

Link to this file:

https://docs.google.com/document/d/1PuaGbsq-tfQrv1zzJjXac4k2JUx9aDcdwj5gj_bpcfI/edit?usp=sharing

Analysis Format:

Data Label

Similar words to specific keywords using Word2Vec and Doc2Vec in tabular format

List of similar words to 'black' using Word2Vec with the Gensim Phrases Package

List of dissimilar words to 'black' using Word2Vec with the Gensim Phrases Package

Plot of dissimilar words to 'black' using Word2Vec with the Gensim Phrases Package

(distance between any two words on the plot = dissimilarity between those two words)

Boston Globe 2018:

Word	Notable Similar Words from Word2Vec	Notable Similar Words from Doc2Vec
black	bullying, unaccountable	young
jamaican	---	disputing, autistic
dominican	emigrated	---
republic	emigrated	---
mano guay (Spanish slang)	outcountry, oozing, duality	---
emigrated	Sumbul (Muslim name), Mogadishu (capital of Somalia)	---

Below are the most similar words according to Word2Vec analysis using the Gensim Phrases Package (this package automatically detects common phrases used in a document); as opposed to earlier, when the models were trained on individual words, this time, the Word2Vec

model was trained on the most common phrases appearing in the text data.

```
#look up a list of the most similar words from keyword, "black"  
w2v_model.wv.most_similar(positive=['black'])
```

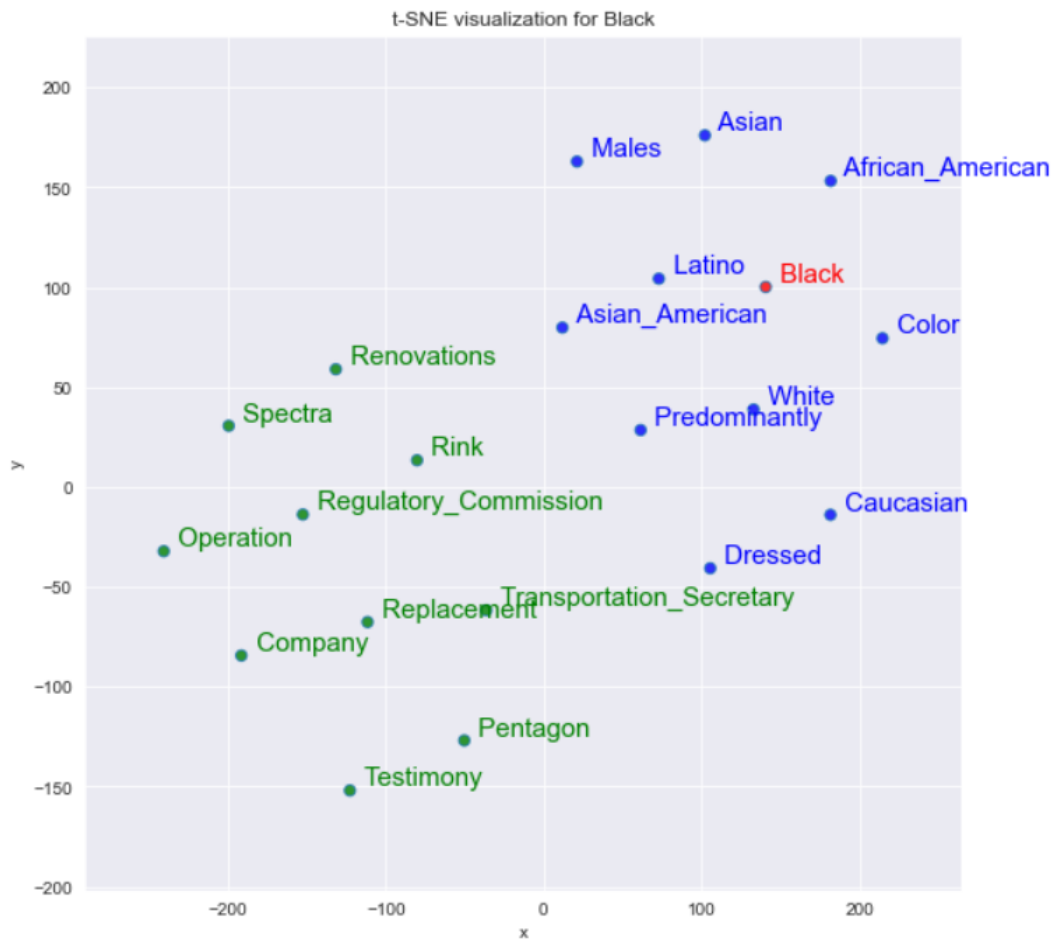
```
[('white', 0.812944769859314),  
 ('african_american', 0.7083898782730103),  
 ('latino', 0.5846176743507385),  
 ('color', 0.5713448524475098),  
 ('young', 0.5659607648849487),  
 ('asian', 0.5410647392272949),  
 ('women', 0.5286791324615479),  
 ('asian_american', 0.5229591727256775),  
 ('hispanic', 0.5162984132766724),  
 ('male', 0.5044015049934387)]
```

Yet another interesting find using the Gensim Phrases package trained Word2Vec model were the least similar words.

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

```
[('replacement', 0.3242053985595703),  
 ('regulatory_commission', 0.3223087787628174),  
 ('company', 0.29238665103912354),  
 ('spectra', 0.28767138719558716),  
 ('pentagon', 0.27769172191619873),  
 ('operation', 0.27544352412223816),  
 ('rink', 0.27497997879981995),  
 ('transportation_secretary', 0.27428847551345825),  
 ('testimony', 0.2742670774459839),  
 ('renovations', 0.2684938907623291)]
```

```
# 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'])])
```



Boston Globe 2017:

Word	Notable Similar Words from Word2Vec	Notable Similar Words from Doc2Vec
black	unarmed, vigilante, protagonist, white, deray	young, female
haitian	newness, learners, bold	modernist, champ
caribbean	gassed	stacked, holy
jamaican	immortalized	sciacca [city in Italy], lesbian
dominican	---	marching

Most similar words according to Word2Vec + Gensim Phrases analysis:

```
#look up a list of the most similar words from keyword, "black"  
w2v_model.wv.most_similar(positive=['black'])
```

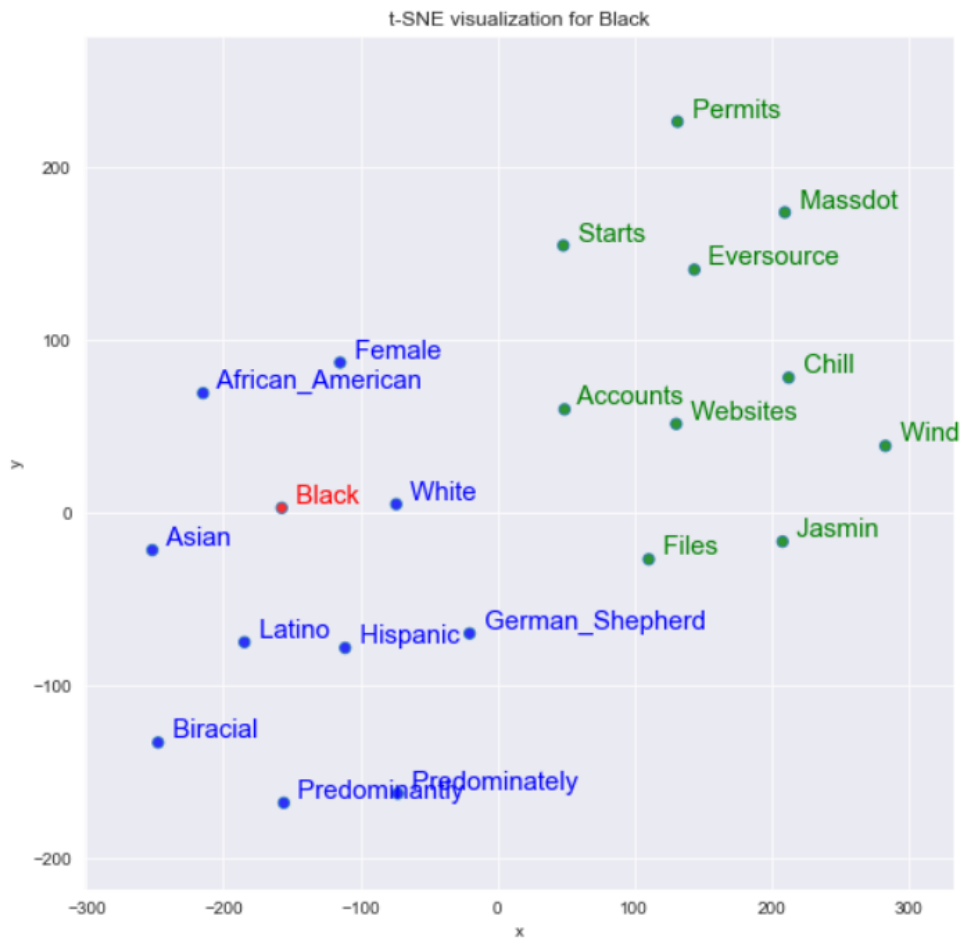
```
[('white', 0.701784610748291),  
 ('african_american', 0.6560036540031433),  
 ('hispanic', 0.6200186014175415),  
 ('biracial', 0.61896812915802),  
 ('asian', 0.6163231730461121),  
 ('predominantly', 0.5821590423583984),  
 ('female', 0.5647924542427063),  
 ('predominately', 0.5605288743972778),  
 ('german_shepherd', 0.555185854434967),  
 ('latino', 0.5543240308761597)]
```

Least similar words according to Word2Vec + Gensim Phrases analysis:

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

```
[('wind', 0.29533684253692627),  
 ('websites', 0.27661919593811035),  
 ('chill', 0.2629413604736328),  
 ('accounts', 0.2627812623977661),  
 ('massdot', 0.25185251235961914),  
 ('eversource', 0.24410991370677948),  
 ('files', 0.24042987823486328),  
 ('permits', 0.23978425562381744),  
 ('starts', 0.23565936088562012),  
 ('jasmin', 0.23508203029632568)]
```

```
# 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'])])
```



Boston Globe 2016:

Word	Notable Similar Words from Word2Vec	Notable Similar Words from Doc2Vec
black	white, retaliated, resists, chosin [USS Chosin/Battle of Chosin Reservoir]	young
haitian	sauntering	immigrant
caribbean	sauntering	academy, poets, episcopal
republic	devolved	---

Most similar words according to Word2Vec + Gensim Phrases analysis:

```
#look up a list of the most similar words from keyword, "black"  
w2v_model.wv.most_similar(positive=['black'])
```

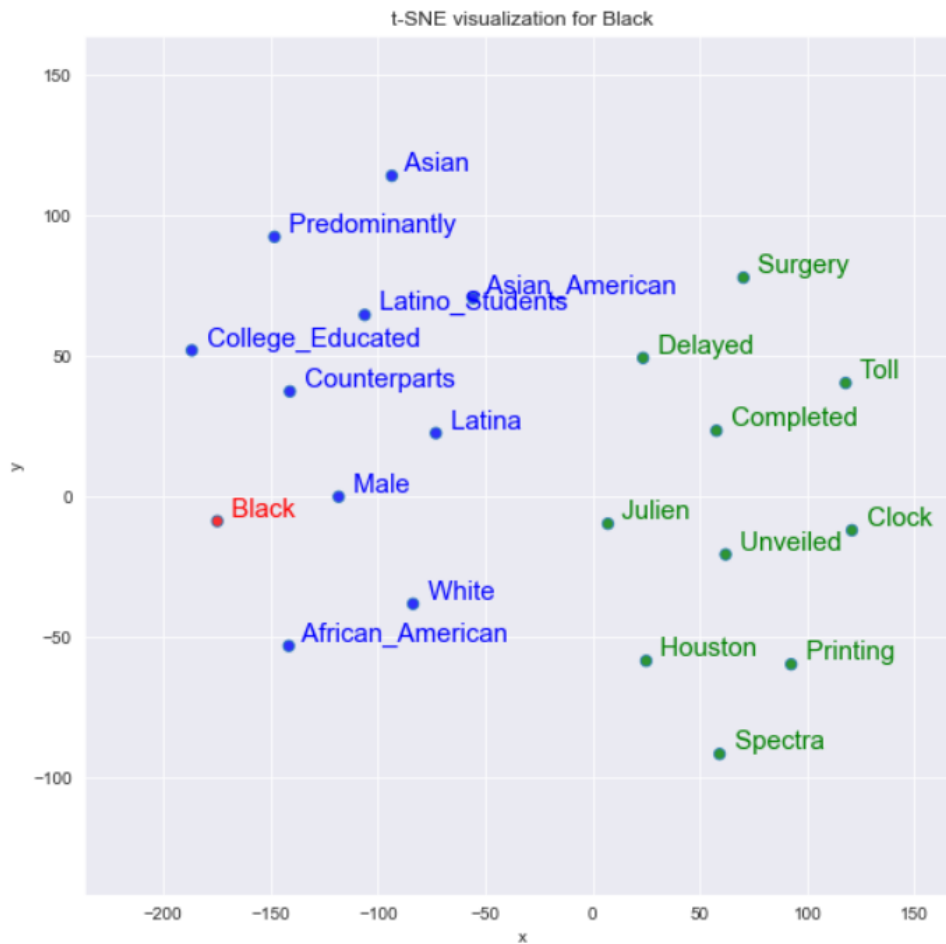
```
[('white', 0.6975640058517456),  
 ('african_american', 0.6343957185745239),  
 ('male', 0.6177483797073364),  
 ('predominantly', 0.6135073900222778),  
 ('counterparts', 0.5948563814163208),  
 ('asian_american', 0.593398928642273),  
 ('college_educated', 0.5785021781921387),  
 ('latina', 0.5748915672302246),  
 ('asian', 0.5665065050125122),  
 ('latino_students', 0.551455020904541)]
```

Least similar words according to Word2Vec + Gensim Phrases analysis:

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

```
[('clock', 0.2532062530517578),  
 ('houston', 0.2531570792198181),  
 ('toll', 0.2389242947101593),  
 ('spectra', 0.2288730889558792),  
 ('printing', 0.21526795625686646),  
 ('surgery', 0.21273605525493622),  
 ('julien', 0.2056630253791809),  
 ('delayed', 0.2054496556520462),  
 ('completed', 0.2043745517730713),  
 ('unveiled', 0.20174488425254822)]
```

```
# 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'])])
```



Boston Globe 2015:

Word	Notable Similar Words from Word2Vec	Notable Similar Words from Doc2Vec
black	white, young, destitute	white, heighten, statistic, caucasian
haitian	bloc, majority, scientology, Armenian	cofounded
caribbean	---	mecca, masquerade, medieval
dominican	jewish, native, oldest	emigrated, journeyed

Most similar words according to Word2Vec + Gensim Phrases analysis:

```
#look up a list of the most similar words from keyword, "black"  
w2v_model.wv.most_similar(positive=['black'])
```

```
[('white', 0.8165690898895264),  
 ('asian', 0.7816506624221802),  
 ('african_american', 0.7715171575546265),  
 ('latino', 0.7585625648498535),  
 ('male', 0.7308614253997803),  
 ('males', 0.716025710105896),  
 ('predominantly', 0.7140694856643677),  
 ('female', 0.6797010898590088),  
 ('overwhelmingly', 0.6713233590126038),  
 ('supremacy', 0.6694000363349915)]
```

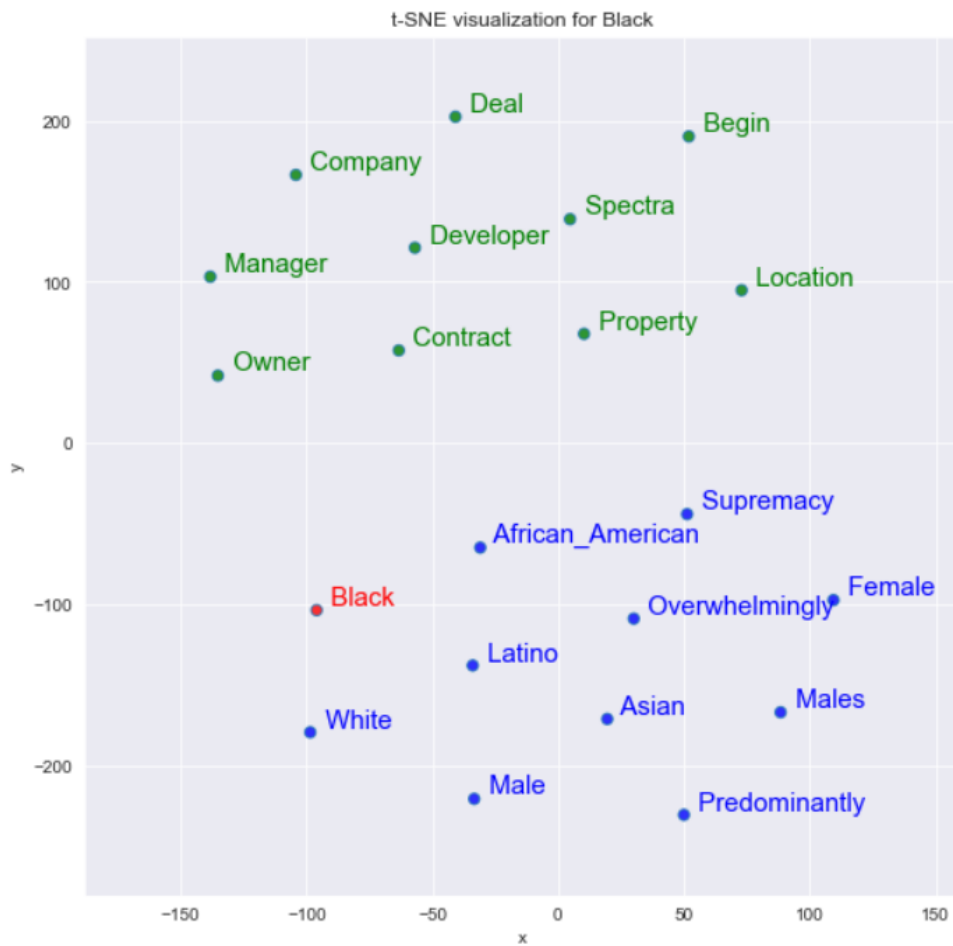
Least similar words according to Word2Vec + Gensim Phrases analysis:

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

```
[('company', 0.3198091983795166),  
 ('deal', 0.2582084834575653),  
 ('property', 0.2535374164581299),  
 ('developer', 0.25218069553375244),  
 ('owner', 0.24039536714553833),  
 ('begin', 0.23259088397026062),  
 ('spectra', 0.21792221069335938),  
 ('contract', 0.2144336700439453),  
 ('manager', 0.21199703216552734),  
 ('location', 0.211470365524292)]
```



```
# 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'])])
```



Boston Globe 2014:

Word	Notable Similar Words from Word2Vec	Notable Similar Words from Doc2Vec
black	white	young, white, transgender, innocent
haitian	immigrant, lifelong	captive
caribbean	cultures	cottage
jamaican	earthy	rum
dominican	---	val

Most similar words according to Word2Vec + Gensim Phrases analysis:

```
#look up a list of the most similar words from keyword, "black"  
w2v_model.wv.most_similar(positive=['black'])
```

```
[('white', 0.9141479730606079),  
 ('male', 0.8686223030090332),  
 ('latino', 0.8194109201431274),  
 ('african_american', 0.8024318218231201),  
 ('predominantly', 0.7746581435203552),  
 ('color', 0.7658802270889282),  
 ('hispanic', 0.7244097590446472),  
 ('asian', 0.7237138748168945),  
 ('muslim', 0.7221528887748718),  
 ('female', 0.7087023258209229)]
```

Least similar words according to Word2Vec + Gensim Phrases analysis:

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

```
[('plan', 0.22196054458618164),  
 ('office', 0.2181970328092575),  
 ('proposed', 0.21777868270874023),  
 ('company', 0.20722068846225739),  
 ('announced', 0.20678149163722992),  
 ('plans', 0.20106220245361328),  
 ('begin', 0.19798430800437927),  
 ('process', 0.1895374059677124),  
 ('move', 0.1832236796617508),  
 ('budget', 0.17687849700450897)]
```

```
# 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'])])
```

