Hacking Human Language

Hendrik Heuer

Word Vector Representations

“You shall know a word
by the company it keeps”

–J. R. Firth 1957

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. In Studies in linguistic analysis, 1–32. Oxford: Blackwell.

“You shall know a word
by the company it keeps”

One of the most successful ideas of modern statistical NLP

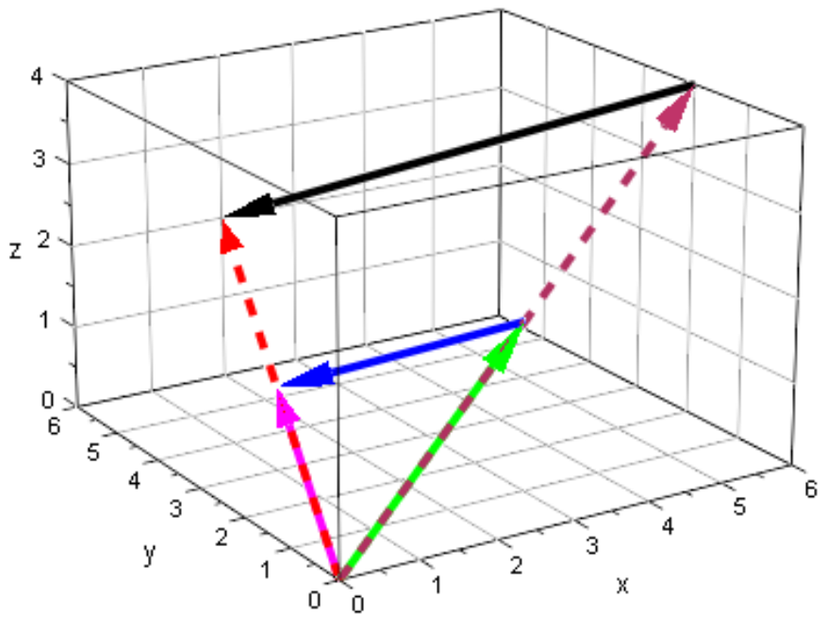
government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

Quoted after Socher

word2vec

Representing a word with a vector



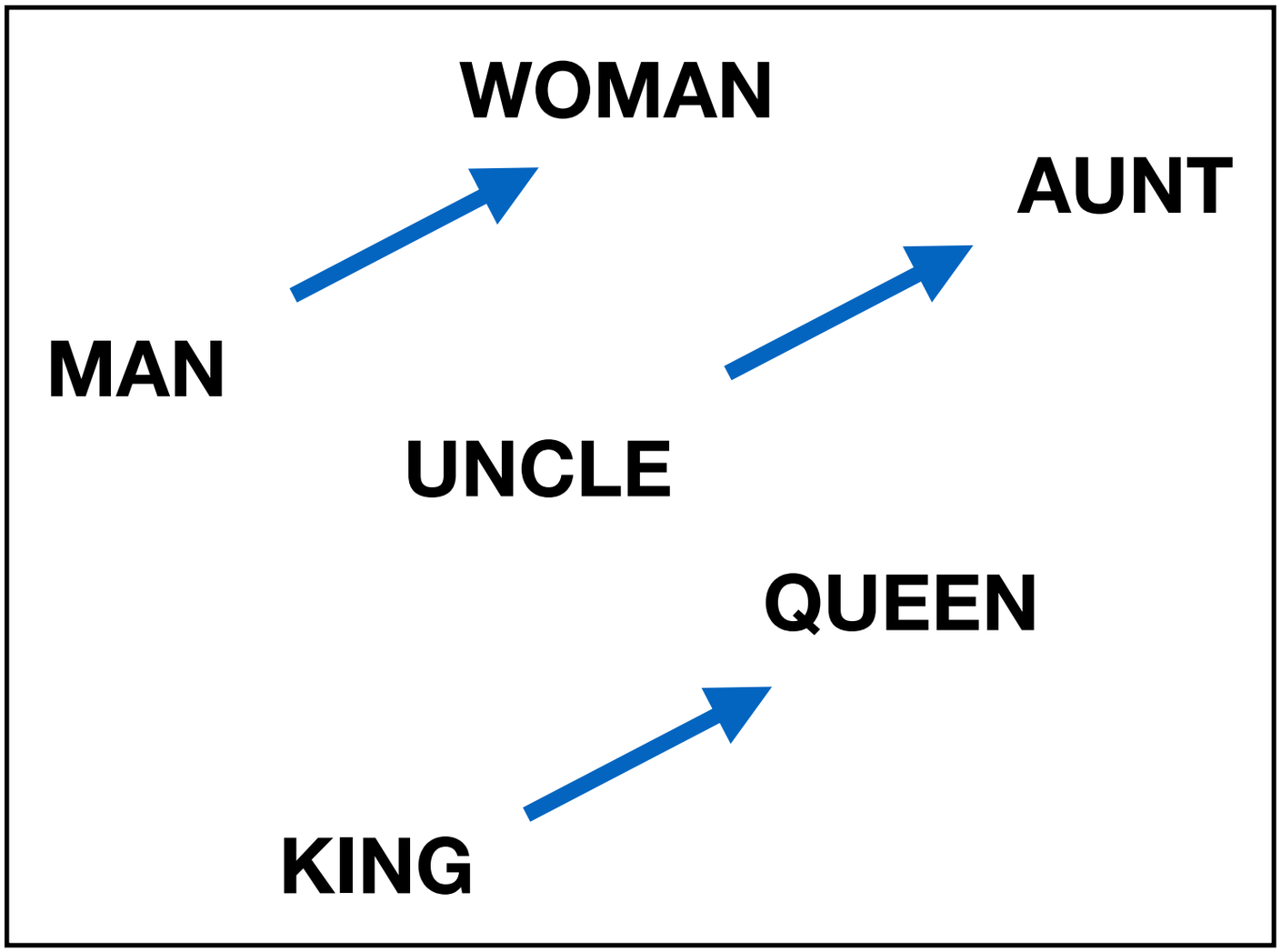
$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Vectors are directions in space

Quoted after Socher

word2vec

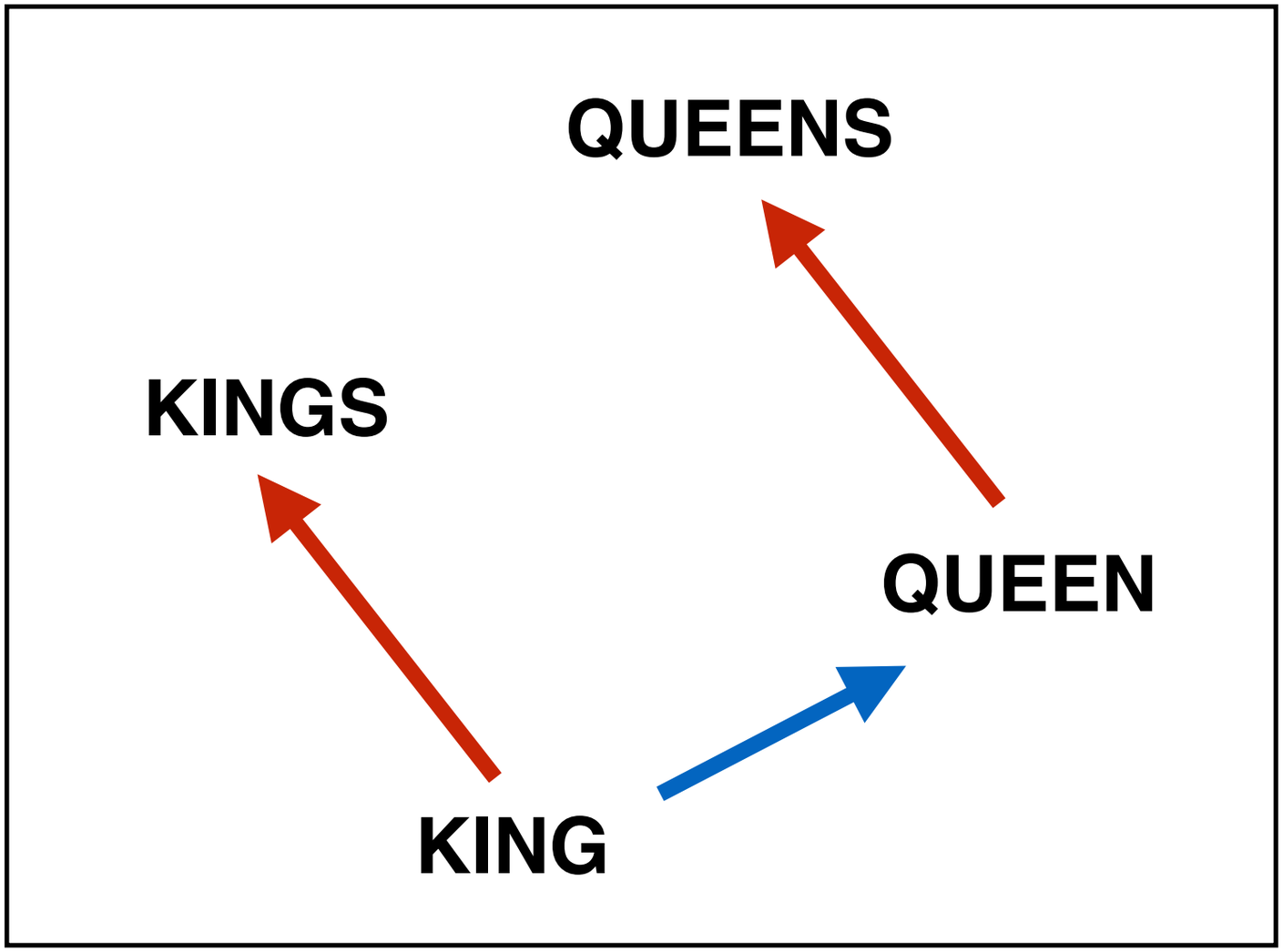
Vectors can encode relationships



T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘Efficient Estimation of Word Representations in Vector Space’, CoRR, vol. abs/1301.3781, 2013 [Online]. Available: http://arxiv.org/abs/1301.3781

word2vec

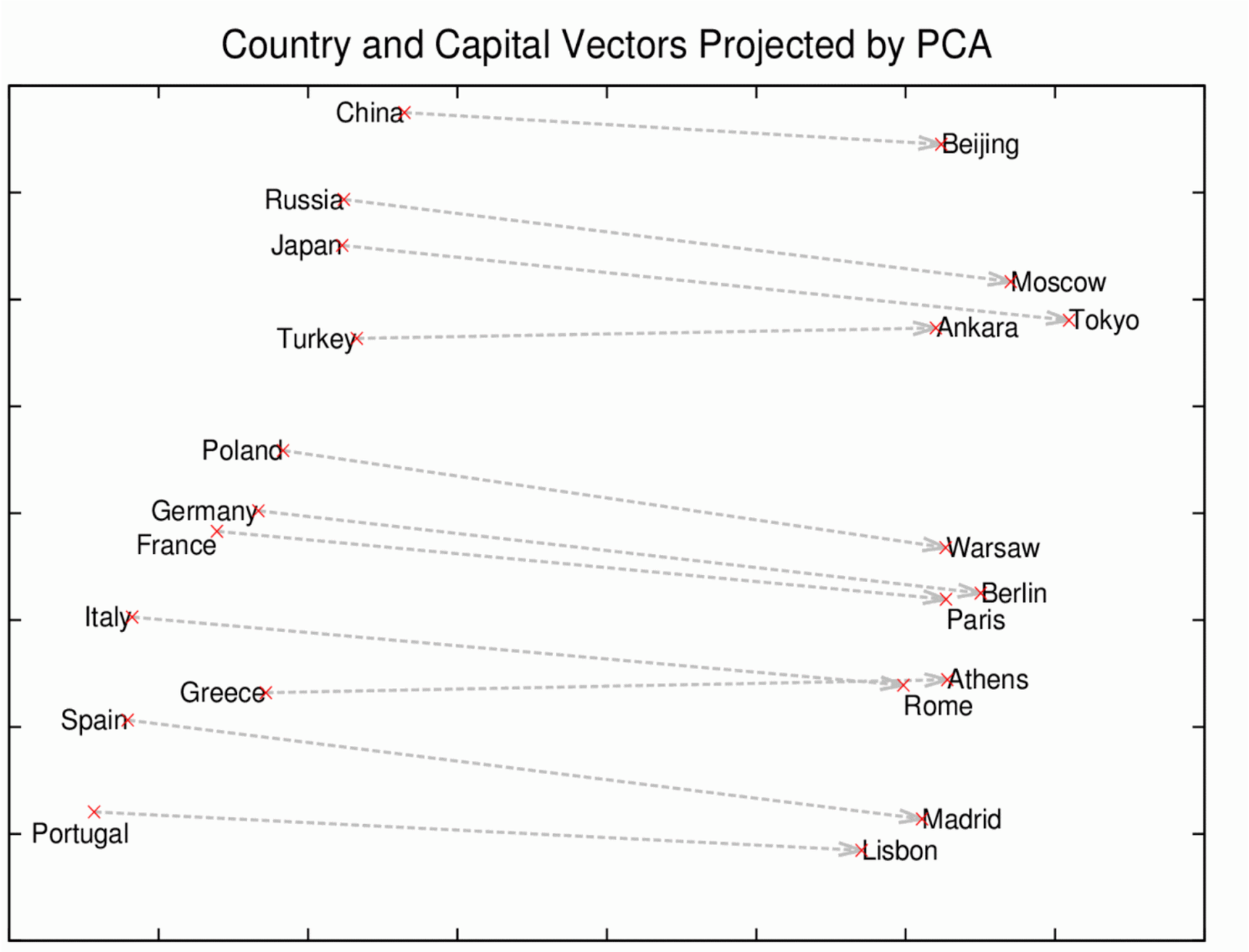
Vectors can encode relationships



T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘Efficient Estimation of Word Representations in Vector Space’, CoRR, vol. abs/1301.3781, 2013 [Online]. Available: http://arxiv.org/abs/1301.3781

word2vec

Vectors can encode relationships



T. Mikolov, K. Chen, G. Corrado, and J. Dean, 'Efficient Estimation of Word Representations in Vector Space', CoRR, vol. abs/1301.3781, 2013 [Online]. Available: http://arxiv.org/abs/1301.3781

word2vec

Analogy puzzles

England is to **London** as
Germany is to ?

598.7ms [["**Berlin**",0.563393235206604],["Dusseldorf",0.5625754594802856],
["Munich",0.5460122227668762],["Budapest",0.528582990169525 1],
["Düsseldorf",0.526650190353393 6]]

England is to **Cameron** as
Germany is to ?

556.8ms [["**Merkel**",0.5016422867774963],["Schroeder",
0.4994197785854339 6],["Klaus",0.4981233477592468],["Schröder",
0.494729697704315 2],["Peer_Nils",0.4926423430442 81]]

Link: http://radimrehurek.com/2014/02/word2vec-tutorial/

word2vec

Analogy puzzles

fast is to **fastest** as
slow is to ?

806.2ms [["**slowest**",0.7025301456451416],["slower",0.6236234307289124],
["slowed",0.5842559337615967],["slowing",0.5462259650230408],["quickest",
0.529043674468941]]

wake is to **woken** as
be is to ?

929.9ms [["**been**",0.4169896841049194 3],["tobe",0.4040281474590301 5],
["are",0.3866569399833679],["being",0.3746173679828644],["notbe",
0.3683787882328033 4]]

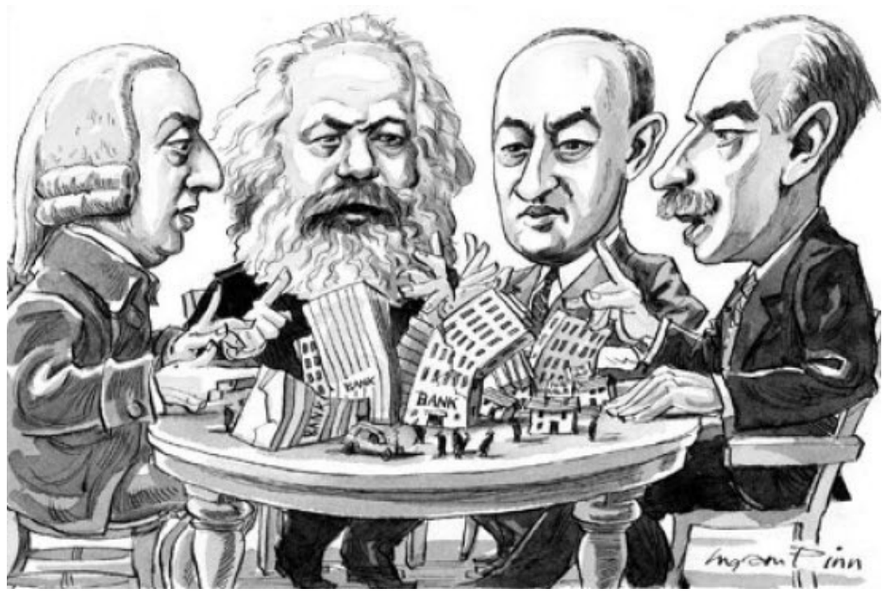
Link: http://radimrehurek.com/2014/02/word2vec-tutorial/

word2vec

Analogy puzzles

communism is to **Karl_Marx** as
capitalism is to ?

544.7ms [["Capitalism",0.5884973406791687],["capitalist",
0.5700926184654236],["**Friedrich_Hayek**",0.5352163314819336],
["**Milton_Friedman**",0.534875512123107 9],["**John_Maynard_Keynes**",
0.5335651636123657]]



Link: http://radimrehurek.com/2014/02/word2vec-tutorial/

Sweden

Most similar words

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408
floorball_federation	0.529570
luxembourg	0.529477
czech_republic	0.528778
slovakia	0.526340
romania	0.524281
kista	0.522488
helsinki_vantaa	0.519936
swedish	0.519901
balrog_ik	0.514556
portugal	0.502495
russia	0.500196
slovakia_slovenia	0.496051
ukraine	0.495712

Sweden

Most similar words

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408
floorball_federation	0.529570
luxembourg	0.529477
czech_republic	0.528778
slovakia	0.526340
romania	0.524281
kista	0.522488
helsinki_vantaa	0.519936
swedish	0.519901
balrog_ik	0.514556
portugal	0.502495
russia	0.500196
slovakia_slovenia	0.496051
ukraine	0.495712

Harvard

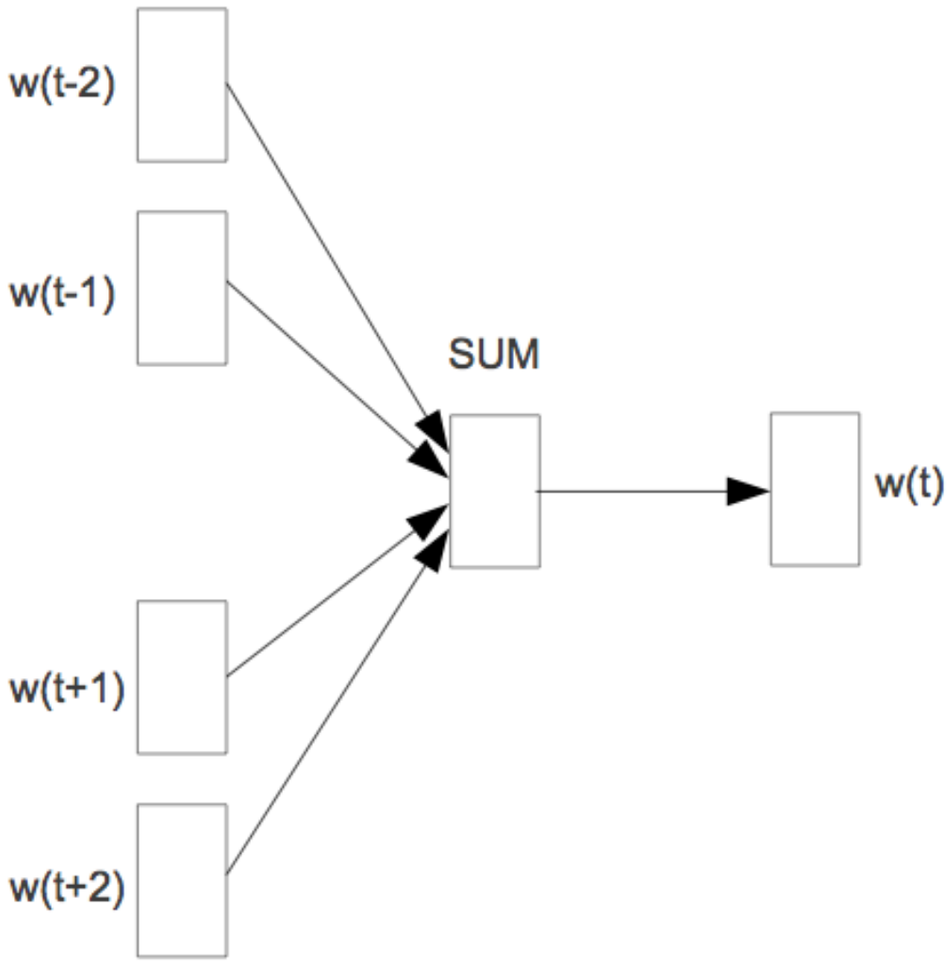
Most similar words

Word	Cosine distance
yale	0.638970
cambridge	0.612665
university	0.597709
faculty	0.588422
harvey_mudd	0.578338
johns_hopkins	0.575645
graduate	0.570294
undergraduate	0.565881
professor	0.563657
mcgill	0.562168
ph_d	0.558665
california_berkeley	0.555539
yale_university	0.550480
harvard_crimson	0.549848
princeton	0.544070
college	0.542838
oxford	0.531948
barnard_college	0.530800
professors	0.529959
princeton_university	0.529763
ucl	0.527395
doctorates	0.526292

word2vec

How it is trained

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



CBOW

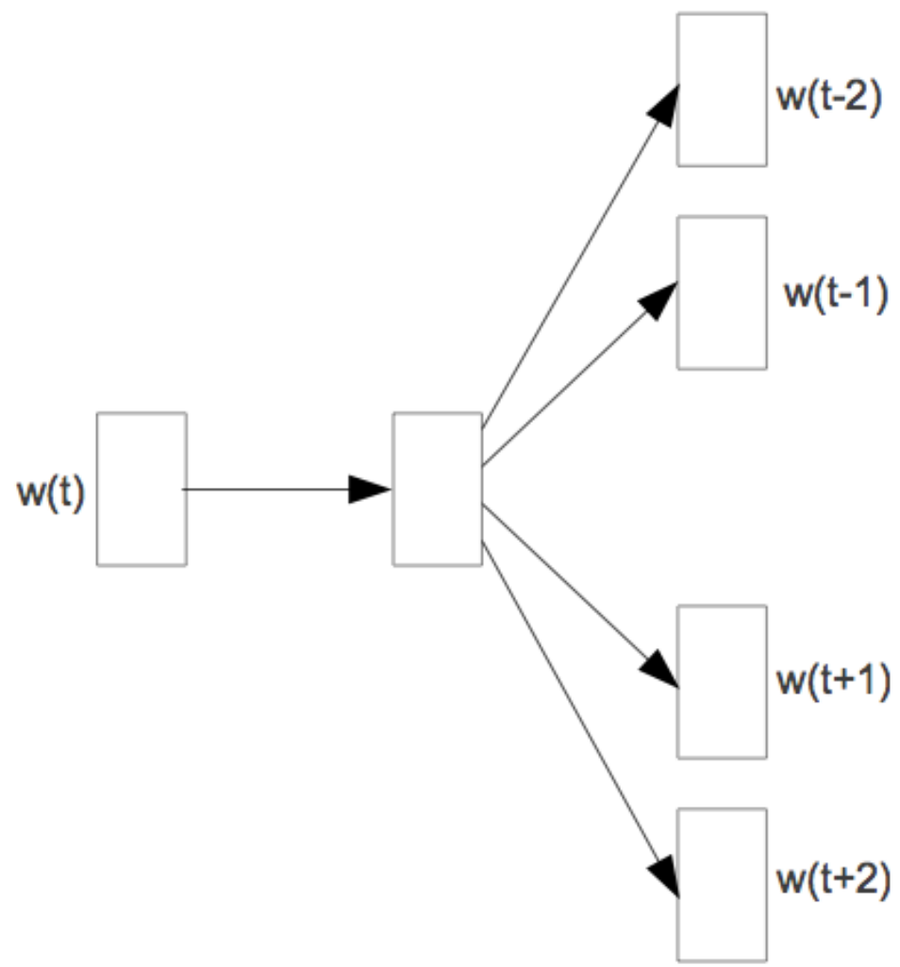
predict the current word

input

w_i-2, w_i-1, w_i+1, w_i+2

output

w_i



Skip-gram

predict the surrounding words

input

w_i

output

w_i-2, w_i-1, w_i +1, w_i +2.