Wortähnlichkeit: Praktische Implementierung

Benjamin Roth

Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilian-Universität München
beroth@cis.uni-muenchen.de

Wort-Ähnlichkeitsmodell

Demo

Wort-Ähnlichkeitsmodell: Schritte

- Welche Schritte müssen gemacht werden, um ein Wortähnlichkeitsmodell zu Implementieren?
- (D.h. um auf Basis eines Korpus die ähnlichsten Wörter zu einem Anfrage-Wort zurückzuliefern)

Wort-Ähnlichkeitsmodell: Schritte

- Vorverarbeitung (Tokenisierung, ...)
- Vokabular definieren, Zuweisung Wort $\Rightarrow$ Numerierung
- Co-Okkurrenzen (Zusammenauftreten) von Wörtern zählen
- Co-Okkurrenz-Matrix X erstellen. Zeile: (Ziel-)Wort, Spalte: (Kontext-)Wort. Wert: Häufigkeit des gemeinsamen Vorkommens.
- Gewichtung der Matrix.
  Positive Pointwise Mutual Information (PPMI): Misst wie sehr Zusammenauftreten von statistisch erwarteter Häufigkeit abweicht.

Wort-Ähnlichkeitsmodell: Schritte

- Singular Value Decomposition (Singulärwertzerlegung; SVD):
  - Die Matrix wird in ein Produkt aus drei Matrizen zerlegt:

$$X = U\Sigma V^T$$

  - Σ ist eine Diagonalmatrix, mit absteigend sortierten, nicht-negativen Werten. Größe eines Wertes $\Leftrightarrow$ Anteil/Wichtigkeit bei der Rekonstruktion von X.
  - U: Matrix. Zeile: Wort, Spalte: Kontext-Repräsentation.
    Die Spalten enthalten die Kontextinformation, komprimiert und nach Wichtigkeit sortiert!
  - V: Matrix. Ebenso optimierte Repräsentation des Kontextes.
- Ähnlichkeit berechnen.
  - Vektor für Anfragewort: Entsprechende Zeile in $U\Sigma$.
  - Kosinus-Ähnlichkeit mit Vektoren für alle anderen Wörter berechnen (d.h. mit allen Wörtern in $U\Sigma$).
  - **Warum nicht nur U verwenden?**

Singulärwertzerlegung: Wiederholung

Singulärwertzerlegung in Python: "Naiver" Ansatz

- Idee bei SVD:
  - Betrachte nur die n (z.B. 50) wichtigsten Singular Vectors.
  - $\Rightarrow$ statistisches "Rauschen" wird entfernt, indem andere Dimensionen ignoriert werden.
  - $\Rightarrow$ besserer Ähnlichkeitsvergleich, weniger Ausreißer.

```
import numpy as np
U, sigma, V = np.linalg.svd(X)
U_trunc = U[:,:n]
sigma_trunc = sigma[:n]
V_trunc = V[:,:n]
```

$\Rightarrow$ Problem des obigen Ansatzes?

Singulärwertzerlegung in Python: “Naiver” Ansatz

```
import numpy as np
U, sigma, V = np.linalg.svd(X)
U_trunc = U[:,:n]
sigma_trunc = sigma[:n]
V_trunc = V[:,:n]
```

$\Rightarrow$ Problem des obigen Ansatzes?
*Die Co-Okkurrenz-Matrix ist **sparse** (99% oder mehr der Einträge sind 0). Die Matrizen U und V sind **dense**, bei ihrer Bechnung wird also ein um den Faktor 100 (oder mehr) größerer Speicherplatz benötigt. Die trunkierten (abgeschnittenen) Matrizen sind zwar wieder sehr klein, weil nur ein Bruchteil der Spalten beibehalten wird; die Berechnung des Zwischenschrittes ist oft aber aus Speicherplatzgründen so nicht möglich.*

Singulärwertzerlegung in Python: Effizienter Ansatz

- Es gibt spezielle Methoden, die nur die n größten Singulärwerte und die dazugehörigen Vektoren berechnen. (Bzw. direkt die trunkierte Matrix $U\Sigma$, die von Interesse ist.)
- Diese Methoden sind auch für die Berechnung von Sparse-Matrizen optimiert.
- Für unsere Zwecke: `sklearn.decomposition.TruncatedSVD`

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=n)
U_sigma_trunc = svd.fit_transform(X)
```

Ähnlichste Wörter: “Naiver” Ansatz

1. Nachschlagen von Vektor für Anfragewort: Zeile in `X` oder `U_sigma_trunc`, je nach dem ob die Co-Okkurrenz-Information mit oder ohne SVD verwendet werden soll.
2. Iterieren über alle Wörter w im Vokabular.
   1. Nachschlagen von Vektor für w.
   2. Kosinus-Ähnlichkeit berechnen, und zusammen mit w an Liste anhängen.
3. Liste nach Ähnlichkeit sortieren.

Ähnlichste Wörter: “Naiver” Ansatz

```
class DenseSimilarityMatrix:
    def __init__(self, U_sigma_matrix, word_to_id):
        self.word_matrix = U_sigma_matrix
        self.word_to_id = word_to_id
        self.id_to_word = {w:i for i,w in self.word_to_id.items()}

    def most_similar_words(self, query, n):
        q_row = self.word_to_id[query]
        q_vec = self.word_matrix[q_row,:]
        dot_q_q = q_vec.dot(q_vec.T)
        sims_words = []
        for w in self.word_to_id:
            w_row = self.word_to_id[w]
            w_vec = self.word_matrix[w_row,:]
            dot_w_w = w_vec.dot(w_vec.T)
            dot_q_w = q_vec.dot(w_vec.T)
            sim = dot_q_w / math.sqrt(dot_q_q * dot_w_w)
            sims_words.append((sim, w))
        return [w for s,w in sorted(sims_words, reverse=True)[:n]]
```

Ähnlichste Wörter: “Naiver” Ansatz

Problem mit obigem Ansatz?

- Aus theoretischer Sicht nicht viel einzuwenden.
- Trotzdem in der Praxis extrem ineffizient (Faktor $>$ 10-100):
  - Erstellen von Einzelobjekten für jeden Vektor.
  - Erweitern der Liste.
  - Kosinus-Berechnung separat für jeden Vektor
- $\Rightarrow$ Matrix-Multiplikation ist eine der am meisten optimierten Operationen in mathematischen Programm-Bibliotheken.
- $\Rightarrow$ Wann immer möglich, sollte man Matrizen als Ganzes multiplizieren!

Ähnlichste Wörter: Effizienter Ansatz

- Vermeide Nachschlagen (Lookup) der Vektoren für Einzelwörter im Vokabular.
- Für Kosinus-Berechnung brauchen wir:
  - Dot-Produkt Vektoren aller Wörter mit Anfrage-Vektor:
    $\Rightarrow$ Effiziente Operation: Multiplikation von Matrix mit Vektor!
  - Dot-Produkt Anfrage-Vektor (Einmalige Operation).
  - Dot-Produkt aller Vektoren im Vokabular.
    $\Rightarrow$ Effiziente Operation:
    - Komponentenweise Multiplikation der Matrix mit sich selbst.
    - Sumierung des Ergebnisses.
    - Beispiel: Whiteboard.
  - komponentenweise Anwendung von Wurzel und Bruchrechnung.

Ähnlichste Wörter: Effizienter Ansatz (für Numpy Dense Arrays)

```
def most_similar_words(self, word, topn):
    row = self.word_to_id[word]
    vec = self.word_matrix[row,:]
    m = self.word_matrix
    dot_m_v = m.dot(vec.T) # vector
    dot_m_m = np.sum(m ** 2, axis=1) # vector
    dot_v_v = vec.dot(vec.T) # float
    sims = dot_m_v / (math.sqrt(dot_v_v) * np.sqrt(dot_m_m))
    return [self.id_to_word[id] for id in (-sims).argsort()[:topn]]
```

- Für Scipy Sparse-Matrizen ist die Syntax etwas verschieden.
- Hinweis: `vec` ist ein Zeilenvektor, `vec.T` ergibt den dazugehörigen Spaltenvektor. (Dot-Produkt ist definiert für Zeilen- mit Spaltenvektor)

Ähnlichste Wörter: Effizienter Ansatz

- `v.argsort()` liefert die Indizes der sortierten Einträge eines Vektors:

```
>>> v=np.array([5,1,1,4])
>>> v.argsort()
array([1, 2, 3, 0])
```

Zusammenfassung

- Singulärwertzerlegung (SVD)
  - Wiederholung
  - Anwendung auf Co-Okkurrenzmatrizen
  - Effiziente Berechnung der trunkierten SVD
- Ähnlichkeitsberechnung mit Matrix-Multiplikation