## 92586 Computational Linguistics

12. Visualisation

Alberto Barrón-Cedeño

Alma Mater Studiorum-Università di Bologna a.barron@unibo.it @\_albarron\_

28/04/2020



Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

1/13 Alberto Barrón-Cedeño (DIT-UniBO)

## Previously

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



## Table of Contents

1 Visualisation

Chapter 6 of Lane et al. (2019)

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

3 / 13

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

#### Visualisation

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

5 / 13

## **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<sup>1</sup>
- We need to map into 2d and try to observe interesting phenomena
- Let us see

<sup>1</sup>I suggest to read Flatland: https://en.wikipedia.org/wiki/Flatland

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

6/13

## **Embeddings Visualisation**

Stuff Available in the Object

Vocab(count:3000000, index:0) </s> Vocab(count:2999999, index:1) in Vocab(count:2999998, index:2) for Vocab(count:2999997, index:3) that Vocab(count:2999996, index:4) is Vocab(count:2999995, index:5) on

Vocab(count:2000000, index:1000000) Starwood\_Hotels\_HOT Vocab(count:1999999, index:1000001) Tammy\_Kilborn Vocab(count:1999998, index:1000002) aortic\_aneurism Vocab(count:1999997, index:1000003) Spragins\_Hall

- Overall counting
- Index

## **Embeddings Visualisation**

Distance Computation

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

#### **Alternatives**

Fuclidean distance

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

Cosine "distance"

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



Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

## **Embeddings Visualisation**

Plotting Cities

#### Getting the cities in the dataset

Alternative 1 Find the top-k most similar vectors to "city" in the space<sup>2</sup>

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



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

<sup>2</sup>This is the book proposal. It doesn't work. You can try wv.most\_similar(positive=['city', 'cities'], topn=10)

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

28/04/2020

9/13

## Alberto Barrón-Cedeño (DIT-UniBO)

### Next

#### doc2vec

**Embeddings Visualisation** 

Memphis, Nashville,

Charlotte, Raleigh, and Atlanta

Columbus, and Philadelphia

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

Houston and Dallas

**US Cities Plot** 

- CNN
- RNN

# **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<sup>3</sup>

 $^3$ en.wikipedia.org/wiki/T-distributed\_stochastic\_neighbor\_embedding

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

Alberto Barrón-Cedeño (DIT-UniBO)

92586 Computational Linguistics

92586 Computational Linguistics

28/04/2020

28/04/2020

10 / 13

Size: population Position: semantics

Color: time zone

 America/Anchorage
America/Indiana/Inciana/In America/Indiana/Indiar
America/Los\_Angeles
America/Boise
Pacific/Honolulu
America/Denver
America/Louver

San Diego LA, SF, and San Jose

# References

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

Alberto Barrón-Cedeño (DIT-UniBO) 92586 Computational Linguistics 28/04/2020 13/13