

**Jan Gołda, Wojciech Basiura - WIET 2018/2019**

**Metody analizy i wizualizacji dużych zbiorów danych**

**Laboratorium 4 - LargeVis**

# Opis zagadnienia

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

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

## 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:

1. T-Shirt / top
2. Trouser
3. Pullover
4. Dress
5. Coat
6. Sandal
7. Shirt
8. Sneaker
9. Bag
10. 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.

# Testowane parametry

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

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

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

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

# Parametr „samples”

|  |  |  |
| --- | --- | --- |
| **Samples** | **Czas [s]** | **Rezultat** |
| 5 | 13.8042 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-0/p280x280/64858132_413468806164348_8658448477241475072_n.png?_nc_cat=109&_nc_ht=scontent-vie1-1.xx&oh=0a4a2e6d704948579d0869ccd94dd810&oe=5DC43EA4 |
| 10 | 16.6009 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-0/p280x280/64350070_1542802475855409_5311339377811521536_n.png?_nc_cat=100&_nc_ht=scontent-vie1-1.xx&oh=53b8475fb29275389b26831bf68efd3d&oe=5D816614 |
| 30 | 23.9173 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-0/p280x280/64574157_682292012216500_1742053471290392576_n.png?_nc_cat=110&_nc_ht=scontent-vie1-1.xx&oh=1f7b3fd7af44cbca1ffe7801e19cfa68&oe=5D8745DB |
| 50 | 32.1755 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-0/p280x280/64548095_2797684070247859_5394026814043586560_n.png?_nc_cat=102&_nc_ht=scontent-vie1-1.xx&oh=250d7a5e8fa356682a8c3d1748089c7a&oe=5D820F70 |
| 100 | 52.0394 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-0/p280x280/64850878_600445393778935_3378520845260423168_n.png?_nc_cat=104&_nc_ht=scontent-vie1-1.xx&oh=7e037bddd76ec2cee1bc576dbf9ad5e1&oe=5DC4DE28 |
| 250 | 113.8634 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-0/p280x280/64962755_605132549978666_6427659647863226368_n.png?_nc_cat=108&_nc_ht=scontent-vie1-1.xx&oh=64110cfd423a759fb33414f31470ca20&oe=5D854DD3 |

## 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.

# Parametr „neighbours”

|  |  |  |
| --- | --- | --- |
| **Neighbours** | **Czas [s]** | **Rezultat** |
| 1 | 365.0772 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/65188282_388533335095661_3273403977412116480_n.png?_nc_cat=109&_nc_ht=scontent-waw1-1.xx&oh=a194aa94612666c72e0a20f0e8f4d99f&oe=5D9629C7 |
| 10 | 388.4651 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64373721_288455278600323_5146185829307121664_n.png?_nc_cat=100&_nc_ht=scontent-waw1-1.xx&oh=952cdc5dce699ea0d5a67c9768d173d0&oe=5DC25B61 |
| 50 | 379.6224 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64599863_607372653104011_336629177199362048_n.png?_nc_cat=110&_nc_ht=scontent-waw1-1.xx&oh=eb6620bf2b08c56914903e08e0bb3edd&oe=5D93F42B |
| 100 | 416.9872 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/65033192_2357611567897948_3529452107780325376_n.png?_nc_cat=109&_nc_ht=scontent-waw1-1.xx&oh=c8f8de1dea477f58e569e7826e9752f1&oe=5D90AE25 |
| 500 | 447.7580 | https://scontent-vie1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64474854_2459256327420327_457751137596997632_n.png?_nc_cat=109&_nc_ht=scontent-vie1-1.xx&oh=d90b725559fd9857e32d482876bd48ab&oe=5DC6346D |

## 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.

# Parametr „perplexity”

|  |  |  |
| --- | --- | --- |
| **Perplexity** | **Czas [s]** | **Rezultat** |
| 10 | 17.4205 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64958145_397159650932663_5614874603634556928_n.png?_nc_cat=100&_nc_ht=scontent-waw1-1.xx&oh=14c72d584055f1d1072dd93ea4f8f0fe&oe=5D944186 |
| 20 | 18.4632 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64763302_2333745720224940_1681935658702077952_n.png?_nc_cat=111&_nc_ht=scontent-waw1-1.xx&oh=223ed6dca0fd08eea3d84ceb38797e4f&oe=5D87E031 |
| 50 | 17.1908 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/65000267_339504080285212_1352922051224010752_n.png?_nc_cat=110&_nc_ht=scontent-waw1-1.xx&oh=c92fb47c076a9f12418f3f7b8c1080b3&oe=5D9395C4 |
| 100 | 17.4334 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64629166_2137033586418610_7480736348387672064_n.png?_nc_cat=105&_nc_ht=scontent-waw1-1.xx&oh=c96c5f5537c2cd9cd2ccb4f42b4054d6&oe=5DC47396 |
| 300 | 18.4589 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64327349_348182439384324_7333383785762783232_n.png?_nc_cat=104&_nc_ht=scontent-waw1-1.xx&oh=6caf79bf276fef060e7ae6175710d818&oe=5DC3D277 |
| 1000 | 18.2168 | https://scontent-waw1-1.xx.fbcdn.net/v/t1.15752-9/s2048x2048/64667482_819225741812379_8482113677166641152_n.png?_nc_cat=100&_nc_ht=scontent-waw1-1.xx&oh=8ca8f0de34fb42c7de9dc45b7340e9d2&oe=5D8883CE |

## 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.

# 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.