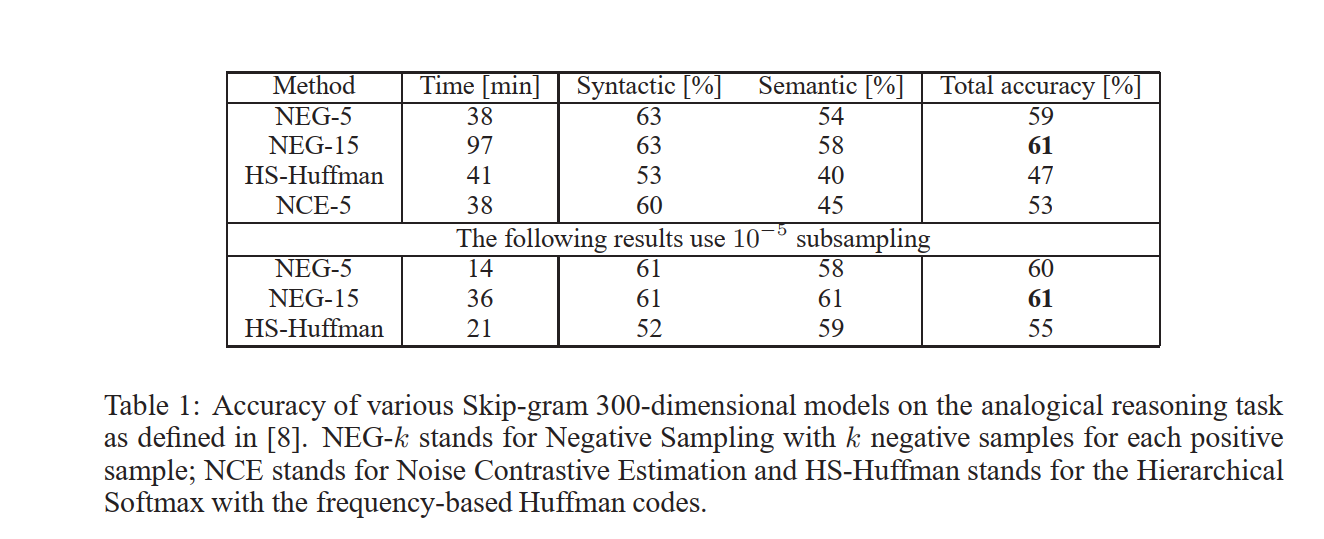
**Assignment 5: Word2Vec with Hierarchical Softmax, Negative Sampling and Subsampling**

Contribution:

Member 1: Implementation of HS, NS and Subsampling, Running Tests, Filling in word2vec.py, etc.  
Member 2: Implementation of NS and Subsampling, Running Tests, Exploratory Data Analysis, Current Report, etc.

Part 1: Training Results

With reference to Mikolov et al., The Skip-gram model with Negative Sampling (with 15 negative samples) had the best performance (cf. Table 1, taken from the original paper), along with a significant boost in training speed with subsampling (t=0.00001). 

Our observations were generally in line with the results above (cf. Excel file for additional details on the effects of HS, NS and Subsampling).

However, even after 400 thousand iterations spanning across over 8 hours, the model failed to “learn” all 72 thousand words in the vocabulary. In fact, with a similarity threshold of 0.3, approximately 1 in 10 words appeared to have similar words (cf. word embeddings).

Various types of relationships (i.e. both syntactic and semantic) were captured by the model.

102 [('english', 1.0), ('remus', 0.3043549656867981), ('b', 0.2582004964351654), ('uninspired', 0.24655044078826904), ('sverdrup', 0.24431198835372925)]

115 [('m', 1.0), ('airports', 0.3257398009300232), ('runways', 0.32356324791908264), ('unpaved', 0.2820527255535126), ('constancy', 0.2513360381126404)]

125 [('x', 1.0), ('phi', 0.3828175663948059), ('operatorname', 0.3536424934864044), ('f', 0.3469811975955963), ('inv', 0.32366570830345154)]

130 [('year', 1.0), ('gregorian', 0.30452853441238403), ('jak', 0.2613445222377777), ('photoshop', 0.24458502233028412), ('utilitarian', 0.23544666171073914)]

136 [('day', 1.0), ('days', 0.3041568100452423), ('february', 0.2847047448158264), ('june', 0.2807992994785309), ('gregorian', 0.26320794224739075)]

146 [('national', 1.0), ('khama', 0.3028440773487091), ('accademia', 0.2741481363773346), ('assembly', 0.24432127177715302), ('inspecting', 0.23662158846855164)]

156 [('n', 1.0), ('cdots', 0.32165664434432983), ('ln', 0.3131163418292999), ('f', 0.3014013171195984), ('x', 0.28816506266593933)]

169 [('f', 1.0), ('x', 0.3469811975955963), ('real', 0.32879266142845154), ('n', 0.3014013171195984), ('y', 0.29584723711013794)]

i: 200

217 [('population', 1.0), ('median', 0.3135295808315277), ('total', 0.2502443194389343), ('aceh', 0.2473580688238144), ('rate', 0.24657630920410156)]

244 [('right', 1.0), ('ddot', 0.30629605054855347), ('crc', 0.24723438918590546), ('buttressed', 0.24465663731098175), ('cellos', 0.24189850687980652)]

253 [('times', 1.0), ('nabla', 0.3079168498516083), ('cdot', 0.24228878319263458), ('ddot', 0.23854069411754608), ('indonesia', 0.2359069138765335)]

269 [('countries', 1.0), ('europe', 0.3111449182033539), ('aldershot', 0.23685471713542938), ('past', 0.22902019321918488), ('sister', 0.2246236652135849)]

271 [('isbn', 1.0), ('hardcover', 0.3349415361881256), ('press', 0.3086699843406677), ('york', 0.29377517104148865), ('reprint', 0.292360782623291)]

284 [('europe', 1.0), ('countries', 0.3111449182033539), ('semicircular', 0.2728331387042999), ('italy', 0.25698578357696533), ('nations', 0.25296908617019653)]

286 [('central', 1.0), ('playoffs', 0.31868743896484375), ('dominican', 0.2581820785999298), ('honeybees', 0.2525707185268402), ('township', 0.24741430580615997)]

Figure: Excerpt of word embeddings with at least 1 element greater than 0.3

For example, “day” and “days” are considered similar by the model (syntactic relationship). In addition, the rest of the words are also closely related to the word “day”. In fact, even contractions of months show a clear relationship, with closer months demonstrating a closer relationship.

8083 [('oct', 1.0), ('sep', 0.5100699663162231), ('nov', 0.40637466311454773), ('dec', 0.28848734498023987), ('palladium', 0.28481239080429077)]

11362 [('jun', 1.0), ('jul', 0.5365533232688904), ('unconquered', 0.5020674467086792), ('extraneous', 0.2854079306125641), ('apr', 0.2825978696346283)]

Figure: ‘oct’, ‘sep’, ‘nov’, ‘dec’ are considered similar by the model; so are ‘jun’, ‘jul’ and ‘apr’

Mathematical terms possess a particularly close relationship with other mathematical terms, owing to their domain-specificity (semantic relationships):

156 [('n', 1.0), ('cdots', 0.32165664434432983), ('ln', 0.3131163418292999), ('f', 0.3014013171195984), ('x', 0.28816506266593933)]

169 [('f', 1.0), ('x', 0.3469811975955963), ('real', 0.32879266142845154), ('n', 0.3014013171195984), ('y', 0.29584723711013794)]

11363 [('manifolds', 1.0), ('homeomorphic', 0.458295613527298), ('torus', 0.4130508303642273), ('manifold', 0.33729088306427), ('finsler', 0.32902905344963074)]

Some words show strong similarity with their syntactic alterations:

11364 [('interpolation', 1.0), ('interpolating', 0.34822654724121094), ('interpolant', 0.27820655703544617), ('bilinear', 0.25005751848220825), ('kink', 0.24544501304626465)]

Alternatively, they share many similarities with their semantic neighbours:

1219 [('remaining', 1.0), ('gregorian', 0.36349162459373474), ('leap', 0.30902472138404846), ('days', 0.3054875433444977), ('calendars', 0.2817561626434326)]

1220 [('apple', 1.0), ('toolbox', 0.3096642792224884), ('gui', 0.2851232886314392), ('interface', 0.2770882546901703), ('intel', 0.26354286074638367)]

1225 [('am', 1.0), ('shortwave', 0.45776307582855225), ('radios', 0.4251428246498108), ('stations', 0.32980877161026), ('fm', 0.32301852107048035)]

1247 [('greece', 1.0), ('macedonia', 0.309558629989624), ('paros', 0.2782817482948303), ('serbia', 0.253755122423172), ('mycenaean', 0.245902419090271)]

At times, they show similarities for both:

536 [('radio', 1.0), ('shortwave', 0.4824630618095398), ('fm', 0.45974451303482056), ('radios', 0.3516915738582611), ('stations', 0.3356715738773346)]

Many such relationships tend to be “associative” or “transitive” (i.e. two or more words which are similar to a certain word also tend to be similar to each other). Collocations such as “tortured soul”, “rome (and) juliet”, “whipped cream” also show a strong “similarity” with each other (i.e. the model learns phrases without additional supervision).

Such lists are hardly exhaustive; refer to the similarities calculated for a general overview on the various “word neighbours”.

Part 2: Training the Model

The word embeddings above were produced through a training regime lasting 400 thousand iterations.

The learning rate and subsampling threshold were changed a few times throughout the training process.

In particular, exploratory data/corpus analysis shows the effects of t=1e-4 and t=1e-5 as compared to no subsampling at all.

(i.e.

"common" words:

work: 139

great: 174

works: 311

greater: 750

read: 962

(Words above this point are subsampled aggressively.)

brother: 1030

reading: 1192

possibly: 1396

think: 1810

sister: 1932

easy: 1975

speak: 1978

"average" words:

apparently: 2045

rapidly: 2339

apparent: 2487

rapid: 2492

dollar: 2730

mouse: 2799

thinking: 2853

ethical: 3374

dollars: 3642

mice: 5319

walking: 5531

grandson: 5606

speaks: 5661

swimming: 7198

tough: 8134

walked: 9529

lucky: 9782

“rare” words:

granddaughter: 17147

easiest: 17665

tougher: 22856

unethical: 23101

swam: 37138

impossibly: 55729

luckiest: 70209)

The learning rate of 1/0.5 (mixed usage) was selected after comparing the training speeds and “accuracies” on preliminary trials.

The threshold of 0.0001 was selected as thresholds of 0.001 and lower do not have as significant an effect (cf. image reflecting the effects of the subsampling threshold)

(e.g. top 10K comprises approx. 80% instead of almost 90+% with 1e-4)

(t=1e-5 reduces the proportion of "more common" words from >80% to <70%)

(cf. subsampling.txt for the effect of threshold on the subsampling result.)

(also cf. multiple.png for a more general visualisation)

Note: t=1e-5 actually increases the proportion of words of rank 2K to 10K, where most of the question words are

(from <20% without subsampling to >50% with subsampling)

(i.e. t=1e-5 is actually better than no subsampling or t=1e-4 for training question words with "average" frequency)

(for training words of rank <10K, t=1e-4 seems to be a good balance; 52% for top 2K and 83% for top 10k)

|  |  |  |
| --- | --- | --- |
| no subsampling:  length of corpus: 16718844  top 2K percentage: 75.829 %  top 10K percentage: 91.322 % | t=1e-4:  length of corpus: 8430060  top 2K percentage: 52.063 %  top 10K percentage: 82.790 % | t=1e-5:  length of corpus: 4671420  top 2K percentage: 31.031 %  top 10K percentage: 68.943 % |

Even with a similarity threshold of 0.3, the model has barely learnt any of the question words, except perhaps "dollar", after at most 300K iterations (approx. 6h).

However, the model has learnt many other useful associations, some of which allow for analogical reasoning tasks (e.g. the “franc” vs “midwife” example involving syntactic reasoning in the source code).

Refer to various additional files for further visualisation/analysis.