SW06 - Gruppe 1

Natural Language Processing Basics 4

M. Nebroj, S. Hauri, S. Ineichen, R. Schwarzentruber

2019-11-06

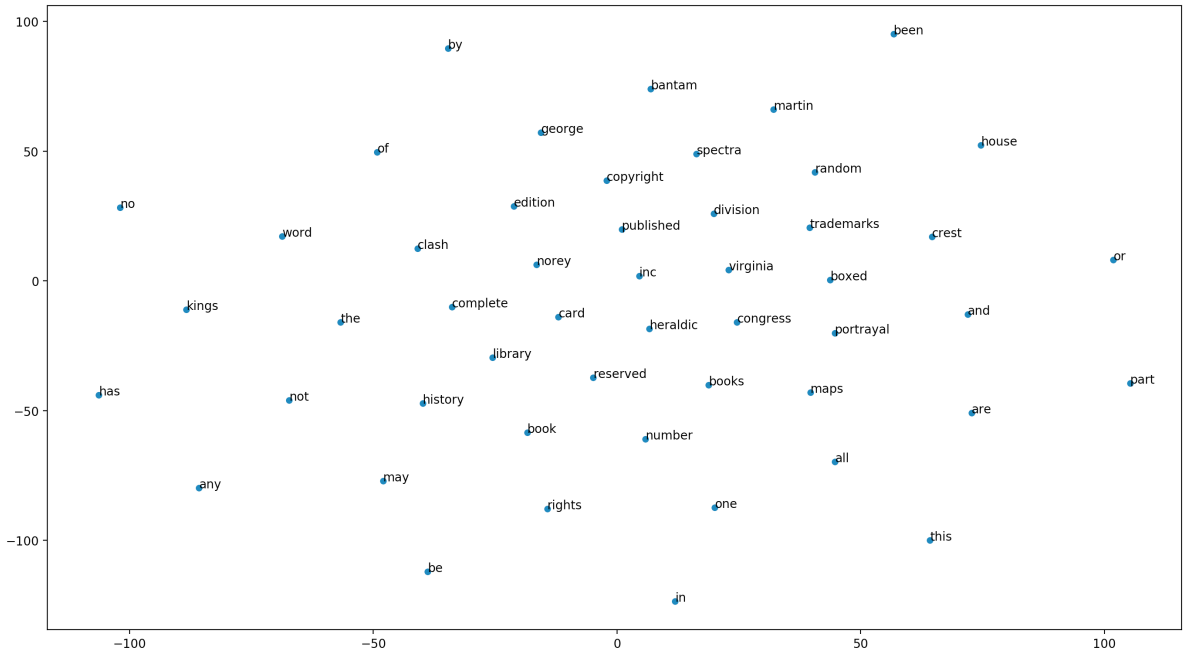
Testatübung SW06

Aufgabe 1

Erstellen Sie für das GOT Model eine 2D Visualisierung von 50 Begriffen

Lösung

Durch python generiertes Bild:



Python Code

Der genutzte Python Code setzt voraus, dass das GOT-vectors.w2v model generiert wurde (Aufgabe SW05)

```
import os
import sys
import gensim
from nltk.data import find

root_dir = os.path.dirname(os.path.dirname(os.path.realpath(__file__)))
got_model = os.path.join(root_dir, "data", "models", "GOT-vectors.w2v")

if not os.path.exists(got_model):
    print("FATAL: GOT-vectors.w2v not found - please run Aufgabe_01.py from SW05 first to
     ↪  generate the model!")
    sys.exit(1)

model = gensim.models.Word2Vec.load(got_model)

import numpy as np
labels = []
count = 0
max_count = 50
X = np.zeros(shape=(max_count, len(model['dog'])))

for term in model.wv.vocab:
    X[count] = model[term]
    labels.append(term)
    count += 1
    if count >= max_count: break

#it's recommended to use PCA first to reduce to ~50 dimensions
from sklearn.decomposition import PCA
pca = PCA(n_components=50)
X_50 = pca.fit_transform(X)

#using TSNE to further reduce to 2 dimensions
from sklearn.manifold import TSNE
model_tsne = TSNE(n_components=2, random_state=0)
Y = model_tsne.fit_transform(X_50)

#show scatter plot
import matplotlib.pyplot as plt
plt.scatter(Y[:,0], Y[:,1], 20)

#add label
for label, x, y in zip(labels, Y[:, 0], Y[:, 1]):
    plt.annotate(label, xy = (x,y), xytext = (0, 0), textcoords = 'offset points', size = 10)

plt.show()
```

Aufgabe 2

Für Game of Thrones soll eine spezifische Wissensdatenbank zu den genutzten Waffen erstellt werden. Wie würden sie die Konzeptextraktion vornehmen? (Vorgehen kurz erläutern)

Lösung

Idee / Vorgehen

Um die genutzten Waffen in Game of Thrones (GoT) herauszufinden, wäre es sinnvoll, zuerst die Verben, welche in dem Kontext einer Waffen genutzt werden, zu definieren. Zum Beispiel: *"X tötet Z mit Y"* wobei in diesem Satz *Y* die Waffe ist. Bei diesem simplen Beispiel ist das Verb *"töten"* interessant. Weitere Verben könnten zum Beispiel *schlagen*, *tragen* oder *würgen* sein. Die Verben fungieren dann als Features um eine Graph Database zu durchsuchen. Dabei können alle Knoten, die keine Kanten zu einem der definierten Verben haben, gelöscht werden. Somit würde man einen Graph erhalten indem die Waffen ersichtlich sind. Dieses Verfahren hat natürlich auch Probleme, weil gewisse Verben auch in einem anderen Kontext Sinn machen. Nützlich wäre es noch den Graphen mit einer Waffen Datenbank abzugleichen um den Graphen weiter zu filtern. Zusätzlich könnte man versuchen noch eine Relation zu den Personen herzustellen, da Waffen zwischen Personen/Gruppen eingesetzt werden (Wie im ersten Beispiel wo *X* und *Z* involvierte Personen sind). Personen und Gruppen können meist einfach aus Text extrahiert werden.