Metody analizy i wizualizacji dużych zbiorów danych
Laboratorium 3 – Metody t-SNE i bh-SNE

Jan Gołda, Wojciech Basiura - WIET 2018/2019

Opis zagadnienia

Laboratorium 3 poświęcone było wizualizacji średnich zbiorów danych przy użyciu metod t-SNE – z aproksymacją Barnes-Hut i dokładną. Porównywaliśmy również wyniki z poprzednio poznanymi algorytmami, szczególnie ze względu na metryki k-nearest neighbors i neighbours hit.

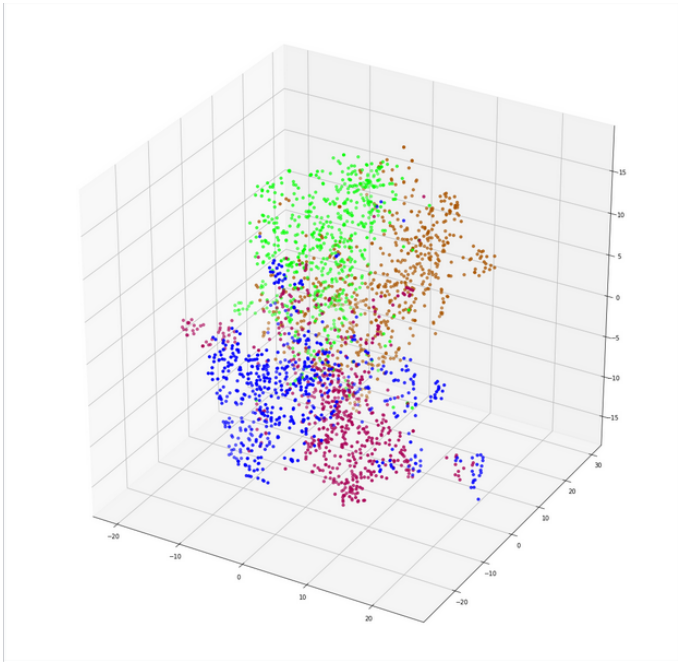
Zadaniem była wizualizacja czterech grup ze zbioru "20 Newsgroups" wspomnianymi metodami, porównanie wyników wizualizacji t-SNE z różnymi wartościami perple

Zadanie pierwsze

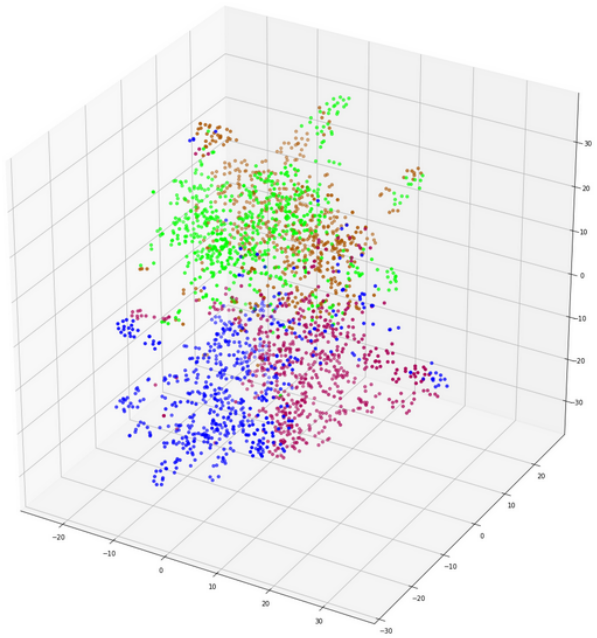
1. Podzbiór „20 Newsgroups”

Wybraliśmy kategorie rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey. Uznaliśmy, że mają wspólny mianownik, ale również powinny być pomiędzy nimi wyraźne różnice.

2. Metody t-SNE i bh-SNE

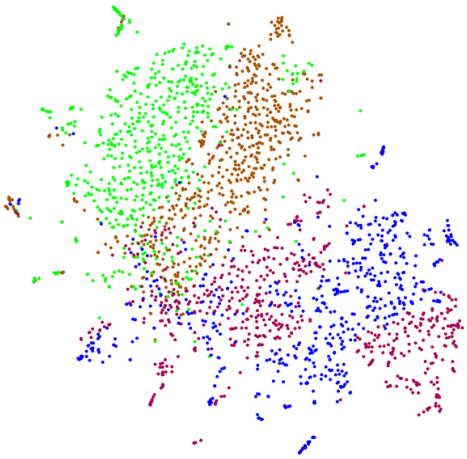


bh-SNE, perplex=30

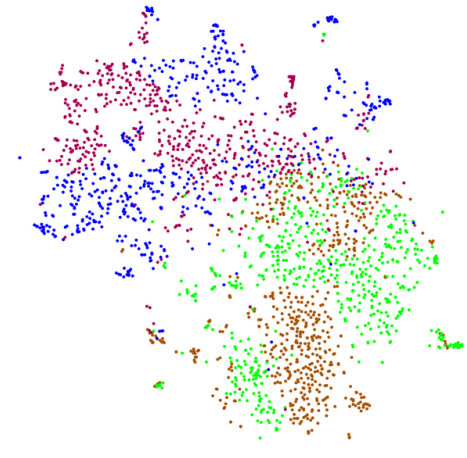


t-SNE exact, perplex=30

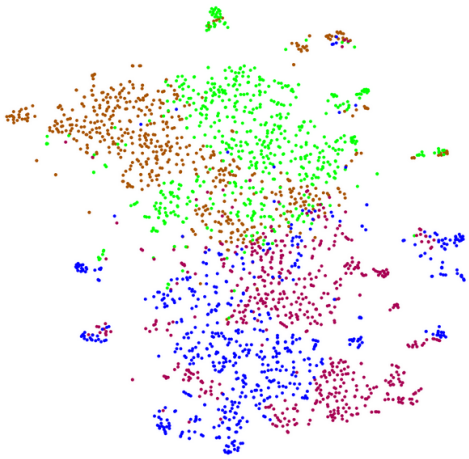
Wyniki dla różnych wartości perplex:



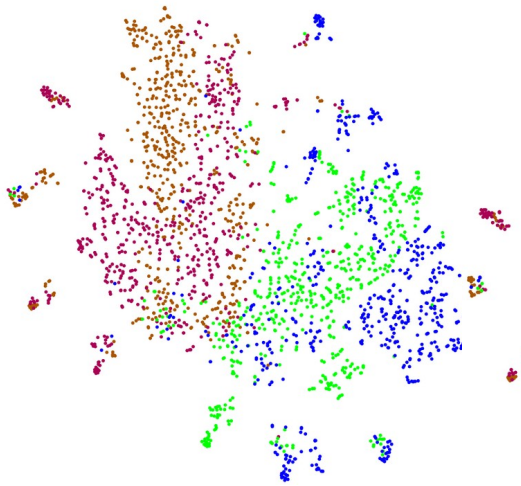
perplex=150.0



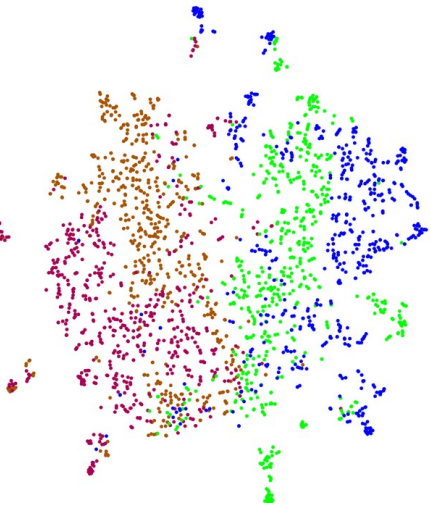
perplex=100.0



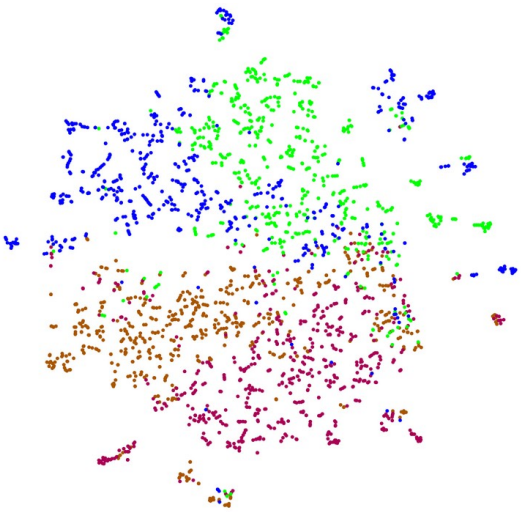
perplex=50.0



perplex=30.0



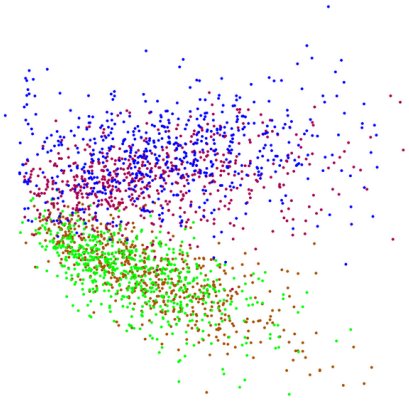
perplex=20.0



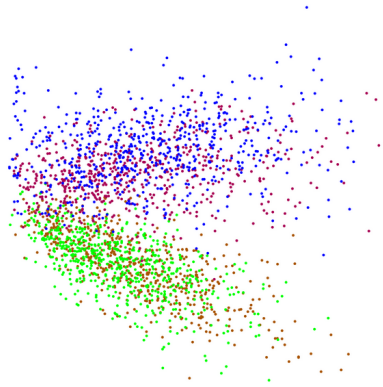
perplex=10.0

Przy niższych perplex tworzą się grupki punktów zlewające się ze sobą, natomiast przy wyższych grupy są bardziej rozrzucone.

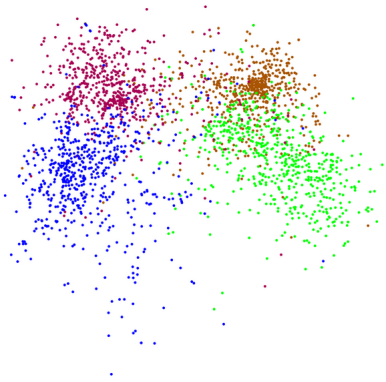
3. Wizualizacje metodami z poprzednich laboratoriów



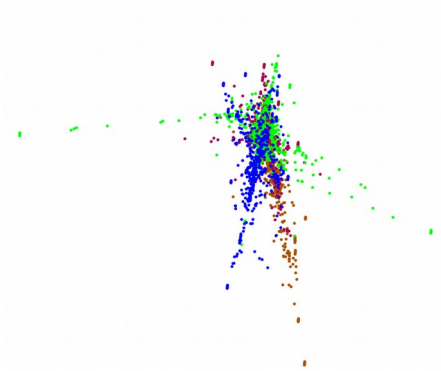
PCA



Kernel PCA, kernel=”cosine”



Isomap, nn=5



LLE, nn=4

Wypadają znacznie gorzej niż t-SNE. Wyraźnie widać dwie pary grup mieszające się ze sobą. Prawdopodobnie jedna z nich dotyczy motoryzacji (motory i samochody), a druga sportu (hokej i baseball). Są one rozróżnialne jako całości (PCA Kernel PCA), ale nie da się ich rozdzielić. Dobrze poradził sobie algorytm Isomap. Na jego rysunku wyraźnie widać cztery grupy.

4. Porównanie w zadanych metrykach

	bh-SNE	t-SNE
hit-to-all	83%	84%
Neighbors-hit (1)	6.015	6.018

Występuje znaczna różnica złożoności czasowej przy znikomym zysku w obu metrykach.

	PCA	Kernel PCA	Isomap	LLE
hit-to-all	57%	57%	80%	61%
Neighbors-hit (1)	1.582	1.583	1.356	1.956

(1) out of 10

Wyniki neighbors hit są ponad 3 razy lepsze przy użyciu t-SNE. Wyniki hit-to-all są też dużo lepsze, za wyjątkiem Isomap której wynik jest porównywalny. Żadna z poznanych wcześniej metod nie ma dobrych wyników w obu metrykach. t-SNE wypada zdecydowanie najlepiej.