CS2611 Lab1 and Basic Assignment Report
Mohammad Mahdi Abdollah Pour - student number: 1006888628
Feb 2022
Code is at https://github.com/mahdiabdollahpour/CSC2611

Basic Assignment
Step 2.

5 most and least common words :

```
Most: [('the', 7258), ('i', 5161), ('one', 3292), ('he', 2982), ('would', 2714)]
Least: [('daytime', 19), ('amen', 19), ('dim', 19), ('distances', 19), ('puzzled', 19)]
```

Step 8. Pearson correlation

```
pearsonr for S and M1 (0.31734305508228516, 0.010000522446091661)
pearsonr for S and M1+ (0.19293985349723838, 0.12359532984089408)
pearsonr for S and M2_10 (0.16557610543548068, 0.1874585436217429)
pearsonr for S and M2_100 (0.3939898505682903, 0.0011653090899399875)
pearsonr for S and M2_300 (0.3842580246736232, 0.0015768604988334147)
```

Lab 1

Synchronic word embedding

Step 3.

Pearson correlation

```
pearsonr for S and W2V (0.7720616125197682, 5.091065805872837e-14)
```

Comparison

The score is more than all previous models, because W2V is a stronger embedding model possibly due to larger context window

Step 4.

The Analogy Test

```
type : semantic --- family ,w2v --> 0.9111111111111111 ,M2_300 --> 0.14444444444444443
type : syntactic --- gram1-adjective-to-adverb ,w2v --> 0.38421052631578945 ,M2_300 --> 0.002631578947368421
type : syntactic --- gram2-opposite ,w2v --> 0.35 ,M2_300 --> 0.0
type : syntactic --- gram3-comparative ,w2v --> 0.8088235294117647 ,M2_300 --> 0.08088235294117647
type : syntactic --- gram4-superlative ,w2v --> 0.9047619047619048 ,M2_300 --> 0.09523809523809523
type : syntactic --- gram5-present-participle ,w2v --> 0.8088235294117647 ,M2_300 --> 0.06985294117647059
type : syntactic --- gram7-past-tense ,w2v --> 0.7583333333333333 ,M2_300 --> 0.17833333333333334
type : syntactic --- gram8-plural ,w2v --> 0.8921568627450981 ,M2_300 --> 0.016339869281045753
type : syntactic --- gram9-plural-verbs ,w2v --> 0.8409090909090909 ,M2_300 --> 0.11363636363636363
---------------
W2V overall syntactic 0.7262845849802372
W2V overall semantic 0.9111111111111111
---------------
M2_300 overall syntactic 0.08547430830039526
M2_300 overall semantic 0.14444444444444443
```

W2V does much better in the analogy test again because it is a stronger embedding model. A larger context window has been used to train W2V but for M2_300 it is based on bigram counts that has context window of one. Beside window size, W2V model is stronger than counts, it considers words as vectors in space but M2_300 first extracts the counts and then builds the vectors in space through PCA.

Step 5.

Suggest a way to improve

Adding a loss function in training the model such that this loss function contrasts the vector V

V = model[words[2]] - model[words[0]] + model[words[1]]

with the words[3] (and random words for the negative pairs of contrastion)

It means it should maximize cosine_sim(V,model[3]) and minimize cosine_sim(V,model[random_word]) in this new loss function. Wo we have

Loss_new =

Loss_prev + cosine_sim(V,model[3]) - Sigma (over k random words)[cosine_sim(V,model[random_word])]

Diachronic word embedding

Step 2.

Method1: Cosine distance between first and last decade for each word

most_changes ['film', 'shift', 'berkeley', 'patterns', 'perspective', 'impact', 'media', 'shri', 'van', 'approach', 'goals', 'sector', 'radio', 'computer', 'objectives', 'programs', 'techniques', 'ml', 'skills', 'mcgraw']
least_changes ['april', 'june', 'november', 'february', 'years', 'october', 'increase', 'january', 'century', 'months', 'daughter', 'december', 'god', 'september', 'feet', 'week', 'evening', 'door', 'payment', 'miles']

Method2 :Average cosine distance of consecetive decades

most_changes ['haven', 'goals', 'johnson', 'therapy', 'adams', 'wilson', 'princeton', 'martin', 'baltimore', 'wiley', 'berkeley', 'techniques', 'sector', 'ml', 'jones', 'harper', 'mcgraw', 'skills', 'computer', 'shri']
least_changes ['april', 'miles', 'november', 'september', 'january', 'december', 'february', 'university', 'vessels', 'trees', 'cent', 'solution', 'july', 'decrease', 'october', 'temperature', 'buildings', 'june', 'patients', 'blood']

Method3 : Maximum distance of consecetive decades

most_changes ['jones', 'radio', 'implications', 'variables', 'jobs', 'procedures', 'wiley', 'therapy', 'input', 'evaluation', 'programs', 'sector', 'objectives', 'goals', 'skills', 'shri', 'mcgraw', 'ml', 'computer', 'techniques']
least_changes ['april', 'november', 'december', 'january', 'september', 'trees', 'miles', 'solution', 'feet', 'june', 'february', 'vessels', 'century', 'duties', 'cent', 'blood', 'evening', 'buildings', 'decrease', 'july']

Pearson correlations of methods

```
[[0.         0.6932991  0.68926445]
 [0.6932991  0.         0.84506762]
 [0.68926445 0.84506762 0.        ]]
```

Method two and method three match better.

Step 3

Evaluation Method: We do not have labels to see if a word had semantic change or not, so I rely on context to check how different set of K=10 closest words have been for each word in each decade. If set of closest words change so mush the word had semantic change. To be specific, semantic change is the number of common closest words between decade a and decade b divided by K. (a score between 0 to 1). I compute the correlation of this evaluation type with three methods before.

```
[[0.         0.6932991  0.68926445 0.43578751]
 [0.6932991  0.         0.84506762 0.21989351]
 [0.68926445 0.84506762 0.         0.23627123]
 [0.43578751 0.21989351 0.23627123 0.        ]]
```

Evaluation matches better with method one.

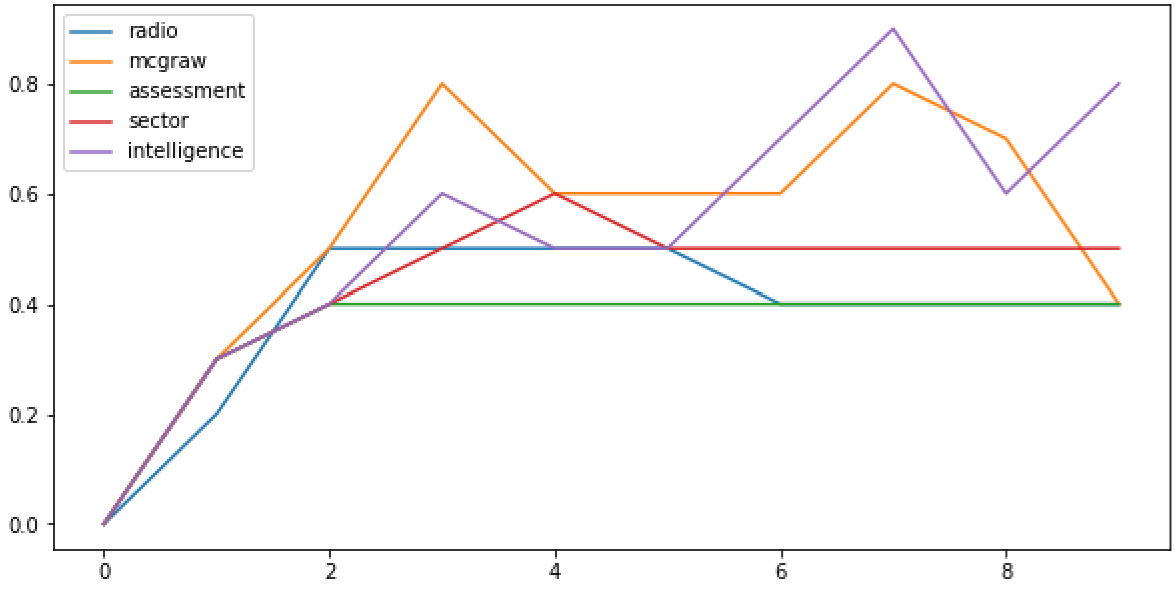
Step 4

most_changes ['radio', 'mcgraw', 'assessment', 'sector', 'intelligence']

Detecting point of change by cosine distance to the embedding of the first decade.

Word usually changed meaning in decades 1 and 2.

The x-axis is the decades (0 to 9) and the y-axis is the cosine distance to the embedding of the first decade.


radio
mcgraw
assessment
sector
intelligence