manipal PROlearn

Text Analytics

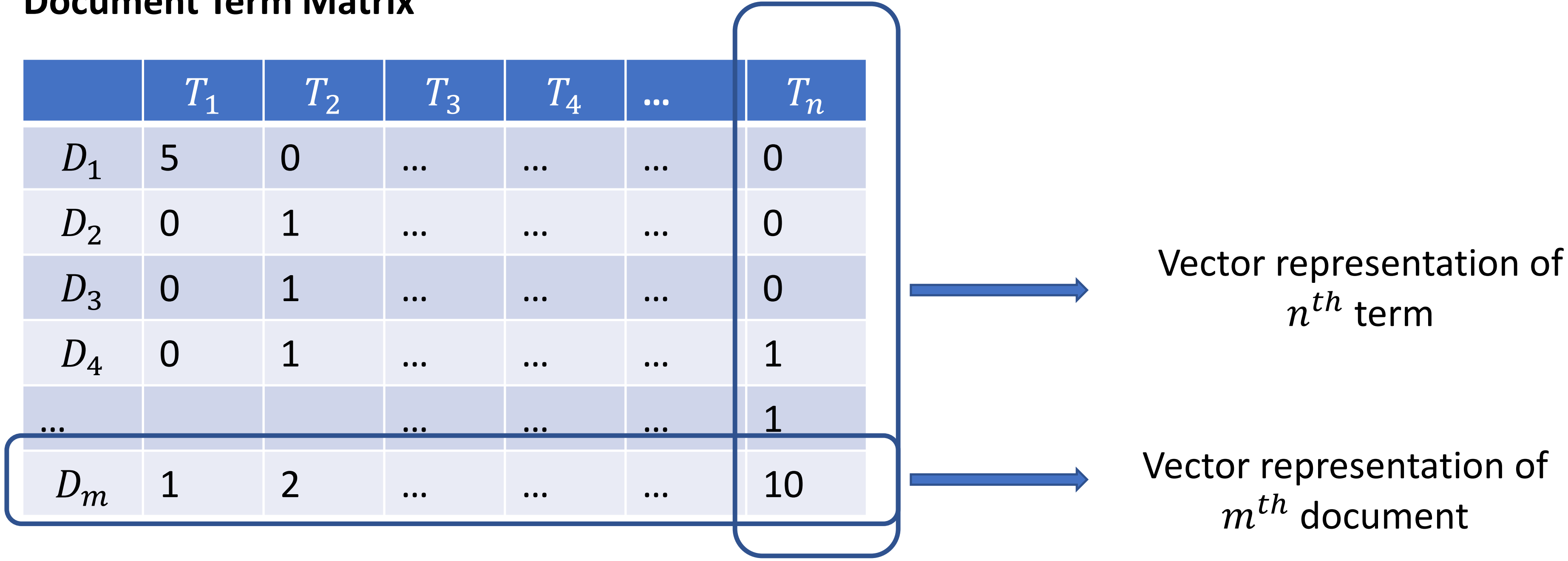
Introduction to word2vec

By

Kathirmani Sukumar

Classic Vector Representations

Document Term Matrix



	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1
…			…	…	…	1
D_m	1	2	…	…	…	10

Vector representation of n^{th} term

Vector representation of m^{th} document

Limitations

- Sparse matrix
  - Very less non-zero values. Mostly 90%-95% values are zero
- High dimension
  - For each word vector dimension is equal the number of documents
- Weak relationship between terms
  - Context between words are loosely represented

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	2	…	…	…	0
D_2	2	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	1	1	…	…	…	1
…			…	…	…	1
D_m	1	2	…	…	…	10

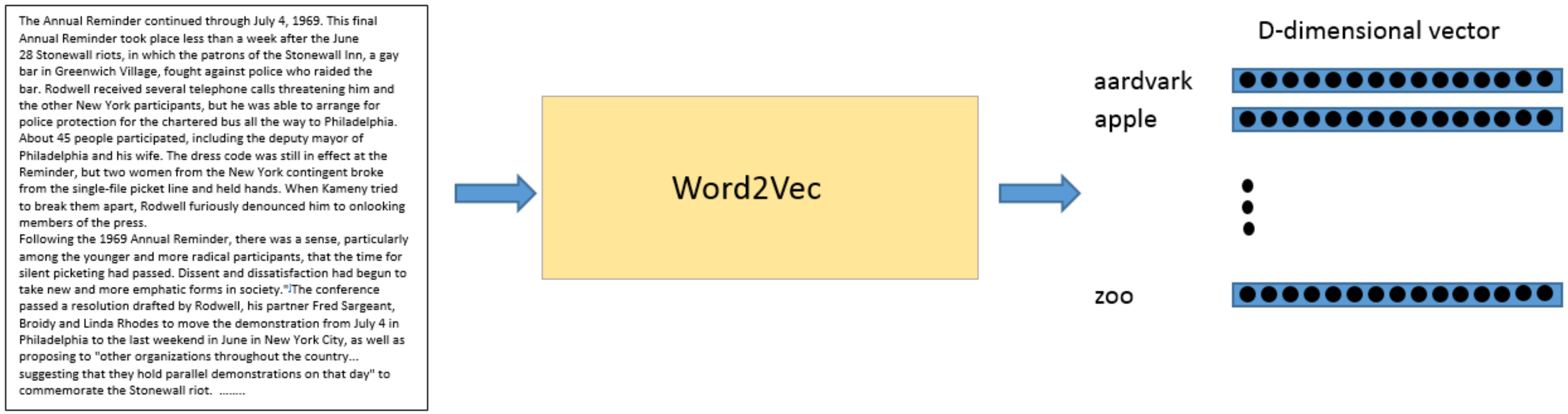
T1 & T2 has appeared together across many documents. But did they appear next to each other??

Word Embeddings

- Word embeddings is a collective name used for those techniques where words are translated in to dense low dimensional vectors instead of sparse high dimensional vectors
- These techniques are usually driven by neural network based model compared to traditional frequency based models (like LSA)
- Word embeddings techniques
  - Word2vec (developed by Google)
  - Global Vectors for Word Representation - GloVe (developed by Standford)
  - FastText

word2vec

- Word2vec is a computationally-efficient predictive model for learning word embeddings from raw text.



The Annual Reminder continued through July 4, 1969. This final Annual Reminder took place less than a week after the June 28 Stonewall riots, in which the patrons of the Stonewall Inn, a gay bar in Greenwich Village, fought against police who raided the bar. Rodwell received several telephone calls threatening him and the other New York participants, but he was able to arrange for police protection for the chartered bus all the way to Philadelphia. About 45 people participated, including the deputy mayor of Philadelphia and his wife. The dress code was still in effect at the Reminder, but two women from the New York contingent broke from the single-file picket line and held hands. When Kameny tried to break them apart, Rodwell furiously denounced him to onlooking members of the press.

Following the 1969 Annual Reminder, there was a sense, particularly among the younger and more radical participants, that the time for silent picketing had passed. Dissent and dissatisfaction had begun to take new and more emphatic forms in society."The conference passed a resolution drafted by Rodwell, his partner Fred Sargeant, Broidy and Linda Rhodes to move the demonstration from July 4 in Philadelphia to the last weekend in June in New York City, as well as proposing to "other organizations throughout the country... suggesting that they hold parallel demonstrations on that day" to commemorate the Stonewall riot. ....

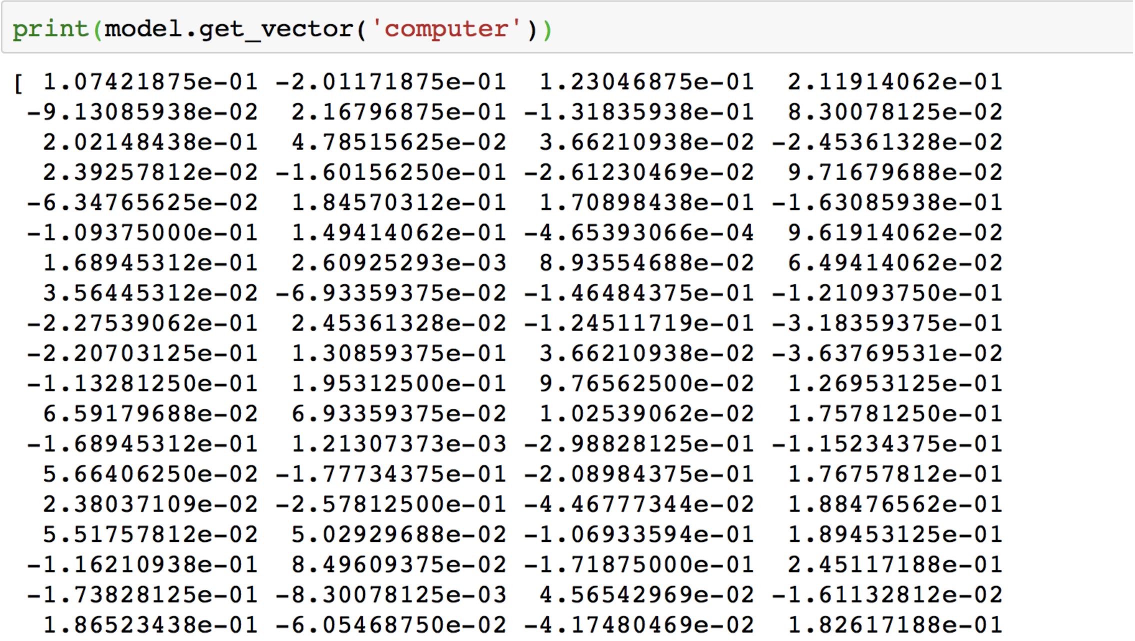
Input: Raw Text

Output: D – Dimensional Dense Vector for each word

Source: orielly.com

Google Pre-Trained Model

- Difficult to train our own model for large text. Need high end machines
- Input: Google has used Google News data set (about 100 billion words) to train a word2vec model.
- Output: Each word is represented using 300 dimension



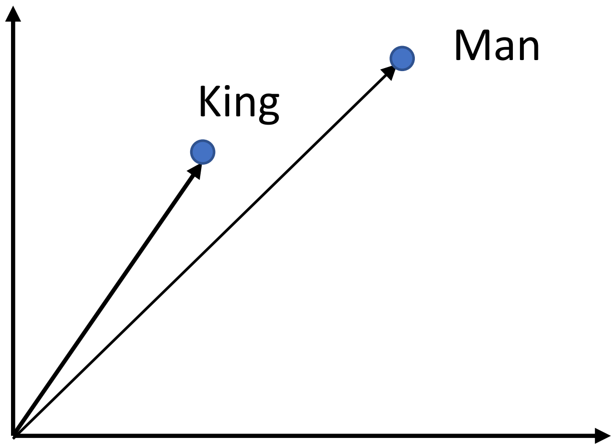
```
print(model.get_vector('computer'))
```

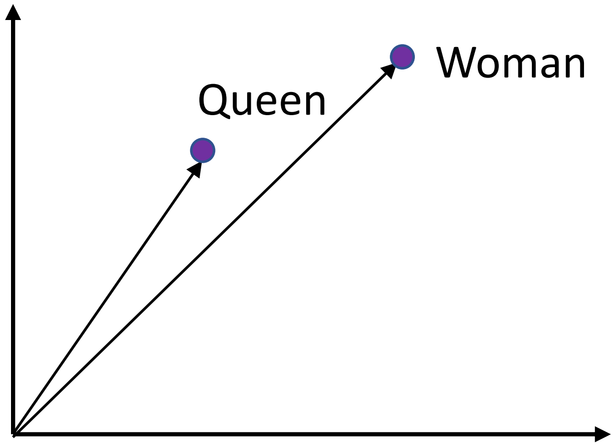
```
[ 1.07421875e-01 -2.01171875e-01  1.23046875e-01  2.11914062e-01
 -9.13085938e-02  2.16796875e-01 -1.31835938e-01  8.30078125e-02
  2.02148438e-01  4.78515625e-02  3.66210938e-02 -2.45361328e-02
  2.39257812e-02 -1.60156250e-01 -2.61230469e-02  9.71679688e-02
 -6.34765625e-02  1.84570312e-01  1.70898438e-01 -1.63085938e-01
 -1.09375000e-01  1.49414062e-01 -4.65393066e-04  9.61914062e-02
  1.68945312e-01  2.60925293e-03  8.93554688e-02  6.49414062e-02
  3.56445312e-02 -6.93359375e-02 -1.46484375e-01 -1.21093750e-01
 -2.27539062e-01  2.45361328e-02 -1.24511719e-01 -3.18359375e-01
 -2.20703125e-01  1.30859375e-01  3.66210938e-02 -3.63769531e-02
 -1.13281250e-01  1.95312500e-01  9.76562500e-02  1.26953125e-01
  6.59179688e-02  6.93359375e-02  1.02539062e-02  1.75781250e-01
 -1.68945312e-01  1.21307373e-03 -2.98828125e-01 -1.15234375e-01
  5.66406250e-02 -1.77734375e-01 -2.08984375e-01  1.76757812e-01
  2.38037109e-02 -2.57812500e-01 -4.46777344e-02  1.88476562e-01
  5.51757812e-02  5.02929688e-02 -1.06933594e-01  1.89453125e-01
 -1.16210938e-01  8.49609375e-02 -1.71875000e-01  2.45117188e-01
 -1.73828125e-01 -8.30078125e-03  4.56542969e-02 -1.61132812e-02
  1.86523438e-01 -6.05468750e-02 -4.17480469e-02  1.82617188e-01
```

Link to download the model: https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

Characteristic

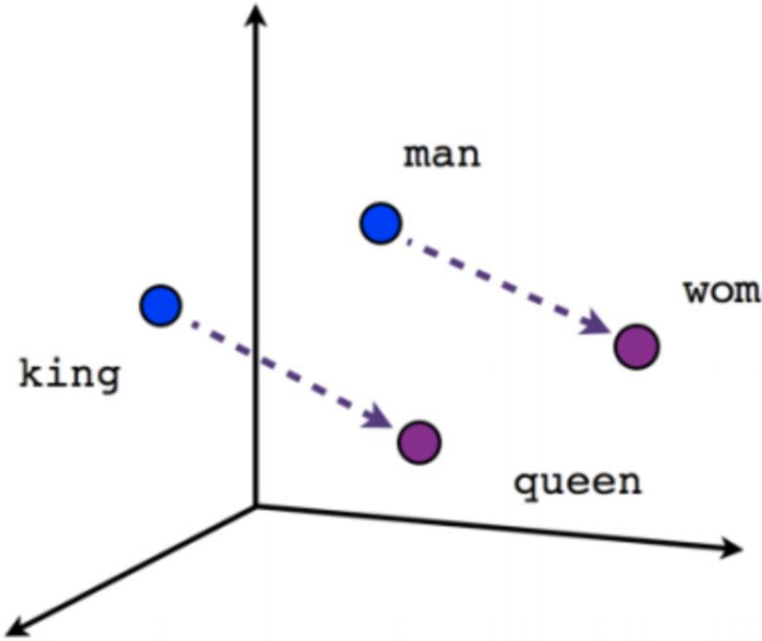
- Similar words have similar vector representation





Characteristic

- Similar words have similar vector representation



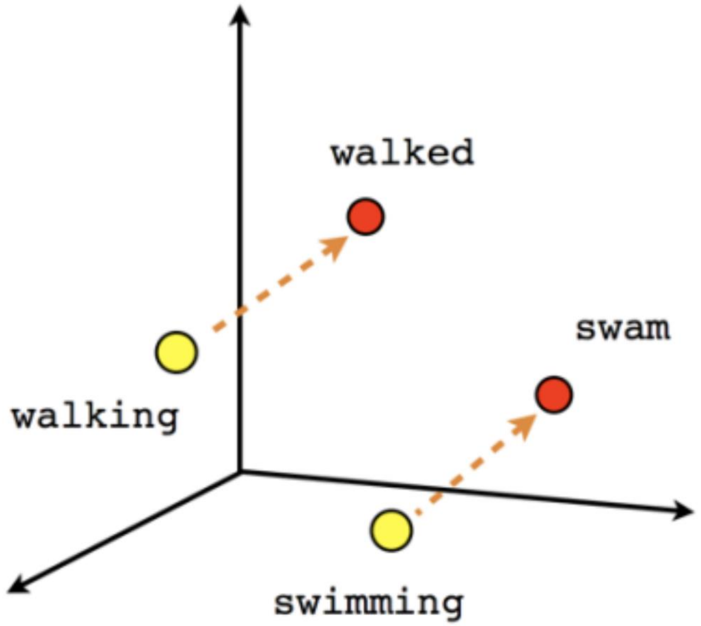
King – Man + Woman = Queen

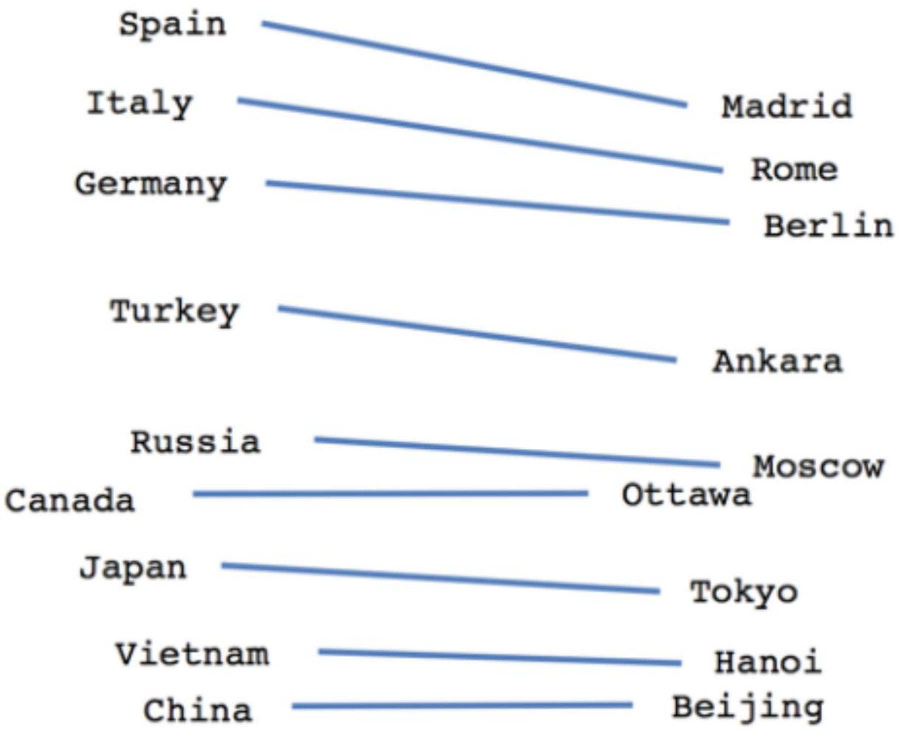
King – Man = Queen - Woman

Source: https://www.tensorflow.org/tutorials/representation/word2vec

Characteristic

- Similar words have similar vector representation





Source: https://www.tensorflow.org/tutorials/representation/word2vec

Applications

- Word similarity
  - Thesaurus: Given a word, we can identify its related words
- Stemming
- Parts of Speech
- Named Entity Recognition
- Sentiment Analysis
- Recommendation Engines

manipal PROlearn

THANK YOU

All product details and company names used or referred in this work are copyright and trademarks or registered trademarks of their respective holders. Use of them in this work does not imply any affiliation with or endorsement by them.