Professor: Dieval Guizelini

Estudante: Bruno Moreira Ribas

Atividade 03: Representação de texto usando PCA

Escolha um texto pequeno e qualquer modelo para transformá-lo em uma representação vetorial. Plot os vetores com PCA, demonstrando as orações ou as palavras representadas pelo vetor.

Instalação e importação das bibliotecas

```
!wget https://nlp.stanford.edu/data/glove.6B.zip
!unzip glove.6B.zip
!wget https://www.gutenberg.org/cache/epub/19033/pg19033.txt
!ls
```

```
--2024-07-01 14:59:48--  https://nlp.stanford.edu/data/glove.6B.zip
Resolving nlp.stanford.edu (nlp.stanford.edu)... 171.64.67.140
Connecting to nlp.stanford.edu (nlp.stanford.edu)|171.64.67.140|:443... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: https://downloads.cs.stanford.edu/nlp/data/glove.6B.zip [following]
--2024-07-01 14:59:49--  https://downloads.cs.stanford.edu/nlp/data/glove.6B.zip
Resolving downloads.cs.stanford.edu (downloads.cs.stanford.edu)... 171.64.64.22
Connecting to downloads.cs.stanford.edu (downloads.cs.stanford.edu)|171.64.64.22|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 862182613 (822M) [application/zip]
Saving to: ‘glove.6B.zip.1’

glove.6B.zip.1      100%[===================>] 822.24M  5.02MB/s    in 2m 39s

2024-07-01 15:02:28 (5.17 MB/s) - ‘glove.6B.zip.1’ saved [862182613/862182613]

Archive:  glove.6B.zip
replace glove.6B.50d.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
replace glove.6B.100d.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
replace glove.6B.200d.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
replace glove.6B.300d.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
--2024-07-01 15:09:08--  https://www.gutenberg.org/cache/epub/19033/pg19033.txt
Resolving www.gutenberg.org (www.gutenberg.org)... 152.19.134.47, 2610:28:3090:3000:0:bad:cafe:47
Connecting to www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 74930 (73K) [text/plain]
Saving to: ‘pg19033.txt’

pg19033.txt         100%[===================>]  73.17K  --.-KB/s    in 0.05s

2024-07-01 15:09:08 (1.49 MB/s) - ‘pg19033.txt’ saved [74930/74930]

glove.6B.100d.txt  glove.6B.300d.txt  glove.6B.zip    pg19033.txt
glove.6B.200d.txt  glove.6B.50d.txt   glove.6B.zip.1  sample_data
```

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
import numpy as np
import pickle
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

Importando texto

```
with open ('pg19033.txt') as f:
  book = f.read()
```

Criação de métodos

Converter texto em representação vetorial, somando os vetores de cada palavra

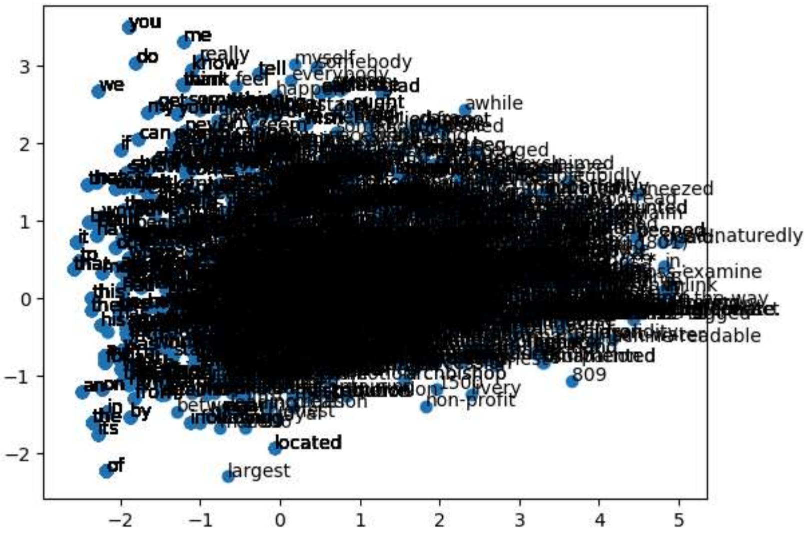
```
def load_glove_model(file_path):
    with open(file_path, 'rb') as f:
        model = {}
        for line in f:
            values = line.split()
            word = values[0].decode('utf-8')
            coefs = np.asarray(values[1:], dtype='float32')
            model[word] = coefs
    return model

modelo_glove = load_glove_model('glove.6B.100d.txt')
```

Aplicação

```
words = book.split()
vectors = np.zeros((len(words), 100))

for i, word in enumerate(words):
    if word in modelo_glove:
        vectors[i, :] = modelo_glove[word]
    else:
        vectors[i, :] = np.zeros(100)

vectorial_text = np.mean(vectors, axis=0)
```

Nuvem com as palavras do Vetor

```
pca = PCA(n_components=2)
vectorial_text = pca.fit_transform(vectors)
plt.scatter(vectorial_text[:, 0], vectorial_text[:, 1])

for i, palavra in enumerate(words):
    plt.annotate(palavra, (vectorial_text[i, 0], vectorial_text[i, 1]))

plt.show()
```


you
me
do
really
know
tell
myself
somebody
everybody
we
feel
awhile
if
can
sneezed
naturedly
examine
understandable
archbishop
every
non-profit
an
on
in
by
the
its
of
located
largest