**Institute of Business Administration**

**Introduction to Text Analytics**

**Assignment 02 – K-Means Clustering Assessment**

**Name: Zuha Aqib ID: 26106**

Report each experiment’s detail and scores for k = 5, 9, and 13. You are required to perform ten experiments for each ‘k’ (number of clusters). Please set random seed value to your ERP ID for each K-Means clustering experiment.

\*The first four entries in the table are provided for reference only. Hence, the scores do not interpret anything and have been entered randomly. Replace these entries while submitting.

**PLEASE NOTE**

Due to the excessive cases, the readability of the table is difficult. Thus I have attached the excel sheet (which also contains the case numbers) I maintained to my submission. I have removed the case numbers here. I have also maintained a github repository with the notebooks committed for EACH CASE.

You can view them here: <https://github.com/z-aqib/text-analytics.git>

**THE YELLOW HIGHLIGHTED IS THE BEST SILL/WSS FOR THAT VECTORIZER**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **k (Number of clusters)** | **Vectorizer Type and Details** | **Stemming** | **Lemmatization** | **N-Grams Utilized** | **Stop words** | **Silhouette** | **WSS** |
| **(Yes/No)** | **(Yes/No)** | **(Yes/No)** | **Score** | **Score** |
| **5** | BOW: CountVectorizer (TP) | Yes | No | unigram | Yes | 0.013282 | 3473.43 |
| BOW: CountVectorizer (TP) | Yes | No | unigram | No | 0.016872 | 4258.04 |
| BOW: CountVectorizer (TP) | Yes | No | bigram | Yes | -0.03306 | 3152.58 |
| BOW: CountVectorizer (TP) | Yes | No | bigram | No | -0.07444 | 4153.87 |
| BOW: CountVectorizer (TP) | No | Yes | unigram | Yes | 0.001199 | 3477.07 |
| BOW: CountVectorizer (TP) | No | Yes | unigram | No | 0.015732 | 4203.41 |
| BOW: CountVectorizer (TP) | No | Yes | bigram | Yes | -0.03326 | 3127.11 |
| BOW: CountVectorizer (TP) | No | Yes | bigram | No | -0.07196 | 4098.04 |
| BOW: CountVectorizer (TF) | Yes | No | unigram | Yes | 0.009557 | 3548.97 |
| BOW: CountVectorizer (TF) | Yes | No | unigram | No | 0.025306 | 4388.59 |
| BOW: CountVectorizer (TF) | Yes | No | bigram | Yes | -0.03325 | 3155.58 |
| BOW: CountVectorizer (TF) | Yes | No | bigram | No | -0.07469 | 4159.86 |
| BOW: CountVectorizer (TF) | No | Yes | unigram | Yes | 0.014238 | 3501.15 |
| BOW: CountVectorizer (TF) | No | Yes | unigram | No | 0.005123 | 4363.45 |
| BOW: CountVectorizer (TF) | No | Yes | bigram | Yes | -0.03326 | 3127.11 |
| BOW: CountVectorizer (TF) | No | Yes | bigram | No | 0.000191 | 4093.68 |
| BOW: TF-IDF | Yes | No | unigram | Yes | 0.00393 | 440.559 |
| BOW: TF-IDF | Yes | No | unigram | No | 0.003238 | 438.559 |
| BOW: TF-IDF | Yes | No | bigram | Yes | 0.00485 | 445.19 |
| BOW: TF-IDF | Yes | No | bigram | No | 0.002056 | 446.498 |
| BOW: TF-IDF | No | Yes | unigram | Yes | 0.003872 | 441.06 |
| BOW: TF-IDF | No | Yes | unigram | No | 0.003327 | 432.287 |
| BOW: TF-IDF | No | Yes | bigram | Yes | 0.004559 | 445.258 |
| BOW: TF-IDF | No | Yes | bigram | No | 0.002796 | 446.139 |
| TruncatedSVD (n\_components = 50) | Yes | No | unigram | Yes | 0.152272 | 103.365 |
| TruncatedSVD (n\_components = 50) | Yes | No | unigram | No | 0.034948 | 106.905 |
| TruncatedSVD (n\_components = 50) | Yes | No | bigram | Yes | 0.453867 | 66.6053 |
| TruncatedSVD (n\_components = 50) | Yes | No | bigram | No | 0.554426 | 66.826 |
| TruncatedSVD (n\_components = 50) | No | Yes | unigram | Yes | 0.054367 | 102.359 |
| TruncatedSVD (n\_components = 50) | No | Yes | unigram | No | 0.054062 | 104.212 |
| TruncatedSVD (n\_components = 50) | No | Yes | bigram | Yes | 0.611437 | 66.04 |
| TruncatedSVD (n\_components = 50) | No | Yes | bigram | No | 0.495226 | 66.1592 |
| TruncatedSVD (n\_components = 100) | Yes | No | unigram | Yes | 0.058368 | 176.812 |
| TruncatedSVD (n\_components = 100) | Yes | No | unigram | No | 0.010665 | 180.852 |
| TruncatedSVD (n\_components = 100) | Yes | No | bigram | Yes | -0.00459 | 120.402 |
| TruncatedSVD (n\_components = 100) | Yes | No | bigram | No | 0.321782 | 122.915 |
| TruncatedSVD (n\_components = 100) | No | Yes | unigram | Yes | 0.005729 | 175.121 |
| TruncatedSVD (n\_components = 100) | No | Yes | unigram | No | 0.025373 | 177.703 |
| TruncatedSVD (n\_components = 100) | No | Yes | bigram | Yes | -0.01228 | 120.295 |
| TruncatedSVD (n\_components = 100) | No | Yes | bigram | No | 0.263086 | 123.649 |
| TruncatedSVD (n\_components = 200) | Yes | No | unigram | Yes | 0.002308 | 287.157 |
| TruncatedSVD (n\_components = 200) | Yes | No | unigram | No | 0.003627 | 289.395 |
| TruncatedSVD (n\_components = 200) | Yes | No | bigram | Yes | -0.0162 | 221.752 |
| TruncatedSVD (n\_components = 200) | Yes | No | bigram | No | -0.00372 | 225.001 |
| TruncatedSVD (n\_components = 200) | No | Yes | unigram | Yes | -0.0147 | 284.989 |
| TruncatedSVD (n\_components = 200) | No | Yes | unigram | No | 0.010122 | 284.571 |
| TruncatedSVD (n\_components = 200) | No | Yes | bigram | Yes | -0.01698 | 221.746 |
| TruncatedSVD (n\_components = 200) | No | Yes | bigram | No | 0.000446 | 224.948 |
| **9** | BOW: CountVectorizer (TP) | Yes | No | unigram | Yes | -0.0036 | 3457.41 |
| BOW: CountVectorizer (TP) | Yes | No | unigram | No | 0.016511 | 4125.94 |
| BOW: CountVectorizer (TP) | Yes | No | bigram | Yes | -0.03111 | 3117.21 |
| BOW: CountVectorizer (TP) | Yes | No | bigram | No | -0.07107 | 4098.74 |
| BOW: CountVectorizer (TP) | No | Yes | unigram | Yes | -0.00902 | 3426.18 |
| BOW: CountVectorizer (TP) | No | Yes | unigram | No | 0.00829 | 4085.34 |
| BOW: CountVectorizer (TP) | No | Yes | bigram | Yes | -0.04772 | 3082.5 |
| BOW: CountVectorizer (TP) | No | Yes | bigram | No | -0.06917 | 4043.13 |
| BOW: CountVectorizer (TF) | Yes | No | unigram | Yes | 0.002473 | 3477.09 |
| BOW: CountVectorizer (TF) | Yes | No | unigram | No | -0.00307 | 4269.57 |
| BOW: CountVectorizer (TF) | Yes | No | bigram | Yes | -0.0325 | 3117.39 |
| BOW: CountVectorizer (TF) | Yes | No | bigram | No | -0.07133 | 4104.73 |
| BOW: CountVectorizer (TF) | No | Yes | unigram | Yes | 0.007316 | 3476.03 |
| BOW: CountVectorizer (TF) | No | Yes | unigram | No | 0.003109 | 4277.8 |
| BOW: CountVectorizer (TF) | No | Yes | bigram | Yes | -0.04772 | 3082.5 |
| BOW: CountVectorizer (TF) | No | Yes | bigram | No | -0.06943 | 4049.11 |
| BOW: TF-IDF | Yes | No | unigram | Yes | 0.004994 | 433.455 |
| BOW: TF-IDF | Yes | No | unigram | No | 0.004518 | 426.227 |
| BOW: TF-IDF | Yes | No | bigram | Yes | 0.005695 | 440.075 |
| BOW: TF-IDF | Yes | No | bigram | No | 0.004948 | 440.601 |
| BOW: TF-IDF | No | Yes | unigram | Yes | 0.005264 | 433.752 |
| BOW: TF-IDF | No | Yes | unigram | No | 0.004401 | 426.32 |
| BOW: TF-IDF | No | Yes | bigram | Yes | 0.008202 | 439.087 |
| BOW: TF-IDF | No | Yes | bigram | No | 0.005882 | 440.325 |
| TruncatedSVD (n\_components = 50) | Yes | No | unigram | Yes | 0.100574 | 94.459 |
| TruncatedSVD (n\_components = 50) | Yes | No | unigram | No | 0.101465 | 101.311 |
| TruncatedSVD (n\_components = 50) | Yes | No | bigram | Yes | 0.418285 | 59.524 |
| TruncatedSVD (n\_components = 50) | Yes | No | bigram | No | 0.561402 | 59.8563 |
| TruncatedSVD (n\_components = 50) | No | Yes | unigram | Yes | 0.077142 | 93.9194 |
| TruncatedSVD (n\_components = 50) | No | Yes | unigram | No | 0.056393 | 96.8025 |
| TruncatedSVD (n\_components = 50) | No | Yes | bigram | Yes | 0.535079 | 59.3088 |
| TruncatedSVD (n\_components = 50) | No | Yes | bigram | No | 0.485763 | 60.247 |
| TruncatedSVD (n\_components = 100) | Yes | No | unigram | Yes | 0.017778 | 170.097 |
| TruncatedSVD (n\_components = 100) | Yes | No | unigram | No | 0.010862 | 173.8 |
| TruncatedSVD (n\_components = 100) | Yes | No | bigram | Yes | 0.017498 | 114.278 |
| TruncatedSVD (n\_components = 100) | Yes | No | bigram | No | 0.258557 | 117.456 |
| TruncatedSVD (n\_components = 100) | No | Yes | unigram | Yes | 0.042785 | 166.474 |
| TruncatedSVD (n\_components = 100) | No | Yes | unigram | No | 4.08E-05 | 171.309 |
| TruncatedSVD (n\_components = 100) | No | Yes | bigram | Yes | -0.05811 | 114.596 |
| TruncatedSVD (n\_components = 100) | No | Yes | bigram | No | 0.346014 | 116.539 |
| TruncatedSVD (n\_components = 200) | Yes | No | unigram | Yes | 0.005805 | 280.828 |
| TruncatedSVD (n\_components = 200) | Yes | No | unigram | No | 0.007101 | 280.68 |
| TruncatedSVD (n\_components = 200) | Yes | No | bigram | Yes | -0.00569 | 217.876 |
| TruncatedSVD (n\_components = 200) | Yes | No | bigram | No | -0.03617 | 221.3 |
| TruncatedSVD (n\_components = 200) | No | Yes | unigram | Yes | -0.00557 | 278.34 |
| TruncatedSVD (n\_components = 200) | No | Yes | unigram | No | -0.0017 | 278.288 |
| TruncatedSVD (n\_components = 200) | No | Yes | bigram | Yes | 0.000785 | 216.63 |
| TruncatedSVD (n\_components = 200) | No | Yes | bigram | No | -0.0198 | 221.329 |
| **13** | BOW: CountVectorizer (TP) | Yes | No | unigram | Yes | -0.00456 | 3407.49 |
| BOW: CountVectorizer (TP) | Yes | No | unigram | No | 0.017478 | 4045.31 |
| BOW: CountVectorizer (TP) | Yes | No | bigram | Yes | -0.04603 | 3078.22 |
| BOW: CountVectorizer (TP) | Yes | No | bigram | No | -0.05229 | 4033.53 |
| BOW: CountVectorizer (TP) | No | Yes | unigram | Yes | -0.00799 | 3345.04 |
| BOW: CountVectorizer (TP) | No | Yes | unigram | No | 0.007993 | 4012.53 |
| BOW: CountVectorizer (TP) | No | Yes | bigram | Yes | -0.0454 | 3045.55 |
| BOW: CountVectorizer (TP) | No | Yes | bigram | No | -0.05317 | 3999.18 |
| BOW: CountVectorizer (TF) | Yes | No | unigram | Yes | 0.002217 | 3414.45 |
| BOW: CountVectorizer (TF) | Yes | No | unigram | No | -0.00142 | 4182.39 |
| BOW: CountVectorizer (TF) | Yes | No | bigram | Yes | -0.10687 | 3082.15 |
| BOW: CountVectorizer (TF) | Yes | No | bigram | No | -0.0538 | 4045.44 |
| BOW: CountVectorizer (TF) | No | Yes | unigram | Yes | -0.01178 | 3416.53 |
| BOW: CountVectorizer (TF) | No | Yes | unigram | No | -0.00081 | 4212.12 |
| BOW: CountVectorizer (TF) | No | Yes | bigram | Yes | -0.0454 | 3045.55 |
| BOW: CountVectorizer (TF) | No | Yes | bigram | No | -0.05344 | 4005.17 |
| BOW: TF-IDF | Yes | No | unigram | Yes | 0.006173 | 426.959 |
| BOW: TF-IDF | Yes | No | unigram | No | 0.004518 | 426.227 |
| BOW: TF-IDF | Yes | No | bigram | Yes | 0.010892 | 434.539 |
| BOW: TF-IDF | Yes | No | bigram | No | 0.00726 | 434.88 |
| BOW: TF-IDF | No | Yes | unigram | Yes | 0.006527 | 427.368 |
| BOW: TF-IDF | No | Yes | unigram | No | 0.004401 | 426.32 |
| BOW: TF-IDF | No | Yes | bigram | Yes | 0.009156 | 434.682 |
| BOW: TF-IDF | No | Yes | bigram | No | 0.008489 | 434.212 |
| TruncatedSVD (n\_components = 50) | Yes | No | unigram | Yes | 0.11596 | 86.9302 |
| TruncatedSVD (n\_components = 50) | Yes | No | unigram | No | 0.080295 | 94.1874 |
| TruncatedSVD (n\_components = 50) | Yes | No | bigram | Yes | 0.4655 | 52.537 |
| TruncatedSVD (n\_components = 50) | Yes | No | bigram | No | 0.587086 | 53.4737 |
| TruncatedSVD (n\_components = 50) | No | Yes | unigram | Yes | 0.112176 | 85.6851 |
| TruncatedSVD (n\_components = 50) | No | Yes | unigram | No | 0.079903 | 88.9609 |
| TruncatedSVD (n\_components = 50) | No | Yes | bigram | Yes | 0.525883 | 52.3797 |
| TruncatedSVD (n\_components = 50) | No | Yes | bigram | No | 0.414504 | 53.9709 |
| TruncatedSVD (n\_components = 100) | Yes | No | unigram | Yes | 0.026357 | 162.914 |
| TruncatedSVD (n\_components = 100) | Yes | No | unigram | No | 0.016508 | 166.313 |
| TruncatedSVD (n\_components = 100) | Yes | No | bigram | Yes | 0.039934 | 108.439 |
| TruncatedSVD (n\_components = 100) | Yes | No | bigram | No | 0.281876 | 111.899 |
| TruncatedSVD (n\_components = 100) | No | Yes | unigram | Yes | 0.034725 | 160.228 |
| TruncatedSVD (n\_components = 100) | No | Yes | unigram | No | 0.011416 | 165.463 |
| TruncatedSVD (n\_components = 100) | No | Yes | bigram | Yes | -0.04799 | 108.721 |
| TruncatedSVD (n\_components = 100) | No | Yes | bigram | No | 0.270533 | 110.034 |
| TruncatedSVD (n\_components = 200) | Yes | No | unigram | Yes | 0.005824 | 275.258 |
| TruncatedSVD (n\_components = 200) | Yes | No | unigram | No | 0.011844 | 274.21 |
| TruncatedSVD (n\_components = 200) | Yes | No | bigram | Yes | -0.02986 | 212.562 |
| TruncatedSVD (n\_components = 200) | Yes | No | bigram | No | -0.0229 | 215.924 |
| TruncatedSVD (n\_components = 200) | No | Yes | unigram | Yes | 0.005614 | 270.774 |
| TruncatedSVD (n\_components = 200) | No | Yes | unigram | No | 0.003762 | 272.886 |
| TruncatedSVD (n\_components = 200) | No | Yes | bigram | Yes | -0.01827 | 212.411 |
| TruncatedSVD (n\_components = 200) | No | Yes | bigram | No | -0.01596 | 216.737 |

**Analysis & Interpretation:**

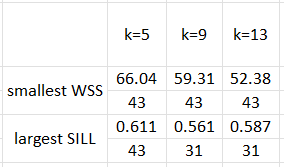
**o Identify which embedding technique resulted in the best clustering.**

**o Discuss how preprocessing choices impacted the results.**

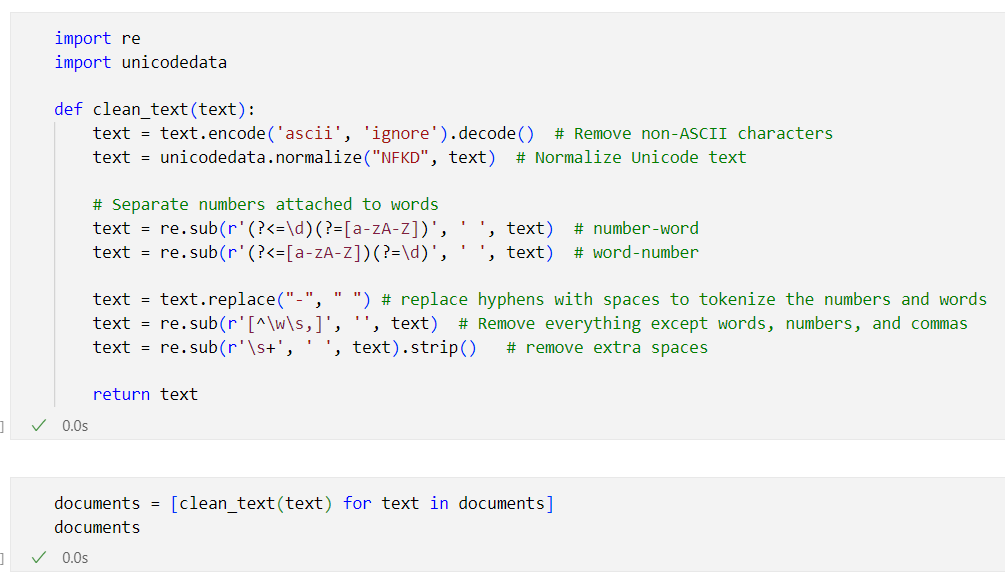
**o Provide sample headlines from different clusters to analyze coherence.**

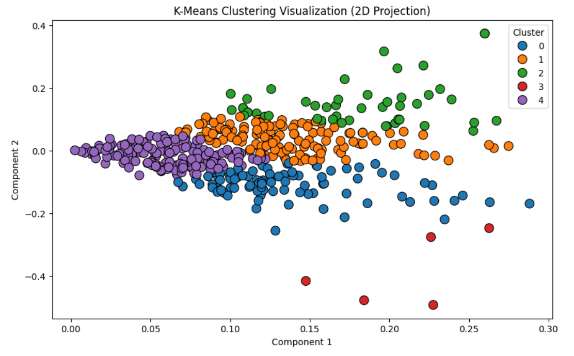
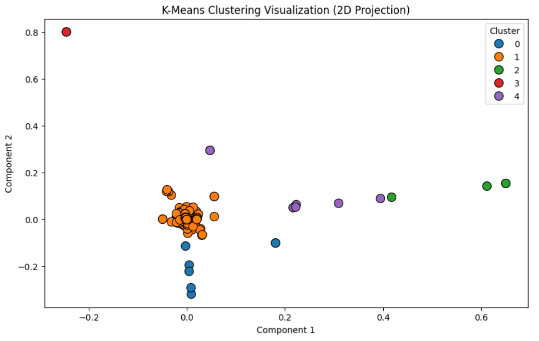
|  |
| --- |
|  |

So I did around 48 for each k, from that these are the best sillhoutte scores and WSS,

This table shows that the smallest WSS was found in case 43 for all k’s, (best was k=13). That case was: **LSA 50, lemmatization, bigrams and removed stop words**.

The best silhouette score was found again in cases 43 and 31. Again, the best was k=5 in case 43. Case 31 was **LSA 50, stemming, bigrams, and not removing stop words**.

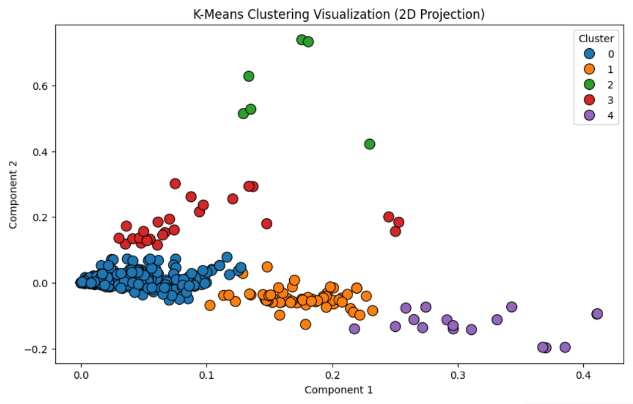
* So the best embedding technique was hands down LSA 50.
  + LSA 50 gave WSS in 50-100, whereas
  + BOW was in 3500-4000.
  + LSA 100 was 100-200 and
  + LSA 200 was 200-300.
  + TFIDF was 400-500.
  + So the best was LSA 50. The silhouette scores were also the highest in LSA 50.
* I noticed that when we applied bigrams, we got more columns, and silhouette falled into a negative FOR BOW. However for the rest bigrams performed better, it gave a higher silhouette and lower WSS
* However removing stop words was ALWAYS good, whenever I didn’t remove, I had more columns and very high WSS and low silhouette
* Lemmatization performed slightly better than stemming
* LSA 50 was faster than LSA 200. It took slightly longer to run LSA 200.
* Values were much better after the data was cleaned. Here is the cleaning code: 
* I ran it on my own laptop on VS code and it did not have any errors or problems, running time for each was an average of 1-2 seconds, sometimes even less than 1 second
* I generated the graph for each iteration, and this is what I found:



This is for ALL words, UNI gram

This is for ALL words, BI gram

A graph with colored dots

AI-generated content may be incorrect.

This is for STOP words removed, UNI gram

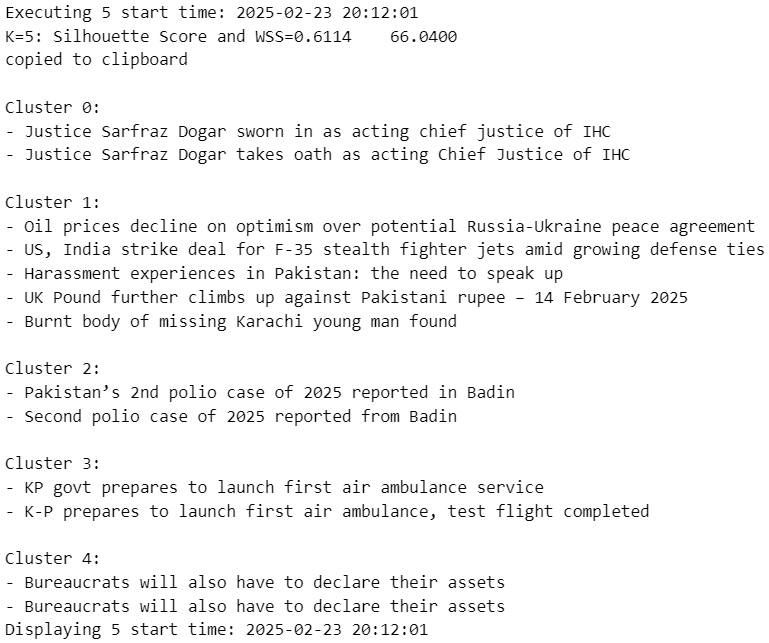
This is for STOP words removed, BI gram

So from this we can see that unigrams have a more spreadout values while bigrams have more concise values with same context

This was the graph generation code:



Here are some sample headlines:





So from these cluster headlines (for each case I have generated 5 headlines for every k. this is headlines for k=5 and case 43 (THE BEST CASE)):

* I can see the first cluster is of politics/chief justice
* Second cluster is a bit mixed, mostly for world affairs
* Third cluster is of polio cases
* Forth cluster is of air ambulance
* Fifth is of bereaucrats

So we can say there is some sense of coherence in the clusters.

SO OVERALL the best is

* LSA 50
* Bigrams
* Lemmatization
* And stop words removed