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 [1]:

```
# 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 [5]:

```
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 [6]:

```
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 [7]:

```
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), ('an
d', 416629), ('one', 411764)]
Sample data [5239, 3084, 12, 6, 195, 2, 3137, 46, 59, 156]
```

Let's display the internal variables to better understand their structure:

In [14]:

```
print(data[:10])
print(count[:10])
print(dictionary.items()[:10])
print(reverse_dictionary.items()[:10])
```

```
[5239, 3084, 12, 6, 195, 2, 3137, 46, 59, 156]
[['UNK', 418391], ('the', 1061396), ('of', 593677), ('and', 416629), ('one', 41176
4), ('in', 372201), ('a', 325873), ('to', 316376), ('zero', 264975), ('nine', 2504
30)]
[('fawn', 45848), ('homomorphism', 9648), ('nordisk', 39343), ('nunnery', 36075),
 ('chthonic', 33554), ('sowell', 40562), ('sonja', 38175), ('showa', 32906), ('woo
ds', 6263), ('hsv', 44222)]
[(0, 'UNK'), (1, 'the'), (2, 'of'), (3, 'and'), (4, 'one'), (5, 'in'), (6, 'a'),
 (7, 'to'), (8, 'zero'), (9, 'nine')]
```

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

In [8]:

```
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_window=skip_window)
    print('\nwith num_skips = %d and skip_window = %d:' % (num_skips, skip_window))
    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', 'term', 'term']
    labels: ['anarchism', 'as', 'originated', 'a', 'as', 'term', 'a', 'of']

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

Note: the labels is a sliding random value of the word surrounding the words of the batch.

It is not obvious with the output above, but all the data are based on index, and not the word directly.

In [15]:

```
print(batch)
print(labels)
```

```
[12 12 12 12  6  6  6  6]
[[5239]
 [3084]
 [ 195]
 [   6]
 [3084]
 [ 195]
 [  12]
 [   2]]
```

Train a skip-gram model.

In [9]:

```
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(softmax_weights, softmax_biases, embed,
                               train_labels, num_sampled, 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 [11]:

```
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: 7.594299
Nearest to united: cellars, praxis, amoeboid, voter, garrick, tatian, defying, wai
nwright,
Nearest to as: dacian, setbacks, donor, corey, mannered, persuades, retains, expou
nded,
Nearest to on: encore, identifiable, vinyl, thighs, parr, lighters, astride, mourn
ing,
Nearest to over: rally, spacetime, niklaus, ricky, enunciated, whigs, warping, tai
chi,
Nearest to from: endless, assur, gyro, provincial, daria, nansen, sines, rushton,
Nearest to world: ancestral, crudely, akad, lasorda, respects, kurz, atheist, seve
ral,
Nearest to had: cube, domination, approximately, oxidizes, psychiatric, relying, d
ecipher, chanced,
Nearest to all: prevenient, nationally, pitcairn, audio, dealers, abbots, blankin
g, fluffy,
Nearest to these: girth, scalar, hydrographic, tracing, hormonal, brl, arnaud, tro
lling,
Nearest to four: walks, misalignment, peri, displaced, atheist, interwar, carey, i
nadvertently,
Nearest to other: twister, fossilization, sumter, bloodstream, octavius, divorcin
g, vaccines, estimate,
Nearest to the: gale, melito, denominations, warheads, attractiveness, ahl, useful
ly, baroque,
Nearest to s: prophylaxis, sissy, charlie, chico, manic, stopper, humayun, induce,
Nearest to when: qc, linker, parsers, platinum, untersuchungen, resists, floors, i
nsofar,
Nearest to at: minefields, alludes, sunderland, amarna, supers, mgm, plugins, defa
ulted,
Nearest to been: haliotis, cardinal, hamito, ced, braille, letters, koruna, pohnpe
i,
Average loss at step 2000: 4.368795
Average loss at step 4000: 3.854706
Average loss at step 6000: 3.786045
Average loss at step 8000: 3.689547
Average loss at step 10000: 3.614999
Nearest to united: cellars, defying, exert, rags, etch, same, glazed, tatian,
Nearest to as: setbacks, fath, candela, laissez, solos, yog, spotless, by,
Nearest to on: in, peculiarly, cheesemaking, senegalese, inasmuch, listener, from,
between,
Nearest to over: spacetime, expects, theoretical, stapleton, stratofortress, warpi
ng, rally, smolensk,
Nearest to from: at, by, endless, in, stunts, business, on, into,
Nearest to world: first, crudely, respects, gaines, cimeti, ridiculed, garonne, ti
mbre,
Nearest to had: have, has, was, where, became, are, portuguese, maldives,
Nearest to all: fluffy, prevenient, borman, audio, torn, oft, blanking, division,
Nearest to these: menander, danubian, some, load, hormonal, shula, alberta, there,
Nearest to four: six, eight, three, five, seven, two, nine, zero,
Nearest to other: chinese, contestant, conceive, chop, gaia, chopstick, hypertalk,
autonomously,
Nearest to the: a, this, its, his, disappear, an, glacial, wireless,
Nearest to s: his, and, or, exact, constraints, recoveries, serine, the,
Nearest to when: where, concerns, platinum, resists, sava, deserted, untersuchunge
n, spat,
Nearest to at: in, from, yan, alludes, ferdinando, during, afdb, combinatorial,
Nearest to been: haliotis, be, braille, cardinal, pohnpei, ced, who, assignments,
Average loss at step 12000: 3.607973
Average loss at step 14000: 3.572373
Average loss at step 16000: 3.405945
```

```
Average loss at step 18000: 3.455610
Average loss at step 20000: 3.539867
Nearest to united: same, cellars, exert, rags, defying, glazed, tatian, etch,
Nearest to as: gok, ashby, when, yog, terrifying, hyperbole, mans, candela,
Nearest to on: in, upon, caption, auras, voor, meridians, aleksandr, listener,
Nearest to over: jiang, stratofortress, expects, noun, within, lifeson, extensibl
e, rally,
Nearest to from: in, into, during, at, including, after, around, stunts,
Nearest to world: crudely, latter, timbre, cimeti, gaines, respects, dv, kavina,
Nearest to had: has, have, was, were, became, been, maldives, portman,
Nearest to all: many, fluffy, these, audio, processors, prevenient, some, borman,
Nearest to these: many, some, all, such, the, dione, there, trick,
Nearest to four: three, six, eight, seven, five, two, nine, one,
Nearest to other: many, various, some, autonomously, impressionist, chinese, scien
ce, flannery,
Nearest to the: its, their, a, his, wireless, this, another, these,
Nearest to s: charlie, perl, constraints, comically, lomo, exact, spell, electroph
ysiology,
Nearest to when: where, after, until, platinum, if, during, as, behind,
Nearest to at: in, during, inflorescence, from, naps, nis, by, alludes,
Nearest to been: be, become, was, haliotis, had, were, ced, heather,
Average loss at step 22000: 3.498566
Average loss at step 24000: 3.484901
Average loss at step 26000: 3.481182
Average loss at step 28000: 3.476800
Average loss at step 30000: 3.501606
Nearest to united: same, cellars, defying, exert, rags, glazed, etch, praxis,
Nearest to as: groton, gladys, combative, howlin, cavett, marmalade, chicagoans, g
rabbed,
Nearest to on: in, internment, upon, dtv, advertisers, from, through, about,
Nearest to over: within, noun, jiang, leds, anchovies, expects, stratofortress, wh
en,
Nearest to from: into, in, under, through, dont, on, fiesole, bl,
Nearest to world: lyotropic, garonne, serving, respects, kavina, homepage, battle,
timbre,
Nearest to had: have, has, was, were, is, irresponsible, became, been,
Nearest to all: many, some, these, active, fluffy, those, expects, audio,
Nearest to these: some, many, such, their, they, all, are, dione,
Nearest to four: six, five, eight, three, seven, two, nine, one,
Nearest to other: various, different, impressionist, some, autonomously, several,
 including, hypertalk,
Nearest to the: its, their, his, a, some, hardwired, orders, these,
Nearest to s: his, charlie, tectonics, exact, melee, whose, catolica, watershed,
Nearest to when: where, if, after, while, until, during, because, however,
Nearest to at: during, in, near, inflorescence, meditation, kane, flyers, periodi
c,
Nearest to been: become, be, was, were, already, had, assignments, kiribati,
Average loss at step 32000: 3.500963
Average loss at step 34000: 3.496071
Average loss at step 36000: 3.455593
Average loss at step 38000: 3.299795
Average loss at step 40000: 3.424177
Nearest to united: cellars, rags, same, german, defying, etch, exert, supreme,
Nearest to as: sophomore, cavett, groton, howlin, yog, combative, by, when,
Nearest to on: upon, at, persuades, senegalese, advertisers, about, in, listener,
Nearest to over: within, anchovies, leds, expects, noun, dystopian, unemployed, ch
ydenius,
Nearest to from: into, during, through, fiesole, under, of, in, after,
Nearest to world: fifa, subregion, conflict, homepage, lyotropic, battle, civil, t
rojan,
Nearest to had: has, have, were, was, irresponsible, subsequently, been, overloadi
```

```
ng,
Nearest to all: many, these, expects, both, prevenient, two, politely, moc,
Nearest to these: some, many, such, their, they, different, several, all,
Nearest to four: six, five, three, seven, eight, two, nine, one,
Nearest to other: various, different, autonomously, some, including, hypocritical,
many, hypertalk,
Nearest to the: its, their, a, his, this, each, any, interquartile,
Nearest to s: his, lomo, homoerotic, catolica, ukiyo, vero, exact, gleichschaltun
g,
Nearest to when: if, where, after, before, while, because, during, until,
Nearest to at: during, on, near, azad, kane, from, against, inflorescence,
Nearest to been: become, be, were, was, already, had, assignments, imposition,
Average loss at step 42000: 3.436665
Average loss at step 44000: 3.452460
Average loss at step 46000: 3.451998
Average loss at step 48000: 3.350378
Average loss at step 50000: 3.379551
Nearest to united: cellars, defying, etch, rags, same, praxis, german, environment
alists,
Nearest to as: candela, laissez, plentiful, bye, dread, caveat, usemodwiki, gradin
g,
Nearest to on: upon, senegalese, in, persuades, caption, walid, under, through,
Nearest to over: within, leds, anchovies, sda, unemployed, noun, chydenius, mctagg
art,
Nearest to from: into, under, in, after, during, schleicher, through, until,
Nearest to world: battle, subregion, constants, debian, merman, fittest, homepage,
conflict,
Nearest to had: has, have, was, were, having, subsequently, been, overloading,
Nearest to all: many, both, every, icosahedron, expects, lughnasadh, knockout, flu
ffy,
Nearest to these: some, many, such, those, several, both, their, different,
Nearest to four: six, five, seven, eight, three, nine, zero, two,
Nearest to other: various, different, many, some, flannery, postulate, food, centr
alisation,
Nearest to the: its, their, his, some, this, aphorism, murdoc, many,
Nearest to s: and, villiers, charlie, uda, his, highways, honky, lomo,
Nearest to when: if, where, while, after, before, during, until, however,
Nearest to at: during, near, in, ethelred, strikers, on, stosunku, of,
Nearest to been: become, be, was, already, had, imposition, were, assignments,
Average loss at step 52000: 3.431731
Average loss at step 54000: 3.422862
Average loss at step 56000: 3.434937
Average loss at step 58000: 3.395922
Average loss at step 60000: 3.397169
Nearest to united: cellars, rags, supreme, defying, etch, praxis, same, watered,
Nearest to as: when, caveat, sophomore, chicagoans, arsacid, candela, husbands, be
came,
Nearest to on: upon, in, through, persuades, before, about, gdb, around,
Nearest to over: within, anchovies, through, in, jingles, birthstone, without, nou
n,
Nearest to from: into, through, including, schleicher, fiesole, under, in, during,
Nearest to world: homepage, subregion, debian, nibelungenlied, fifa, fittest, conf
lict, merman,
Nearest to had: has, have, was, having, were, been, subsequently, failed,
Nearest to all: many, both, icosahedron, every, those, these, knockout, some,
Nearest to these: many, some, such, several, those, both, their, different,
Nearest to four: five, six, eight, three, seven, nine, zero, two,
Nearest to other: various, different, many, some, several, more, food, including,
Nearest to the: its, their, a, some, this, any, each, his,
Nearest to s: whose, catolica, mppc, villiers, tra, mountaineer, leicester, concil
iator,
```

```
Nearest to when: if, after, before, where, while, during, although, though,
Nearest to at: in, near, during, belichick, ret, contradicted, nutshell, ethelred,
Nearest to been: become, be, was, already, imposition, had, were, grantham,
Average loss at step 62000: 3.241358
Average loss at step 64000: 3.249666
Average loss at step 66000: 3.397908
Average loss at step 68000: 3.394693
Average loss at step 70000: 3.356550
Nearest to united: cellars, supreme, rags, same, defying, uk, etch, environmentali
sts,
Nearest to as: like, terrifying, marketplace, grounded, groton, when, ashby, harmo
nising,
Nearest to on: upon, in, through, gdb, at, about, persuades, under,
Nearest to over: about, anchovies, within, birthstone, noun, leds, confirmed, mmx,
Nearest to from: through, into, in, under, during, terminology, before, after,
Nearest to world: homepage, subregion, fittest, nibelungenlied, title, debian, ano
nymously, syndicates,
Nearest to had: has, have, was, having, were, been, subsequently, failed,
Nearest to all: some, many, both, various, every, icosahedron, those, any,
Nearest to these: such, some, many, several, are, their, those, different,
Nearest to four: six, three, five, seven, eight, two, nine, zero,
Nearest to other: various, different, some, many, food, autonomously, including, f
lannery,
Nearest to the: their, this, any, its, a, some, these, his,
Nearest to s: whose, catolica, cumings, my, charlie, stonehenge, chorale, should,
Nearest to when: if, while, where, before, though, because, after, although,
Nearest to at: near, on, dispersing, in, during, bpp, from, ethelred,
Nearest to been: become, be, was, imposition, already, were, had, enigmas,
Average loss at step 72000: 3.376100
Average loss at step 74000: 3.347918
Average loss at step 76000: 3.313139
Average loss at step 78000: 3.352952
Average loss at step 80000: 3.381708
Nearest to united: cellars, rags, supreme, defying, uk, praxis, backlit, environme
ntalists,
Nearest to as: terrifying, gok, chicagoans, caveat, swabian, venetians, jacobsen,
 bonnie,
Nearest to on: upon, in, through, persuades, appel, dtv, advertisers, visconti,
Nearest to over: within, noun, about, anchovies, sy, around, unrwa, answerable,
Nearest to from: through, into, under, before, including, schleicher, after, durin
g,
Nearest to world: subregion, fittest, homepage, title, nibelungenlied, minimalisti
c, banquo, angola,
Nearest to had: has, have, having, were, was, saw, comnenus, never,
Nearest to all: both, various, many, every, any, several, some, icosahedron,
Nearest to these: several, many, those, both, various, some, such, their,
Nearest to four: five, six, three, seven, eight, two, nine, zero,
Nearest to other: various, different, many, food, some, headroom, pepys, baptize,
Nearest to the: its, a, this, their, his, dunwich, meats, wolsey,
Nearest to s: whose, grohl, abdullah, reportage, evidenced, catolica, usurper, my,
Nearest to when: before, if, after, while, though, during, where, although,
Nearest to at: ethelred, near, during, in, belichick, stosunku, dispersing, newsgr
oups,
Nearest to been: become, be, already, was, imposition, enigmas, christy, had,
Average loss at step 82000: 3.406282
Average loss at step 84000: 3.412886
Average loss at step 86000: 3.390965
Average loss at step 88000: 3.348504
Average loss at step 90000: 3.362904
Nearest to united: supreme, cellars, neighbouring, rags, constitution, defying, u
k, aston,
```

```
Nearest to as: bye, howlin, arsacid, marketplace, kamen, cavett, grading, cryopres
ervation,
Nearest to on: upon, under, zapotec, walid, appel, about, visconti, at,
Nearest to over: within, about, lifeson, anchovies, sda, around, noun, through,
Nearest to from: through, into, during, across, before, schleicher, under, in,
Nearest to world: subregion, nibelungenlied, xia, banquo, title, fittest, reclinin
g, mosque,
Nearest to had: has, have, was, were, having, failed, comnenus, never,
Nearest to all: both, many, several, various, every, some, icosahedron, any,
Nearest to these: several, some, many, such, both, those, various, are,
Nearest to four: seven, five, three, six, eight, two, nine, one,
Nearest to other: various, different, individual, autonomously, tycoon, semigroup,
publicized, including,
Nearest to the: its, this, his, their, a, any, arbitrator, lensing,
Nearest to s: catolica, whose, isbn, his, wiseman, thursdays, charlie, mahuad,
Nearest to when: before, if, while, where, after, during, though, although,
Nearest to at: during, near, until, bela, ns, on, risotto, after,
Nearest to been: become, be, already, was, imposition, being, had, remained,
Average loss at step 92000: 3.397071
Average loss at step 94000: 3.256800
Average loss at step 96000: 3.354685
Average loss at step 98000: 3.238157
Average loss at step 100000: 3.353688
Nearest to united: cellars, supreme, neighbouring, constitution, rags, uk, aston,
 baltic,
Nearest to as: cavett, like, when, mazzola, kamen, stratford, colossal, bye,
Nearest to on: upon, in, through, around, under, at, persuades, visconti,
Nearest to over: within, anchovies, about, around, unrwa, sy, couturat, noun,
Nearest to from: in, into, schleicher, including, through, during, after, gyro,
Nearest to world: angola, homepage, season, country, mosque, subregion, fittest, t
itle,
Nearest to had: has, have, having, was, comnenus, were, overloading, failed,
Nearest to all: every, both, many, several, these, various, those, any,
Nearest to these: several, some, many, those, various, different, all, are,
Nearest to four: six, five, seven, eight, two, three, nine, zero,
Nearest to other: various, food, others, different, individual, caesars, smaller,
 including,
Nearest to the: their, his, its, your, a, her, each, this,
Nearest to s: whose, his, dunfermline, gleichschaltung, isbn, my, giscard, leicest
er,
Nearest to when: if, while, before, where, though, although, after, during,
Nearest to at: near, during, in, until, on, after, from, stosunku,
Nearest to been: become, be, already, was, enigmas, imposition, remained, being,
```

This is what an embedding looks like:

In [16]:

```
print(final_embeddings[0])
```

```
[  1.06621169e-01   6.47641346e-02   1.10693499e-01   6.95466325e-02
  -1.20505832e-01  -1.71028391e-01   2.86890212e-02   3.50132920e-02
   5.71503080e-02   3.59160639e-02   8.46984237e-03  -3.42145860e-02
  -1.29389673e-01   8.97347033e-02  -1.26064569e-01  -6.87386692e-02
   1.69990957e-02  -3.38512510e-02   4.35074605e-02   1.17898435e-05
  -3.18375677e-02   1.27275288e-01  -2.05253400e-02  -2.02035736e-02
   4.40550297e-02  -1.07321709e-01  -1.51172578e-01   9.22299400e-02
   1.44088134e-01   1.00372277e-01  -1.63126945e-01   3.33458697e-03
   2.09417641e-02  -9.32454765e-02  -2.02740654e-01   7.63537139e-02
   1.98085401e-02  -7.20573217e-02   2.26487741e-02  -4.53140922e-02
  -5.43340901e-03  -7.94578791e-02  -1.51506528e-01  -5.58510348e-02
   1.06505699e-01   1.05670445e-01   8.74138772e-02   6.65049851e-02
   1.98716670e-02   3.70842330e-02  -6.81178644e-02  -9.46649089e-02
  -1.10908963e-01  -1.59632470e-02  -8.60161483e-02  -2.44216453e-02
   4.98696715e-02  -7.00593144e-02  -1.12661168e-01   1.01890631e-01
   1.74634047e-02  -2.53298357e-02   4.17647809e-02   8.86055753e-02
   6.34797588e-02   4.22846451e-02   2.88530774e-02   1.80510789e-01
   1.09651521e-01  -1.89188287e-01  -5.14636934e-02  -1.13803849e-01
  -1.03230894e-01  -1.15770828e-02   5.58652654e-02   4.78474945e-02
   3.87796434e-03  -7.87318498e-03  -2.49171723e-02  -7.75614232e-02
   8.86766911e-02  -9.03864130e-02   1.21276230e-02  -8.28527063e-02
  -1.27150878e-01  -4.45805304e-02  -1.34781331e-01   8.57374519e-02
  -1.13219917e-01  -6.81594238e-02  -9.72415283e-02   1.36315688e-01
   1.82366461e-01  -6.89850524e-02   4.64664176e-02  -6.11557178e-02
  -6.32885844e-03  -6.17034175e-02  -9.65924561e-02   1.27425000e-01
  -1.89186204e-02  -1.66289762e-01  -1.98210075e-01   5.26272841e-02
   4.40752320e-02  -4.85853590e-02   1.25283822e-01  -3.90184969e-02
  -2.33972166e-02   1.60162896e-01   5.77416196e-02   1.19781204e-01
  -6.72247037e-02  -1.90483108e-02   3.18378024e-02  -5.08362837e-02
  -1.17624931e-01   8.62420648e-02  -9.86490175e-02  -3.50846201e-02
  -4.87034470e-02   3.81041616e-02   1.35818228e-01   1.18673407e-01
   1.99699122e-02  -8.61866027e-02   1.97994877e-02   8.15893486e-02]
```

All the values are abstract, there is practical meaning of the them. Moreover, the final embeddings are normalized as you can see here:

In [17]:

```
print(np.sum(np.square(final_embeddings[0])))
```

```
1.0
```

In [12]:

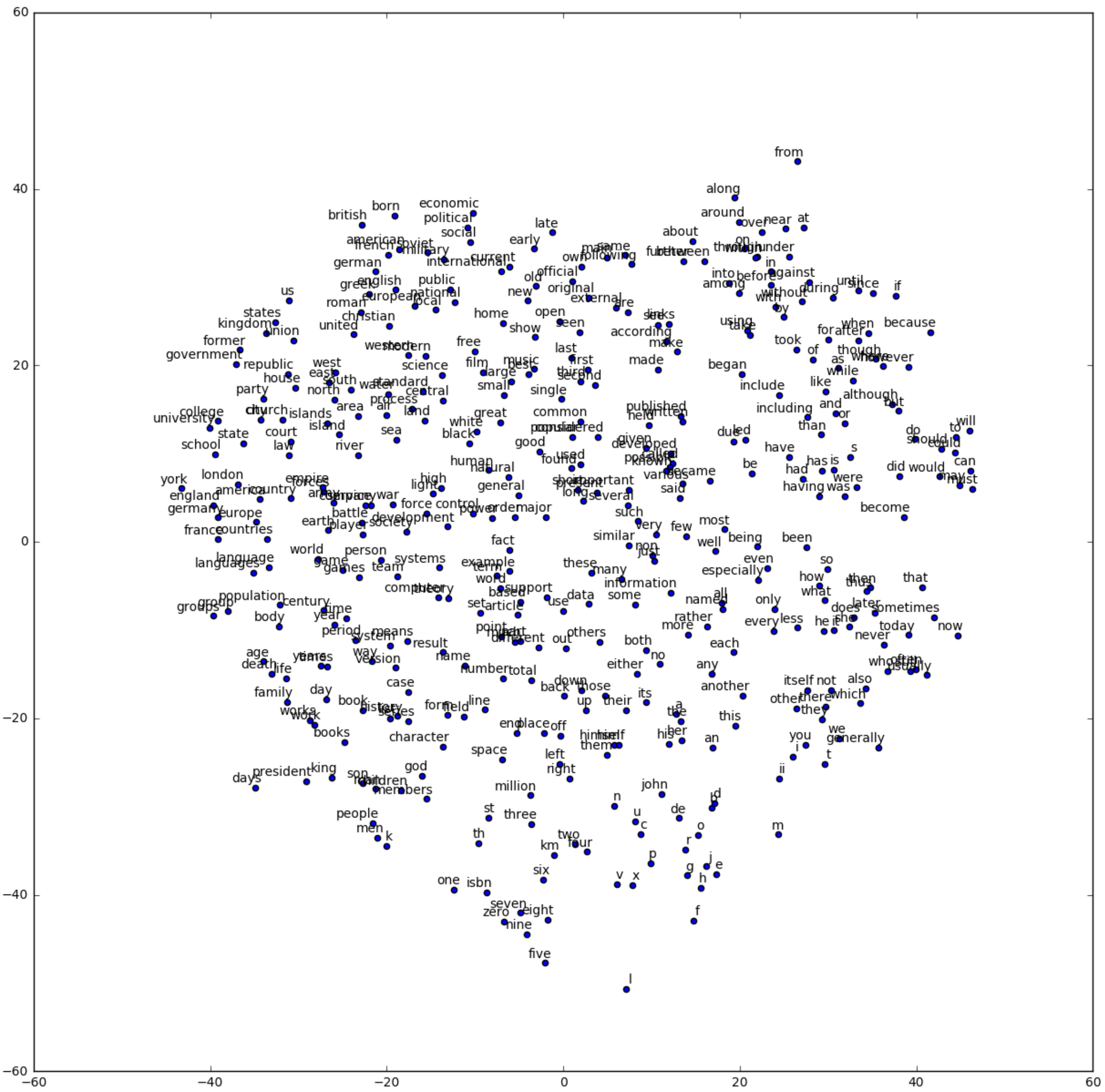
```
num_points = 400

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 [13]:

```
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.

For the continuous bag of words, the train inputs are slightly different from the skip-gram:

In [18]:

```
data_index = 0

def generate_batch(batch_size, bag_window):
  global data_index
  span = 2 * bag_window + 1 # [ bag_window target bag_window ]
  batch = np.ndarray(shape=(batch_size, span - 1), dtype=np.int32)
  labels = np.ndarray(shape=(batch_size, 1), dtype=np.int32)
  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):
    # just for testing
    buffer_list = list(buffer)
    labels[i, 0] = buffer_list.pop(bag_window)
    batch[i] = buffer_list
    # iterate to the next buffer
    buffer.append(data[data_index])
    data_index = (data_index + 1) % len(data)
  return batch, labels

print('data:', [reverse_dictionary[di] for di in data[:16]])

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

```
data: ['anarchism', 'originated', 'as', 'a', 'term', 'of', 'abuse', 'first', 'use
d', 'against', 'early', 'working', 'class', 'radicals', 'including', 'the']

with bag_window = 1:
    batch: [['anarchism', 'as'], ['originated', 'a'], ['as', 'term'], ['a', 'of']]
    labels: ['originated', 'as', 'a', 'term']

with bag_window = 2:
    batch: [['anarchism', 'originated', 'a', 'term'], ['originated', 'as', 'term',
'of'], ['as', 'a', 'of', 'abuse'], ['a', 'term', 'abuse', 'first']]
    labels: ['as', 'a', 'term', 'of']
```

Note the instruction change on the loss function, with reduce_sum to sum the word vectors in the context:

In [19]:

```
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.
bag_window = 2 # How many words to consider left and right.
# 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, bag_window * 2])
  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.
  embeds = 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(softmax_weights, softmax_biases, tf.reduce_sum(embeds, 1),
                               train_labels, num_sampled, vocabulary_size))

  # Optimizer.
  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 [20]:

```
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, bag_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: 8.162153
Nearest to it: named, conduction, negotiating, violently, accessories, zoroastria
n, scarab, shen,
Nearest to will: titicaca, civilis, abendana, calculating, shemini, shaken, gang,
 generate,
Nearest to he: tijuana, allocated, alternated, ginza, captive, embellished, cherri
es, vere,
Nearest to his: quests, qin, sharks, luddites, fic, teschen, notable, aztecs,
Nearest to not: salaam, rest, keeper, ensign, mennonite, dumpster, fullmetal, paul
y,
Nearest to would: chisinau, tte, cartridges, gish, pheasants, croix, baptize, chat
eau,
Nearest to their: mammary, connexion, buchan, reformer, buteo, bipm, santorini, ba
mar,
Nearest to by: tricky, does, sprays, curry, selznick, bn, cervical, coin,
Nearest to six: job, authorship, seine, tsushima, distressing, galilee, give, cons
ubstantiation,
Nearest to is: zur, insists, ao, shun, sima, distrusted, perpetrator, ts,
Nearest to the: separating, yog, ici, springer, xyz, genesis, admitted, spline,
Nearest to after: rovere, rebellious, compensation, ramon, mildly, threading, dor
y, shunning,
Nearest to two: mcduck, tasting, bandanese, rationalist, bitters, cryptozoology, t
axonomists, contravening,
Nearest to three: paullus, ala, neuroscientist, hartman, disavowed, quixtar, conce
ived, euphemistic,
Nearest to or: greedy, witt, paz, internal, quadratic, edwin, contexts, mobility,
Nearest to new: rigid, down, coarse, dbms, paraphyletic, cvs, reverses, debt,
Average loss at step 2000: 7.800249
Average loss at step 4000: 5.551911
Average loss at step 6000: 5.370345
Average loss at step 8000: 4.319964
Average loss at step 10000: 4.238056
Nearest to it: carnap, devil, four, ordinances, arianism, six, channels, commodor
e,
Nearest to will: would, may, could, can, should, might, to, must,
Nearest to he: they, she, it, gael, anu, who, immigrate, spectacled,
Nearest to his: their, its, her, the, dicaprio, our, descendent, your,
Nearest to not: bakongo, nanoscale, undergoing, micrometres, there, nash, motivat
e, universit,
Nearest to would: will, could, may, can, should, might, must, to,
Nearest to their: arianism, oh, valve, devil, both, ordinances, trinity, four,
Nearest to by: who, grossman, through, have, sweeps, deathbed, shower, turbulence,
Nearest to six: four, trinity, devil, arianism, commodore, oh, law, valve,
Nearest to is: was, are, has, dowland, trousers, beau, deborah, ur,
Nearest to the: a, any, his, simeon, its, this, silas, an,
Nearest to after: before, orbiting, footy, parr, during, fighter, modell, songwrit
er,
Nearest to two: three, zero, erasable, five, nine, injustice, snakes, filename,
Nearest to three: two, five, eight, seven, nucleation, earthbound, millikan, zero,
Nearest to or: and, than, hif, nrc, wills, pei, allston, aneurin,
Nearest to new: deductive, hostility, volcanism, caeiro, bibliographies, territor
y, festive, reverses,
Average loss at step 12000: 3.878344
Average loss at step 14000: 3.810855
Average loss at step 16000: 3.816695
Average loss at step 18000: 3.728244
Average loss at step 20000: 3.478109
Nearest to it: he, she, this, there, musicbrainz, thinning, kwajalein, salinas,
Nearest to will: would, could, can, may, should, must, might, did,
Nearest to he: she, it, who, they, khosrau, disseminate, anxiolytics, charters,
```

```
Nearest to his: their, her, its, the, our, s, your, my,
Nearest to not: never, so, still, upc, amara, almost, sicilies, sulfates,
Nearest to would: could, will, may, should, can, might, must, did,
Nearest to their: its, his, her, the, our, chien, fringed, all,
Nearest to by: trigonometry, when, who, fta, naci, deathbed, sweeps, pompidou,
Nearest to six: five, eight, four, seven, three, zero, maharashtri, tropes,
Nearest to is: was, are, became, were, has, makes, means, exists,
Nearest to the: their, his, its, a, an, another, weidman, countermeasure,
Nearest to after: before, when, during, following, fighter, for, mov, claypool,
Nearest to two: three, five, four, eight, zero, yalta, six, vertebra,
Nearest to three: four, six, five, seven, two, eight, vaccines, stretches,
Nearest to or: and, than, canids, meshech, like, overbeck, vinny, peshitta,
Nearest to new: deductive, residential, prerequisite, cosmologists, old, riff, par
aphyletic, vo,
Average loss at step 22000: 3.644970
Average loss at step 24000: 3.476662
Average loss at step 26000: 3.450097
Average loss at step 28000: 3.536635
Average loss at step 30000: 3.404492
Nearest to it: she, he, this, there, musicbrainz, mingled, still, wastewater,
Nearest to will: would, must, could, can, may, should, might, did,
Nearest to he: she, it, they, andres, portsmouth, disseminate, who, previously,
Nearest to his: her, their, its, my, our, the, your, s,
Nearest to not: still, never, sicilies, also, magnetosphere, huggins, astrakhan, k
yrgyzstan,
Nearest to would: could, will, should, may, might, can, must, did,
Nearest to their: its, his, her, our, the, both, fringed, meditator,
Nearest to by: trigonometry, trichloride, who, fta, thus, skip, penetrates, when,
Nearest to six: five, qquad, b, two, variability, ear, motorola, devil,
Nearest to is: was, are, makes, became, remains, gave, has, becomes,
Nearest to the: its, his, a, our, their, another, cirth, menthol,
Nearest to after: before, during, when, following, although, fighter, through, aga
inst,
Nearest to two: ear, variability, qquad, weekly, piercing, afrikaans, devil, arama
ic,
Nearest to three: five, six, four, two, temperature, supposedly, mandible, x,
Nearest to or: and, troyes, than, powiat, dekker, maghreb, hansard, bijection,
Nearest to new: old, cosmologists, hostility, deductive, paraphyletic, residentia
l, utah, separate,
Average loss at step 32000: 3.087801
Average loss at step 34000: 3.620939
Average loss at step 36000: 3.380226
Average loss at step 38000: 3.320057
Average loss at step 40000: 3.320940
Nearest to it: he, she, this, never, largely, ordinances, girls, constantinople,
Nearest to will: could, would, must, should, may, might, cannot, did,
Nearest to he: she, it, they, eventually, longer, clarity, decoration, never,
Nearest to his: her, their, its, your, my, the, our, nrsv,
Nearest to not: still, almost, now, usually, peabody, toilets, also, dab,
Nearest to would: will, should, could, might, may, cannot, can, must,
Nearest to their: its, her, his, your, our, chien, my, them,
Nearest to by: trigonometry, penetrates, trichloride, who, swedes, ping, among, sa
yers,
Nearest to six: five, qquad, carbon, inquiry, afrikaans, survey, variability, comm
odore,
Nearest to is: remains, qquad, survey, piercing, phoenician, clausewitz, alicante,
ear,
Nearest to the: a, this, his, bagpipe, unwanted, deut, leicester, jena,
Nearest to after: before, following, during, within, despite, when, until, dacia,
Nearest to two: four, three, zero, marburg, tron, maxillofacial, psychometrics, ta
jiks,
```

```
Nearest to three: seven, eight, four, two, six, five, zero, strunk,
Nearest to or: piercing, isbn, qquad, human, afrikaans, ear, carl, survey,
Nearest to new: vo, poets, fiddler, vila, rankings, boxer, separate, full,
Average loss at step 42000: 3.355122
Average loss at step 44000: 3.280298
Average loss at step 46000: 3.288660
Average loss at step 48000: 3.487177
Average loss at step 50000: 3.217888
Nearest to it: she, he, this, there, musicbrainz, even, itch, still,
Nearest to will: would, should, can, might, could, may, must, cannot,
Nearest to he: she, it, they, absurdities, who, even, formally, there,
Nearest to his: her, their, its, our, the, my, your, him,
Nearest to not: never, still, almost, now, grundgesetz, endorse, taxing, mostly,
Nearest to would: could, might, will, should, may, can, must, did,
Nearest to their: its, his, her, our, the, your, fringed, my,
Nearest to by: penetrates, trigonometry, sayers, trichloride, among, typeface, pin
g, claire,
Nearest to six: seven, five, zero, four, three, edvard, program, chapter,
Nearest to is: was, are, includes, became, makes, were, contains, becomes,
Nearest to the: his, their, a, its, our, your, mordecai, holberg,
Nearest to after: before, during, following, within, geologist, despite, until, si
nce,
Nearest to two: five, three, mackintosh, four, six, reservation, tron, etiology,
Nearest to three: five, four, seven, eight, six, two, reusable, nighttime,
Nearest to or: and, tn, than, octopussy, abandoning, workplaces, tar, agadir,
Nearest to new: modern, different, old, utah, former, naturalis, fiddler, vo,
Average loss at step 52000: 3.548038
Average loss at step 54000: 3.249379
Average loss at step 56000: 3.028839
Average loss at step 58000: 3.105691
Average loss at step 60000: 3.111861
Nearest to it: he, she, this, there, cuba, itself, increasingly, still,
Nearest to will: would, should, must, might, could, may, can, cannot,
Nearest to he: she, it, they, influenza, anxiolytics, spectacled, lak, who,
Nearest to his: her, their, its, our, s, your, my, the,
Nearest to not: never, still, almost, now, mostly, grundgesetz, rarely, plankton,
Nearest to would: might, will, could, should, must, may, can, did,
Nearest to their: its, his, our, her, the, your, whose, fringed,
Nearest to by: under, compulsion, among, without, quotients, bonewits, ratifying,
 typeface,
Nearest to six: four, five, seven, eight, three, zero, nine, two,
Nearest to is: was, are, includes, became, has, makes, remains, contains,
Nearest to the: its, their, a, his, an, another, any, this,
Nearest to after: before, during, following, since, despite, thereafter, when, lat
er,
Nearest to two: three, five, four, zero, eight, six, erasable, one,
Nearest to three: four, five, six, seven, eight, zero, two, nestor,
Nearest to or: and, than, while, confidentiality, tracy, like, santer, arpanet,
Nearest to new: former, vo, raiding, separate, different, streaming, particular, r
iff,
Average loss at step 62000: 3.080276
Average loss at step 64000: 3.088795
Average loss at step 66000: 3.021331
Average loss at step 68000: 3.028283
Average loss at step 70000: 3.076700
Nearest to it: he, she, this, there, they, probably, cuba, even,
Nearest to will: would, could, must, can, might, should, may, cannot,
Nearest to he: she, it, they, who, charters, andrews, emilio, decimus,
Nearest to his: their, her, its, our, your, my, the, s,
Nearest to not: now, never, still, rarely, grundgesetz, toilets, mostly, torn,
Nearest to would: could, will, might, may, should, must, can, did,
```

```
Nearest to their: its, his, our, her, your, the, these, my,
Nearest to by: via, naci, when, trigonometry, among, damper, ping, typeface,
Nearest to six: seven, five, three, eight, four, zero, two, ng,
Nearest to is: was, makes, are, remains, includes, requires, exists, becomes,
Nearest to the: their, its, a, his, weidman, our, an, another,
Nearest to after: before, during, following, despite, through, thereafter, since,
 afterwards,
Nearest to two: three, five, four, six, zero, seven, eight, various,
Nearest to three: five, six, seven, two, four, eight, zero, kcal,
Nearest to or: and, than, asatru, objecting, but, etc, vinyl, citing,
Nearest to new: separate, hybrids, different, special, riff, webpage, vo, small,
Average loss at step 72000: 2.993144
Average loss at step 74000: 2.944780
Average loss at step 76000: 3.057956
Average loss at step 78000: 3.069739
Average loss at step 80000: 2.927056
Nearest to it: he, coffee, afghanistan, many, boy, ordinances, game, z,
Nearest to will: would, could, might, must, should, may, cannot, shall,
Nearest to he: it, she, they, leto, ln, andrews, override, clausewitz,
Nearest to his: her, its, their, my, our, your, s, the,
Nearest to not: never, still, almost, now, rajputs, either, nothing, dab,
Nearest to would: will, might, could, should, must, may, did, cannot,
Nearest to their: its, her, your, his, our, the, nrsv, many,
Nearest to by: trigonometry, via, naci, taxonomic, typeface, trichloride, without,
when,
Nearest to six: seven, five, eight, four, three, zero, nine, powerpc,
Nearest to is: was, makes, are, becomes, includes, provides, remains, contains,
Nearest to the: their, any, a, another, our, its, your, his,
Nearest to after: during, before, afterwards, thereafter, giancana, since, upon, w
hen,
Nearest to two: three, five, seven, four, erasable, milligrams, zero, combs,
Nearest to three: four, seven, five, two, triple, forty, hummer, many,
Nearest to or: and, than, without, hurting, but, zoot, tossed, ligamentum,
Nearest to new: separate, concocted, small, residential, hybrids, different, marsh
land, smiley,
Average loss at step 82000: 2.996220
Average loss at step 84000: 2.948041
Average loss at step 86000: 2.987175
Average loss at step 88000: 3.041709
Average loss at step 90000: 2.903274
Nearest to it: he, she, this, there, itself, itch, they, kare,
Nearest to will: would, could, should, must, can, might, cannot, shall,
Nearest to he: she, it, they, enrich, nitrate, radioactivity, emilio, lipstick,
Nearest to his: her, its, their, my, your, our, whose, the,
Nearest to not: never, still, rarely, grundgesetz, also, nothing, rajputs, almost,
Nearest to would: will, might, could, should, can, must, cannot, did,
Nearest to their: its, his, your, nrsv, variability, bourgeois, resistors, our,
Nearest to by: trigonometry, among, via, typeface, ratifying, through, ping, when,
Nearest to six: four, five, seven, three, zero, eight, forty, p,
Nearest to is: was, becomes, are, remains, makes, has, exists, provides,
Nearest to the: a, any, its, their, our, his, an, another,
Nearest to after: before, during, following, despite, when, within, afterwards, th
ereafter,
Nearest to two: three, erasable, five, four, one, six, twelve, injustice,
Nearest to three: five, four, six, seven, two, eight, lise, zero,
Nearest to or: and, than, while, bro, etc, containing, emblematic, dekker,
Nearest to new: separate, concocted, previous, old, hybrids, lyman, macha, vo,
Average loss at step 92000: 2.963882
Average loss at step 94000: 2.930725
Average loss at step 96000: 2.780458
Average loss at step 98000: 2.507631
```

```
Average loss at step 100000: 2.760214
Nearest to it: he, she, there, this, valyria, finland, ordinances, ball,
Nearest to will: must, would, could, should, shall, might, can, cannot,
Nearest to he: she, they, it, lipstick, there, nest, typographic, absurdities,
Nearest to his: her, their, its, our, your, my, s, the,
Nearest to not: never, indeed, almost, now, nothing, identifiable, rarely, still,
Nearest to would: might, could, will, should, must, can, may, cannot,
Nearest to their: its, his, her, your, our, my, these, segmented,
Nearest to by: trigonometry, trichloride, through, completely, from, orbis, quebec
ers, when,
Nearest to six: seven, five, four, eight, three, zero, nine, two,
Nearest to is: was, provides, are, exists, makes, includes, becomes, remains,
Nearest to the: its, a, his, her, another, any, billet, this,
Nearest to after: before, following, despite, afterwards, without, during, thereaf
ter, survived,
Nearest to two: three, four, five, seven, honeys, zero, six, speller,
Nearest to three: four, seven, six, five, eight, two, zero, carl,
Nearest to or: and, than, but, etc, santer, vs, though, exogenous,
Nearest to new: old, separate, vo, fiddler, hybrids, previous, lyman, khanty,
```

In [22]:

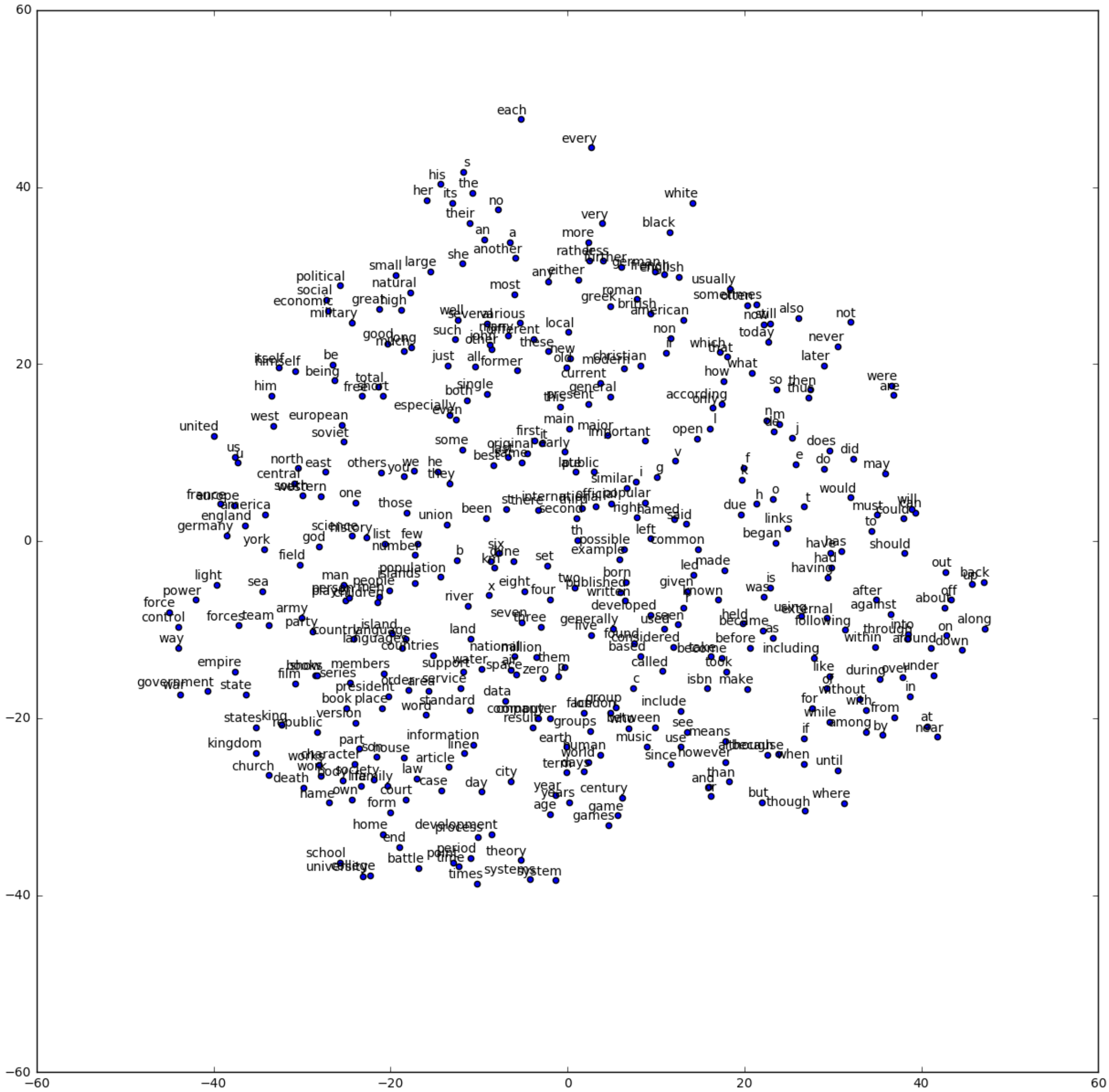
```
num_points = 400

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 [23]:

```
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)
```



Some clusters are less obvious (like the standalone characters), but it clearly totaly works!

How does your CBOW model perform compared to the given Word2Vec model? (to be answered)

At the first sight, they look similar. The CBOW is a more compact.

In [ ]: