```
import gensim
import numpy as np
import pandas as pd
import os
from nltk import sent_tokenize
from gensim.utils import simple_preprocess
import plotly.express as px
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))

story = []
for filename in os.listdir('Harry Potter Books'):
    with open(os.path.join('Harry Potter Books', filename),
encoding='utf-8') as f:
        corpus = f.read()
        raw_sent = sent_tokenize(corpus)
        for sent in raw_sent:
            tokens = simple_preprocess(sent)
            filtered_tokens = [word for word in tokens if word not in
stop_words]
            story.append(filtered_tokens)

len(story)
```

```
70952
```

```
model =
gensim.models.Word2Vec(window=10,vector_size=300,min_count=2,workers=4
)

model.build_vocab(story)

# finding unique words out of my corpus

model.train(story,
            total_examples=model.corpus_count,
            epochs=100
           )
```

```
(52095966, 57207200)
```

```
model.wv.most_similar('hermoine')
model.wv.most_similar('harry')
model.wv.most_similar('hagrid')
model.wv.most_similar('dumbledore')
```

```
[('headmaster', 0.39291825890541077),
 ('snape', 0.366092324256897),
 ('dippet', 0.3538476228713989),
 ('severus', 0.31403183937072754),
 ('know', 0.3071868419647217),
```

```
 ('voldemort', 0.3031572699546814),
 ('doge', 0.298662930727005),
 ('lupin', 0.2857092022895813),
 ('power', 0.284101277589798),
 ('harry', 0.2789573669433594)]
```

```
model.wv.doesnt_match(['harry potter','hermione granger','ron
weasly','voldemort','professor albus dumbledore'])
```

```
'voldemort'
```

```
# checking vector of a word
```

```
model.wv['hagrid'].shape
```

```
(300,)
```

```
model.wv.similarity('harry','hermione')
model.wv.similarity('harry','ron')
```

```
0.49368906
```

```
# plotting graph out of these vectors
```

```
y = model.wv.index_to_key
```

```
from sklearn.decomposition import PCA
pca = PCA(n_components=3)
pca
```

```
PCA(n_components=3)
```

```
x = pca.fit_transform(model.wv.get_normed_vectors())
```

```
x.shape
```

```
(14576, 3)
```

```
px.scatter_3d(x[:100],x=0,y=1,z=2,color=y[:100])
```

Follow Me on LinkedIn

Connect with me on LinkedIn for more updates, projects, and insights!