CSE 628 - Assignment 1: Word Vectors - Report

1. Hyper-parameters explored w/ a description of what each parameter controls and how varying each is likely to affect the training

Experiment with different sized context window:

configuraion1:

```
################################################################################
#######
  # Hyper Parameters to config
  batch_size = 256
  embedding_size = 256  # Dimension of the embedding vector.
  skip_window = 4       # How many words to consider left and right.
  num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.120215079049

Nearest to first: the, of, UNK, one, and, in, zero, to, two, eight, a, s, seven, as, are, nine, was, or, he, his,
Nearest to american: of, the, a, to, two, one, nine, and, s, as, zero, four, eight, UNK, five, in, an, was, that, or,
Nearest to would: the, to, and, that, of, two, three, by, zero, one, a, four, nine, for, it, six, also, in, as, UNK,

CE loss - Average loss at step 200000 : 5.53326544495

Nearest to first: assignments, forcing, tylenol, fearless, estad, triskelion, sident, posthuman, altiplano, personalize, vat, julee, forerunners, cryptographer, ircana, liddy, projection, ounce, genius, landsteiner,
Nearest to american: midland, unimaginative, intrusions, vola, reactivity, tamura, hingis, monegasque, cymbals, robby, furore, heartbroken, walford, evonne, erling, strips, sbc, tradeable, nominations, romanorum,
Nearest to would: aoste, capriati, bergomi, desk, oxonian, was, girlfriends, baulieu, arrows, circ, messene, alerts, sandwiches, scivias, mingus, merritt, distract, affine, jason, sensibility,

configuraion2:

```
################################################################################
#######
```

```
 # Hyper Parameters to config
 batch_size = 512
 embedding_size = 512  # Dimension of the embedding vector.
 skip_window = 6       # How many words to consider left and right.
 num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.115739194609

Nearest to first: of, the, with, and, in, s, one, a, nine, eight, four, is, by, UNK, zero, was, to, two, at, also,
Nearest to american: in, the, nine, zero, one, and, from, of, as, UNK, two, by, a, for, four, with, to, that, s, it,
Nearest to would: the, of, was, and, zero, one, seven, UNK, four, is, in, six, to, three, nine, are, for, two, by, s,

CE loss - Average loss at step 200000 : 6.25468926373

Nearest to first: fluted, shrank, misrepresent, waldseem, carissimi, ephron, kitcher, flatland, tolkien, gaffer, concluding, schoolly, electrolysis, ellipsoids, botulinum, agustin, jassi, dialects, negotiations, dxf,
Nearest to american: glaucous, noch, blinkenlights, idolizes, frugal, wisecrack, horsemen, prodigies, gentis, pvc, covetous, neustrians, sweeten, shield, lule, oses, hathor, reflectivity, negotiate, markov,
Nearest to would: oop, confounded, appearing, transfered, execute, binary, nijhoff, piloto, fools, coughed, bourses, mandioca, tetris, rohypnol, furnace, tricking, vallarta, snorted, alternation, maghrib,

configuraion3:

```
################################################################################
 # Hyper Parameters to config
 batch_size = 64
 embedding_size = 64  # Dimension of the embedding vector.
 skip_window = 6       # How many words to consider left and right.
 num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.126460549086

Nearest to first: accumulated, that, on, six, its, if, three, by, zero, was, his, origins, taoism, kropotkin, UNK, nine, would, seven, which, is,
Nearest to american: zero, five, one, was, seven, three, eight, the, two, s, and, UNK, nine, of, world, by, in, on, four, it,
Nearest to would: its, resources, six, first, s, by, has, seven, which, that, two, accumulated, decreed, three, from, he, with, association, UNK, one,

CE loss - Average loss at step 200000 : 4.14847928534

Nearest to first: pahawh, institutional, spiracles, superstructure, adriatic, latveria, christianizing, poplar, remo, erina, beebe, hotman, starches, muff, zarz, ostwald, zf, nickname, extremist, arish,
Nearest to american: vibrant, heights, excessively, chalons, panentheists, eulogized, fangio, northney, fille, craic, constituency, gags, ulti, skirmished, whore, outliers, telstra, estis, vanes, hastened,
Nearest to would: kaczynski, morphine, sulfate, cielobuio, ichikawa, primarily, christy, lusatia, fracas, suva, dener, falcons, highwind, erudition, doig, hamito, simulate, melian, monoculture, phalanx,

configuraion4:

```
################################################################################
#######
  # Hyper Parameters to config
  batch_size = 32
  embedding_size = 32  # Dimension of the embedding vector.
  skip_window = 6       # How many words to consider left and right.
  num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.156957978113

Nearest to first: such, by, it, two, his, zero, nine, four, this, he, not, s, with, one, and, its, three, at, UNK, until,
Nearest to american: other, for, has, six, d, two, his, their, when, five, be, known, western, f, book, three, but, he, its, an,
Nearest to would: nine, one, four, two, and, in, only, at, this, were, an, zero, was, with, s, seven, by, from, but, UNK,

CE loss - Average loss at step 200000 : 3.46036002703

Nearest to first: anaphora, insecurities, voltage, hoth, underdogs, les, newstead, zaires, querelle, signer, coinages, secularist, rely, drale, palatalization, breviaries, bengtsson, bipeds, olympias, underseas,
Nearest to american: fable, bengals, wpan, magazines, clays, ganda, pollution, belasco, circulations, buzzsaw, belief, venizelos, areas, ratz, lima, tentacle, lago, gruy, heideggerian, googolplex,
Nearest to would: psd, computed, pleas, bupropion, reedited, inversions, zki, justly, scenario, bostik, kufra, encountering, cartoons, dhimmi, jalalabad, diverged, vol, typological, simony, hitless,

configuraion5:

```
##################################################################################
#######
  # Hyper Parameters to config
  batch_size = 32
  embedding_size = 32  # Dimension of the embedding vector.
  skip_window = 2       # How many words to consider left and right.
  num_skips = 4         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.125708107245

Nearest to first: were, s, vary, or, from, he, an, about, they, two, on, eight, UNK, seven, then, six, zero, and, with, one,
Nearest to american: up, six, seven, is, zero, their, england, two, are, and, three, UNK, eight, nine, s, one, from, means, these, that,
Nearest to would: nine, under, north, is, some, was, with, zero, three, they, UNK, found, six, two, for, early, s, see, american, one,

CE loss - Average loss at step 200000 : 3.39686465039

Nearest to first: heritable, chalcogens, tavola, antigen, parastatistics, gplv, sonet, asn, cholula, downarrow, khz, detecting, sucka, epd, leftmost, galactica, collip, noteworthy, widower, westbam,
Nearest to american: consummate, bouman, re, usflag, mock, swan, quadrant, hermetical, slovenia, britney, constantia, scenario, disputed, administrations, jenin, ordained, faneuil, khaganate, hostnames, swimwear,
Nearest to would: could, may, did, can, will, should, cloete, had, miyako, hikers, does, aladdin, fallopian, cased, huevos, mycorrhizae, hamdi, translucent, must, fiorentina,

Experiement with negative sample and unigram probability values:

configuraion 6(default given):

```
##################################################################################
#######
 learning rate 1, pretrained checkpoint,

  # Hyper Parameters to config
  batch_size = 128
  embedding_size = 128  # Dimension of the embedding vector.
  skip_window = 4       # How many words to consider left and right.
  num_skips = 8         # 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.random.choice(valid_window, valid_size, replace=False)
  valid_examples = np.append(valid_examples, [dictionary["first"], dictionary["american"],
dictionary["would"]])
  num_sampled = 64    # Number of negative examples to sample.

  # summary_path = './summary_%s'%(loss_model)
  pretrained_model_path = './pretrained/'

  checkpoint_model_path = './checkpoints_%s/'%(loss_model)
  model_path = './models'


  # maximum training step
  max_num_steps  = 200001
  checkpoint_step = 50000
```

NCE loss - Average loss at step 200000 : 0.0709739996528

Nearest to first: he, during, three, at, one, seven, zero, time, UNK, was, were, an, had, his, on, since, nine, eight, with, use,
Nearest to american: british, mollusca, nisbet, vivipara, skarsg, polanyi, houari, lacombe, lawmaker, louse, overshoes, km, rudis, dealbata, five, german, monas, ugliest, other, oresund,
Nearest to would: not, i, him, eight, four, so, only, it, such, this, for, six, their, two, some, s, many, called, her, more,

CE loss - Average loss at step 200000 : 4.82234705238

Nearest to first: last, most, same, original, continued, following, latter, end, a, main, best, largest, due, because, left, before, coast, much, until, west,
Nearest to american: german, british, its, his, english, russian, french, european, canadian, italian, trade, international, transpiration, united, extension, reject, outputs, carmen, their, projectiles,
Nearest to would: will, could, might, must, did, may, can, was, does, should, is, began, had, do, seems, came, appears, we, t, are,

<u>configuraion 7:</u>

```
##############################################################################
#######
  # Hyper Parameters to config
  batch_size = 32
  embedding_size = 32  # Dimension of the embedding vector.
  skip_window = 4       # How many words to consider left and right.
  num_skips = 8         # How many times to reuse an input to generate a label.
```

num_sampled = 32   # Number of negative examples to sample.

NCE loss - Average loss at step 200000 : 0.116317770059

Nearest to first: used, as, three, one, s, a, he, two, zero, has, in, and, for, five, UNK, world, eight, at, only, nine,
Nearest to american: three, eight, for, he, an, two, zero, as, new, a, one, was, UNK, in, s, nine, there, which, from, such,
Nearest to would: not, as, one, from, s, been, its, in, a, which, called, history, that, may, government, zero, seven, UNK, also, nine,

CE loss - Average loss at step 200000 : 3.43155877233

Nearest to first: baptize, reynaud, ferroni, vulva, kodokan, culbert, bed, remain, mboxn, passacaglia, lignin, sterilize, quadrupedal, barbarous, cancellation, basilio, lsr, asia, rebellions, binalshibh,
Nearest to american: mille, nyungan, hegelianism, jean, andesite, davydenko, carefree, particularly, ceux, bankrupting, pictographs, perdido, guangxi, stockpiled, corresponds, ansett, steadfastly, configurable, mellon, based,
Nearest to would: lim, fist, solving, chalcogen, abiogenic, fazel, brass, arakan, announce, bastards, instillation, epona, abandoned, koutoubia, festivity, redactional, migraine, shelby, uproot, pies,

configuraion 8:

```
##################################################################################
#######
 learning rate 1
  # Hyper Parameters to config
  batch_size = 32
  embedding_size = 32  # Dimension of the embedding vector.
  skip_window = 4       # How many words to consider left and right.
  num_skips = 8         # How many times to reuse an input to generate a label.

  num_sampled = 16    # Number of negative examples to sample.
```

NCE loss - Average loss at step 200000 : 0.105297995041

Nearest to first: a, is, the, eight, UNK, of, zero, that, on, and, with, in, was, than, or, from, to, nine, by, be,
Nearest to american: for, as, zero, two, which, used, when, had, would, states, of, the, one, in, six, nine, has, and, UNK, s,

Nearest to would: s, some, his, american, had, zero, book, used, was, two, about, english, took, against, system, an, than, five, UNK, south,

CE loss - Average loss at step 200000 : 3.43076278262

Nearest to first: nav, bike, suzerainty, asner, nsg, cna, vub, fawn, papa, backported, freed, mythical, multicellular, neris, mme, tubing, zizka, warburton, ongoing, blobby,
Nearest to american: kondrashin, kentish, telef, circumference, toth, draftsmen, proto, mwali, soleil, fathoms, alloy, luria, taras, soundboard, desron, grime, growl, iskandariyah, bisbal, harian,
Nearest to would: sre, gourde, axeinos, estonians, snowdonia, necropolis, gris, gorgosaurus, meditating, ctenophora, alciphron, wren, durga, camelids, jotuns, republique, petzold, karmic, insurrectionary, esdras,

configuraion 9:

```
################################################################################
#######
 learning rate 1

  # Hyper Parameters to config
  batch_size = 32
  embedding_size = 32  # Dimension of the embedding vector.
  skip_window = 6       # How many words to consider left and right.
  num_skips = 8         # How many times to reuse an input to generate a label.

  num_sampled = 8    # Number of negative examples to sample.
```

NCE loss - Average loss at step 200000 : 0.0736719462202

Nearest to first: four, on, end, at, nine, five, have, three, was, zero, in, s, a, by, two, he, UNK, eight, for, one,
Nearest to american: in, a, of, one, UNK, nine, and, be, was, for, the, as, six, been, zero, his, five, on, or, to,
Nearest to would: that, seven, it, with, was, he, they, in, play, of, on, this, their, at, and, its, which, the, two, nine,

CE loss - Average loss at step 200000 : 3.46020747638

Nearest to first: delights, orsi, kee, grandstands, shocklee, optimized, discreet, dared, nibley, bytb, garageband, salmonella, afterlife, croce, predictably, urus, huyuk, tallies, ciphertext, aussie,
Nearest to american: vagaries, crystallization, flagstaff, acorns, bonde, arktos, resembles, stubbornness, prasad, kypria, exons, hedonistic, other, frescobaldi, sonderkommandos, pentium, eoka, rosin, catalyst, bates,
Nearest to would: oblige, ohci, cyclone, capac, mercantilism, devastation, sailed, humoral, cheney, steamer, nehamas, quarrelling, piles, restricts, recasts, autoconfiguration, moo, chromaticity, schwangyu, abol,

configuraion 10:

##
#######
learning rate 1

```
# Hyper Parameters to config
batch_size = 32
embedding_size = 32  # Dimension of the embedding vector.
skip_window = 6       # How many words to consider left and right.
num_skips = 8         # How many times to reuse an input to generate a label.

num_sampled = 128    # Number of negative examples to sample.
```

NCE loss - Average loss at step 200000 : nan

Nearest to first: castilians, percepts, forbear, nkhotakota, blush, gaffer, overbuilt, khat, khar, yngvi, whine, nane, eyelash, wombats, riddel, proclivities, meningococcal, lod, spaak, artin,
Nearest to american: castilians, percepts, forbear, nkhotakota, blush, gaffer, overbuilt, khat, khar, yngvi, whine, nane, eyelash, wombats, riddel, proclivities, meningococcal, lod, spaak, artin,
Nearest to would: castilians, percepts, forbear, nkhotakota, blush, gaffer, overbuilt, khat, khar, yngvi, whine, nane, eyelash, wombats, riddel, proclivities, meningococcal, lod, spaak, artin,

CE loss - Average loss at step 200000 : 3.46051120896

Nearest to first: nasr, thurgood, dept, knitter, caplet, relate, desk, aplonis, stuntmen, tomography, learnings, kallisti, perquisites, wuthering, misrule, bhagavad, interferons, spirituality, centimeter, phones,
Nearest to american: yamuna, prosthetics, replicability, sabellianism, adullam, conclave, greene, flagpole, immanence, diels, helper, doin, laraine, bhosle, ura, dunkel, moby, redhead, wy, pyrex,
Nearest to would: tin, juzaburo, intervertebral, hist, utd, nachmanides, conceptually, axiomatized, mouth, unlawfully, foregrounds, artin, truetype, adgb, wavetable, taxicab, firecracker, miatas, paratime, searcher,

Experiment with learning rate and number of epoch:

configuraion 11:

##
#######

learning rate 1.5

```
# Hyper Parameters to config
batch_size = 64
embedding_size = 64  # Dimension of the embedding vector.
skip_window = 4       # How many words to consider left and right.
num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.110414775851

Nearest to first: one, it, was, zero, at, be, s, nine, eight, for, UNK, most, which, their, five, seven, on, the, into, have,
Nearest to american: state, five, UNK, zero, on, while, six, nine, the, for, that, were, its, it, two, eight, from, one, was, later,
Nearest to would: with, some, civil, an, i, UNK, was, it, s, have, his, for, one, on, three, two, be, that, nine, find,

CE loss - Average loss at step 200000 : 4.13454262438

Nearest to first: repetitious, kangaroos, kspread, mesogens, addu, riposte, baqir, wags, distinction, halakhah, protects, pfe, sanskrit, hayling, zoonoses, gauntlets, savart, punters, keenest, worldly,
Nearest to american: benelli, grammatical, vijay, surveying, steuben, epilogue, rosemary, bohr, elion, justicia, artbooks, jcf, moluccas, shura, sylheti, enfranchised, conversing, racing, dongell, bayqongyr,
Nearest to would: unhallowed, has, akermanis, lorrin, cgas, finalized, statement, contradanza, edda, commissions, petherton, theol, chiffons, alencar, bottomless, was, innings, presentable, nyse, invitational,

<u>configuraion 12:</u>

```
################################################################################
#######
 learning rate 1.5

 # Hyper Parameters to config
 batch_size = 64
 embedding_size = 64  # Dimension of the embedding vector.
 skip_window = 4       # How many words to consider left and right.
 num_skips = 8         # How many times to reuse an input to generate a label.

 Number of epoch : 4
```

NCE loss - Average loss at step 200000 : 0.107766077326

Nearest to first: was, seven, that, history, UNK, nine, how, from, has, ideas, zero, and, association, the, five, of, one, two, according, also,

Nearest to american: seven, nine, that, for, which, one, was, and, five, an, at, eight, s, his, the, be, zero, they, six, four,
Nearest to would: he, UNK, first, also, from, of, six, seven, s, and, zero, nine, which, the, four, his, be, one, eight, two,

CE loss - Average loss at step 200000 : 4.13467730338

Nearest to first: ugueth, holism, inflexible, qnx, properties, antenatal, stauffenberg, smirnoff, edinburghers, burkina, tents, elizabethtown, overturn, grovelling, kiick, firefighter, mattingly, dernier, hardiness, berghof,
Nearest to american: overthrew, cowboy, hypothesizing, inerrant, randal, mezzogiorno, lojban, fulltime, governance, carte, reelected, whitmire, cal, landvetter, invite, sinc, necessity, prophecy, tempe, hirschfeld,
Nearest to would: xeon, hellman, lodgepole, meddle, tayside, palliative, interleaf, fangs, logout, were, cardiovascular, renal, nadph, sharp, holocaust, vasco, was, soekershof, emotion, badoglio,

configuraion 13:

##
learning rate 0.5

```
# Hyper Parameters to config
batch_size = 64
embedding_size = 64  # Dimension of the embedding vector.
skip_window = 4       # How many words to consider left and right.
num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.211744055485

Nearest to first: one, his, s, eight, the, which, their, also, a, to, zero, four, two, over, in, of, nine, and, is, UNK,
Nearest to american: the, two, of, nine, in, also, to, and, zero, at, is, that, UNK, a, as, one, by, five, from, not,
Nearest to would: seven, one, zero, two, three, on, eight, six, and, of, UNK, to, be, was, the, with, a, area, it, that,

CE loss - Average loss at step 200000 : 4.12829264288

Nearest to first: yoshihisa, deposits, lightness, providential, miramar, peuple, descendent, wrestlemania, experimental, cinematography, bentsen, evaluate, swieten, cortina, fiedler, changes, subcritical, cna, filiation, krestinsky,
Nearest to american: kabir, eke, lans, spelt, biondo, tschermak, prez, mosin, gragn, jevons, mountbatten, lawyer, spahn, dilutions, gaylord, ussuri, lookout, strip, canguilhem, pundits,

Nearest to would: goslin, inteligencia, lagadha, gy, pcs, gladius, generalizations, govenor, proud, communion, driscoll, eleusinian, diehard, dielectrics, bosniaks, leeds, diomede, reimagined, tami, gilbertese,

configuraion 14:

```
#################################################################################
#######
 learning rate 0.5

  # Hyper Parameters to config
  batch_size = 128
  embedding_size = 128  # Dimension of the embedding vector.
  skip_window = 4       # How many words to consider left and right.
  num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.190918545416

Nearest to first: the, one, was, seven, six, nine, of, three, to, with, four, UNK, not, in, zero, is, on, a, two, for,
Nearest to american: was, the, eight, two, s, zero, UNK, in, four, and, three, of, a, they, one, international, nine, for, his, use,
Nearest to would: he, seven, it, for, were, s, which, two, they, and, at, can, three, one, five, this, zero, eight, in, was,

CE loss - Average loss at step 200000 : 4.82938805981

Nearest to first: taxanes, ovum, philbin, virgo, waxed, candaules, sentences, sebasti, devonport, malipiero, luckiest, tawantinsuyu, anguilla, hazrat, euouae, manco, putnam, tanya, tutsi, antheil,
Nearest to american: sundance, spoofed, foramen, eilean, deaf, deprecation, compounds, ketoacidosis, pakenham, moniker, granules, listname, komet, interconnections, zeist, whispered, vorder, molluscs, nuages, pilgrimage,
Nearest to would: drake, reform, vital, seniors, lipophilic, swellings, lion, lester, schl, lland, trimming, calvins, bornholm, azerbaijanis, eqworld, hewitt, mistresses, slav, deen, mankiewicz,

configuraion 15:

```
#################################################################################
#######
 learning rate 0.8,  pretrained checkpoint,

  # Hyper Parameters to config
  batch_size = 128
  embedding_size = 128  # Dimension of the embedding vector.
```

```
skip_window = 4       # How many words to consider left and right.
num_skips = 8         # How many times to reuse an input to generate a label.
```

NCE loss - Average loss at step 200000 : 0.0706022326474

Nearest to first: during, he, at, three, one, seven, zero, was, time, UNK, were, use, an, on, eight, since, had, his, nine, before,
Nearest to american: of, british, mollusca, the, houari, skarsg, nisbet, vivipara, km, polanyi, five, redondo, rudis, lacombe, overshoes, between, systemically, micha, german, louse,
Nearest to would: not, i, him, eight, four, only, it, so, this, such, for, some, will, their, s, six, two, called, more, many,

CE loss - Average loss at step 200000 : 4.82120571852

Nearest to first: last, most, same, original, continued, following, latter, end, best, a, main, largest, due, until, because, coast, before, much, ismailis, left,
Nearest to american: german, british, its, his, english, french, russian, european, canadian, italian, trade, international, transpiration, reject, offense, carmen, projectiles, ww, extension, united,
Nearest to would: will, could, might, must, did, does, may, can, should, was, is, began, do, seems, had, we, said, appears, came, wanted,

2. Best Configuration and accuracy with explanation:

From the given experiments details given above the best configuration for which, I have achieved the best accuracy in the DEV data(word_analogy_dev.txt) is :

```
################################################################################
#######
 learning rate 1, pretrained checkpoint,

 # Hyper Parameters to config
 batch_size = 128
 embedding_size = 128  # Dimension of the embedding vector.
 skip_window = 4       # How many words to consider left and right.
 num_skips = 8         # 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.random.choice(valid_window, valid_size, replace=False)
```

```
  valid_examples = np.append(valid_examples, [dictionary["first"], dictionary["american"],
dictionary["would"]])
  num_sampled = 64    # Number of negative examples to sample.

  # summary_path = './summary_%s'%(loss_model)
  pretrained_model_path = './pretrained/'

  checkpoint_model_path = './checkpoints_%s/'%(loss_model)
  model_path = './models'


  # maximum training step
  max_num_steps  = 200001
  checkpoint_step = 50000
```

Accuracy for Cross-entropy loss:

```
Generated by:                          score_maxdiff.pl
Mechanical Turk File:                  word_analogy_dev_mturk_answers.txt
Test File:                             predictionOutput_cross_entropy.txt
Number of MaxDiff Questions:           914
Number of Least Illustrative Guessed Correctly:   349
Number of Least Illustrative Guessed Incorrectly: 565
Accuracy of Least Illustrative Guesses:    38.2%
Number of Most Illustrative Guessed Correctly:    307
Number of Most Illustrative Guessed Incorrectly:  607
Accuracy of Most Illustrative Guesses:     33.6%
Overall Accuracy:                      35.9%
```

Accuracy for NCE loss:

```
Generated by:                          score_maxdiff.pl
Mechanical Turk File:                  word_analogy_dev_mturk_answers.txt
Test File:                             predictionOutput_nce.txt
Number of MaxDiff Questions:           914
Number of Least Illustrative Guessed Correctly:   351
Number of Least Illustrative Guessed Incorrectly: 563
Accuracy of Least Illustrative Guesses:    38.4%
Number of Most Illustrative Guessed Correctly:    287
Number of Most Illustrative Guessed Incorrectly:  627
Accuracy of Most Illustrative Guesses:     31.4%
Overall Accuracy:                      34.9%
```

Explanation with experiments:

For the first 1 - 5 experiments, I have increased the context vector dimension from 128 –> 256 –> 512 and I have observed that:

- Model takes more time to train
- Loss is decreasing significantly for both cross-entropy and NCE
- The similar words are becoming more noisy

For the 6 – 10 experiments I have mostly changed negative samples and skip-window size by fixing the context vector dimension:

- There is a decrease observed in the loss function values
- Similar words are less noisy
- For a big negative sample value(128) the loss became 'nan'

For the experiment 11 – 15, I have varied the learning rate and epoch size and nothing much change observed compared the previous experiment.

3. Top 20 similar words according to your NCE and cross entropy model for the words (first, american, would)

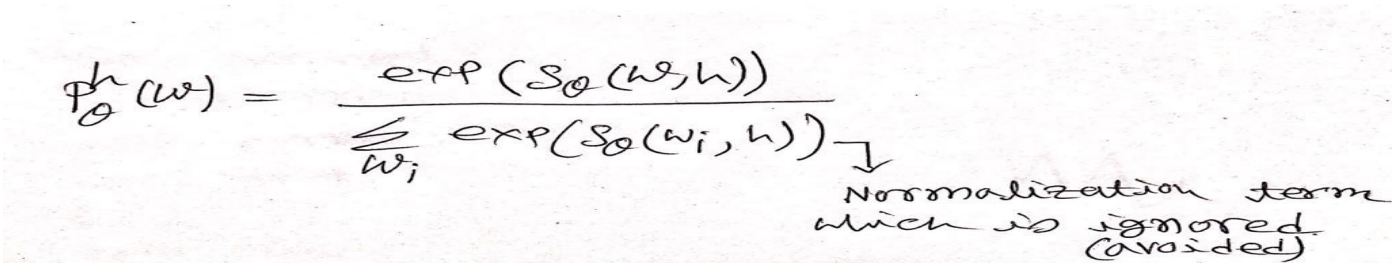
Cross-entropy loss	Top 20 similar words
Nearest to first	last, most, same, original, continued, following, latter, end, a, main, best, largest, due, because, left, before, coast, much, until, west,
Nearest to american	german, british, its, his, english, russian, french, european, canadian, italian, trade, international, transpiration, united, extension, reject, outputs, carmen, their, projectiles,
Nearest to would	will, could, might, must, did, may, can, was, does, should, is, began, had, do, seems, came, appears, we, t, are,

NCE Loss	Top 20 similar words
Nearest to first	he, during, three, at, one, seven, zero, time, UNK, was, were, an, had, his, on, since, nine, eight, with, use,
Nearest to american	british, mollusca, nisbet, vivipara, skarsg, polanyi, houari, lacombe, lawmaker, louse, overshoes, km, rudis, dealbata, five, german, monas, ugliest, other, oresund,
Nearest to would	not, i, him, eight, four, so, only, it, such, this, for, six, their, two, some, s, many, called, her, more,

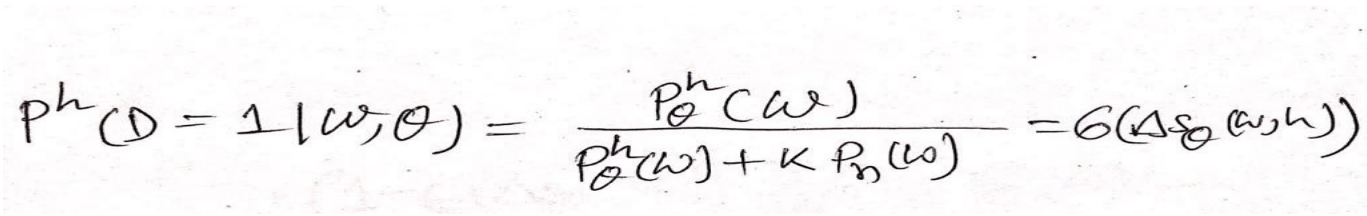
4. summary of the justification behind the NCE method loss

Noise-contrastive estimation (NCE) works on the mechanism that it tries to reduce the density estimation to probabilistic binary classifier. So, we try this method to fit our un-normalized models. Using the probability ratio of sample to the noise distribution, this NCE method train a logistic regression classifier which distinguish between samples from a data distribution and samples from a noise distribution. The objective of NCE is to take a model with approximately normalized and recovers a perfectly normalized model if the model class contains data distribution.

One noticeable advantage is that this method is that this makes the training time independent of the underlying vocabulary size by avoiding the normalized factor in the below equation:

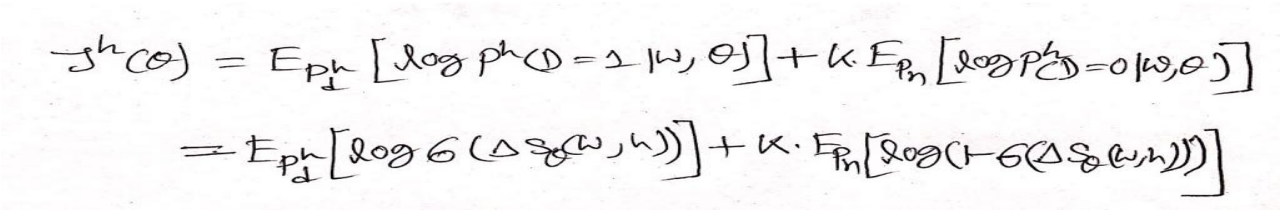


Our goal is to learn the distribution of words for some specific context word. To get that, an auxiliary binary classification needs to be created with negative samples from noise distribution and input training data as positive samples. While selecting the noise distribution we need to make sure that it does not assign any zero probabilities to any word while computing probabilities. To make sure to avoid the zero probabilities for noise distribution we will use well known unigram distribution of the input data as the noise distribution. We will estimate these probabilities using the below equation:

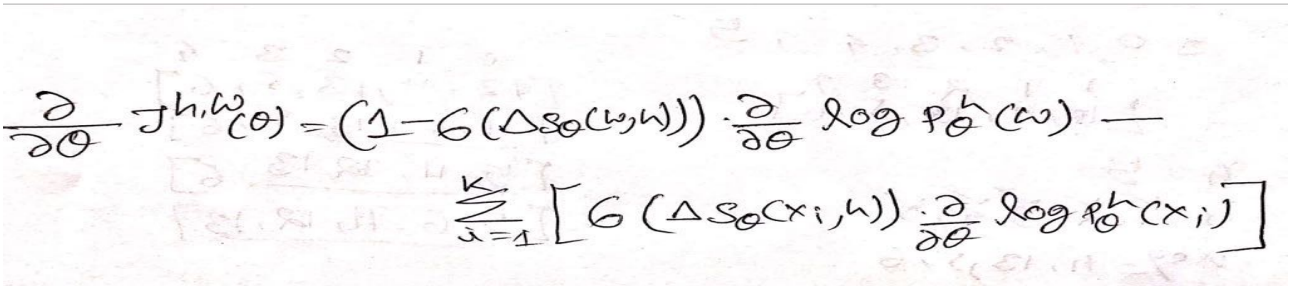


Where, $P^h(w)$ = probability of some context word h,
Pn(w) = noise distribution from negative samples to choose from,
D = correct labels,
σ(x) is the logistic function
k = scaling factor and
$\Delta s\theta(w, h) = s\theta(w, h) - \log(kPn(w))$ = difference in the scores of word w under the model and the (scaled) noise distribution

Then our aim is to fit the model and so we need to maximize the log-posterior probability over correct labels D and averaged over provided data and noise samples. Below equation expresses this:



Finally, we will be using gradient of the achieved expectation over the noise distribution my means of sampling. Below equation shows the same over k noise samples {xi}:



$$\frac{\partial}{\partial\theta} J^{h,w}(\theta) = \left(1 - \sigma\left(\Delta s_\theta(w,h)\right)\right) \cdot \frac{\partial}{\partial\theta} \log p_\theta^h(w) - \sum_{i=1}^{k} \left[\sigma\left(\Delta s_\theta(x_i,h)\right) \frac{\partial}{\partial\theta} \log p_\theta^h(x_i)\right]$$

We use sum over k noise samples instead of sum over entire vocabulary which makes the training time of NCE linear in number of noise samples and independent of vocabulary size.