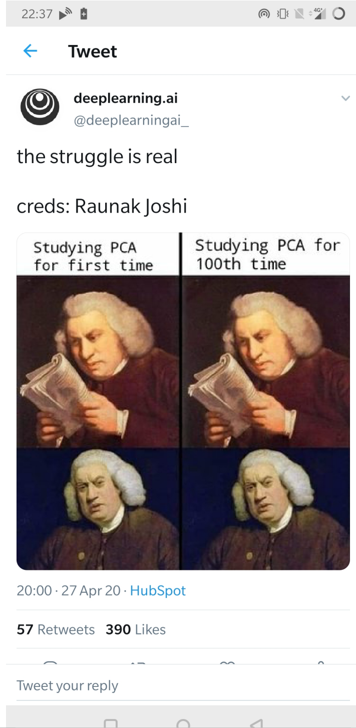
92586 Computational Linguistics

12. Visualisation

Alberto Barrón-Cedeño

Alma Mater Studiorum-Università di Bologna
a.barron@unibo.it @_albarron_

28/04/2020



Previously

- Pre-trained embeddings
- Gensim
- Model construction
- Embedding alternatives

Table of Contents

Chapter 6 of Lane et al. (2019)

Visualisation

Embeddings Visualisation

- Embeddings are in a high-dimensional space (e.g., 300d), which is impossible to visualise
- The human being can visualise up to 3d only[1]
- We need to map into 2d and try to observe interesting phenomena

Let us see

[1] I suggest to read Flatland: `https://en.wikipedia.org/wiki/Flatland`

Embeddings Visualisation

Stuff Available in the Object

```
</s>        Vocab(count:3000000, index:0)
in          Vocab(count:2999999, index:1)
for         Vocab(count:2999998, index:2)
that        Vocab(count:2999997, index:3)
is          Vocab(count:2999996, index:4)
on          Vocab(count:2999995, index:5)
...
Starwood_Hotels_HOT    Vocab(count:2000000, index:1000000)
Tammy_Kilborn          Vocab(count:1999999, index:1000001)
aortic_aneurism        Vocab(count:1999998, index:1000002)
Spragins_Hall          Vocab(count:1999997, index:1000003)
```

- Overall counting
- Index

Embeddings Visualisation

Distance Computation

1. Get the vector representation for w_1 and w_2
2. Compute the distance

Alternatives

- Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

- Cosine "distance"

$$d(p,q) = 1 - \frac{\sum_{i=1}^{n}(p_i\, q_i)}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

Let us see

Embeddings Visualisation

Plotting Cities

Getting the cities in the dataset

Alternative 1 Find the top-k most similar vectors to "city" in the space[2]

Alternative 2 Grab a number of lists from a dictionary

Computing the vectors and plot

- Get sure the items exist in the vocabulary
- Get them and add additional information (e.g., state)
- Reduce the dimension, using PCA

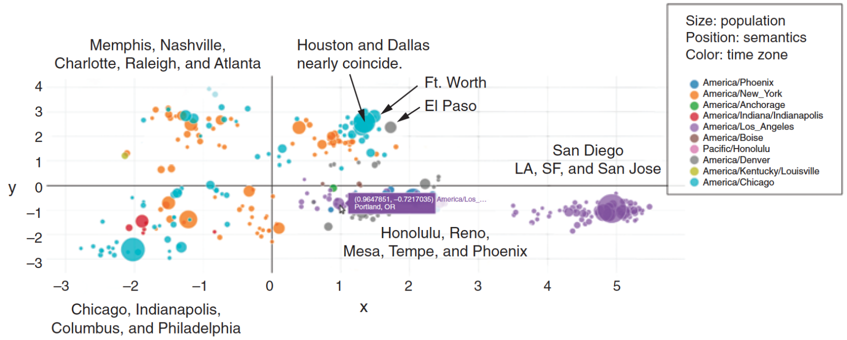
Let us see

(Lane et al., 2019, p. 209)

[2]This is the book proposal. It doesn't work. You can try
`wv.most_similar(positive=['city', 'cities'], topn=10)`

Embeddings Visualisation

US Cities Plot



(Lane et al., 2019, p. 212)

Embeddings Visualisation

Considerations

- PCA works well in this case because we are targeting a limited space
- t-SNE is a better alternative for more diverse vectors[3]

[3]`en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding`

Next

- doc2vec
- CNN
- RNN

References

Lane, H., C. Howard, and H. Hapkem
2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publication Co.