69) = 1.69$$

Rare Pair: (data, cryptozoology)

$c_{data} = 10,000$, $c_{crypto} = 2$
Raw Joint Count $= 1$ (Fluke)

$$P_\alpha(\text{crypto}) \approx \frac{2^{0.75}}{1,000^{0.75} + \ldots} \approx 0.00003$$

$$P(\text{data}) = 0.01$$

$$P(\text{joint}) = 0.000001$$

$$PMI_\alpha = \log_2\left(\frac{0.000001}{0.01 \times 0.00003}\right) = -1.58$$

$$PPMI = \max(0, -1.58) = 0$$

Conclusion: Raising rare counts to $\alpha = 0.75$ inflates their expected baseline probability, forcing the PMI due to uncommon occurrence to negative so PPMI safely squashes it to 0.

SIMILARITY MEASURES

A similarity measure is a real-valued function that quantifies the similarity between two objects - in this case words. Some of the similarity measures are given below.

$$\textbf{Euclidean Distance} - \mathbb{E}(\vec{w_1}, \vec{w_2}) = \sqrt{w_1^2 - w_2^2} \tag{11}$$

$$\underset{\in[-1,1]}{\textbf{Cosine Similarity}} = \frac{\vec{w_1}.\vec{w_2}}{\|\vec{w_1}\| \, \|\vec{w_2}\|} \tag{12}$$

$$\underset{\in[0,1]}{\textbf{Cosine Distance}} = 1 - \textbf{Cosine Similarity} \tag{13}$$

WORD VECTOR EXAMPLES

Similar words for $\overrightarrow{\text{apple}}$

Term	Score
apple	0.000
iphone	0.266
ipad	0.287
apples	0.356
blackberry	0.361
ipod	0.365
macbook	0.383
mac	0.391
android	0.391
google	0.395
microsoft	0.418
ios	0.433
iphones	0.445
touch	0.446
sony	0.447

Similar words for $\overrightarrow{\text{american}}$

Word	Score
american	0.000
america	0.255
americans	0.312
u.s.	0.320
british	0.323
canadian	0.329
history	0.356
national	0.364
african	0.374
society	0.375
states	0.386
european	0.387
world	0.394
nation	0.399
us	0.399

VECTOR DIFFERENCE BETWEEN TWO WORDS

Table: Word Similarity Scores

$\overrightarrow{\text{Word}}$	$\overrightarrow{\text{Score}}$
raisin	0.5745
pecan	0.5761
cranberry	0.5840
butternut	0.5882
cider	0.5911
apricot	0.6037
tomato	0.6074
rosemary	0.6151
rhubarb	0.6158
feta	0.6183
apples	0.6226
avocado	0.6235
fennel	0.6306
chutney	0.6313
spiced	0.6328

VECTOR ARITHMETIC ON WORD VECTORS...

840B words and 300 elements word vectors used for this computation

$\overrightarrow{\text{apple}}$	$\overrightarrow{\text{apple}} - \overrightarrow{\text{iphone}}$	$\overrightarrow{\text{apple}} - \overrightarrow{\text{fruit}}$
('apple', 0)	('apples', 0.39)	('ipad', 0.412)
('apples', 0.25)	('fruit', 0.43)	('iphone', 0.433)
('blackberry', 0.31)	('grape', 0.44)	('macbook', 0.435)
('Apple', 0.35)	('tomato', 0.44)	('ipod', 0.445))
('iphone', 0.37)	('pecan', 0.45)	('imac', 0.465
('fruit', 0.37)	('rhubarb', 0.45)	('3gs', 0.473)
('blueberry', 0.38)	('pears', 0.45)	('Ipad', 0.490)
('strawberry', 0.38)	('cranberry', 0.452)	('itouch', 0.512)
('ipad', 0.39)	('raisin', 0.453)	('ipad2', 0.514)
('pineapple', 0.39)	('apricot', 0.459)	('Iphone', 0.514)
('pear', 0.39)	('carrot', 0.461)	('ios', 0.520)
('cider', 0.39)	('candied', 0.462)	('Macbook', 0.524)
('mango', 0.40)	('blueberry', 0.463)	('ibook', 0.534)
('ipod', 0.40)	('apricots', 0.466)	('IPhone', 0.541)
('raspberry', 0.40)	('tomatoes', 0.466)	('32gb', 0.545)

REFERENCES I

[1] George Kingsley Zipf. “The psychology of language”. In: *Encyclopedia of psychology*. Philosophical Library, 1946, pp. 332–341.