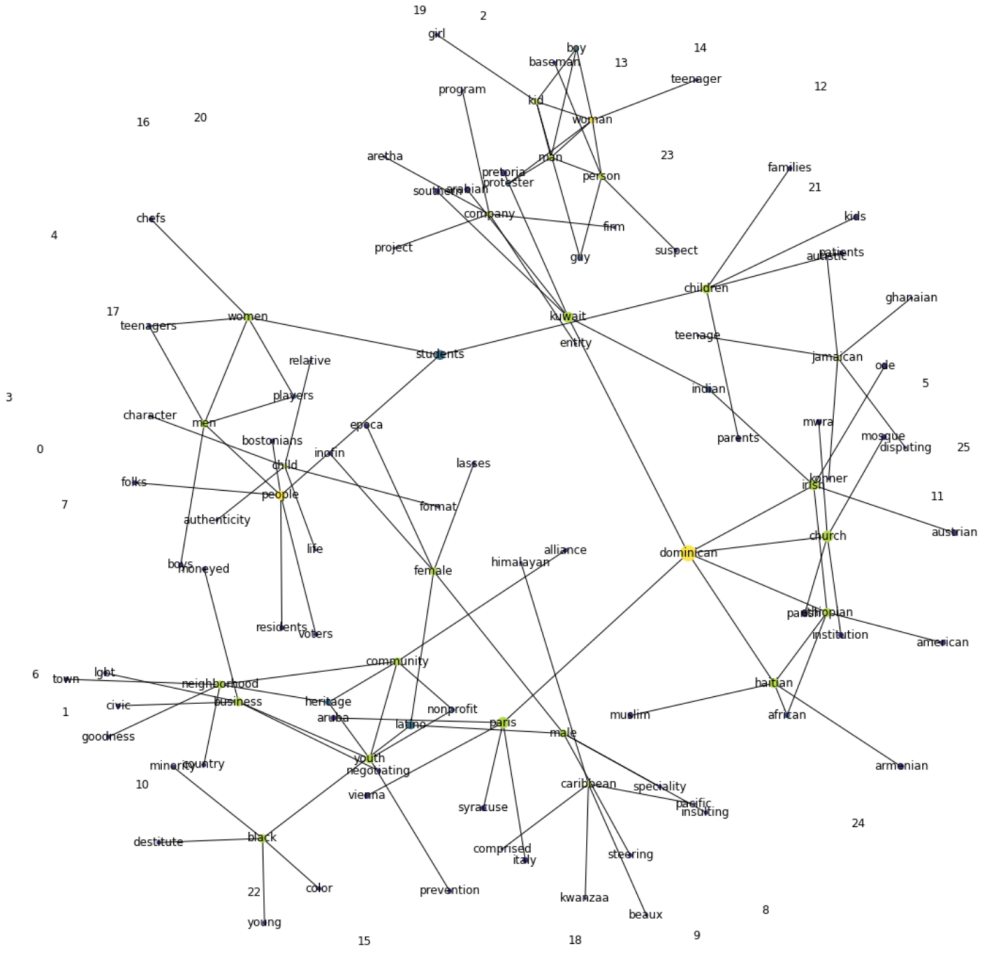
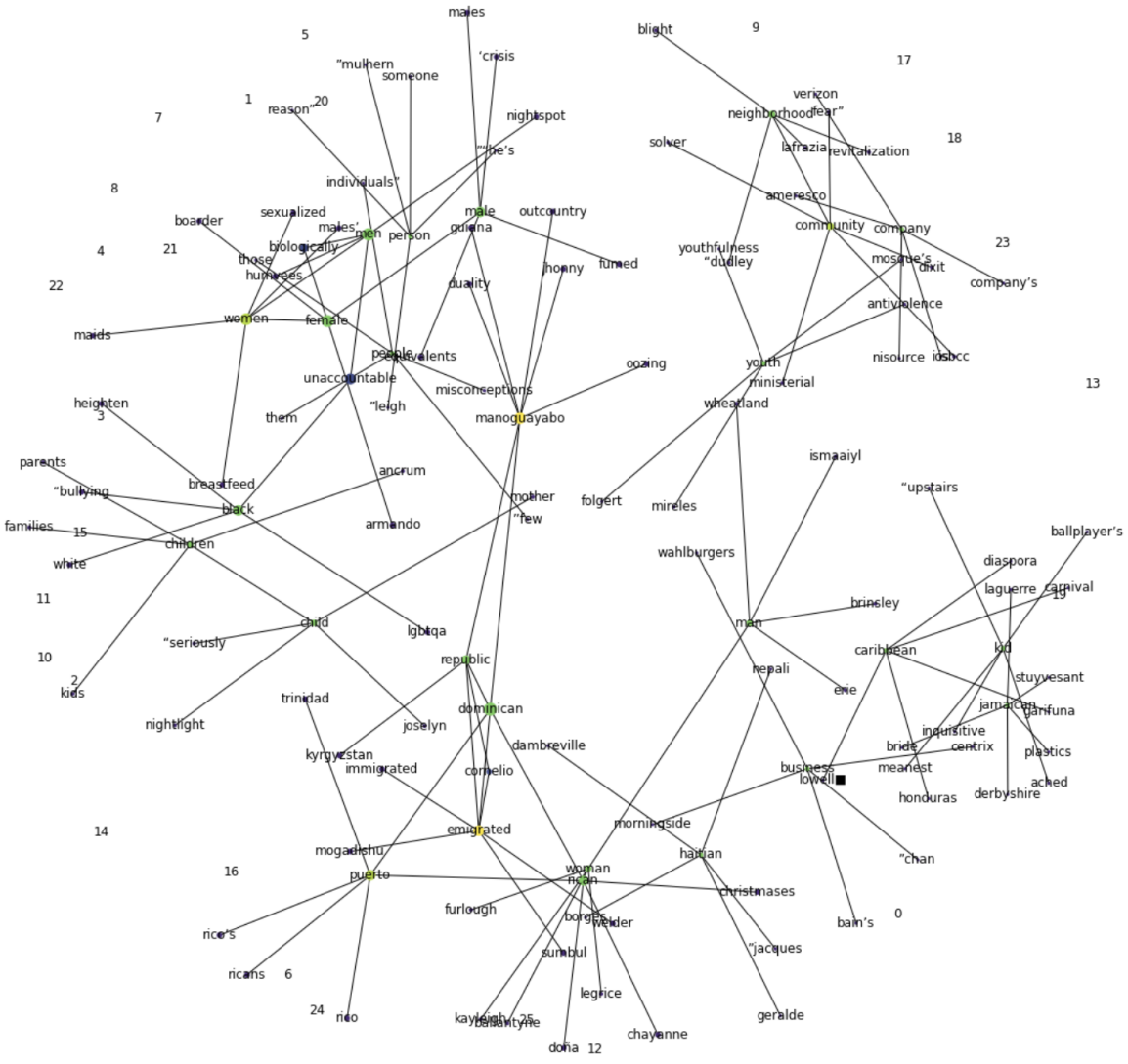
https://docs.google.com/document/d/1rW765Ta24LeK1i1pnD_P_nG1wcl4bAg-U3WHHDokKTw/edit?usp=sharing

Data set: Boston Globe 2018

Word Cloud for Doc2Vec:



Word Cloud for Word2Vec:



Gensim Phrases package to automatically detect common phrases (bigrams) from a list of sentences.

#look up a list of the most similar words from keyword,"black"

Code:

```
w2v_model.wv.most_similar(positive=['black'])
```

Output:

```
[('white', 0.812944769859314),
 ('african_american', 0.7083898782730103),
```

('latino', 0.5846176743507385),

('color', 0.5713448524475098),

('young', 0.5659607648849487),

('asian', 0.5410647392272949),

('women', 0.5286791324615479),

('asian_american', 0.5229591727256775),

('hispanic', 0.5162984132766724),

('male', 0.5044015049934387)]

#measure the similarity between any 2 words

Code:

```
w2v_model.wv.similarity('black', 'green')
```

Output

0.1462823

#Analogy difference

#Which word is to "lation" as "black" is to "african_american"?

Code:

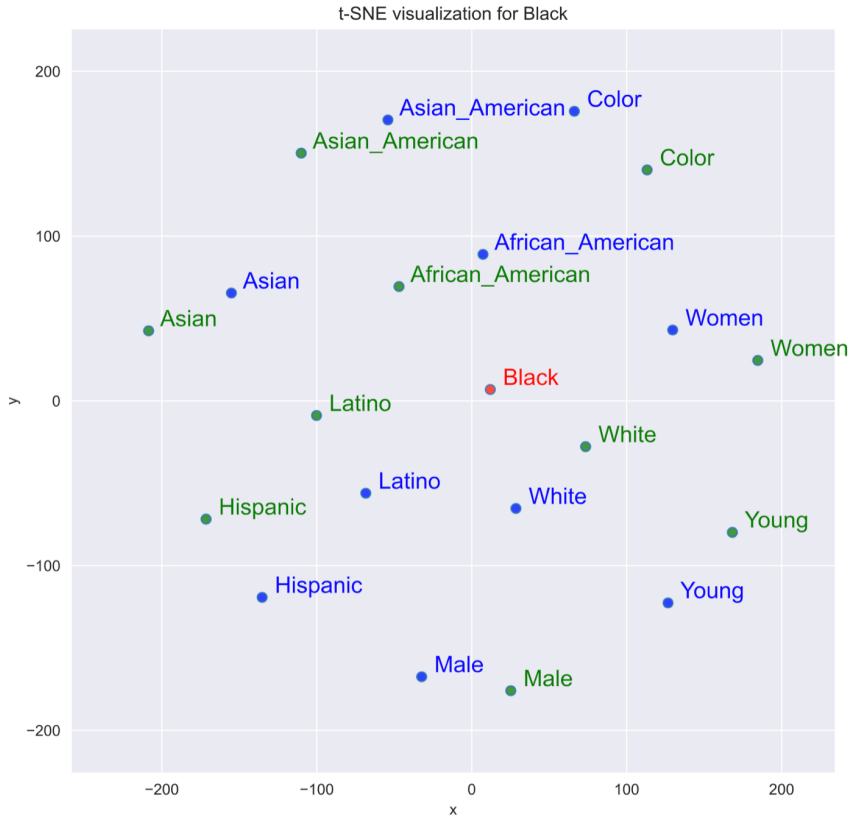
```
w2v_model.wv.most_similar(positive=["latino", "black"], negative=["african_american"], topn=3)
```

Output:

[('white', 0.5755782723426819),
 ('hispanic', 0.5324521660804749),
 ('minority', 0.5324294567108154)]

#t-SNE visualizations:==========

```
# the vector representation of "black" and 10 most similar words lies in a 2D graph.
tsnescatterplot(w2v_model, 'black', ['white', 'african_american', 'latino', 'color', 'young', 'asian', 'women', 'asian_americ
an','hispanic','male'])
```



```
# the vector representation of "black" and 10 least similar words lies in a 2D graph.
tsnescatterplot(w2v_model, 'black', [i[0] for i in w2v_model.wv.most_similar(negative=['black'])])
```