Jak działa findAssocs() krok po kroku

Funkcja `findAssocs()` służy do znajdowania słów, które są **najbardziej skorelowane z wybranym słowem (terminem)** w macierzy typu TDM lub DTM na podstawie **współczynnika korelacji Pearsona.**

1. Wejście: TDM lub DTM

TermDocumentMatrix (TDM):

- **wiersze**: terminy (słowa)
- **kolumny**: dokumenty (np. wypowiedzi, akapity)
- **wartości**: liczba wystąpień danego słowa w danym dokumencie

Przykład:

```
          doc1  doc2  doc3
word1        1     0     2
word2        3     1     0
word3        0     1     4
```

2. `findAssocs(tdm, "word1", corlimit = 0.5)`

- Funkcja pobiera **wektor słowa "word1"**, np. `c(1, 0, 2)`, czyli ile razy "word1" występuje w każdym dokumencie.
- Następnie **oblicza korelację Pearsona** tego wektora z każdym innym słowem (czyli z każdym innym wektorem).

Korelacja Pearsona to miara współzmienności, obliczana jako:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

3. Zwracany wynik

Funkcja zwraca listę słów, których współczynnik korelacji z wybranym słowem przekracza zadany próg (`corlimit`).

0-0,3 to słaba korelacja
0,3-0,5 to korelacja umiarkowana
0,5-0,7 to korelacja silna
0,7-1 to korelacja bardzo silna

Przykład:

```
findAssocs(tdm, "climate", corlimit = 0.7)
```

Funkcja może zwrócić np. taki wynik:

```
$`climate`
     change    carbon   energy
      0.85      0.78      0.71
```

Interpretacja wyniku:

W obrębie dokumentów reprezentowanych w macierzy TermDocumentMatrix (TDM), termin **"climate"** wykazuje bardzo silny **poziom liniowej współzależności** z terminami:

- "change" (r = 0.85),
- "carbon" (r = 0.78),
- "energy" (r = 0.71),

przy zastosowaniu współczynnika korelacji Pearsona, liczonego między wektorem częstości występowania słowa "climate" a wektorami odpowiadającymi pozostałym terminom w tej samej macierzy.

Wartości współczynnika w przedziale powyżej 0.7 oznaczają, że słowa te mają **zbliżone rozkłady częstości** w dokumentach korpusu (tzn. często występują **w tych samych dokumentach** lub w dokumentach o podobnym profilu tematycznym). Oznacza to również, że termin „climate" **współwystępuje w sposób systematyczny** z terminami „change", „carbon" i „energy" w ramach analizowanego zbioru tekstów.

Wynik ten można interpretować jako **empiryczny sygnał współtematyczności**, wskazujący na potencjalną przynależność terminów do wspólnego pola semantycznego (np. „climate change", „climate carbon", „climate energy".

Wyniki findAssocs() dostarczają statystycznie uzasadnionych relacji między słowami, które służą najczęściej jako punkt wyjścia do bardziej zaawansowanej analizy semantycznej: do grupowania silnie powiązanych terminów w **tematy** (wstępne zbiory tematyczne) oraz wstępnej interpretacji tego, jakiego rodzaju **tematy mogą występować** w danych.

Ważne:

- Zarówno w DTM, jak i TDM funkcja findAssocs() bada **relacje między terminami** (wektorami słów). W DTM korelacja liczona jest między kolumnami, a w TDM - między wierszami. Rezultat interpretujemy tak samo: lista terminów powiązanych ze wskazanym terminem poprzez korelację współwystępowania.
- Znajdowanie asocjacji działa najlepiej dla **liczbowych danych** (częstości wystąpień)
- **Przy ekstremalnie rzadkich słowach** korelacja może być **zawyżona** (mało danych = niestabilny wynik)
- Nie jest to model uczenia maszynowego, tylko **statystyka oparta na współczynniku Pearsona**.