Fake News Detection

Overview

- With an increasing number of people across the world relying on social media as their primary source of news, there is a need to tackle the underlying problem of fake news on those platforms. This project aims to do that by employing techniques of Natural Language Processing (NLP) and cross-references the piece of information entered by the user with the news articles scraped from renowned news websites around a topic.
- The dataset for this project was scraped from https://abcnews.go.com/ around three pertinent topics namely, Coronavirus, Donald Trump and US Elections.
- For demonstration purposes, I have only scraped around 18k articles. But, more exhaustive the dataset is, the better 'word2vec' will get trained and in-turn yield better results. Ofcourse, training the model will take more time accordingly.

Features

- Features used in the dataset-

```
i)   Link      - URL Link to the relevant news articles
ii)  Title     - Title of the news articles
iii) Content   - The actual content of the news articles
iv)  Category  - Categories like International, Health, US, Sports, Politics, etc
into which the articles are classified into
```

Summary of the code

- Importing dependencies and acquiring data
- Performing Data Pre-processing and tokenizing the 'Content' feature
- Training 'word2vec' and running t-sne for multiple values of *perplexity* and *n_iter*
- Based on the user input, the list of relevant articles, words most relating to the ones entered by the user as well as a map showing how closely words in the input corpus relate to each other are displayed.

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import re
import math
from nltk.corpus import stopwords
import pdb
import warnings
from collections import OrderedDict
import multiprocessing
import nltk
from collections import Counter
import random
import gensim.models.word2vec as w2v
from sklearn.feature_extraction.text import CountVectorizer
import sklearn.manifold
```

In [2]:
```
nltk.download("stopwords")
```

```
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[2]:
```
True
```

In [3]:
```
from google.colab import files


uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving master_scrapped_data.csv to master_scrapped_data.csv
```

In [4]:
```
data = pd.read_csv('master_scrapped_data.csv')
```

In [5]:
```
data.head()
```

Out[5]:

	Link	Title	Content	Category
0	https://abcnews.go.com/Health/texas-front-line...	Texas front-line doctor dies of COVID-19 compl...	A Texas doctor who treated COVID-19 patients d...	Health
1	https://abcnews.go.com/Politics/pompeo-invites...	Pompeo invites hundreds to indoor State Dept. ...	Despite warnings by top health officials to av...	Politics
2	https://abcnews.go.com/US/york-autonomous-zone...	Protest over shutdown of New York's 'autonomou...	Protesters waving American flags and slamming ...	US
3	https://abcnews.go.com/US/104-year-wwii-vetera...	104-year-old WWII veteran survives COVID-19, r...	A World War II veteran defeated COVID-19 just ...	US
4	https://abcnews.go.com/Politics/cdc-encourage-...	CDC to encourage 'Vaccinated for COVID-19' but...	If there are "I Voted" stickers, why not "Vacc...	Politics

Data Pre-processing

In [6]:
```
data.isna().sum()
```

Out[6]:
```
Link        1
Title       8
Content     8
Category    2
dtype: int64
```

In [7]:
```
data[data.isna().any(axis=1)]
```

Out[7]:

	Link	Title	Content	Category
164	https://abcnews.go.com/Health/covid-free-count...	NaN	NaN	Health
192	https://abcnews.go.com/US/faces-coronavirus-pa...	Faces of some of the more than 254,000 lives l...	The novel coronavirus pandemic has left an ind...	NaN
193	NaN	US	NaN	NaN
1304	https://abcnews.go.com/International/putin-don...	NaN	NaN	International
3729	https://abcnews.go.com/Health/frenchman-	NaN	NaN	Health

	dies-s...			
3730	https://abcnews.go.com/Politics/high-dollar-tr...	NaN	NaN	Politics
3731	https://abcnews.go.com/US/trumps-quest-dominat...	NaN	NaN	US
6592	https://abcnews.go.com/Sports/paul-george-clip...	NaN	NaN	Sports
12525	https://abcnews.go.com/Politics/trump-ignores-...	NaN	NaN	Politics
12624	https://abcnews.go.com/Politics/biden-launch-3...	NaN	Former Vice President Joe Biden launched the t...	Politics

In [8]:
```
data.dropna(inplace=True)
data.drop_duplicates(subset='Link',inplace=True)
data.reset_index(drop=True,inplace=True)
data.isna().sum()
```

Out[8]:
```
Link        0
Title       0
Content     0
Category    0
dtype: int64
```

In [9]:
```
data.shape
```

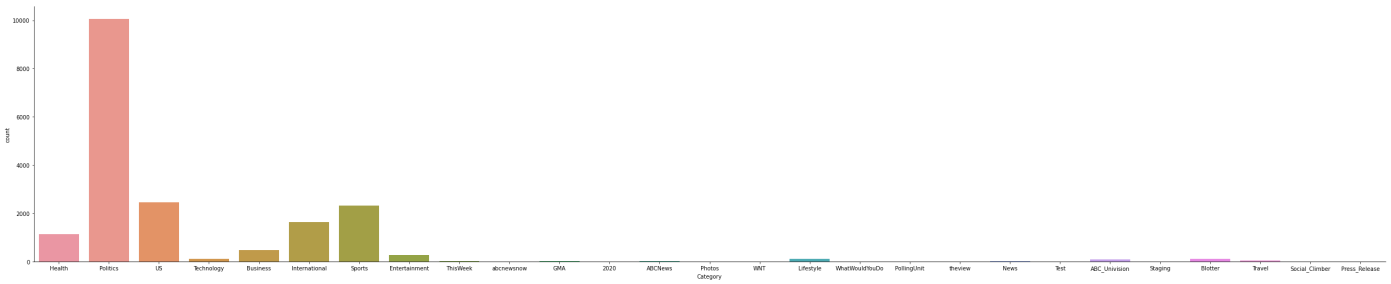
Out[9]:
```
(18916, 4)
```

In [10]:
```
data.Category.unique()
```

Out[10]:
```
array(['Health', 'Politics', 'US', 'Technology', 'Business',
       'International', 'Sports', 'Entertainment', 'ThisWeek',
       'abcnewsnow', 'GMA', '2020', 'ABCNews', 'Photos', 'WNT',
       'Lifestyle', 'WhatWouldYouDo', 'PollingUnit', 'theview', 'News',
       'Test', 'ABC_Univision', 'Staging', 'Blotter', 'Travel',
       'Social_Climber', 'Press_Release'], dtype=object)
```

In [11]:
```
sns.catplot(x="Category", kind="count", data=data, height = 7, aspect = 5.0)
```

Out[11]:
```
<seaborn.axisgrid.FacetGrid at 0x7f960872aef0>
```



In [12]:
```
# loading stop words from nltk library
stop_words = set(stopwords.words('english'))

# referenced from https://stackoverflow.com/questions/23996118/replace-special-character
# and https://pythonexamples.org/python-replace-multiple-spaces-with-single-space-in-tex

def sentence_split(text):
    """
    The sentences following 'MORE:' in the 'Content' feature are backlinks for other art
    irrelevant for that particular news article and so we remove it. The sent_tokenize t
    into indivisual snetences.
```

```
    """
    text = re.sub(r'(?=MORE:)(.*)','',text)
    text = nltk.sent_tokenize(text)
    return text

def sentence_cleaner(text):
    new_text = []
    for sentence in text:
        sentence = sentence.lower()
        sentence = re.sub("((\S+)?(http(S)?)(\S+))|((\S+)?(www)(\S+))|((\S+)?(\@)(\S+)?)
        sentence = re.sub('[^a-zA-Z0-9\n]',' ', sentence)
        sentence = nltk.word_tokenize(sentence)
        sentence = [word for word in sentence if word not in stop_words and len(word)>1]
        new_text.append(sentence)
    return new_text

def final_tokenizer(text):
    return sentence_cleaner(sentence_split(text))
```

In [13]:
```
data['tokenized_words'] = data['Content'].apply(final_tokenizer)
```

In [14]:
```
data.head()
```

Out[14]:

	Link	Title	Content	Category	tokenized_words
0	https://abcnews.go.com/Health/texas-front-line...	Texas front-line doctor dies of COVID-19 compl...	A Texas doctor who treated COVID-19 patients d...	Health	[[texas, doctor, treated, covid, 19, patients,...
1	https://abcnews.go.com/Politics/pompeo-invites...	Pompeo invites hundreds to indoor State Dept. ...	Despite warnings by top health officials to av...	Politics	[[despite, warnings, top, health, officials, a...
2	https://abcnews.go.com/US/york-autonomous-zone...	Protest over shutdown of New York's 'autonomou...	Protesters waving American flags and slamming ...	US	[[protesters, waving, american, flags, slammin...
3	https://abcnews.go.com/US/104-year-wwii-vetera...	104-year-old WWII veteran survives COVID-19, r...	A World War II veteran defeated COVID-19 just ...	US	[[world, war, ii, veteran, defeated, covid, 19...
4	https://abcnews.go.com/Politics/cdc-encourage-...	CDC to encourage 'Vaccinated for COVID-19' but...	If there are "I Voted" stickers, why not "Vacc...	Politics	[[voted, stickers, vaccinated, covid, 19, butt...

In [15]:
```
all_words = list(data['tokenized_words'])
all_words = [subitem for item in all_words for subitem in item]
```

Training the *word2vec*

In [16]:
```
num_features = 300
min_word_count = 10
num_workers = multiprocessing.cpu_count()
window_size = 10
downsampling = 1e-3
seed = 1
```

In [17]:
```
word2vec = w2v.Word2Vec(
    sg=1,
    seed=seed,
    workers=num_workers,
    size=num_features,
```

```
    min_count=min_word_count,
    window=window_size,
    sample=downsampling
)
```

In [18]:
```
word2vec.build_vocab(all_words)
```

In [19]:
```
print("Word2Vec vocabulary length:", len(word2vec.wv.vocab))
```

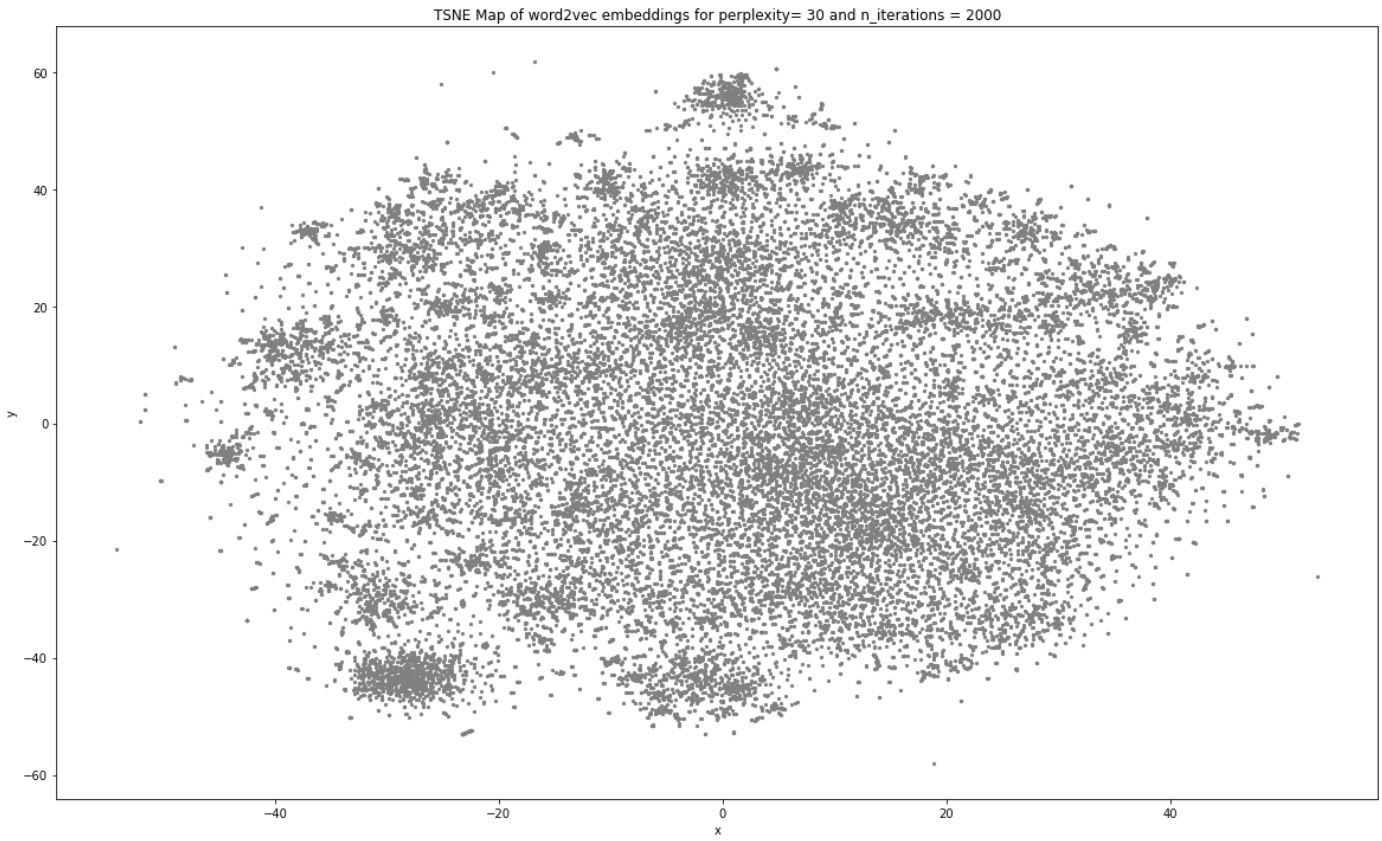
```
Word2Vec vocabulary length: 27430
```

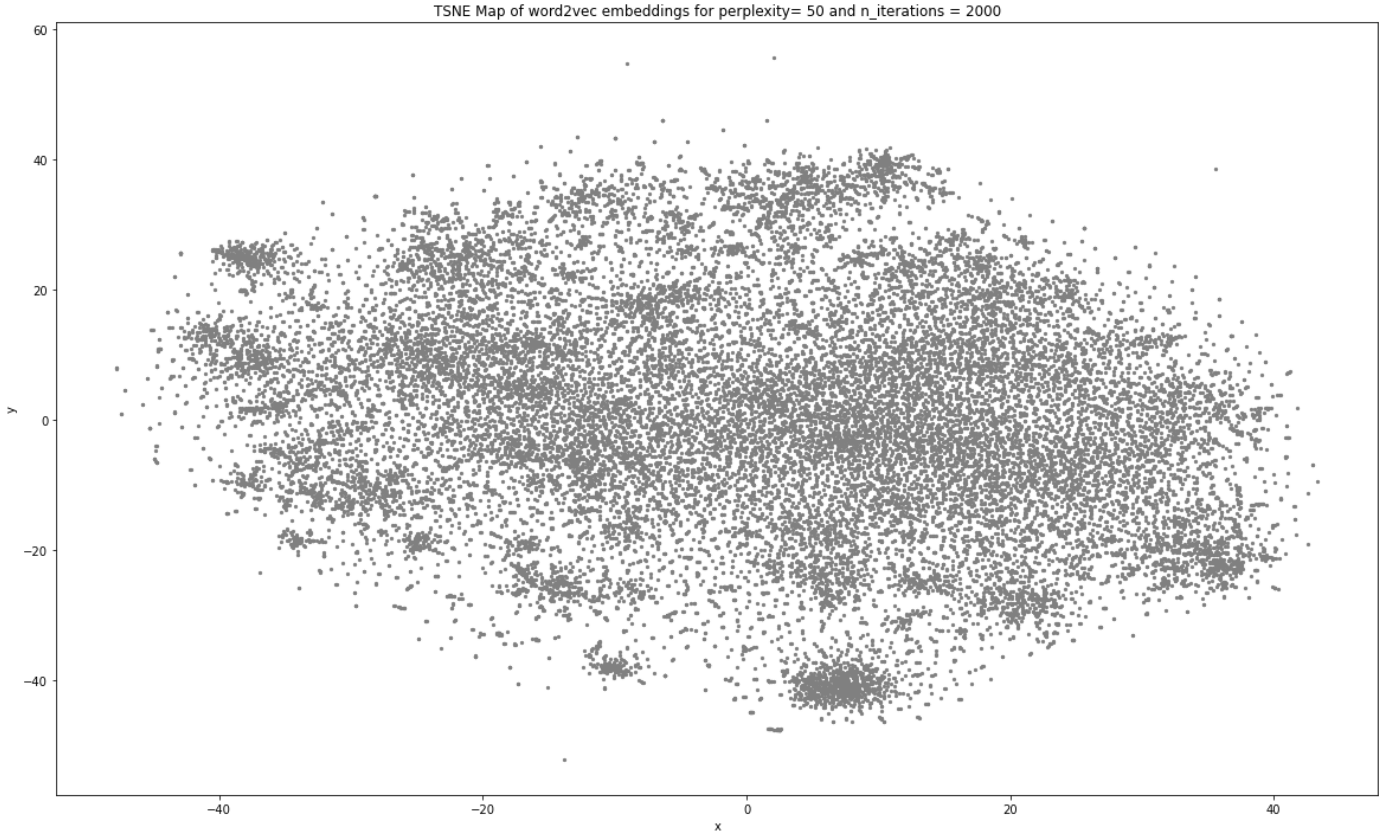
In [20]:
```
word2vec.train(all_words, total_examples=word2vec.corpus_count, epochs=word2vec.epochs)
```

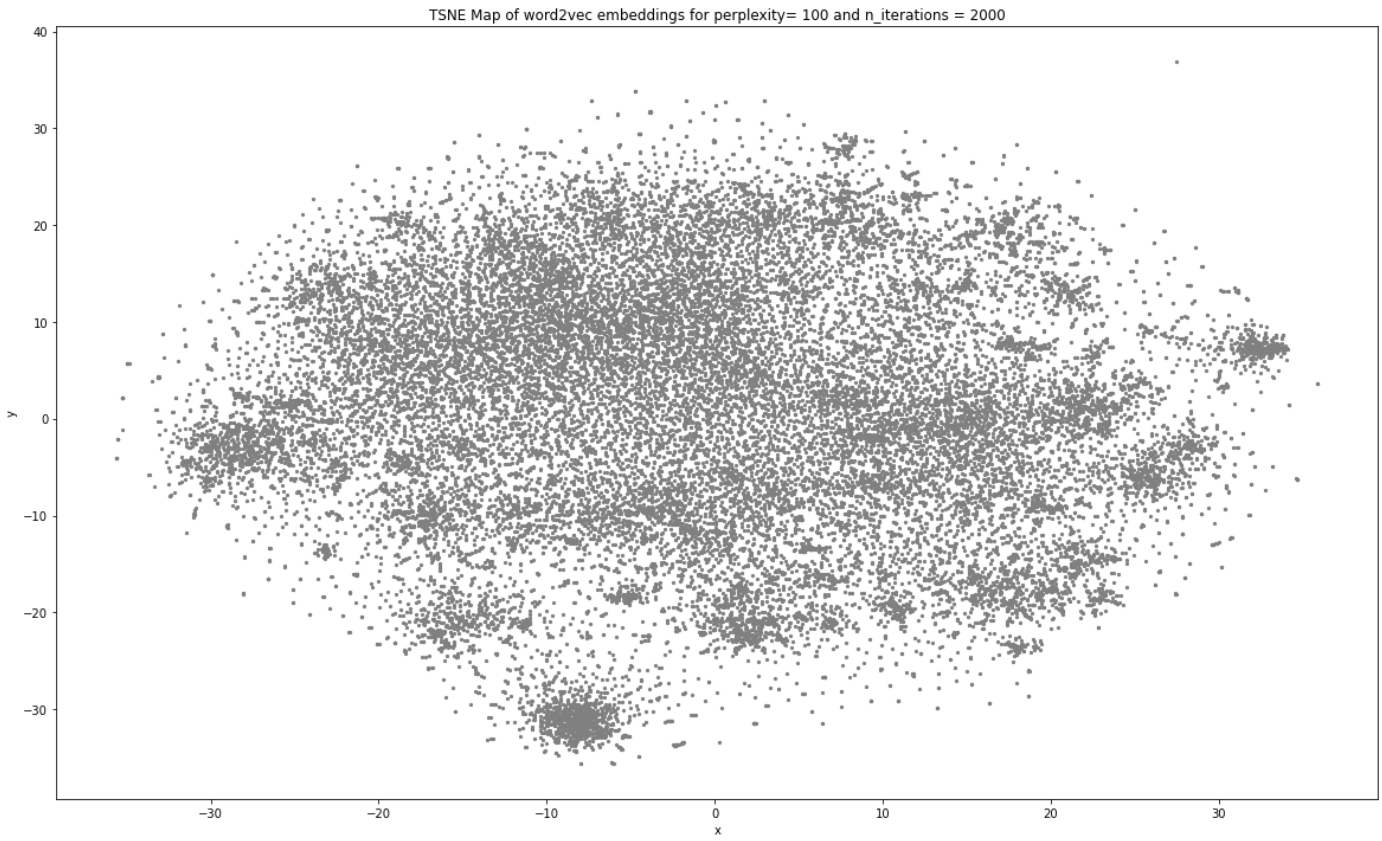
Out[20]:
```
(38878841, 40835200)
```

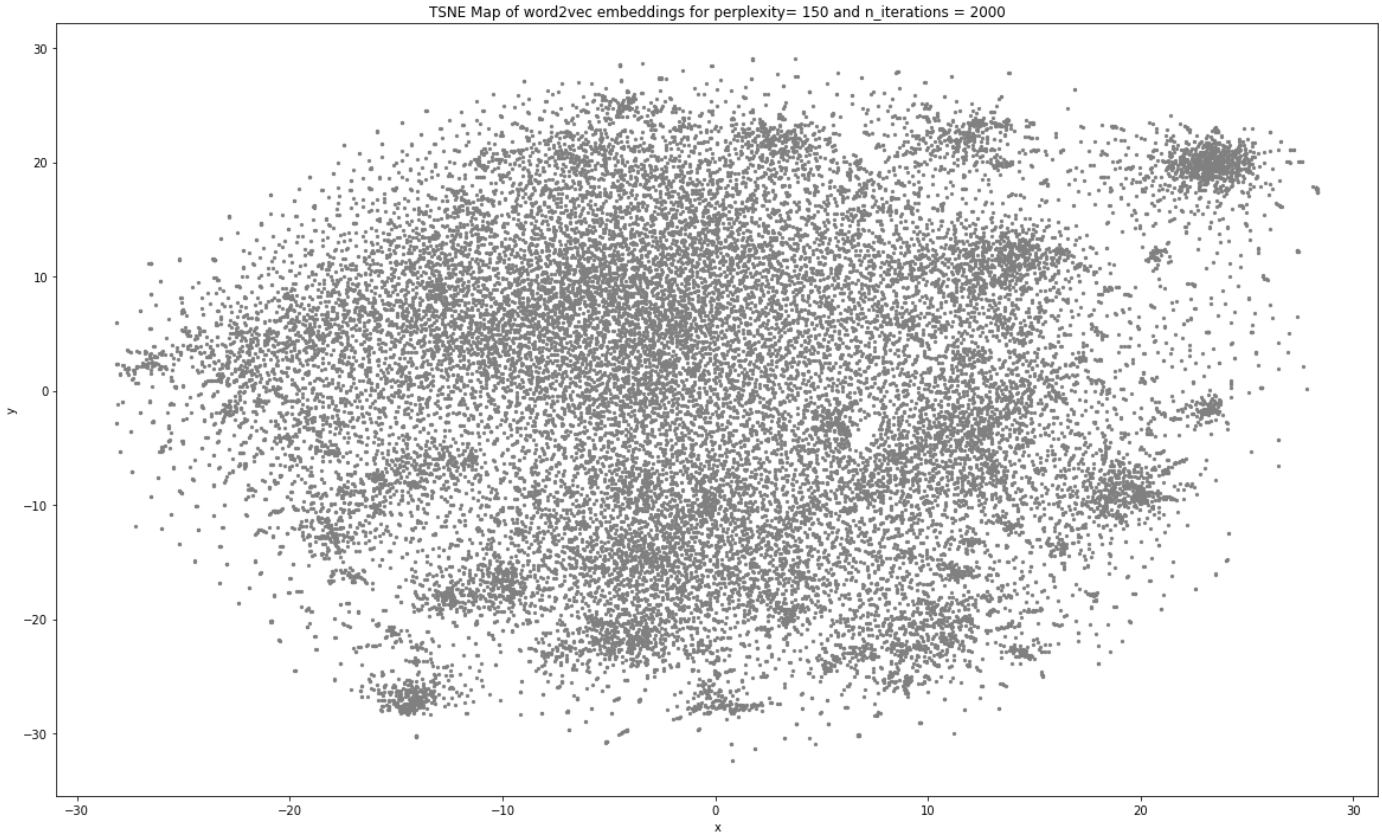
- Training t-SNE over multiple perplexity and n_iter values to find the optimum parameters.

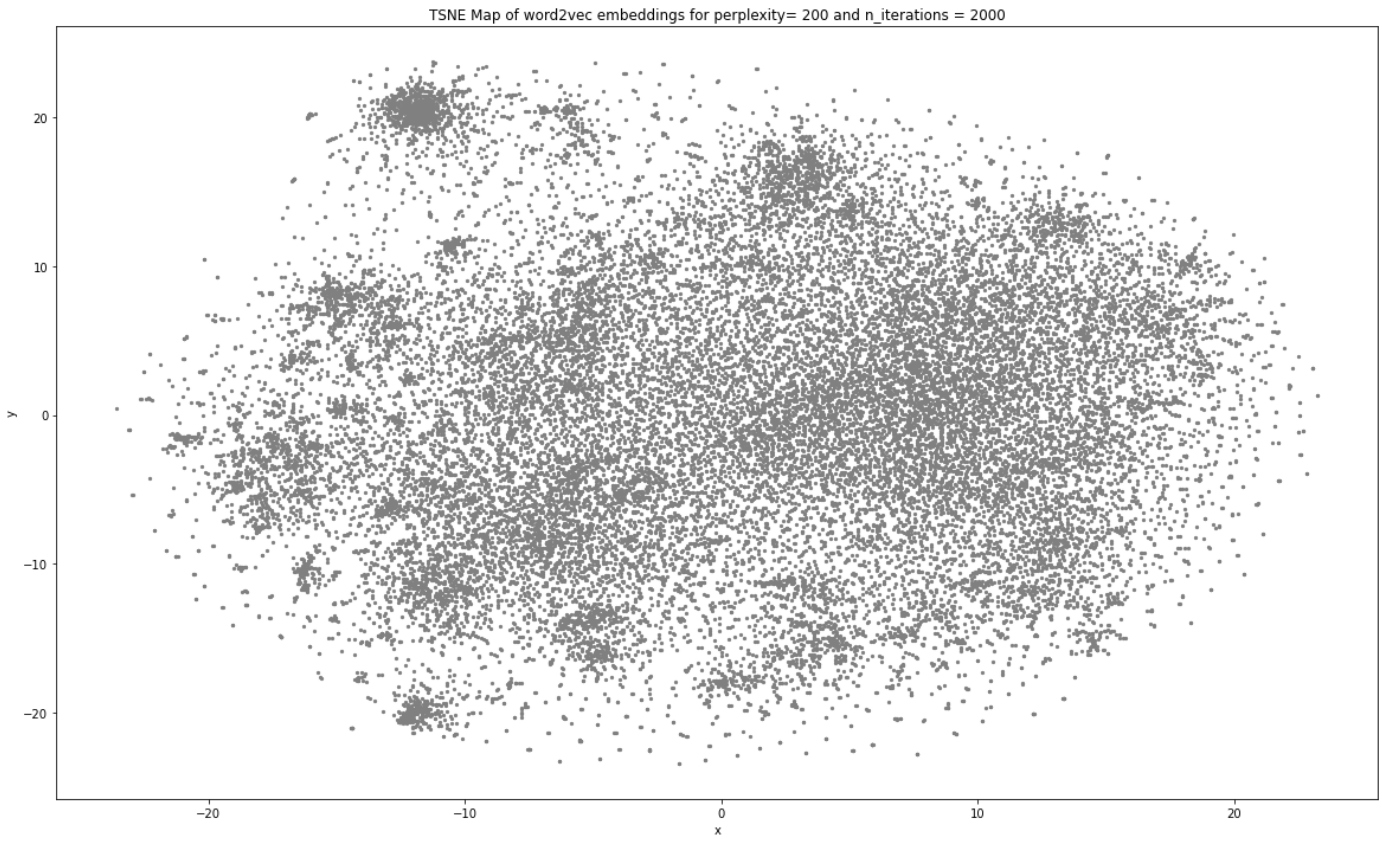
In [21]:
```
all_word_vectors_matrix = word2vec.wv.vectors

perplex = [30,50,100,150,200,300]
for i in perplex:

    tsne = sklearn.manifold.TSNE(n_components=2, random_state=0, perplexity=i, n_iter=10
    all_word_vectors_matrix_2d = tsne.fit_transform(all_word_vectors_matrix)

    points = pd.DataFrame(
        [
            (word, coords[0], coords[1])
            for word, coords in [
                (word, all_word_vectors_matrix_2d[word2vec.wv.vocab[word].index])
                for word in word2vec.wv.vocab
            ]
        ],
        columns=["word", "x", "y"]
    )


    title = 'TSNE Map of word2vec embeddings for perplexity= '+str(i)+' and n_iterations
    points.plot.scatter("x", "y", s=5, figsize=(20, 12), title=title, c='grey')
```

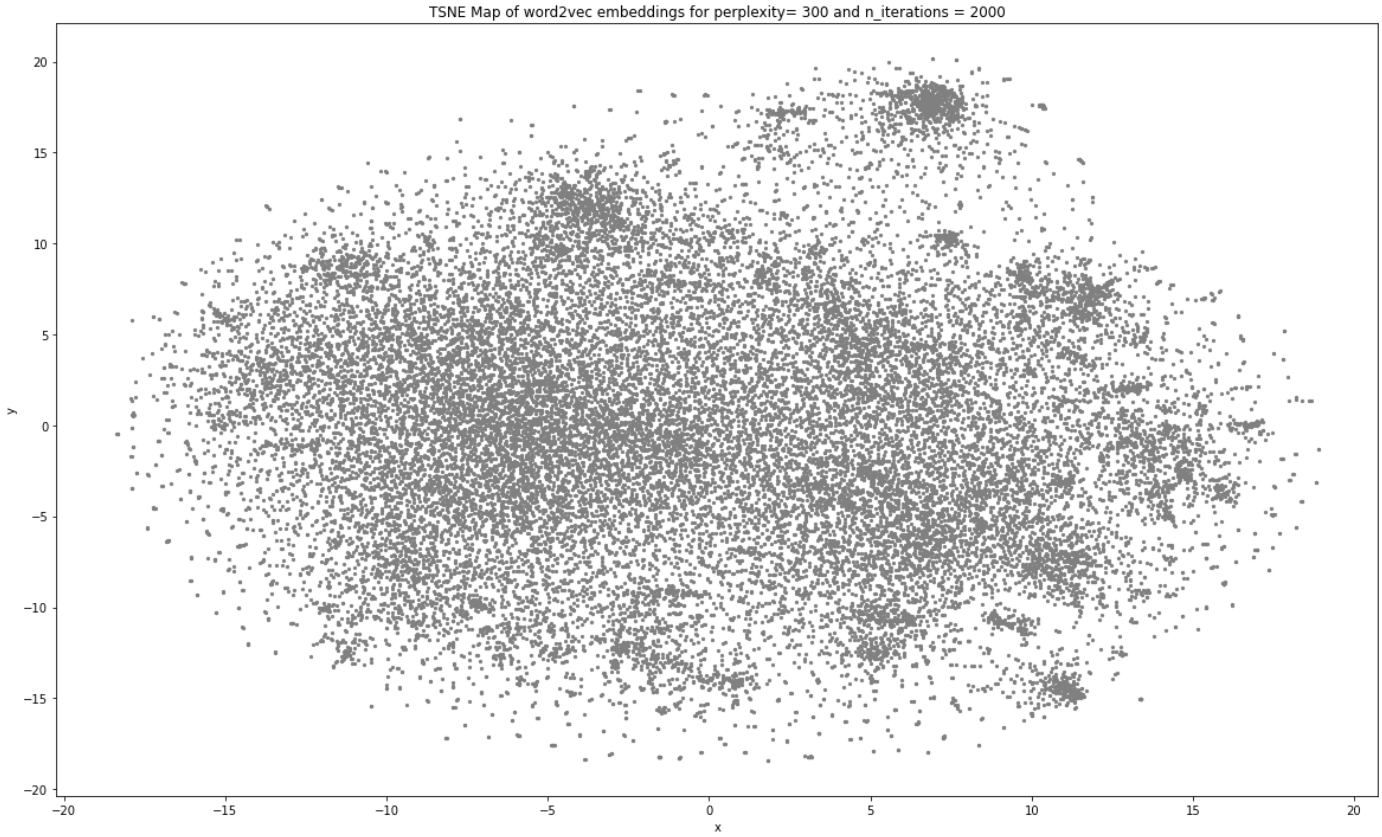

TSNE Map of word2vec embeddings for perplexity= 30 and n_iterations = 2000
60
40
20
0
−20
−40
−60
y
−40
−20
0
20
40
x


TSNE Map of word2vec embeddings for perplexity= 50 and n_iterations = 2000
60
40
20
0
−20
−40
y
−40
−20
0
20
40
x


TSNE Map of word2vec embeddings for perplexity= 100 and n_iterations = 2000
40
30
20
10
0
−10
−20
−30
y
−30
−20
−10
0
10
20
30
x


TSNE Map of word2vec embeddings for perplexity= 150 and n_iterations = 2000
30
20
10
0
−10
−20
−30
y
−30
−20
−10
0
10
20
30
x

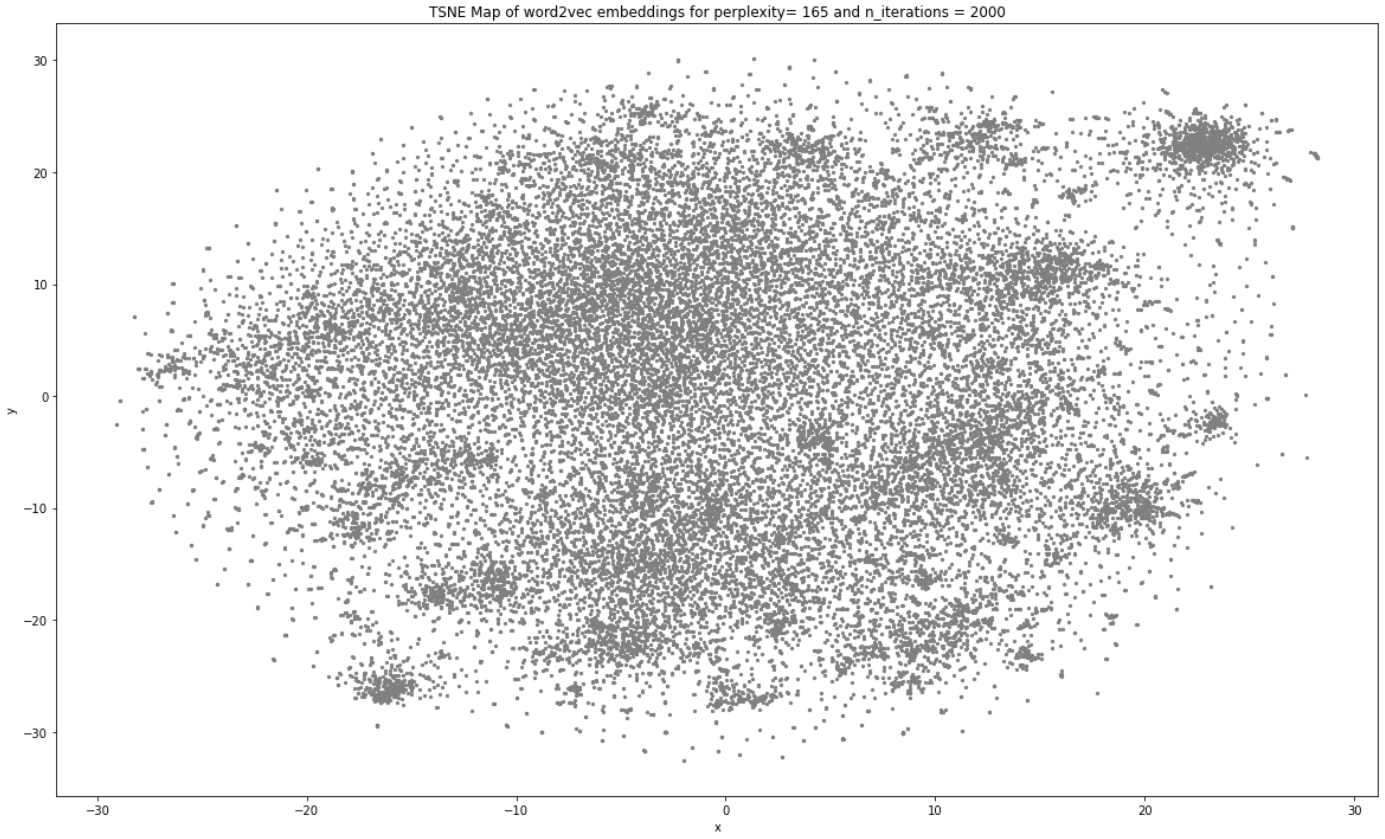


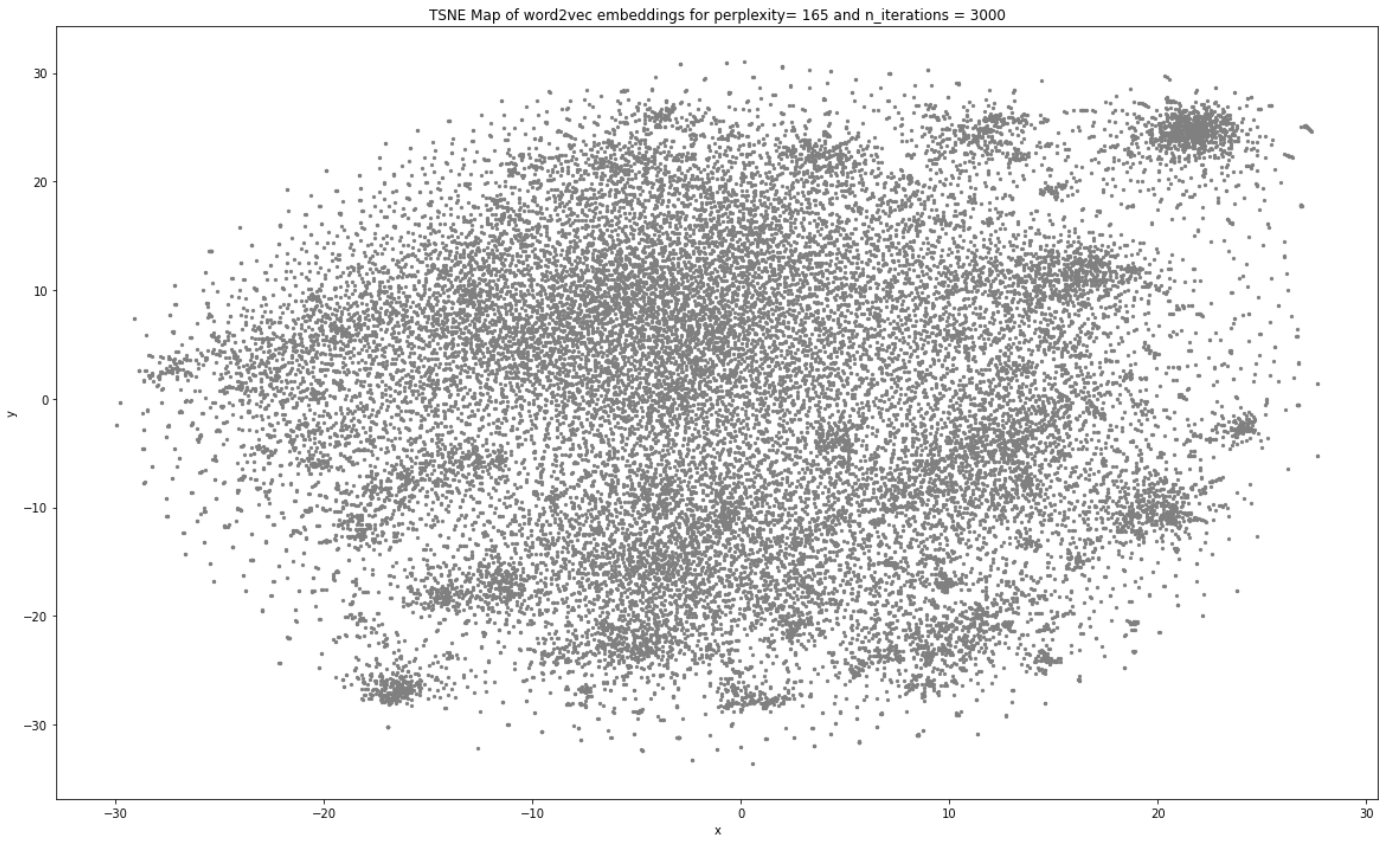


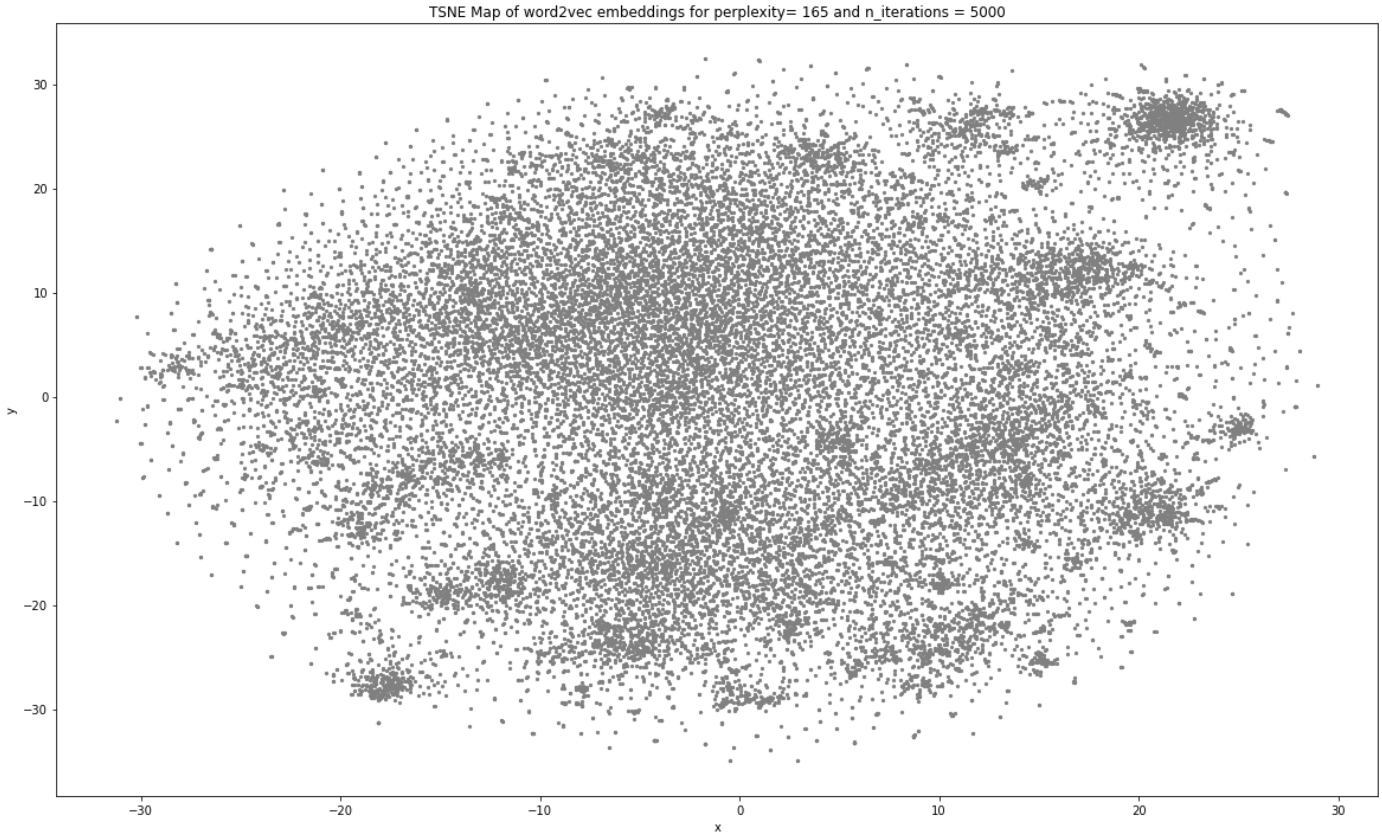
As we can see above that for perplexity ranging from 100-200 the tsne map does not show much change. Hence, in order to find out the optimum perplexity value, I decided to go ahead with 'perplexity = sqrt(N)' i.e. Perplexity = 165 (referenced from https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868). Now we'll be increasing the iterations to check if the results improve further.

In [23]:
```
iter = [2000,3000,5000]

for i in iter:

    tsne = sklearn.manifold.TSNE(n_components=2, random_state=0, perplexity=165, n_iter=
```

```
all_word_vectors_matrix_2d = tsne.fit_transform(all_word_vectors_matrix)

points = pd.DataFrame(
    [
        (word, coords[0], coords[1])
        for word, coords in [
            (word, all_word_vectors_matrix_2d[word2vec.wv.vocab[word].index])
            for word in word2vec.wv.vocab
        ]
    ],
    columns=["word", "x", "y"]
)
title = 'TSNE Map of word2vec embeddings for perplexity= 165 and n_iterations = '+st
points.plot.scatter("x", "y", s=5, figsize=(20, 12), title=title, c='grey')
```







There is no notable difference when increasing the number of iterations. Hence, we can infer that tsne is perfectly stable at perplexity = 165.

In [24]:
```
b = []
for k,v in data.iterrows():
    a = re.sub(r'(?=MORE:)(.*)','',v.Content)
    a = a.lower()
    a = re.sub("((\S+)?(http(S)?)(\S+))|((\S+)?(www)(\S+))|((\S+)?(\@)(\S+)?)", " ", a)
    a = re.sub('[^a-zA-Z0-9\n]',' ', a)
    a = nltk.word_tokenize(a)
```

```
    a = dict(Counter(a))
    b.append(a)

data['word_count'] = b
```

In [25]:
```
data.head()
```

Out[25]:

	Link	Title	Content	Category	tokenized_words	word_count
0	https://abcnews.go.com/Health/texas-front-line...	Texas front-line doctor dies of COVID-19 compl...	A Texas doctor who treated COVID-19 patients d...	Health	[[texas, doctor, treated, covid, 19, patients,...	{'a': 7, 'texas': 2, 'doctor': 3, 'who': 1, 't...
1	https://abcnews.go.com/Politics/pompeo-invites...	Pompeo invites hundreds to indoor State Dept. ...	Despite warnings by top health officials to av...	Politics	[[despite, warnings, top, health, officials, a...	{'despite': 1, 'warnings': 1, 'by': 5, 'top': ...
2	https://abcnews.go.com/US/york-autonomous-zone...	Protest over shutdown of New York's 'autonomou...	Protesters waving American flags and slamming ...	US	[[protesters, waving, american, flags, slammin...	{'protesters': 2, 'waving': 1, 'american': 1, ...
3	https://abcnews.go.com/US/104-year-wwii-vetera...	104-year-old WWII veteran survives COVID-19, r...	A World War II veteran defeated COVID-19 just ...	US	[[world, war, ii, veteran, defeated, covid, 19...	{'a': 4, 'world': 4, 'war': 3, 'ii': 2, 'veter...
4	https://abcnews.go.com/Politics/cdc-encourage-...	CDC to encourage 'Vaccinated for COVID-19' but...	If there are "I Voted" stickers, why not "Vacc...	Politics	[[voted, stickers, vaccinated, covid, 19, butt...	{'if': 3, 'there': 1, 'are': 2, 'i': 2, 'voted...

In [26]:
```
# This function returns a list of top 50 articles relating to the string entered by the

def search_index(c):

    pd.set_option('display.max_colwidth', -1)

    c = c.lower()
    c = re.sub("((\S+)?(http(S)?)(\S+))|((\S+)?(www)(\S+))|((\S+)?(\@)(\S+)?)", " ", c)
    c = re.sub('[^a-zA-Z0-9\n]',' ',c)

    d = list(c.split())
    input_str = list(set([word for word in d if word not in stop_words]))
    print(input_str)
    e = OrderedDict()
    f = []
    temp=[]

# The outer loop for iterating over each row in main dataset
    for index,content in data.iterrows():
        prime_count = 0
        count = 0
        points = 0

'''
Now iterating over the words from the input string to check if they're present in the ar
```

```
them in the order of number of words present in the 'Title' (prime_count) > Max no. of w
Total number of times the words were present in the 'Content' corpus (points).
'''
        for i in input_str:
            if i in content['Title']:
                prime_count +=1
                if i in (content['word_count']).keys():
                    #pdb.set_trace()
                    count += 1                              # considering how many words fr
                    points += content['word_count'][i]    # considering the number of tim
            else:
                if i in (content['word_count']).keys():
                    #pdb.set_trace()
                    count += 1
                    points += content['word_count'][i]

        if len(input_str)<=3:
            if count>=1:
                e[index] = {'prime_count':prime_count,'count':count,'points':points}

        elif 3 < len(input_str) <=5:
            if count>=2:
                e[index] = {'prime_count':prime_count,'count':count,'points':points}

        elif len(input_str) > 5:
            if count>=3:
                e[index] = {'prime_count':prime_count,'count':count,'points':points}


    # the lambda function first sorts the dictionary based on 'prime_count', 'count' and

    for key,val in sorted(e.items(), key=lambda kv: (kv[1]['prime_count'],kv[1]['count']
        f.append(key)
        #print(key,val)


    print('\nTotal number of results: ',len(f))

    if len(f)>50 :
        temp=f[:50]
        #print(temp)
        print('\nTop 50 results:\n')
        a = data.iloc[temp,[0,1,2,3]].copy()
    else:
        a = data.iloc[f,[0,1,2,3]].copy()

    #print(a,'\n')
    print(a)
```

In [27]:

```
'''
This function returns three of the following things:
  1. Top 5 words from the input corpus and list of words related to them.
  2. Pairs of words from within the input corpus which are most related.
  3. A 'factual score' of the piece of content entered by the user.

'''
def reality_check(c):

    w = nltk.sent_tokenize(c)
    w = sentence_cleaner(w)
    print(w)
    # w = [words for words in inp_list if words in word2vec.wv.vocab]
    sum_points = 0
    counter = 0
    top5 = 5
```

```
x = OrderedDict()
y = OrderedDict()
similar = OrderedDict()

# So in cases of multiple sentences entered by the user, the outer loop is to iterat
for i in w:

  # This loop is the iterate over the words inside the sentence
  for f in i:

    # Extracts the count of a word if it is present inside the word2vec vocab, else
    try:
      temp_count = word2vec.wv.vocab[f]
      y[f] = temp_count.count
    except:
      continue

  # Calculates the similarity score of a word with its neighbours within a window si
  if len(i)>=3:

    for j in range(len(i)-2):
      try:
        a = word2vec.wv.similarity(i[j],i[j+1])
        x[a] = [i[j],i[j+1]]
        sum_points += a
        counter += 1
        a = word2vec.wv.similarity(i[j],i[j+2])
        x[a] = [i[j],i[j+2]]
        sum_points += a
        counter += 1
      except:
        continue

    # The loop above iterates till len(i)-2. Hence, we are still left to calculate t
    try:
      a = word2vec.wv.similarity(i[len(i)-2],i[len(i)-1])
      x[a] = [i[len(i)-2],i[len(i)-1]]
      sum_points += a
      counter += 1
    except:
      continue

  elif len(i)==2:
    a = word2vec.wv.similarity(i[0],i[1])
    x[a] = [i[0],i[1]]
    sum_points += a
    counter += 1

  else:
    sim = word2vec.wv.most_similar(i[0],topn=20)
    print('\n','-'*200,'\n',f'The top 20 words related to {w} are:\n')
    for i in sim:
        print(i)
    print('\n','-'*200)

# So for input corpus which have more than 1 word, will have their 'counter' value g
if counter>=1:
  similar = sorted(y.items(), key=lambda kv:(kv[1]), reverse=True)
  print('-'*200)
  print('The following is a list of most common words found in your input corpus and
  if len(similar)>5:
    for i in range(0,5):
      print(f'{similar[i][0]}:')
      print(word2vec.wv.most_similar(similar[i][0],topn=5),'\n')

  else:
```

```
        for i in range(len(similar)):
          print(f'{similar[i][0]}:')
          print(word2vec.wv.most_similar(similar[i][0],topn=5),'\n')


      print('-'*200)

      # This assorts the pairs of values within the input corpus which are highly relate
      similarity = sorted(x.items(), key=lambda kv:(kv[0]), reverse=True)
      if len(similarity)>5:
        print('\n\nTop 5 similar words in the input corpus are:')
        for i in range(0,5):
          print(similarity[i])
      else:
        print('\n\nTop similar words in the input corpus are:')
        for i in range(len(similarity)):
          print(similarity[i])

      print('\n')
      avg = sum_points/counter
      print('-'*200)
      print(f'On a scale of [0-1] the score of your content is: {avg}')
      print('-'*200)
```

In [28]:
```
'''
Based on whether the user just wants to research about a specific topic or wants to fact
returns the appropriate output.
'''


def user_ask():
    a = input('Please choose your desired path-\n[0]: Know more about a topic\n[1]: Fact

    if int(a) == 0:
        b = input('Enter the topic you''re interested in:')
        search_index(b)

    elif int(a) == 1:
        b = input('Enter the content you want to fact check:')
        articles = search_index(b)
        articles
        reality_check(b)
        print('\n')
        trial = points[points['word'].isin(input_str)].copy()
        plt.figure(figsize=[20,12])
        plt.scatter(points.x, points.y, s=5, c='grey')
        plt.scatter(trial.x, trial.y,s=30, c='red')
        for i, point in trial.iterrows():
            a = point.x
            b = point.y
            plt.annotate(point.word, xy=(a+0.5,b+0.5))
```

In [29]:
```
user_ask()
```

```
Please choose your desired path-
[0]: Know more about a topic
[1]: Fact check a piece of content

1
Enter the content you want to fact check:Doctors have repeatedly suggested to wear mask
s. Still some people refuse to wear one.
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:3: FutureWarning: Passing a
negative integer is deprecated in version 1.0 and will not be supported in future versio
```

n. Instead, use None to not limit the column width.
  This is separate from the ipykernel package so we can avoid doing imports until
['wear', 'one', 'people', 'masks', 'still', 'suggested', 'doctors', 'refuse', 'repeatedl
y']

Total number of results: 5141

Top 50 results:

	Link	...	Category
4298	https://abcnews.go.com/Politics/airlines-ban-passengers-refuse-wear-masks/story?id=71279833	...	Politics
2038	https://abcnews.go.com/Politics/trump-masks-wearing-public/story?id=71561790	...	Politics
517	https://abcnews.go.com/Health/diy-masks-protect-covid-19-smart-wear/story?id=69957443	...	Health
1185	https://abcnews.go.com/US/officials-pushback-states-counties-require-people-wear-masks/story?id=71452988	...	US
1121	https://abcnews.go.com/US/covid-19-controlled-months-people-wear-masks-cdc/story?id=71783194	...	US
3665	https://abcnews.go.com/Business/delta-adds-460-people-fly-list-refusing-wear/story?id=73815553	...	Business
3529	https://abcnews.go.com/Politics/33-states-voters-wear-masks-polling-stations/story?id=73375662	...	Politics
3055	https://abcnews.go.com/Politics/texas-rep-louie-gohmert-refused-wear-mask-tests/story?id=72052207	...	Politics
2213	https://abcnews.go.com/Politics/trump-eschews-face-covering-ignoring-rubio-urging-wear/story?id=71451795	...	Politics
3208	https://abcnews.go.com/Politics/trump-downplaying-virus-mocked-wearing-masks-months/story?id=73392694	...	Politics
2129	https://abcnews.go.com/Politics/coronavirus-government-response-updates-coming-close-trump-rapid/story?id=69957759	...	Politics
110	https://abcnews.go.com/Politics/trumps-secret-service-agents-required-wear-masks-sources/story?id=73640288	...	Politics
1622	https://abcnews.go.com/Health/nursing-homes-masks-official/story?id=71266250	...	Health
1791	https://abcnews.go.com/Health/airport-coronavirus-testing-people-flying-experts/story?id=70632838	...	Health
3294	https://abcnews.go.com/Politics/democratic-women-launch-campaign-wear-pink-masks-capitol/story?id=71898191	...	Politics
2659	https://abcnews.go.com/US/florida-sheriff-forbids-staff-visitors-wearing-masks/story?id=72325542	...	US
1155	https://abcnews.go.com/Politics/89-americans-wear-masks-public-coronavirus-pandemic-persists/story?id=71455062	...	Politics
2416	https://abcnews.go.com/US/walmart-sams-club-require-customers-wear-masks/story?id=71796964	...	US
3082	https://abcnews.go.com/US/security-guard-charged-murder-fighting-man-refused-wear/story?id=71690133	...	US
3094	https://abcnews.go.com/Technology/edit-tweets-convince-wear-mask/story?id=71596962	...	Technology
944	https://abcnews.go.com/International/coronavirus-china-deploys-drones-cameras-loudhailers-chastise-people/story?id=68746989	...	International
887	https://abcnews.go.com/Politics/half-americans-wear-masks-coronavirus-normal-takes-hold/story?id=70073942	...	Politics
15756	https://abcnews.go.com/Entertainment/loved-confidants-remember-farrah-fawcetts-final-days-worst/story?id=63061673	...	Entertainment
3288	https://abcnews.go.com/Politics/money-americans-waiting-covid-19-stimulus-boost/story?id=70066813	...	Politics
1224	https://abcnews.go.com/Business/amc-theatres-backtracks-guidelines-require-movie-goers-wear/story?id=71348519	...	Business
1665	https://abcnews.go.com/US/oregon-governor-slams-state-troopers-refusing-wear-masks/story?id=71595504	...	US
6810	https://abcnews.go.com/Politics/melania-trump-bullied-people-world-distrusts-west-wing/story?id=58419018	...	Politics
4799	https://abcnews.go.com/Sports/pittsburgh-panthers-virginia-tech-hokies-wear-masks		

```
-field/story?id=74322510                                ...  Sports
1769   https://abcnews.go.com/US/fauci-david-muir-universal-wearing-masks-essential-comb
at/story?id=72294374                                    ...  US
4563   https://abcnews.go.com/Sports/pittsburgh-panthers-virginia-tech-hokies-required-w
ear-masks/story?id=74326639                             ...  Sports
6429   https://abcnews.go.com/Sports/dallas-cowboys-evp-stephen-jones-injured-qb-dak/sto
ry?id=73572205                                          ...  Sports
1655   https://abcnews.go.com/Politics/coronavirus-government-response-updates-trump-gui
delines-matter-life/story?id=69894694                   ...  Politics
2489   https://abcnews.go.com/US/coronavirus-live-updates-illinois-gov-jb-pritzker-file
s/story?id=71836502                                     ...  US
584    https://abcnews.go.com/Politics/cuomo-york-hospitals-reject-trumps-claim-masks-ba
ck/story?id=69877209                                    ...  Politics
3271   https://abcnews.go.com/Politics/schools-latest-battleground-fight-mandatory-mask
s/story?id=71840119                                     ...  Politics
1997   https://abcnews.go.com/US/people-flouting-coronavirus-social-distancing-precautio
ns-save-lives/story?id=70880349                         ...  US
2395   https://abcnews.go.com/US/coronavirus-updates-cdc-forecast-predicts-death-toll-20
0000/story?id=72367904                                  ...  US
2729   https://abcnews.go.com/Politics/medical-reasons-wearing-face-mask/story?id=720209
29                                                      ...  Politics
949    https://abcnews.go.com/Politics/tested-covid-19-subject-doctors-orders-pence/stor
y?id=69366725                                           ...  Politics
930    https://abcnews.go.com/US/leaving-hospital-tears-arizona-doctors-worry-icus-fill/
story?id=71519097                                       ...  US
3111   https://abcnews.go.com/Health/talk-covid-masks/story?id=74324421
                                                        ...  Health
1208   https://abcnews.go.com/Politics/white-house-trump-called-nih-terrible-questioned-
vaccines/story?id=70001201                              ...  Politics
1408   https://abcnews.go.com/Health/beards-n95-face-masks/story?id=69916196
                                                        ...  Health
2492   https://abcnews.go.com/Politics/trump-realistic-tone-coronavirus-briefing-repeats
-false-claims/story?id=71923191                         ...  Politics
2458   https://abcnews.go.com/Health/200-cases-doctors-warn-reports-rare-coronavirus-lin
ked/story?id=70703314                                   ...  Health
2957   https://abcnews.go.com/Politics/trump-reverses-convention-masks-briefings/story?i
d=71963747                                              ...  Politics
5261   https://abcnews.go.com/Sports/las-vegas-raiders-players-masks-charity-event/stor
y?id=73329905                                           ...  Sports
3548   https://abcnews.go.com/Politics/note-months-pandemic-mixed-signals-dominate-washi
ngton/story?id=72057090                                 ...  Politics
6307   https://abcnews.go.com/Sports/ohio-state-iowa-football-parents-speak-outraged-bi
g/story?id=72386675                                     ...  Sports
880    https://abcnews.go.com/Politics/string-attacks-doctors-experts-trump-takes-aim-sc
ience/story?id=72170408                                 ...  Politics

[50 rows x 4 columns]
[['doctors', 'repeatedly', 'suggested', 'wear', 'masks'], ['still', 'people', 'refuse',
'wear', 'one']]
--------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------
------------------------
The following is a list of most common words found in your input corpus and other words
which are most similar to them:

people:
[('victimized', 0.566369891166687), ('citizens', 0.499575138092041), ('americans', 0.496
60080671310425), ('mexicans', 0.4863411486148834), ('folks', 0.474070072174072 27)]

one:
[('two', 0.4925777316093445), ('another', 0.47101202607154846), ('tangling', 0.451512008
90541077), ('lingered', 0.451229989528656), ('dynasties', 0.44971001148223877)]

still:
[('however', 0.5285410284996033), ('remains', 0.5256170034408569), ('although', 0.513238
4300231934), ('remain', 0.5003204345703125), ('though', 0.4724016785621643)]
```

```
masks:
[('wear', 0.7374045848846436), ('mask', 0.7109178900718689), ('coverings', 0.7036418318748474), ('wearing', 0.702998161315918), ('gloves', 0.6449150443077087)]

doctors:
[('patients', 0.6398053765296936), ('physicians', 0.6158918142318726), ('clinicians', 0.6107984781265259), ('nurses', 0.5913727283477783), ('surgeons', 0.5746719837188721)]

--------------------------------------------------------------------------------------------------------------------------------------------------------

Top 5 similar words in the input corpus are:
(0.7374046, ['wear', 'masks'])
(0.4226969, ['repeatedly', 'suggested'])
(0.3112364, ['still', 'people'])
(0.23844972, ['still', 'refuse'])
(0.23793782, ['refuse', 'wear'])

--------------------------------------------------------------------------------------------------------------------------------------------------------

On a scale of [0-1] the score of your content is: 0.21872258638697012
--------------------------------------------------------------------------------------------------------------------------------------------------------
```