Deep Learning

Assignment 5

The goal of this assignment is to train a Word2Vec skip-gram model over Text8 (http://mattmahoney.net/dc/textdata) data.

In [0]:

```
# These are all the modules we'll be using later. Make sure you can import them
# before proceeding further.
%matplotlib inline
from __future__ import print_function
import collections
import math
import numpy as np
import os
import random
import tensorflow as tf
import zipfile
from matplotlib import pylab
from six.moves import range
from six.moves.urllib.request import urlretrieve
from sklearn.manifold import TSNE
```

Download the data from the source website if necessary.

In [0]:

```
url = 'http://mattmahoney.net/dc/'

def maybe_download(filename, expected_bytes):
  """Download a file if not present, and make sure it's the right size."""
  if not os.path.exists(filename):
    filename, _ = urlretrieve(url + filename, filename)
  statinfo = os.stat(filename)
  if statinfo.st_size == expected_bytes:
    print('Found and verified %s' % filename)
  else:
    print(statinfo.st_size)
    raise Exception(
      'Failed to verify ' + filename + '. Can you get to it with a browser?')
  return filename

filename = maybe_download('text8.zip', 31344016)
```

```
Found and verified text8.zip
```

Read the data into a string.

In [0]:

```
def read_data(filename):
  """Extract the first file enclosed in a zip file as a list of words"""
  with zipfile.ZipFile(filename) as f:
    data = tf.compat.as_str(f.read(f.namelist()[0])).split()
  return data

words = read_data(filename)
print('Data size %d' % len(words))
```

```
Data size 17005207
```

Build the dictionary and replace rare words with UNK token.

In [0]:

```
vocabulary_size = 50000

def build_dataset(words):
  count = [['UNK', -1]]
  count.extend(collections.Counter(words).most_common(vocabulary_size - 1))
  dictionary = dict()
  for word, _ in count:
    dictionary[word] = len(dictionary)
  data = list()
  unk_count = 0
  for word in words:
    if word in dictionary:
      index = dictionary[word]
    else:
      index = 0  # dictionary['UNK']
      unk_count = unk_count + 1
    data.append(index)
  count[0][1] = unk_count
  reverse_dictionary = dict(zip(dictionary.values(), dictionary.keys()))
  return data, count, dictionary, reverse_dictionary

data, count, dictionary, reverse_dictionary = build_dataset(words)
print('Most common words (+UNK)', count[:5])
print('Sample data', data[:10])
del words  # Hint to reduce memory.
```

```
Most common words (+UNK) [['UNK', 418391], ('the', 1061396), ('of',
 593677), ('and', 416629), ('one', 411764)]
Sample data [5243, 3083, 12, 6, 195, 2, 3136, 46, 59, 156]
```

Function to generate a training batch for the skip-gram model.

In [0]:

```
data_index = 0

def generate_batch(batch_size, num_skips, skip_window):
  global data_index
  assert batch_size % num_skips == 0
  assert num_skips <= 2 * skip_window
  batch = np.ndarray(shape=(batch_size), dtype=np.int32)
  labels = np.ndarray(shape=(batch_size, 1), dtype=np.int32)
  span = 2 * skip_window + 1 # [ skip_window target skip_window ]
  buffer = collections.deque(maxlen=span)
  for _ in range(span):
    buffer.append(data[data_index])
    data_index = (data_index + 1) % len(data)
  for i in range(batch_size // num_skips):
    target = skip_window  # target label at the center of the buffer
    targets_to_avoid = [ skip_window ]
    for j in range(num_skips):
      while target in targets_to_avoid:
        target = random.randint(0, span - 1)
      targets_to_avoid.append(target)
      batch[i * num_skips + j] = buffer[skip_window]
      labels[i * num_skips + j, 0] = buffer[target]
    buffer.append(data[data_index])
    data_index = (data_index + 1) % len(data)
  return batch, labels

print('data:', [reverse_dictionary[di] for di in data[:8]])

for num_skips, skip_window in [(2, 1), (4, 2)]:
    data_index = 0
    batch, labels = generate_batch(batch_size=8, num_skips=num_skips, skip_windo
w=skip_window)
    print('\nwith num_skips = %d and skip_window = %d:' % (num_skips, skip_windo
w))
    print('    batch:', [reverse_dictionary[bi] for bi in batch])
    print('    labels:', [reverse_dictionary[li] for li in labels.reshape(8)])
```

```
data: ['anarchism', 'originated', 'as', 'a', 'term', 'of', 'abuse',
 'first']

with num_skips = 2 and skip_window = 1:
    batch: ['originated', 'originated', 'as', 'as', 'a', 'a', 'ter
m', 'term']
    labels: ['as', 'anarchism', 'a', 'originated', 'term', 'as',
 'a', 'of']

with num_skips = 4 and skip_window = 2:
    batch: ['as', 'as', 'as', 'as', 'a', 'a', 'a', 'a']
    labels: ['anarchism', 'originated', 'term', 'a', 'as', 'of', 'or
iginated', 'term']
```

Train a skip-gram model.

In [0]:

```
batch_size = 128
embedding_size = 128 # Dimension of the embedding vector.
skip_window = 1 # How many words to consider left and right.
num_skips = 2 # How many times to reuse an input to generate a label.
# We pick a random validation set to sample nearest neighbors. here we limit the
# validation samples to the words that have a low numeric ID, which by
# construction are also the most frequent.
valid_size = 16 # Random set of words to evaluate similarity on.
valid_window = 100 # Only pick dev samples in the head of the distribution.
valid_examples = np.array(random.sample(range(valid_window), valid_size))
num_sampled = 64 # Number of negative examples to sample.

graph = tf.Graph()

with graph.as_default(), tf.device('/cpu:0'):

  # Input data.
  train_dataset = tf.placeholder(tf.int32, shape=[batch_size])
  train_labels = tf.placeholder(tf.int32, shape=[batch_size, 1])
  valid_dataset = tf.constant(valid_examples, dtype=tf.int32)

  # Variables.
  embeddings = tf.Variable(
    tf.random_uniform([vocabulary_size, embedding_size], -1.0, 1.0))
  softmax_weights = tf.Variable(
    tf.truncated_normal([vocabulary_size, embedding_size],
                         stddev=1.0 / math.sqrt(embedding_size)))
  softmax_biases = tf.Variable(tf.zeros([vocabulary_size]))

  # Model.
  # Look up embeddings for inputs.
  embed = tf.nn.embedding_lookup(embeddings, train_dataset)
  # Compute the softmax loss, using a sample of the negative labels each time.
  loss = tf.reduce_mean(
    tf.nn.sampled_softmax_loss(weights=softmax_weights, biases=softmax_biases, i
nputs=embed,
                               labels=train_labels, num_sampled=num_sampled, num
_classes=vocabulary_size))

  # Optimizer.
  # Note: The optimizer will optimize the softmax_weights AND the embeddings.
  # This is because the embeddings are defined as a variable quantity and the
  # optimizer's `minimize` method will by default modify all variable quantities

  # that contribute to the tensor it is passed.
  # See docs on `tf.train.Optimizer.minimize()` for more details.
  optimizer = tf.train.AdagradOptimizer(1.0).minimize(loss)

  # Compute the similarity between minibatch examples and all embeddings.
  # We use the cosine distance:
  norm = tf.sqrt(tf.reduce_sum(tf.square(embeddings), 1, keep_dims=True))
  normalized_embeddings = embeddings / norm
  valid_embeddings = tf.nn.embedding_lookup(
    normalized_embeddings, valid_dataset)
  similarity = tf.matmul(valid_embeddings, tf.transpose(normalized_embeddings))
```

In [0]:

```
num_steps = 100001

with tf.Session(graph=graph) as session:
  tf.global_variables_initializer().run()
  print('Initialized')
  average_loss = 0
  for step in range(num_steps):
    batch_data, batch_labels = generate_batch(
      batch_size, num_skips, skip_window)
    feed_dict = {train_dataset : batch_data, train_labels : batch_labels}
    _, l = session.run([optimizer, loss], feed_dict=feed_dict)
    average_loss += l
    if step % 2000 == 0:
      if step > 0:
        average_loss = average_loss / 2000
      # The average loss is an estimate of the loss over the last 2000 batches.
      print('Average loss at step %d: %f' % (step, average_loss))
      average_loss = 0
    # note that this is expensive (~20% slowdown if computed every 500 steps)
    if step % 10000 == 0:
      sim = similarity.eval()
      for i in range(valid_size):
        valid_word = reverse_dictionary[valid_examples[i]]
        top_k = 8 # number of nearest neighbors
        nearest = (-sim[i, :]).argsort()[1:top_k+1]
        log = 'Nearest to %s:' % valid_word
        for k in range(top_k):
          close_word = reverse_dictionary[nearest[k]]
          log = '%s %s,' % (log, close_word)
        print(log)
  final_embeddings = normalized_embeddings.eval()
```

```
Initialized
Average loss at step 0 : 8.58149623871
Nearest to been: unfavourably, marmara, ancestral, legal, bogart, gl
ossaries, worst, rooms,
Nearest to time: conformist, strawberries, sindhi, waterfall, xia, n
ominates, psp, sensitivity,
Nearest to over: overlord, panda, golden, semigroup, rawlings, invol
ved, shreveport, handling,
Nearest to not: hymenoptera, reintroducing, lamiaceae, because, dava
o, omnipotent, combustion, debilitating,
Nearest to three: catalog, koza, gn, braque, holstein, postgresql, l
uddite, justine,
Nearest to if: chilled, vince, fiddler, represented, sandinistas, ha
ppiness, lya, glands,
Nearest to there: coast, photosynthetic, kimmei, legally, inner, ill
yricum, formats, fullmetal,
Nearest to between: chuvash, prinz, suitability, wolfe, guideline, c
omputability, diminutive, paulo,
Nearest to from: tanganyika, workshop, elphinstone, spearhead, resur
rected, kevlar, shangri, loves,
Nearest to state: sextus, wuppertal, glaring, inches, unrounded, cou
rageous, adler, connie,
Nearest to on: gino, phocas, rhine, jg, macrocosm, jackass, jays, th
eorie,
Nearest to and: standings, towed, reyes, willard, equality, jugglin
g, wladislaus, faked,
Nearest to eight: gresham, dogg, moko, tennis, superseded, telegraph
y, scramble, vinod,
Nearest to they: prisons, divisor, coder, ribeira, willingness, fact
ional, nne, lotta,
Nearest to more: blues, fur, sterling, tangier, khwarizmi, discourag
ed, cal, deicide,
Nearest to other: enemies, bogged, brassicaceae, lascaux, dispense,
 alexandrians, crimea, dou,
Average loss at step 2000 : 4.39983723116
Average loss at step 4000 : 3.86921076906
Average loss at step 6000 : 3.72542127335
Average loss at step 8000 : 3.57835536212
Average loss at step 10000 : 3.61056993055
Nearest to been: glossaries, legal, unfavourably, be, hadad, wore, s
carcity, were,
Nearest to time: strawberries, conformist, gleichschaltung, waterfal
l, molality, nominates, baal, dole,
Nearest to over: golden, semigroup, catus, motorways, brick, shehri,
mussolini, overlord,
Nearest to not: hinayana, it, often, they, boots, also, noaa, lindse
y,
Nearest to three: four, seven, six, five, nine, eight, two, zero,
Nearest to if: glands, euros, wallpaper, redefine, toho, confuse, un
sound, shepherd,
Nearest to there: it, they, fullmetal, pace, legally, harpsichord, m
ma, bug,
Nearest to between: chuvash, wandering, from, kirsch, pursuant, euro
cents, suitability, jackie,
Nearest to from: into, in, workshop, to, at, misogynist, elphinston
e, spearhead,
Nearest to state: sextus, glaring, connie, adler, esoteric, didacti
c, handedness, presidents,
Nearest to on: in, at, for, ruminants, wakefulness, torrey, foley, g
ino,
Nearest to and: or, who, but, zelda, of, for, thirst, chisel,
```

```
Nearest to eight: nine, six, seven, five, four, three, zero, two,
Nearest to they: he, prisons, there, we, hydrate, it, not, cumbersom
e,
Nearest to more: skye, blues, trypomastigotes, deicide, most, readab
le, used, sterling,
Nearest to other: trochaic, hush, surveyors, joachim, differentiatio
n, attackers, reverence, attestation,
Average loss at step 12000 : 3.66169466591
Average loss at step 14000 : 3.60342905837
Average loss at step 16000 : 3.57761328053
Average loss at step 18000 : 3.57667332476
Average loss at step 20000 : 3.53310145146
Nearest to been: be, become, was, hadad, unfavourably, were, wore, p
artido,
Nearest to time: gleichschaltung, strawberries, year, nominates, con
formist, etch, admittedly, treasuries,
Nearest to over: golden, semigroup, motorways, rawlings, triangle, t
rey, ustawa, mattingly,
Nearest to not: they, boots, often, dieppe, still, hinayana, nearly,
be,
Nearest to three: two, four, five, seven, eight, six, nine, one,
Nearest to if: wallpaper, euros, before, toho, unsound, so, bg, pfc,
Nearest to there: they, it, he, usually, which, we, not, transaction
s,
Nearest to between: from, with, about, near, reactance, eurocents, w
andering, voltaire,
Nearest to from: into, workshop, by, between, in, on, elphinstone, u
nder,
Nearest to state: glaring, esoteric, succeeding, sextus, vorarlberg,
presidents, depends, connie,
Nearest to on: in, at, upon, during, from, janis, foley, nubian,
Nearest to and: or, thirst, but, where, s, who, pfaff, including,
Nearest to eight: nine, seven, six, five, four, three, zero, one,
Nearest to they: there, he, we, not, it, you, prisons, who,
Nearest to more: less, most, deicide, skye, trypomastigotes, interve
ntionism, toed, drummond,
Nearest to other: such, joachim, hush, attackers, surveyors, trochai
c, differentiation, reverence,
Average loss at step 22000 : 3.59519316927
Average loss at step 24000 : 3.55378576797
Average loss at step 26000 : 3.56455037558
Average loss at step 28000 : 3.5040882225
Average loss at step 30000 : 3.39208897972
Nearest to been: become, be, were, was, spotless, hadad, by, hausdor
ff,
Nearest to time: gleichschaltung, year, day, nominates, jesus, straw
berries, way, admittedly,
Nearest to over: golden, semigroup, motorways, rawlings, interventio
nism, counternarcotics, adaption, brick,
Nearest to not: often, they, it, never, still, nor, boots, pki,
Nearest to three: four, six, two, eight, five, seven, nine, zero,
Nearest to if: when, before, so, should, toho, where, bg, wallpaper,
Nearest to there: they, it, which, usually, he, that, also, now,
Nearest to between: with, from, in, panasonic, presupposes, churchme
n, hijacking, where,
Nearest to from: into, elphinstone, workshop, between, through, spec
ulates, sosa, in,
Nearest to state: esoteric, glaring, presidents, vorarlberg, atmosph
ere, succeeding, lute, connie,
Nearest to on: upon, in, janis, during, torrey, against, infield, ca
talans,
```

```
Nearest to and: or, thirst, in, but, of, sobib, cleaves, including,
Nearest to eight: nine, six, four, seven, three, zero, five, one,
Nearest to they: we, there, he, you, it, these, who, i,
Nearest to more: less, most, deicide, faster, toed, very, skye, toni
c,
Nearest to other: different, attackers, joachim, various, such, man
y, differentiation, these,
Average loss at step 32000 : 3.49501452419
Average loss at step 34000 : 3.48593705952
Average loss at step 36000 : 3.50112806576
Average loss at step 38000 : 3.49244426501
Average loss at step 40000 : 3.3890105716
Nearest to been: become, be, were, was, jolie, hausdorff, spotless,
 had,
Nearest to time: year, way, gleichschaltung, period, day, stanislav,
stage, outcome,
Nearest to over: through, semigroup, rawlings, golden, about, brick,
on, motorways,
Nearest to not: they, radiated, never, pki, still, omnipotent, hinay
ana, really,
Nearest to three: four, six, five, two, seven, eight, one, nine,
Nearest to if: when, before, where, then, bg, because, can, should,
Nearest to there: they, it, he, usually, this, typically, still, oft
en,
Nearest to between: with, in, from, about, against, churchmen, johan
sen, presupposes,
Nearest to from: into, through, elphinstone, in, workshop, between,
 suing, under,
Nearest to state: esoteric, presidents, atmosphere, vorarlberg, lut
e, succeeding, glaring, didactic,
Nearest to on: upon, at, in, during, unitarians, under, catalans, ba
tavians,
Nearest to and: or, but, s, incapacitation, including, while, of, wh
ich,
Nearest to eight: nine, six, seven, four, five, three, one, two,
Nearest to they: we, he, there, you, she, i, not, it,
Nearest to more: less, most, deicide, toed, greater, faster, quite,
 longer,
Nearest to other: various, different, attackers, joachim, clutter, n
z, trochaic, apulia,
Average loss at step 42000 : 3.45294014364
Average loss at step 44000 : 3.47660055941
Average loss at step 46000 : 3.47458503014
Average loss at step 48000 : 3.47261548793
Average loss at step 50000 : 3.45390708435
Nearest to been: become, be, had, was, were, hausdorff, prem, remain
ed,
Nearest to time: way, year, period, stv, day, gleichschaltung, stag
e, outcome,
Nearest to over: through, golden, semigroup, about, brick, counterna
rcotics, theremin, mattingly,
Nearest to not: they, still, never, really, sometimes, it, kiwifrui
t, nearly,
Nearest to three: five, four, six, seven, two, eight, one, nine,
Nearest to if: when, before, where, because, connexion, though, so,
 whether,
Nearest to there: they, it, he, this, now, often, usually, still,
Nearest to between: with, from, fashioned, churchmen, panasonic, exp
lores, within, racial,
Nearest to from: into, through, under, elphinstone, between, worksho
p, circumpolar, idiom,
```

```
Nearest to state: atmosphere, vorarlberg, esoteric, presidents, madh
ya, majority, moulin, bowmen,
Nearest to on: upon, in, catalans, tezuka, minotaurs, wakefulness, b
atavians, guglielmo,
Nearest to and: or, but, thirst, signifier, which, however, includin
g, unattractive,
Nearest to eight: six, nine, seven, five, four, three, zero, two,
Nearest to they: we, there, he, you, it, she, these, not,
Nearest to more: less, most, quite, very, further, faster, toed, dei
cide,
Nearest to other: various, different, many, attackers, are, joachim,
nihilo, reject,
Average loss at step 52000 : 3.43597227755
Average loss at step 54000 : 3.25126817495
Average loss at step 56000 : 3.35102432287
Average loss at step 58000 : 3.44654818082
Average loss at step 60000 : 3.4287913968
Nearest to been: become, be, was, prem, had, remained, hadad, stanis
lavsky,
Nearest to time: year, way, period, stv, barely, name, stage, restor
ing,
Nearest to over: about, through, golden, adaption, counternarcotics,
up, mattingly, brick,
Nearest to not: still, never, nor, kiwifruit, they, nearly, therefor
e, rarely,
Nearest to three: two, five, four, six, seven, eight, one, nine,
Nearest to if: when, though, before, where, although, because, can,
 could,
Nearest to there: they, it, he, still, she, we, this, often,
Nearest to between: with, from, churchmen, among, ethical, within, v
ma, panasonic,
Nearest to from: through, into, under, during, between, in, suing, a
cross,
Nearest to state: atmosphere, infringe, madhya, vorarlberg, governme
nt, bowmen, vargas, republic,
Nearest to on: upon, through, within, ridiculous, janis, in, under,
 over,
Nearest to and: or, while, including, but, of, like, whose, banniste
r,
Nearest to eight: nine, six, five, four, seven, zero, three, two,
Nearest to they: we, there, you, he, it, these, she, prisons,
Nearest to more: less, most, quite, further, toed, very, faster, rat
her,
Nearest to other: different, various, many, nihilo, these, amour, in
cluding, screenplays,
Average loss at step 62000 : 3.38358767056
Average loss at step 64000 : 3.41693099326
Average loss at step 66000 : 3.39588000977
Average loss at step 68000 : 3.35567189544
Average loss at step 70000 : 3.38878934443
Nearest to been: become, be, was, prem, remained, were, being, disco
unts,
Nearest to time: year, way, day, period, barely, ethos, stage, reaso
n,
Nearest to over: about, through, fortunately, semigroup, theremin, o
ff, loudest, up,
Nearest to not: still, nor, never, they, actually, nearly, unelecte
d, therefore,
Nearest to three: five, two, four, six, seven, eight, nine, zero,
Nearest to if: when, though, before, where, because, then, after, si
nce,
```

```
Nearest to there: they, it, he, often, she, we, usually, still,
Nearest to between: among, with, within, from, ethical, churchmen, r
acial, prentice,
Nearest to from: through, into, within, during, under, until, betwee
n, across,
Nearest to state: city, atmosphere, desks, surrounding, preservatio
n, bohr, principal, republic,
Nearest to on: upon, tezuka, through, within, wakefulness, catalans,
at, ingeborg,
Nearest to and: or, but, while, including, thirst, jerzy, massing, a
badan,
Nearest to eight: seven, six, nine, five, four, three, two, zero,
Nearest to they: we, you, he, there, she, it, prisons, who,
Nearest to more: less, most, quite, very, faster, smaller, further,
 larger,
Nearest to other: various, different, some, screenplays, lab, many,
 including, debugging,
Average loss at step 72000 : 3.41103189731
Average loss at step 74000 : 3.44926435578
Average loss at step 76000 : 3.4423020488
Average loss at step 78000 : 3.41976813722
Average loss at step 80000 : 3.39511853886
Nearest to been: become, be, remained, was, grown, were, prem, alrea
dy,
Nearest to time: year, way, period, reason, barely, distance, stage,
day,
Nearest to over: about, fortunately, through, semigroup, further, ma
ttingly, rawlings, golden,
Nearest to not: still, they, nor, never, we, kiwifruit, noaa, reall
y,
Nearest to three: five, two, seven, four, eight, six, nine, zero,
Nearest to if: when, where, though, before, since, because, althoug
h, follows,
Nearest to there: they, it, he, we, she, still, typically, actually,
Nearest to between: with, among, within, in, racial, around, from, s
erapeum,
Nearest to from: into, through, in, within, under, using, during, to
wards,
Nearest to state: city, atmosphere, ferro, vorarlberg, surrounding,
 republic, madhya, national,
Nearest to on: upon, poll, in, from, tezuka, janis, through, within,
Nearest to and: or, but, including, while, s, which, thirst, althoug
h,
Nearest to eight: nine, seven, six, five, four, three, zero, two,
Nearest to they: we, you, there, he, she, it, these, not,
Nearest to more: less, most, smaller, very, faster, quite, rather, l
arger,
Nearest to other: various, different, joachim, including, theos, sma
ller, individual, screenplays,
Average loss at step 82000 : 3.40933967865
Average loss at step 84000 : 3.41618054378
Average loss at step 86000 : 3.31485116804
Average loss at step 88000 : 3.37068593091
Average loss at step 90000 : 3.2785516749
Nearest to been: become, be, was, prem, remained, grown, recently, a
lready,
Nearest to time: year, way, period, day, barely, battle, buds, name,
Nearest to over: through, about, fortunately, off, theremin, semigro
up, extraterrestrial, mattingly,
Nearest to not: nor, still, never, otherwise, generally, separately,
gown, hydrate,
```

```
Nearest to three: four, five, six, two, eight, seven, nine, zero,
Nearest to if: when, where, before, though, because, since, then, wh
ile,
Nearest to there: they, it, he, we, she, still, typically, fiorello,
Nearest to between: with, among, within, from, churchmen, prentice,
 racial, panasonic,
Nearest to from: through, into, across, during, towards, until, at,
 within,
Nearest to state: bohr, city, atmosphere, ferro, bowmen, republic, r
etaliation, vorarlberg,
Nearest to on: upon, in, tezuka, at, during, within, via, catalans,
Nearest to and: or, including, but, while, like, thirst, with, schum
an,
Nearest to eight: seven, nine, six, five, four, three, zero, two,
Nearest to they: we, there, he, you, she, it, prisons, these,
Nearest to more: less, most, very, faster, larger, quite, smaller, b
etter,
Nearest to other: different, various, tamara, prosthetic, including,
individual, failing, restaurants,
Average loss at step 92000 : 3.40355363208
Average loss at step 94000 : 3.35647508007
Average loss at step 96000 : 3.34374570692
Average loss at step 98000 : 3.4230104093
Average loss at step 100000 : 3.36909827
Nearest to been: become, be, grown, was, being, already, remained, p
rem,
Nearest to time: way, year, day, period, years, days, mothersbaugh,
 separators,
Nearest to over: through, about, semigroup, further, fortunately, of
f, into, theremin,
Nearest to not: never, nor, still, dieppe, really, unelected, actual
ly, now,
Nearest to three: four, two, five, seven, six, eight, nine, zero,
Nearest to if: when, though, where, before, is, abe, then, follows,
Nearest to there: they, it, he, we, still, she, typically, often,
Nearest to between: within, with, among, churchmen, around, explore
s, from, reactance,
Nearest to from: into, through, within, across, in, between, using,
 workshop,
Nearest to state: atmosphere, bohr, national, ferro, germ, desks, ci
ty, unpaid,
Nearest to on: upon, in, within, tezuka, janis, batavians, about, ma
crocosm,
Nearest to and: or, but, purview, thirst, sukkot, epr, including, ho
nesty,
Nearest to eight: seven, nine, six, four, five, three, zero, one,
Nearest to they: we, there, you, he, she, prisons, it, these,
Nearest to more: less, most, very, quite, faster, larger, rather, sm
aller,
Nearest to other: various, different, tamara, theos, some, cope, man
y, others,
```

In [0]:

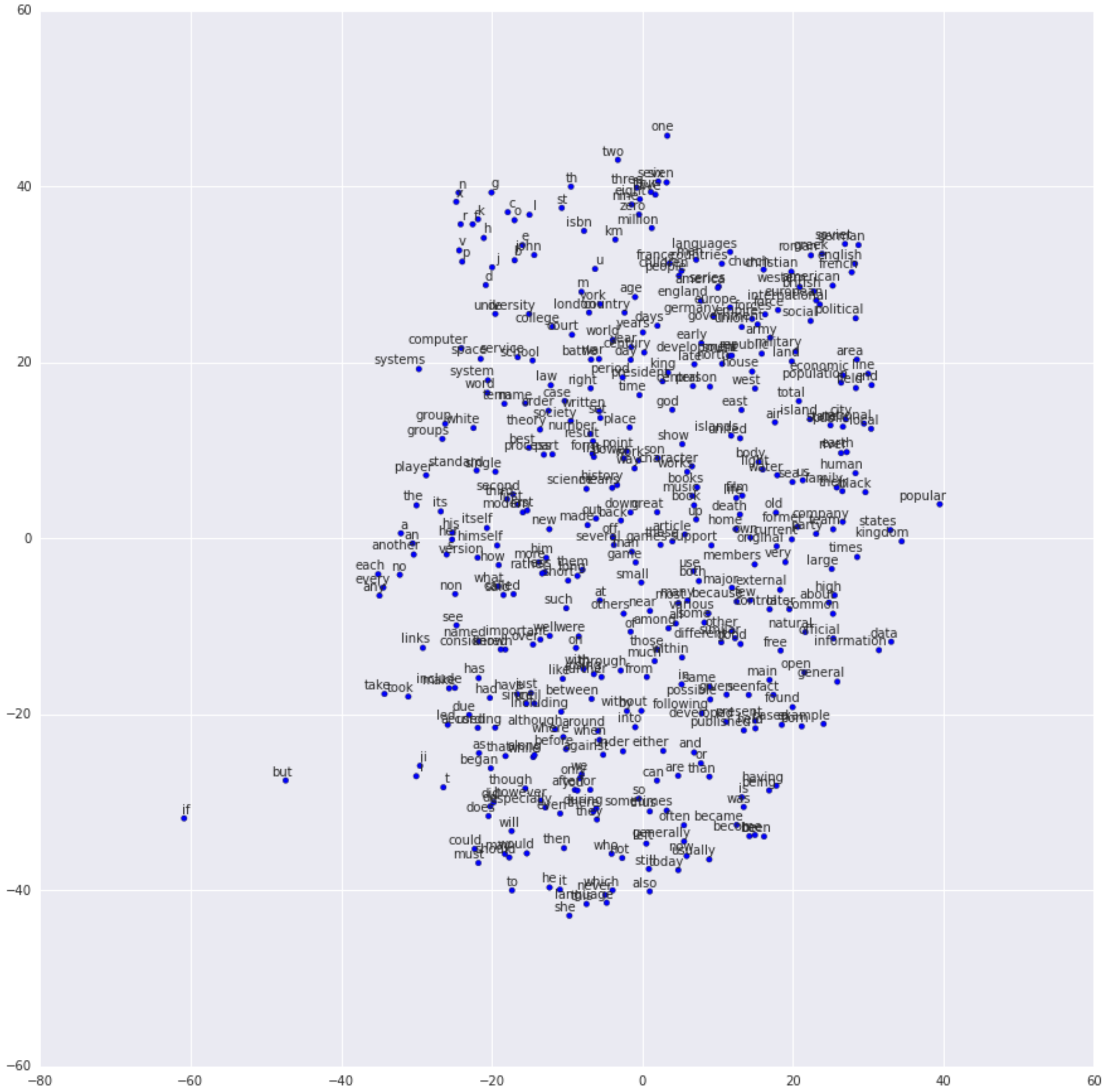
```
num_points = 400

tsne = TSNE(perplexity=30, n_components=2, init='pca', n_iter=5000)
two_d_embeddings = tsne.fit_transform(final_embeddings[1:num_points+1, :])
```

In [0]:

```
def plot(embeddings, labels):
  assert embeddings.shape[0] >= len(labels), 'More labels than embeddings'
  pylab.figure(figsize=(15,15))  # in inches
  for i, label in enumerate(labels):
    x, y = embeddings[i,:]
    pylab.scatter(x, y)
    pylab.annotate(label, xy=(x, y), xytext=(5, 2), textcoords='offset points',
                   ha='right', va='bottom')
  pylab.show()

words = [reverse_dictionary[i] for i in range(1, num_points+1)]
plot(two_d_embeddings, words)
```



Problem

An alternative to skip-gram is another Word2Vec model called CBOW (http://arxiv.org/abs/1301.3781) (Continuous Bag of Words). In the CBOW model, instead of predicting a context word from a word vector, you predict a word from the sum of all the word vectors in its context. Implement and evaluate a CBOW model trained on the text8 dataset.