Cluster Analysis of Airbnb listings

`KMeans().fit(Average word embeddings of listing descriptions)`

Data: Sydney Airbnb Open Data

36,662 rows x 96 columns

Q: Can we use word embeddings to discover clusters in Airbnb listings?

Hypothesis:

The average word embedding of a description...

- H0: cannot be used to cluster listings
- H1: can be used to cluster listings into groups

Metric: Silhouette score

Data Preprocessing

1. Raw

```
listing["description"][0]
```

" Come stay with Vinh & Stuart (Awarded as one of Australia's top hosts by Airbnb CEO Brian Chesky & key shareholder Ashton Kutcher. We're Sydney's #1 reviewed hosts too). Find out why we've been positively reviewed 500+ times [...]

" - 182 words

>>

2. Cleaned

```
listing["description_cleaned"][0]
```

" ceo key shareholder review host positively review message talk reservation request read listing end hint hint know pretty relaxed host appreciate hotel casual host hotelier alternative expensive hotel treat way treat friend fluffy bathrobe hello message reservation request speed thing smooth thing read list way end get confirm

" - 48 words

>>

3. Vectorized

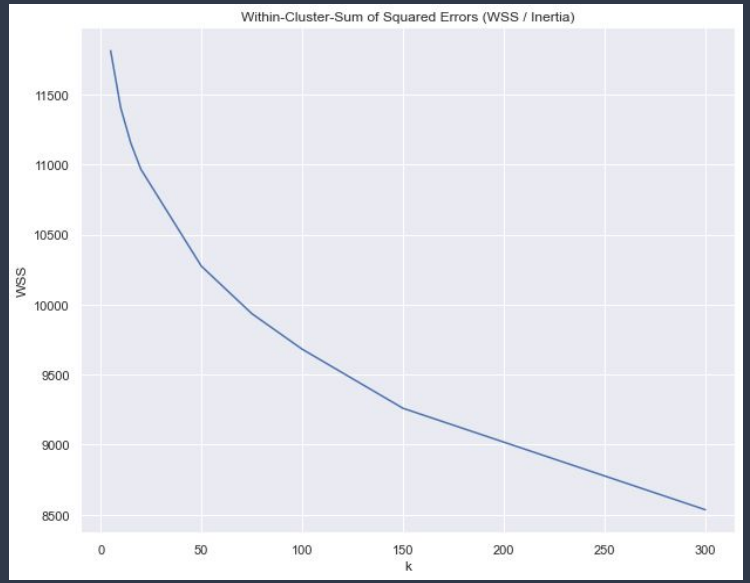
```
vectors[0]
```

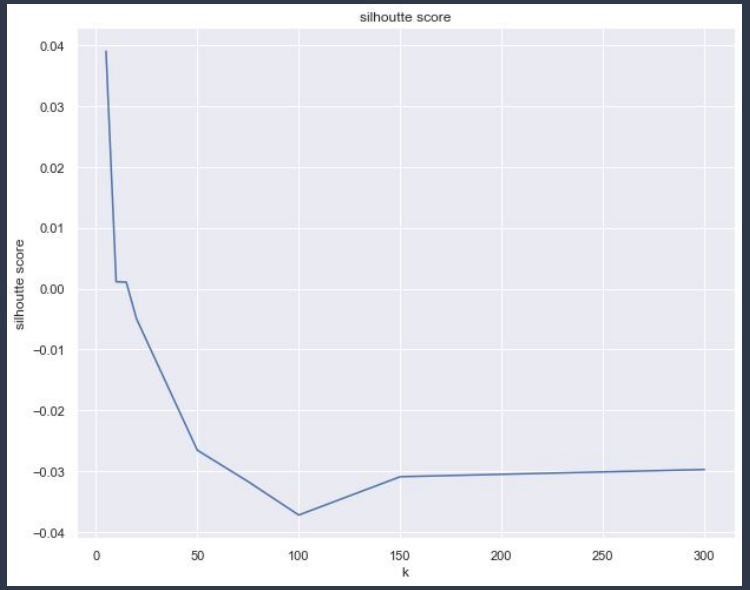
```
>> array([

-2.05833535e-03,  7.92708248e-03,
2.65229214e-02,  7.42500043e-03,
-5.79583226e-03, -1.77062526e-02,
1.11729158e-02, -2.26645861e-02,
1.46166673e-02, -3.38791609e-02,
6.06874982e-03,  1.10104149e-02,
4.22062539e-02, -3.12916785e-02,

...])
```

```
vectors[0].shape
```

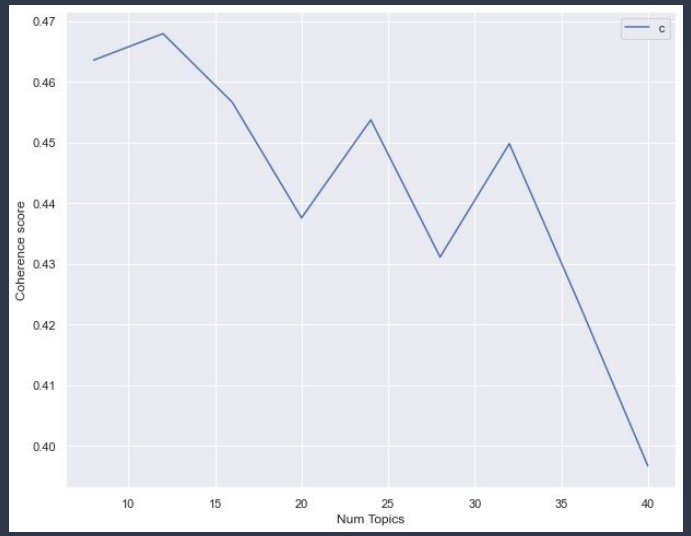
```
>> (300,)
```

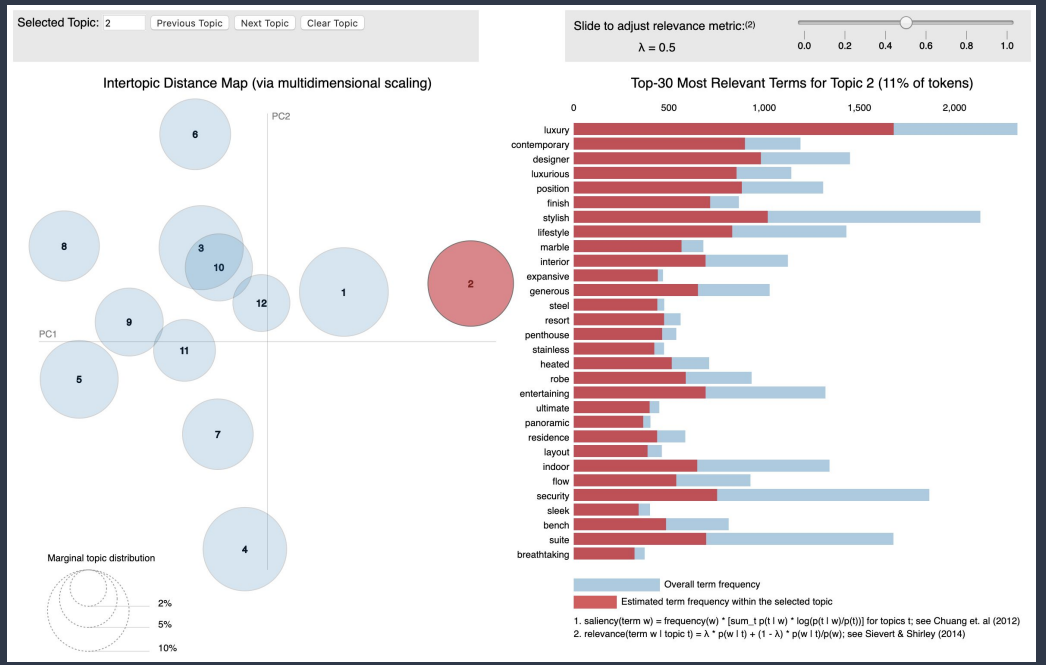
Training KMeans on Word Embeddings





Topic Modelling





Conclusion / Limitation / Suggestion

- Main conclusion: Our approach of using KMeans on word embeddings is more suitable for situations where the text discusses a smaller set of subtopics.
    - E.g. descriptions of over-the-ear headphones is better than descriptions of speakers + earphones + headphones
    - Listing descriptions discuss too many subtopics (Introduction, Interior, Exterior, Neighborhood, Facility, Amenity, Transportation)
- Contrasting the poor clustering result with clean topic modeling results, it is likely that the high number of global optima for the number of topics correlate with poor generalization in the average word embeddings.
- LDA can even be used first to check that there are only a few topics present in the dataset, to determine whether the KMeans approach is viable. If LDA shows that there are many subtopics, the KMeans approach may still produce positive results by training own neural networks to output word embeddings specifically for the dataset.
- LDA was not intended to replace our approach. It was used to provide context to understand why KMeans didn’t work.

Appendix

Top 10 Words that best represent centroid of each cluster

k=1: ['brown', 'clothes', 'everything', 'kind', 'one',
'something', 'stuff', 'thing', 'things', 'wooden']

k=2: ['a', 'charming', 'classic', 'everything', 'kind',
'one', 'something', 'stylish', 'too', 'truly']

**k=3: ['adventure', 'adventurer', 'adventurers', 'adventures',
'adventuring', 'artist', 'expeditions', 'journey', 'musician', 'solo']**

**k=4: ['boat', 'downtown', 'hotel', 'kind', 'one',
'school', 'shops', 'something', 'thing', 'town']**

k=5: ['either', 'go', 'just', 'kind', 'one',
'say', 'something', 'thing', 'too', 'way']

Data Preprocessing

1

```
listing["description"][0]
```

" Come stay with Vinh & Stuart (Awarded as one of Australia's top hosts by Airbnb CEO Brian Chesky & key shareholder Ashton Kutcher. We're Sydney's #1 reviewed hosts too). Find out why we've been positively reviewed 500+ times [...]

" - 182 words

Clean

- Remove stopwords, punctuations
- Casted to lower-case
- Lemmatized (Normalized to base form)
  - "Walking" becomes "walk"
- Removed "Named Entities"
  - "Bondi Beach"
  - "David"

2

```
listing["description_cleaned"][0]
```

" ceo key shareholder review host positively review message talk reservation request read listing end hint hint know pretty relaxed host appreciate hotel casual host hotelier alternative expensive hotel treat way treat friend fluffy bathrobe hello message reservation request speed thing smooth thing read list way end get confirm

" - 48 words

Take the mean of all word vectors from the description

3

```
vectors[0]
>> array([
-2.05833535e-03,  7.92708248e-03,
2.65229214e-02,  7.42500043e-03,
-5.79583226e-03, -1.77062526e-02,
1.11729158e-02, -2.26645861e-02,
1.46166673e-02, -3.38791609e-02,
6.06874982e-03,  1.10104149e-02,
4.22062539e-02, -3.12916785e-02,
...])

vectors[0].shape
>> (300,)
```