Word Vectors Introduction

Course:
INFO-6145 Data Science and Machine Learning

FANSHAWE

Revised by:
Mohammad Noorchenarboo

October 22, 2024

Contents

1. Word Vectors and Semantic Relations in NLP
   - Introduction to Word Vectors
   - Semantic Relations and WordNet®
   - One Hot Vectors
   - Word Vectors and Word2Vec

Current Section

1. Word Vectors and Semantic Relations in NLP
   - Introduction to Word Vectors
   - Semantic Relations and WordNet®
   - One Hot Vectors
   - Word Vectors and Word2Vec

Word Vectors Introduction

Word vectors are a way to represent words numerically so that computers can process their meanings. They map words to vectors in a mathematical space where the distance between the vectors indicates how similar the words are.

Key Concepts

- **Syntax:** The arrangement of words or symbols to create a valid sentence or a computer program.
- **Semantics:** What the words or symbols actually mean.

WordNet®

WordNet is a lexical database of semantic relations between words, linking them into relations such as synonyms, hyponyms, and meronyms.

Definitions

- **Synonym:** Words that mean exactly or nearly the same thing (e.g., happy, joyful).
- **Hyponym:** A word that has a more specific meaning (e.g., fork is a hyponym of cutlery).
- **Meronym:** A word that is part of a whole (e.g., leaf is part of a tree).

WordNet®

WordNet Example

WordNet groups words into "synsets" (sets of synonyms) and provides short definitions and usage examples for each word. It extends the functionality of a dictionary and thesaurus.

One Hot Vectors

One hot vectors are a simple way to represent words in a dictionary. Each word is represented by a binary vector where all the values are zero, except for the position corresponding to the word of interest, which is marked as one.

One Hot Vector Example

Consider the word "monet" (Claude Monet, the French painter) in a dictionary: 0 0 0 1 0 0 ... Only the position for "monet" is marked as one, while all other positions are zero.

Limitations of One Hot Vectors

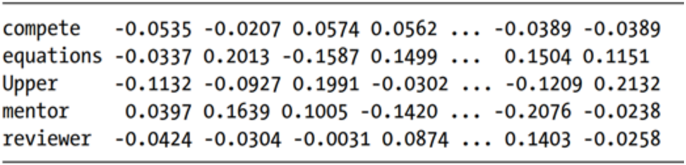
One hot vectors don't capture the meaning or relationships between words. "Monet" and "painting" would have no connection in this representation, even though they are related in meaning.

Word Vectors

Word vectors are a more sophisticated way to represent words, where each word is mapped to a vector in a space where semantically similar words have similar vectors.

Word2Vec

Word2Vec is a model that learns word vectors by processing large amounts of unlabeled text. No one has to manually label the words.



compete	-0.0535	-0.0207	0.0574	0.0562	...	-0.0389	-0.0389
equations	-0.0337	0.2013	-0.1587	0.1499	...	0.1504	0.1151
Upper	-0.1132	-0.0927	0.1991	-0.0302	...	-0.1209	0.2132
mentor	0.0397	0.1639	0.1005	-0.1420	...	-0.2076	-0.0238
reviewer	-0.0424	-0.0304	-0.0031	0.0874	...	0.1403	-0.0258

How Word2Vec Works

Word2Vec learns the meaning of words by analyzing their surrounding context. Words that occur near each other in a sentence are likely to have similar meanings.

Example of Word Context

Consider the sentence: "Claude Monet painted the Grand Canal of Venice in 1806." Word2Vec looks at each word in the context of the previous 2 words and the next 2 words. For example, it associates "Claude Monet" with "painted" and "Grand Canal."

Mathematical Representation

Word2Vec creates word vectors such that vector operations can reflect relationships between words. For example: **King - Man + Woman = Queen**

Example of Actual Word Vectors

Word2Vec generates actual word vectors, which can be used for various tasks such as finding similar words or analogies.

Example of Actual Word Vector Entry

A word vector for "monet" might look something like: monet: [0.02, 0.31, -0.14, 0.08, ...] This vector represents the relationships "monet" has with other words.

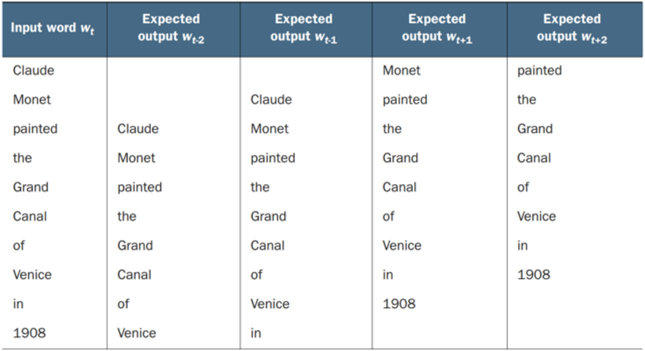
Accessing Pre-trained Word Vectors

Pre-trained word vectors can be downloaded from sources like Facebook's FastText (e.g., `https://fasttext.cc/docs/en/english-vectors.html`).

Examples

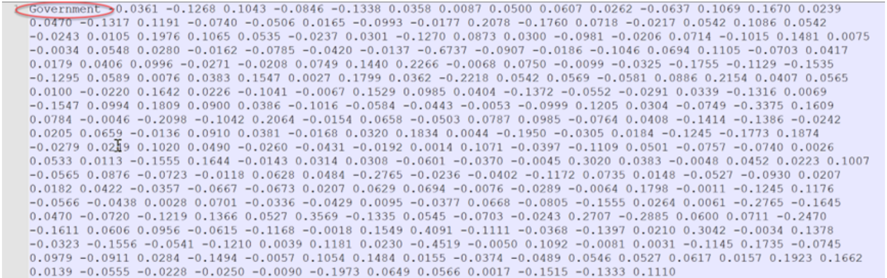
Consider the following sentence:

Claude Monet painted the Grand Canal of Venice in 1806.



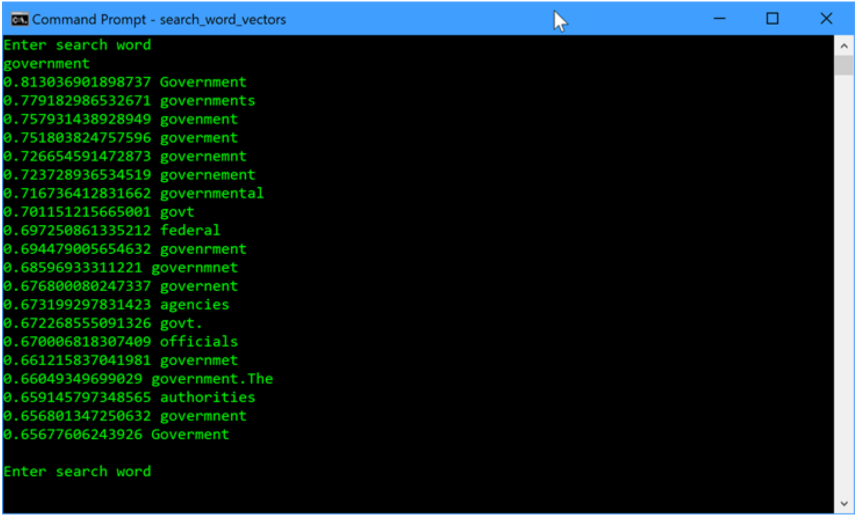
Input word w_t	Expected output w_{t-2}	Expected output w_{t-1}	Expected output w_{t+1}	Expected output w_{t+2}
Claude			Monet	painted
Monet		Claude	painted	the
painted	Claude	Monet	the	Grand
the	Monet	painted	Grand	Canal
Grand	painted	the	Canal	of
Canal	the	Grand	of	Venice
of	Grand	Canal	Venice	in
Venice	Canal	of	in	1908
in	of	Venice	1908	
1908	Venice	in		

Examples



```
Government -0.0361 -0.1268 0.1043 -0.0846 -0.1338 0.0358 0.0087 0.0500 0.0607 0.0262 -0.0637 0.1069 0.1670 0.0239
0.0470 -0.1317 0.1191 -0.0740 -0.0506 0.0165 -0.0993 -0.0177 0.2078 -0.1760 0.0718 -0.0217 0.0542 0.1086 0.0542
-0.0243 0.0105 0.1976 0.1065 0.0535 -0.0237 0.0301 -0.1270 0.0873 0.0300 -0.0981 -0.0206 0.0714 -0.1015 0.1481 0.0075
-0.0034 0.0548 0.0280 -0.0162 -0.0785 -0.0420 -0.0137 -0.6737 -0.0907 -0.0186 -0.1046 0.0694 0.1105 -0.0703 0.0417
0.0179 0.0406 0.0996 -0.0271 -0.0208 0.0749 0.1440 0.2266 -0.0068 0.0750 -0.0099 -0.0325 -0.1755 -0.1129 -0.1535
-0.1295 0.0589 0.0076 0.0383 0.1547 0.0027 0.1799 0.0362 -0.2218 0.0542 0.0569 -0.0581 0.0886 0.2154 0.0407 0.0565
0.0100 -0.0220 0.1642 0.0226 -0.1041 -0.0067 0.1529 0.0985 0.0404 -0.1372 -0.0552 -0.0291 0.0339 -0.1316 0.0069
-0.1547 0.0994 0.1809 0.0900 0.0386 -0.1016 -0.0584 -0.0443 -0.0053 -0.0999 0.1205 0.0304 -0.0749 -0.3375 0.1609
0.0784 -0.0046 -0.2098 -0.1042 0.2064 -0.0154 0.0658 -0.0503 0.0787 0.0985 -0.0764 0.0408 -0.1414 -0.1386 -0.0242
0.0205 0.0659 -0.0136 0.0910 0.0381 -0.0168 0.0320 0.1834 0.0044 -0.1950 -0.0305 0.0184 -0.1245 -0.1773 0.1874
-0.0279 0.0219 0.1020 0.0490 -0.0260 -0.0431 -0.0192 0.0014 0.1071 -0.0397 -0.1109 0.0501 -0.0757 -0.0740 0.0026
0.0533 0.0113 -0.1555 0.1644 -0.0143 0.0314 0.0308 -0.0601 -0.0370 -0.0045 0.3020 0.0383 -0.0048 0.0452 0.0223 0.1007
-0.0565 0.0876 -0.0723 -0.0118 0.0628 0.0484 -0.2765 -0.0236 -0.0402 -0.1172 0.0735 0.0148 -0.0527 -0.0930 0.0207
0.0182 0.0422 -0.0357 -0.0667 -0.0673 0.0207 0.0629 0.0694 -0.0076 -0.0289 -0.0064 0.1798 -0.0011 -0.1245 0.1176
-0.0566 -0.0438 0.0028 0.0701 -0.0336 -0.0429 0.0095 -0.0377 0.0668 -0.0805 -0.1555 0.0264 0.0061 -0.2765 -0.1645
0.0470 -0.0720 -0.1219 0.1366 0.0527 0.3569 -0.1335 0.0545 -0.0703 -0.0243 0.2707 -0.2885 0.0600 0.0711 -0.2470
-0.1611 0.0606 0.0956 -0.0615 -0.1168 -0.0018 0.1549 0.4091 -0.1111 -0.0368 -0.1397 0.0210 0.3042 -0.0034 0.1378
-0.0323 -0.1556 -0.0541 -0.1210 0.0039 0.1181 0.0230 -0.4519 -0.0050 0.1092 -0.0081 0.0031 -0.1145 0.1735 -0.0745
0.0979 -0.0911 0.0284 -0.1494 -0.0057 0.1054 0.1484 0.0155 -0.0374 -0.0489 0.0546 0.0527 0.0617 0.0157 0.1923 0.1662
0.0139 -0.0555 -0.0228 -0.0250 -0.0090 -0.1973 0.0649 0.0566 0.0017 -0.1515 -0.1333 0.1110
```

Examples



```
Enter search word
government
0.813036901898737 Government
0.779182986532671 governments
0.757931438928949 govenment
0.751803824757596 goverment
0.726654591472873 governemnt
0.723728936534519 governement
0.716736412831662 governmental
0.701151215665001 govt
0.697250861335212 federal
0.694479005654632 govenrment
0.68596933311221 governmnet
0.676800080247337 governent
0.673199297831423 agencies
0.672268555091326 govt.
0.670006818307409 officials
0.661215837041981 governmet
0.66049349699029 government.The
0.659145797348565 authorities
0.656801347250632 govermnent
0.65677606243926 Goverment

Enter search word
```