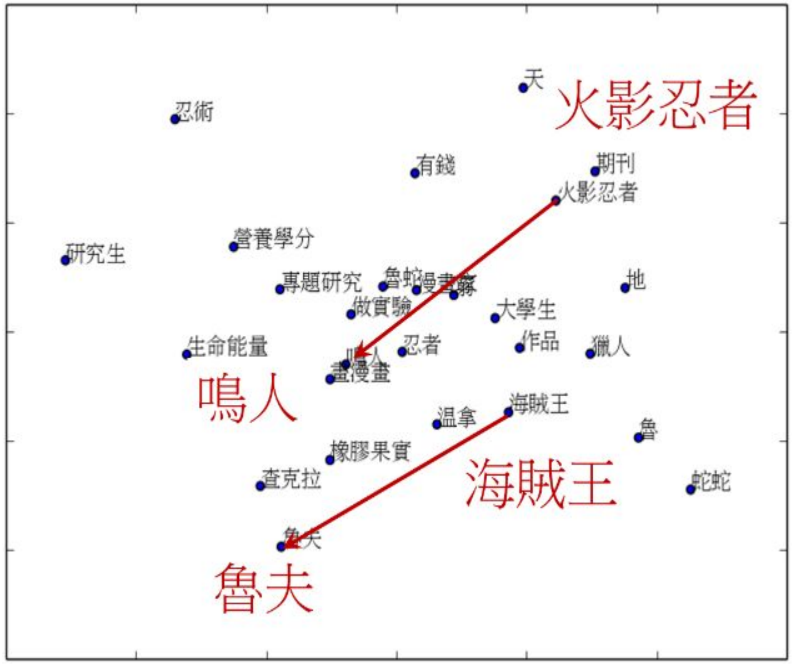
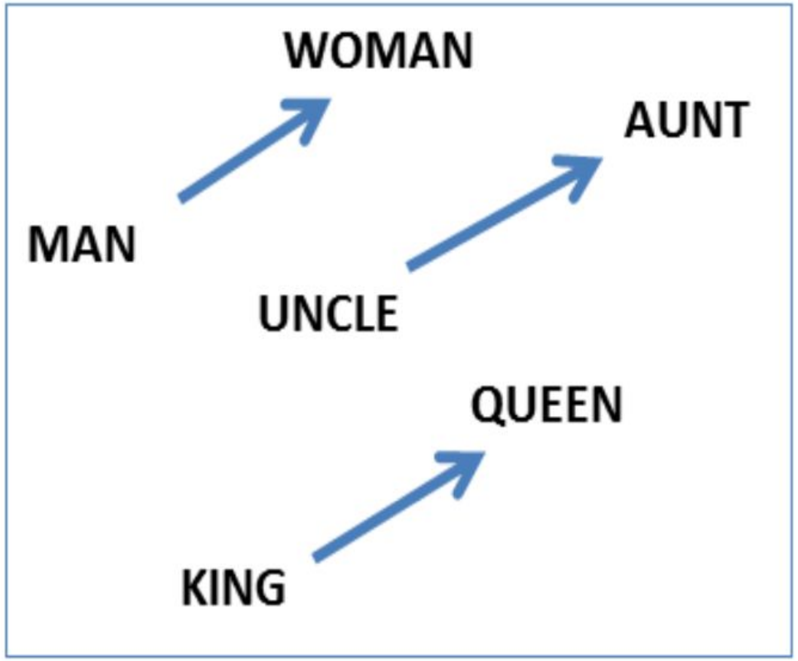
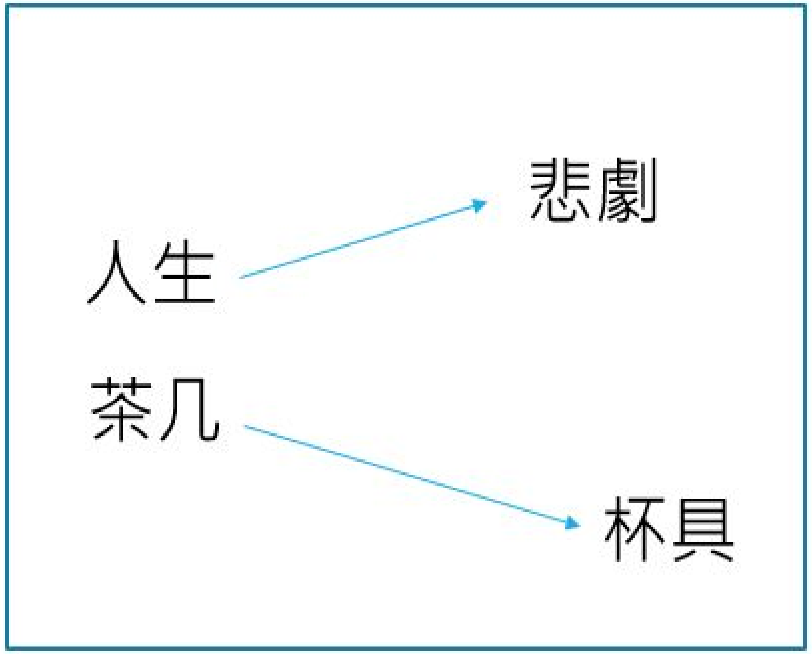
Demo 1: word2vec

dataset: https://drive.google.com/open?id=0B27ghKdkaWv-amZDQWtQRElsUDA
word2vec.py: https://drive.google.com/open?id=0B27ghKdkaWv-ZGlSMjdHSUlPdFU

What is word2vec?







人生像茶几 擺滿了杯具 (悲劇)

vec(woman) - vec(man) ≈ vec(queen) - vec(king)

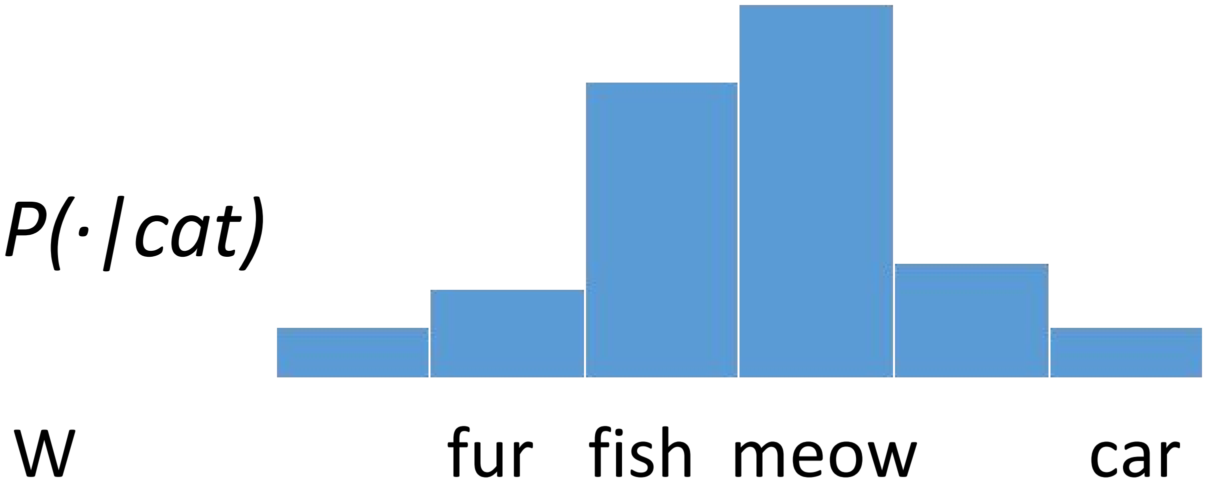
人生= [0.50.7-0.8 0.3-0.4]

Distributional Hypothesis in linguistics:
‘a word is characterized by the company it keeps’.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955

Suppose we read the word “cat”. What is the probability *P(w|cat)* that we’ll read the word w nearby?

word embedding $\Leftrightarrow$ “cat” = $P(\cdot|cat)$



How to train?

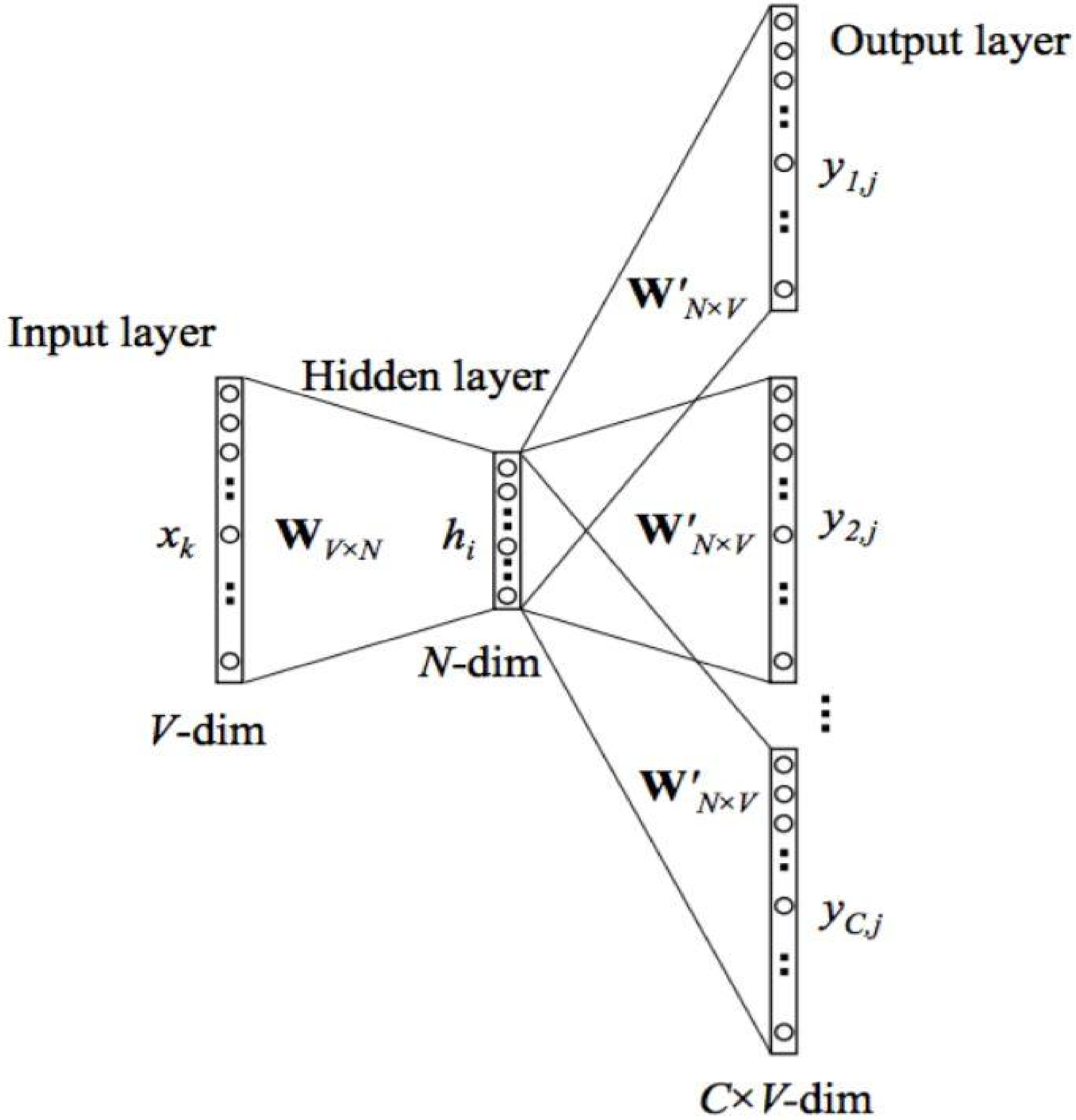
w_0 context

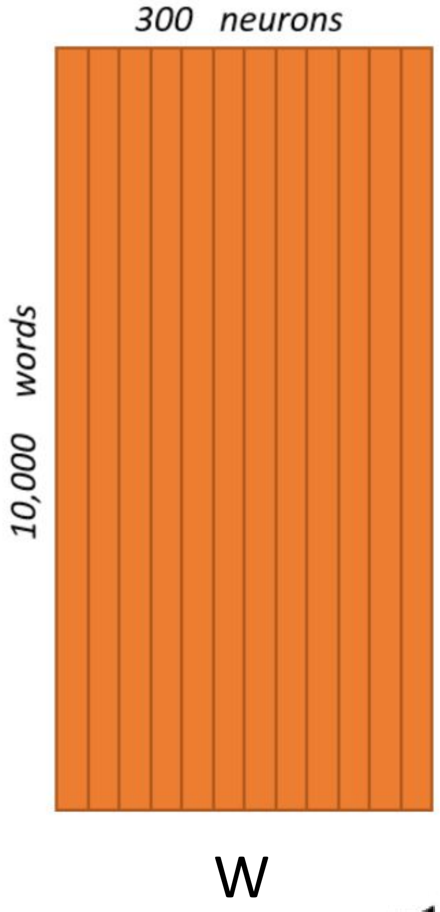
Input: the quick brown fox jumped over the lazy dog

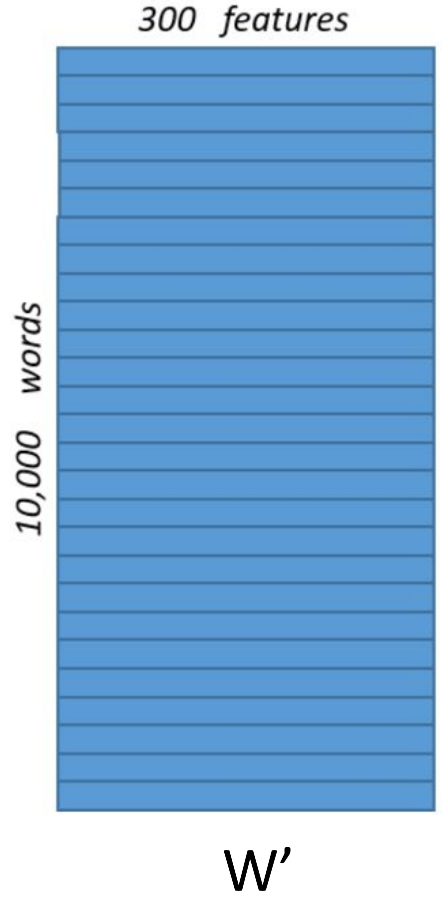
window	w_0	context
1	fox	{brown, jumped}
2	jumped	{brown, fox, over, the}

Predict company context: P(context|w)

Skip-gram model







$$[0 \quad 0 \quad 0 \quad \underset{\text{cat}}{1} \quad 0] \times \underset{\text{W}}{\begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix}} = [10 \quad 12 \quad 19]$$

Tensorflow Review

- Build dataset
  - Step 1: Download the data.
  - Step 2: Build the dictionary(word->int32).
  - Step 3: Function to generate a training batch.
- Define graph — Step 4: Build and train a skip-gram model.
  - Variables
  - Inputs/labels
  - Loss
  - Optimizer
- Run session — Step 5: Begin training.
  - Initial variables
  - Define feed_dict
  - Session.run

Exercise

- 共七題

```
with graph.as_default():

  # Input data.
  # (2)*************************************************
  train_inputs = tf."?"("?", shape=[batch_size])
  train_labels = tf."?"("?", shape=[batch_size, 1])
  valid_dataset = tf.constant(valid_examples, dtype=tf.int32)
```

Run!

Before:

Loss = 295.61

Nearest to seven: ozzy, gringo, idiomatic, hydrolysis….

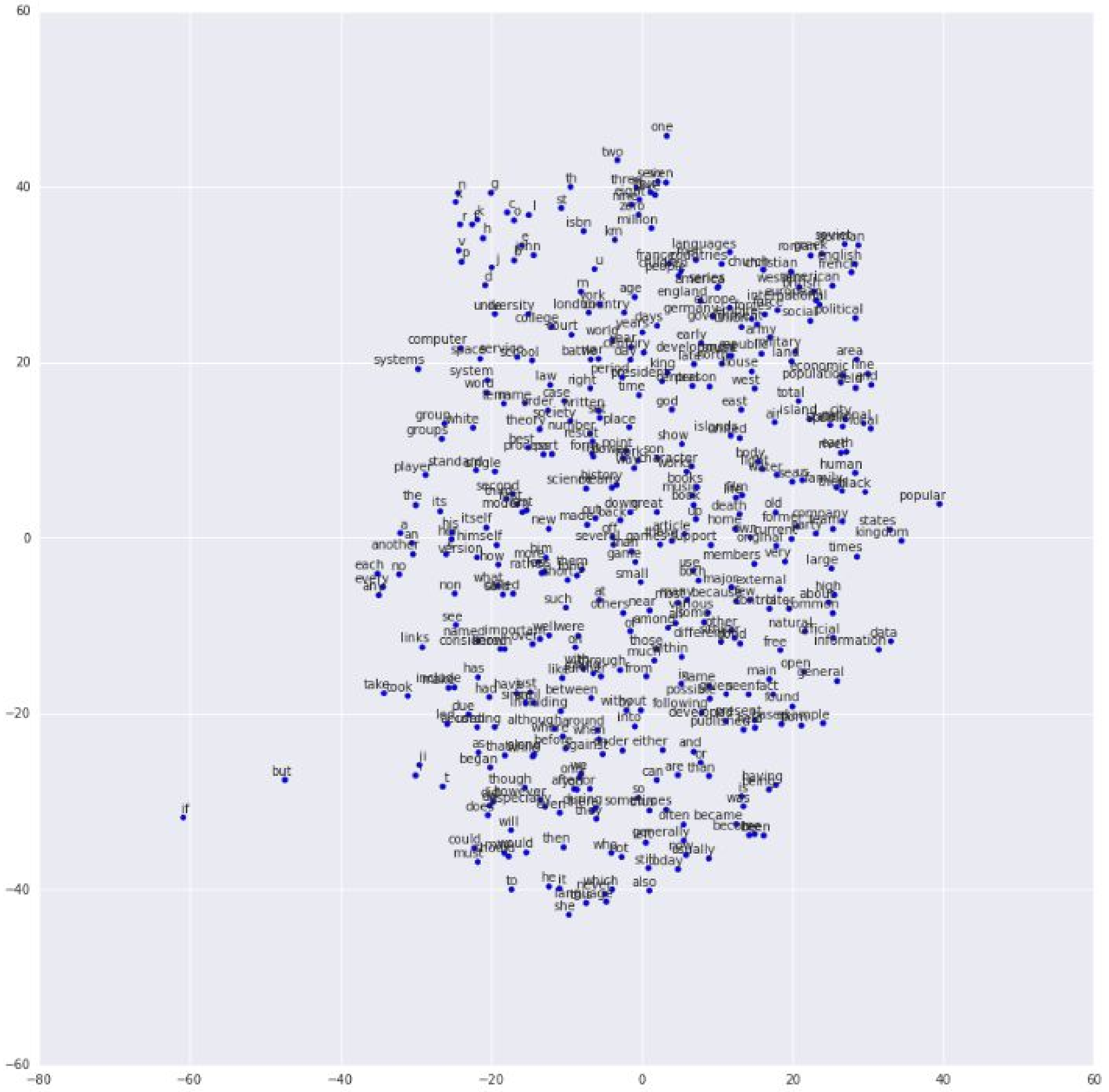
```
Initialized
Average loss at step  0 :  295.614532471
Nearest to seven: ozzy, gringo, idiomatic, hydrolysis, overtly, mainstays, pagos, thinner,
Nearest to four: antiochian, records, mentioned, cookies, hype, pia, fingerboard, eastman,
Nearest to not: segmentation, sinkholes, artistry, vernacular, sodium, clapping, fails, montju,
Nearest to may: mesh, wayside, sharks, ruskin, kronos, fatal, cleverly, bribed,
Nearest to often: billings, chiropractic, defends, yuan, greenblatt, prospectors, degenerate, portrait,
Nearest to is: ogdoad, subsumed, latveria, frits, rook, def, phyllis, rebelled,
Nearest to up: eigenvectors, hybrid, syllogism, presbyter, massoud, trevor, smuggle, unspoken,
Nearest to an: drosophila, analyze, hewson, evocative, regulator, disraeli, mjs, myspace,
Nearest to b: young, phoned, organised, dram, osnabr, marcellinus, semester, urgent,
Nearest to had: phillippe, pressburg, deadliest, bot, oracles, weird, chelsea, kinshasa,
Nearest to many: paradigm, schism, spawning, tritone, sponsor, taklamakan, skills, mentors,
Nearest to be: sensational, ile, battery, beijing, prequels, subduction, doj, supplementing,
Nearest to however: axiomatic, sdp, tumors, guanosine, buprenorphine, mise, raytheon, unreliable,
Nearest to their: blanca, ghali, wrestlers, releases, catalina, servicemen, dvd, standoff,
Nearest to a: bartolomeo, sepia, initiation, absent, learning, confiscation, rsfsr, epicycle,
Nearest to it: billionaire, intervals, billions, wolfman, lonesome, pfp, reptiles, incest,
```

After:

Loss = 4.687

Nearest to five: four, six, thre

```
Average loss at step  92000 :  4.66823846
Average loss at step  94000 :  4.71707212
Average loss at step  96000 :  4.69267583
Average loss at step  98000 :  4.57841269
Average loss at step  100000 :  4.6871345
Nearest to used: referred, known, flames,
Nearest to during: after, in, under, at,
Nearest to as: busan, kapoor, cebus, when
Nearest to by: was, as, be, michelob, rel
Nearest to to: would, thibetanus, will, c
Nearest to from: into, in, through, betwe
Nearest to more: less, most, very, too, c
Nearest to be: been, have, were, by, bein
Nearest to and: or, but, circ, dasyprocta
Nearest to five: four, six, three, seven,
Nearest to other: various, foods, many, s
Nearest to that: which, however, but, thi
Nearest to so: iit, chymotrypsin, now, ag
Nearest to if: when, kapoor, though, befo
Nearest to six: seven, eight, five, four,
Nearest to an: ecological, ursus, simplic
```



Ref

- word2vec_answer.py: https://drive.google.com/open?id=0B27ghKdkaWv-c2lfSUZlM1dwS1k
- Deep Learning, Ali Ghodsi, University of Waterloo https://uwaterloo.ca/data-science/sites/ca.data-science/files/uploads/files/word2vec.pdf
- Tensorflow documentation, Vector Representations of Words https://www.tensorflow.org/tutorials/word2vec
- Demo code modified from: https://www.tensorflow.org/code/tensorflow/examples/tutorials/word2vec/word2vec_basic.py
- Word2Vec Tutorial - The Skip-Gram Model http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
- Word2Vec Tutorial Part I: The Skip- Gram Model