



**AGH**

**AKADEMIA GÓRNICZO-HUTNICZA  
IM. STANISŁAWA STASZICA  
W KRAKOWIE**

**Metody analizy i wizualizacji dużych zbiorów danych**

**Laboratorium 4 - LargeVis**

# 1. Opis zagadnienia

Laboratorium 4 poświęcone było metodzie LargeVis, w szczególności porównaliśmy ją do MDS, t-SNE i bh-SNE.

Zadaniem domowym było przetestowanie jakie wpływ na wyniki mają wartości parametrów **sample**, **neighbours** oraz **perplexity**.

## 1.1. Zbiór danych

Zbiorem użytym do testów będzie Fashion-MNIST dostępny pod adresem:

<https://www.kaggle.com/zalando-research/fashionmnist>

Jest to zbiór czarno-białych obrazków przedstawiających elementy garderoby, które podzielone są na następujące klasy:

0. T-Shirt / top
1. Trouser
2. Pullover
3. Dress
4. Coat
5. Sandal
6. Shirt
7. Sneaker
8. Bag
9. Ankle boot

Pobrany zbiór składa się z 70000 obrazków, lecz w racji dużej ilości obliczeń ograniczyliśmy je do pierwszych 10000 obrazków z czterech wybranych grup.

## 2. Testowane parametry

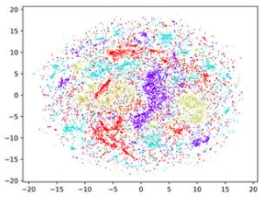
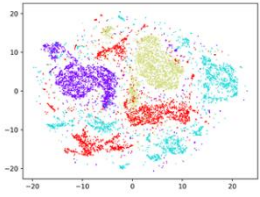
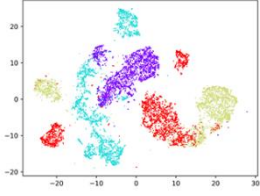
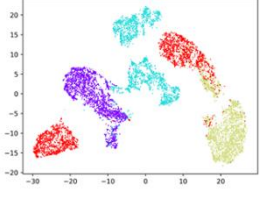
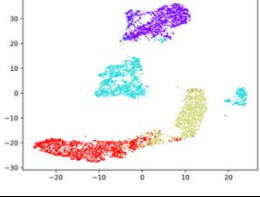
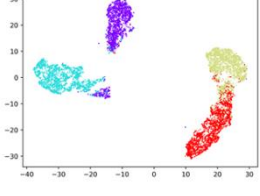
Testy przeprowadzać będziemy dla różnych wartości trzech parametrów (każdy testowany będzie osobno):

**samples** – (default: size/100) ilość próbkowanych krawędzi (w milionach)

**neighbours** – (default: 150) ilość sąsiadów (K) w K-NNG

**perplexity** – (default: 50) definiuje balans pomiędzy lokalnymi i globalnymi aspektami danych

### 3. Parametr „samples”

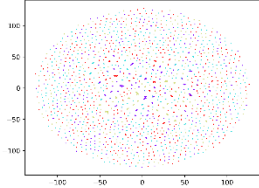
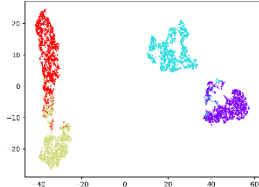
Samples	Czas [s]	Rezultat
5	13.8042	
10	16.6009	
30	23.9173	
50	32.1755	
100	52.0394	
250	113.8634	

#### 3.1. Wnioski

Dla małych wartości sample obliczenia są stosunkowo szybkie lecz mamy do czynienia z szumem, im większa wartość sample tym grupy są bardziej wyraźne, ale rośnie też czas.

Za optymalną wartość możemy przyjąć **50** gdyż klastry już są wyraźnie oddzielone, a czasy jeszcze są nie są bardzo duże.

## 4. Parametr „neighbours”

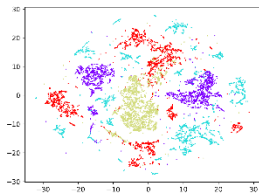
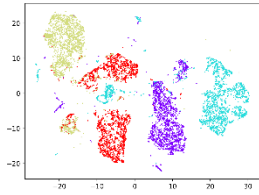
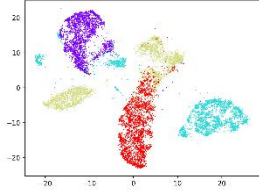
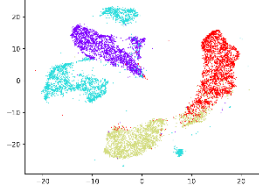
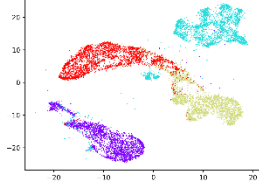
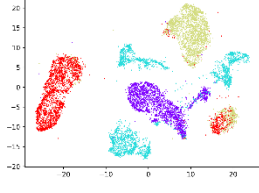
Neighbours	Czas [s]	Rezultat
1	365.0772	
10	388.4651	
50	379.6224	
100	416.9872	
500	447.7580	

### 4.1. Wnioski

Im większa wartość tym lepszy podział na klastry, podczas gdy wzrost czasu nie jest zbyt znaczący.

Warto zauważyć, że domyślną wartością dla tego parametru jest 150, a zważywszy że zarówno dla 50 jak i 100 otrzymujemy całkiem zadowalające wyniki, warto ją dostosowywać w praktyce, jako że 150 to już niepotrzebnie dużo.

## 5. Parametr „perplexity”

Perplexity	Czas [s]	Rezultat
10	17.4205	
20	18.4632	
50	17.1908	
100	17.4334	
300	18.4589	
1000	18.2168	

### 5.1. Wnioski

Zgodnie z tym informacją podaną w opisie LargeVis, parametr perplexity jest dosyć skomplikowany, w tym sensie że trudno jednoznacznie powiązać jego zmianę z wynikami. Warto jednak zauważyć że poniżej 50 mieliśmy do czynienia z szumami, natomiast paramter ten nie miał wpływu na czasy.

Zatem przyjęcie domyślnej wartości (50) lub też trochę wyższej powinno mieć dobry wpływ na wyniki.

## **6. Wnioski**

Metoda LargeVis, jako jedyna z dotychczas testowanych, dobrze radzi sobie z oddzielaniem klastrów dla zbiorów średniego rozmiaru. Nie tylko daje poprawne wyniki, ale też robi to w sensownym czasie.