Assignment 1: Distributional Models of Semantics

–,–

27 January 2016

1 Question 1:

1.1 item c) Cosine simmilarity in frequency space:

What are the similarity scores between house.n, home.n and time.n in the frequency space? Do you think these similarities reflect our intuition that house.n is most similar than home.n?

The similarity scores for this example are the following:

- simmilarity between *home.n* and *house.n* = **0.812743856464**
- simmilarity between *home.n* and *time.n* = **0.818144793373**
- simmilarity between *house.n* and *time.n* = **0.82359219053**

These similarity scores are too similar with the three comparisons. The most similar according to this metric is the items *house.n* and *time.n* that it is not intuitive. In conclusion, these do not reflect our intuition.

1.2 item e) Cosine similarity in tf-idf space:

What are the similarity scores between *house.n*, *home.n* and *time.n* in the tf-idf space? Does yout tf-idf space capture lexical similarity better than the frequency space?

The similarity scores for this example are the following:

- simmilarity between *home.n* and *house.n* = **0.644002976484**
- simmilarity between *home.n* and *time.n* = **0.54386864522**
- simmilarity between *house.n* and *time.n* = **0.57753821538**

The *tf-idf* space capture better the lexical simmilarity with these words. In fact, now we can distinguish that *home.n* and *house.n* are closer than the other two comparisons.

1.3 item f) Best parameters for word2vec

For this question, the best parameters given the tried mesh will be provided. For this, the combinations generated from the following parameters were tried:

Table 1: My caption

Parameter	List
Learning rate	[0.01, 0.03, 0.05]
Downsampling rate	[0.0, 0.01, 0.001, 0.00001
Neg sampling	[0, 5, 10]

And the best combination of parameters was (At least for the tried grid): Learning rate= .05; Downsampling rate = .001 and Neg Sampling = 10. Now, the best parameters are used to train with the full set. The general observations for this grid is that when the learning rate is bigger, better the model. Also, when the negative sampling is bigger, the model is better. For the downsampling rate, when is 0, the model performs bad in terms of accuracy.

1.4 item g) Cosine simmilarities for word2vec

What are the similarity scores between *house.n*, *home.n* and *time.n* in your word2 vec model? Does the word2vec model capture the similarities better than the *fi-idf* space?

The similarity scores for this example are the following:

- simmilarity between *home.n* and *house.n* = **0.579456296125**
- simmilarity between *home.n* and *time.n* = **0.379887119346**
- simmilarity between *house.n* and *time.n* = **0.190818908**

What are the similarity scores between *house.n*, *home.n* and *time.n* in your word2 vec model? Does the word2vec model capture the similarities better than the tf-idf and word2vec spaces?

Although the tf-idf captures better the simmilarity between house and home and can make a difference between house-time and home-time, it still too close. The diffence is of only .10; on the other hand, for the word2vec model, the differences are much bigger (.20 and .38).

1.5 item i) Cosine simmilarities for LDA

The similarity scores for this example are the following:

- simmilarity between *home.n* and *house.n* = **0.799031032467**
- simmilarity between *home.n* and *time.n* = **0.0515955364453**
- simmilarity between *house.n* and *time.n* = **0.0342907459548**

This model is much better than the previous ones, now the simmilarity of house-time and home-time is close to 0. On the other hand, the simmilarity of home-house is close to .8, a very good number in term of the cosine.

1.6 item j)

Going through the topics learned by your LDA model, are there any meaningful ones which you can identify?

In this section we can be meaningful, we have to not consider the most frequent words. If these are considered, the most probable will be verbs like have, can, make.

A interesting topic that I found was related to violence, that contains crime, sex, assault, another one was related to persons, like young, man, woman, people. If everytopic is analysed, then you can obtain interesting words related to them.

2 Question 2:

2.1 item e) Precision and Recall for all the models:

Table 2: My caption

Model	Precision	Recall
Addition - tf-idf	8.98	8.89
Multiplication - tf-idf	13.50	13.37
Addition - word2vec	11.92	11.92
Multiplication - word2vec	6.48	6.48
LDA	12.42	10.68

2.2 item f)

In this section we can see that the multiplicative models work much better in tf-idf. On the other hand, the additive model works much better with the word2 vec and the LDA. This makes sense because in theory the models that works with the complete space (for example, tf-idf or the original frequency space) is better with multiplicative. And LDa and word2vec that works in a "reduced" space is better with addition.