**Assignment 2 COMPSCI 423 Report :**

**Vedant Tewari**

*For this Assignment we use Gensim is a library for topic modelling which includes utilities related to*

*word embeddings.*

*We make use of 2 pre-trained models (word2vec and GloVe) to get measurements of similarity and*

*evaluation of word pairs and word vectors.*

***Task 1:***

For word2vec:

>>> wv.evaluate\_word\_pairs("SimLexA2.txt")

((0.4471705560019418, 3.1602700639092923e-50), SpearmanrResult(correlation=0.43607859778335434, pvalue=1.3896416353352258e-47), 0.10010010010010009)

(The Pearson's Correlation coefficient for vw = 0.4471705560019418)

For GloVe:

>>> glove.evaluate\_word\_pairs("SimLexA2.txt")

((0.387782792068166, 3.6876452795547013e-37), SpearmanrResult(correlation=0.3692121013345344,

pvalue=1.371018956747063e-33), 0.10010010010010009)

(The Pearson's Correlation coefficient for glove = 0.387782792068166)

***Task 2:***

Part 1:

***TABLE # 1***

| WORD PAIR | ASSIGNED SIMILARITY SCORES BETWEEN (0 - 1) |
| --- | --- |
| {dog, cat} | 0.67 |
| {football, soccer} | 0.74 |
| {car, bike} | 0.67 |
| {internet, cable} | 0.53 |
| {key, lock} | 0.50 |
| {table, top} | 0.30 |
| {toilet, bathroom} | 0.81 |
| {latitude, longitude} | 0.65 |
| {shoe, lace} | 0.49 |
| {lamp, bulb} | 0.59 |
| {picture, frame} | 0.45 |
| {outlet, charger} | 0.33 |

***TABLE #2***

For word2vec:

| WORD PAIR | EVALUATED SIMILARITY SCORES BETWEEN (0 - 1) |
| --- | --- |
| {dog, cat} | 0.76094574 |
| {football, soccer} | 0.73135483 |
| {car, bike} | 0.5854154 |
| {internet, cable} | 0.30422905 |
| {key, lock} | 0.21806347 |
| {table, top} | 0.2694193 |
| {toilet, bathroom} | 0.68147415 |
| {latitude, longitude} | 0.2722402 |
| {shoe, lace} | 0.40895918 |
| {lamp, bulb} | 0.5665534 |
| {picture, frame} | 0.3207433 |
| {outlet, charger} | 0.24482216 |

***TABLE #3***

For GloVe:

| WORD PAIR | EVALUATED SIMILARITY SCORES BETWEEN (0 - 1) |
| --- | --- |
| {dog, cat} | 0.68167466 |
| {football, soccer} | 0.76825917 |
| {car, bike} | 0.46721223 |
| {internet, cable} | 0.5288797 |
| {key, lock} | 0.25479427 |
| {table, top} | 0.39873493 |
| {toilet, bathroom} | 0.68817955 |
| {latitude, longitude} | 0.8111733 |
| {shoe, lace} | 0.32098332 |
| {lamp, bulb} | 0.56800234 |
| {picture, frame} | 0.3207433 |
| {outlet, charger} | 0.11521664 |

Part 2:

Accuracies and Pearson correlation coefficients obtained using each of the two embeddings:

*Part 2:*

*word2vec:*

*>>> wv.evaluate\_word\_pairs("qs2.txt")*

*((0.29803049681354815, 0.4029443853484936), SpearmanrResult(correlation=0.406060606060606, pvalue=0.24428229408662638), 0.0)*

(The Pearson's Correlation coefficient for vw based on ‘qs2.txt’ dataset = *0.29803049681354815*)

*GloVe:*

*>>> glove.evaluate\_word\_pairs("qs2.txt")*

*((0.48456903356157566, 0.15579466008760465), SpearmanrResult(correlation=0.5757575757575757, pvalue=0.08155281477260236), 0.0)*

(The Pearson's Correlation coefficient for vw based on ‘qs2.txt’ dataset = *0.48456903356157566*)

*Comments and Observations based on results:*

*The Similarity scores provided by me as a rough estimate make sure I have a good range of scores, in comparison to my similarity scores with the scores obtained by running the dataset using the predefined-vector models, I seem to get a much better estimate to the similarities obtained using the GloVe Model, rather than the word2vec results. One possible reason could be due to the human error in assigning scores, and the context of being parsed in.*

*The correlation value closer to 1 would result in a perfect correlation, and according to the results and for values between 0.5 and 1 result in strong correlation, hence GloVe Model provides a much better estimate to similarity and better correlation for data.*

***Task 3:***

*Accuracies obtained by the 2 embeddings:*

*For word2vec:*

*>>> a = wv.evaluate\_word\_analogies("qs3.txt")*

*>>> a[0]*

*1.0 (Accuracy)*

*>>> a[1]*

*[{'section': 'basketball', 'correct': [('DETROIT', 'DETROIT\_PISTONS', 'HOUSTON', 'HOUSTON\_ROCKETS'), ('HOUSTON', 'HOUSTON\_ROCKETS', 'MILWAUKEE', 'MILWAUKEE\_BUCKS'), ('CHARLOTTE', 'CHARLOTTE\_BOBCATS', 'SACRAMENTO', 'SACRAMENTO\_KINGS'), ('CHICAGO', 'CHICAGO\_BULLS', 'PHOENIX', 'PHOENIX\_SUNS'), ('INDIANA', 'INDIANA\_PACERS', 'DALLAS', 'DALLAS\_MAVERICKS'), ('MEMPHIS', 'MEMPHIS\_GRIZZLIES', 'PORTLAND', 'PORTLAND\_TRAIL\_BLAZERS'), ('PHOENIX', 'PHOENIX\_SUNS', 'TORONTO', 'TORONTO\_RAPTORS')], 'incorrect': []}, {'section': 'Total accuracy', 'correct': [('DETROIT', 'DETROIT\_PISTONS', 'HOUSTON', 'HOUSTON\_ROCKETS'), ('HOUSTON', 'HOUSTON\_ROCKETS', 'MILWAUKEE', 'MILWAUKEE\_BUCKS'), ('CHARLOTTE', 'CHARLOTTE\_BOBCATS', 'SACRAMENTO', 'SACRAMENTO\_KINGS'), ('CHICAGO', 'CHICAGO\_BULLS', 'PHOENIX', 'PHOENIX\_SUNS'), ('INDIANA', 'INDIANA\_PACERS', 'DALLAS', 'DALLAS\_MAVERICKS'), ('MEMPHIS', 'MEMPHIS\_GRIZZLIES', 'PORTLAND', 'PORTLAND\_TRAIL\_BLAZERS'), ('PHOENIX', 'PHOENIX\_SUNS', 'TORONTO', 'TORONTO\_RAPTORS')], 'incorrect': []}]*

*For GloVe:*

*>>> b = glove.evaluate\_word\_analogies("qs3.txt")*

*>>> b[0]*

*0.0 (Accuracy)*

*>>> b[1]*

*[{'section': 'basketball', 'correct': [], 'incorrect': []}, {'section': 'Total accuracy', 'correct': [], 'incorrect': []}]*

*Comments and Observations based on obtained results:*

*Since the dataset used in the task was most probably part of the pre-defined \word2vec and hence it was able to recognize and provide a high accuracy of results on all the correct analogies in the given dataset, for the GloVe Model the explanation for a zero accuracy on a given dataset analyzed through the model which was neither trained on the dataset or has seen the following analogies.*

***Task 4:***

***For word2vec:***

| ***Word Chosen*** | ***Ranked by Similarity Score (in Descending Order)***  ***(‘word’, similarity score)*** |
| --- | --- |
| ***raccoon*** | ***('raccoons', 0.7092068791389465), ('rabid\_raccoon', 0.671484649181366), ('bobcat', 0.6711269617080688), ('squirrel', 0.6657680869102478), ('coyote', 0.6650583744049072)*** |
| ***soldier*** | ***('solider', 0.9117935299873352), ('serviceman', 0.7837359309196472), ('soldiers', 0.7634838223457336), ('airman', 0.6886240839958191), ('guardsman', 0.6794973015785217)*** |
| ***guardian*** | ***('guardians', 0.7329328656196594), ('guardianship', 0.5965357422828674), ('played\_Holger\_Palmgren', 0.5946720242500305), ('Leslie\_Andino', 0.5900617837905884), ('ad\_Litem', 0.5346183776855469)*** |
| ***angel*** | ***('angels', 0.7340543270111084), ('luminous\_cocoon', 0.554248571395874), ('guardian\_angel', 0.5442337989807129), ('Seliethia\_Parker', 0.5338423848152161), ('Clarence\_Odbody', 0.5276748538017273)*** |
| ***grass*** | ***('grasses', 0.6635476350784302), ('Bermuda\_grass', 0.6476073265075684), ('bermuda\_grass', 0.6298290491104126), ('rye\_grass', 0.6293388605117798), ('lawns', 0.6126667261123657)*** |
| ***laptop*** | ***('laptops', 0.805374026298523), ('laptop\_computer', 0.7848465442657471), ('notebook', 0.67857825756073), ('netbook', 0.6707929372787476), ('computer', 0.6640493273735046)*** |
| ***lamp*** | ***('lamps', 0.743071436882019), ('Rubbery\_pizza\_languishing', 0.5924029350280762), ('tealight', 0.5723157525062561), ('candle', 0.5694510340690613), ('lantern', 0.5688028931617737)*** |
| ***globe*** | ***('world', 0.6945998072624207), ('worldwide', 0.647681474685669), ('continents', 0.6263391971588135), ('continent', 0.6154399514198303), ('globally', 0.5908562541007996)*** |
| ***bird*** | ***('birds', 0.8141971230506897), ('raptor', 0.6830927729606628), ('owl', 0.6825829148292542), ('squirrel', 0.6653631329536438), ('falcon', 0.6649249196052551)*** |
| ***bee*** | ***('bees', 0.7053181529045105), ('honeybee', 0.6075024604797363), ('spelling\_bee', 0.5671892166137695), ('honey\_bee', 0.5634711384773254), ('honey\_bees', 0.5585135221481323)*** |

***For GloVe:***

| ***Word Chosen*** | ***Ranked by Similarity Score (in Descending Order)*** |
| --- | --- |
| ***raccoon*** | ***('raccoons', 0.5740074515342712), ('squirrel', 0.5537731647491455), ('coyote', 0.5170983672142029), ('boar', 0.5077536702156067), ('mink', 0.5049654245376587)*** |
| ***soldier*** | ***('soldiers', 0.7162267565727234), ('wounded', 0.6503106355667114), ('policeman', 0.6371234655380249), ('army', 0.5879426598548889), ('killed', 0.5516754984855652)*** |
| ***guardian*** | ***('guardians', 0.4964340627193451), ('newspaper', 0.47482213377952576), ('reviewer', 0.42936545610427856), ('herald', 0.42828378081321716), ('editorial', 0.4262792766094208)*** |
| ***angel*** | ***('miguel', 0.5479671359062195), ('jimenez', 0.5176248550415039), ('gabriel', 0.47134312987327576), ('jose', 0.47126585245132446), ('lopez', 0.470060795545578)*** |
| ***grass*** | ***('grasses', 0.5605592727661133), ('lawn', 0.5430611968040466), ('pasture', 0.4857950210571289), ('lawns', 0.4798297584056854)*** |
| ***laptop*** | ***('laptops', 0.7956499457359314), ('computers', 0.6733037233352661), ('phones', 0.599344789981842), ('computer', 0.5955509543418884), ('portable', 0.5796298384666443)*** |
| ***lamp*** | ***('lamps', 0.8024932146072388), ('bulb', 0.5680024027824402), ('candle', 0.5575706958770752), ('incandescent', 0.5573796629905701), ('fluorescent', 0.5541897416114807)*** |
| ***globe*** | ***('columnist', 0.4608248472213745), ('times', 0.42327189445495605), ('boston', 0.40685129165649414), ('cox', 0.39769992232322693), ('mail', 0.39709267020225525)*** |
| ***bird*** | ***('birds', 0.7303513288497925), ('flu', 0.7103857398033142), ('avian', 0.6787645220756531), ('h5n1', 0.6514254212379456), ('influenza', 0.58339524269104)*** |
| ***bee*** | ***('bees', 0.5616937279701233), ('gees', 0.5614164471626282), ('gee', 0.46755173802375793), ('honey', 0.45776888728141785), ('hive', 0.4324120879173279)*** |

*Comments and Observations based on obtained results:*

*In most cases the most similar words for almost all 10 words seem to be expected results, although there are a few observed inconsistencies in the results outputted by both the models, for word2vec, the word lamp appears to have a close similarity score the word* ***‘Rubbery\_pizza\_languishing',*** *which could be a result of contextual misinterpretation while parsing the data. Another interesting observation can be made for the word angel, in which case word2vec vector model does a much better job producing close to reliable results whereas GloVe interprets the word ‘angel’ corresponding to names of people; similarly in the last case as already commented upon GloVe does a better at predicting the context of the word. The similarity values for GloVe and word2vec are different for a lot of words where one has a closer score to the one, the other does not. Overall, both models make predictions context of the given word to a reliable and accurate degree.*