Wariacje na temat algorytmu k-średnich

Natalia Potocka

05/25/2015

Dwa słowa przypomnienia

- wczytano 1 075 568 artykułów z polskiej Wikipedii
- razem to 2 806 765 różnych słów...
- ... z czego poklastrowano 1 707 683 słów w 186 942 grup
- pozostałe słowa wystąpiły raz we wszystkich tekstach i je pominięto

Następnym krokiem jest kategoryzacja tesktów, gdzie kryterium podziału to częstości występowania grup słów w artykułach.

Algorytm k-średnich

Wmetodzie k-średnich minimalizujemy

$$\sum_{x \in X} ||f(C, x) - x||^2,$$

gdzie X tozbiór wektorów cech $x \in \mathbb{R}^m$, C tozbiór środków klastrów $c \in \mathbb{R}^m$, gdzie $|C| = k$, a f zwraca najbliższy wektorowi x środek $c \in C$, używając odległości Euklidesowej.