System Setup and Configuration:

CPU : Ryzen 7

Memory: 16 gb GPU: Nvidia gtx1060 Library Used:

NLTK, Pandas, SKLEARN, MATPLOTLIB Training Dataset:

UCI ML Drug Review dataset:

<https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018> Includes Train and Test Dataset separately.

Used Train dataset for LSA and Test Dataset for LDA. Since LDA is computation intensive and system hanging issue, trained with Test Dataset alone.

1. Data Preprocessing:

Performed - tokenization, lowercasing, stemming and stop word removal.

Sample –

Before preprocess:

"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"

After preprocess:

side effect take combination Bystolic Mg Fish Oil

1. Document Representation:

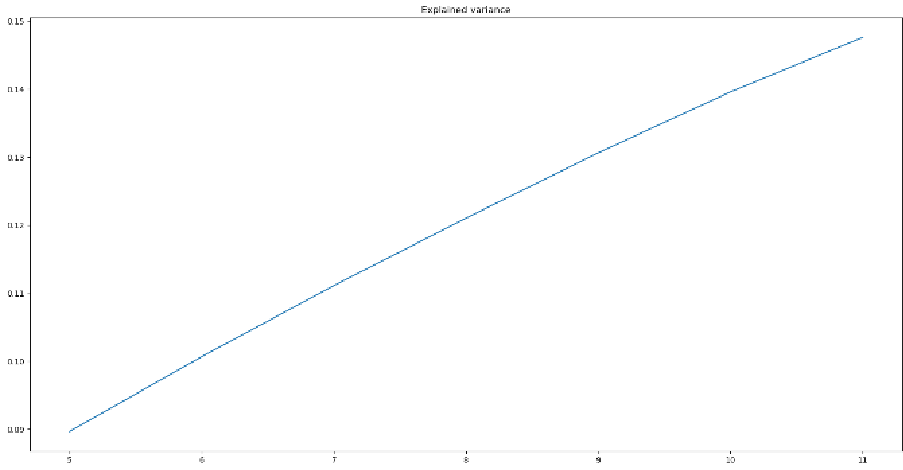
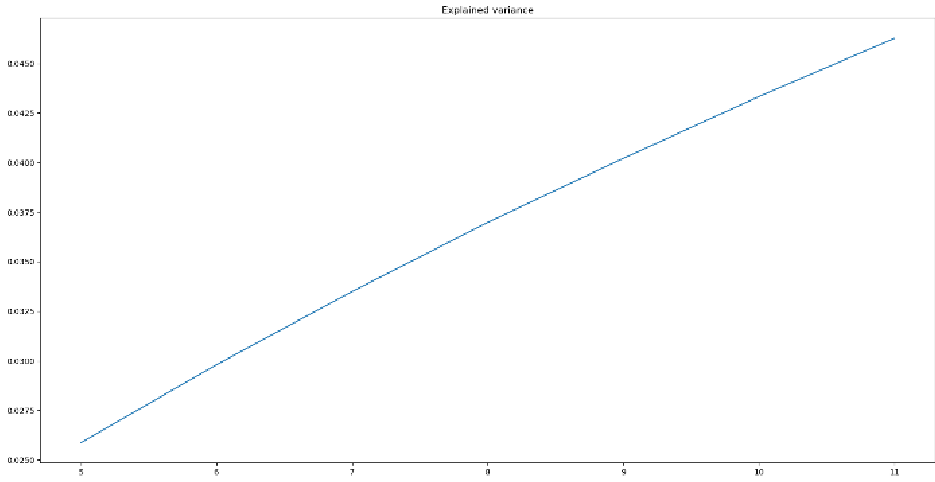
LSA Model:

* 1. Finding number of components:

No of components can be chosen by trial and error method and with the help of explained variance and elbow curve method we can find the explained variance [1].

LSA with:

TFIDF vectorizer BOW vectorizer



No of topics : 8 No of topics : 9

(Please zoom in for the Explained variance values, TFIDF as low values and Bow as high values)

* 1. Top 10 Topics with words extracted:

|  |  |
| --- | --- |
| TF-IDF | BOW |
| ['day','month', 'period', 'im', 'week', 'year', 'pil l', 'ive', 'effect', 'side', 'taking', 'pain', 'time ', 'mg', 'first', 'take', 'get', 'started', 'like',  'feel'] | ['day', 'month', 'year', 'week', 'im', 'effect', 't ime', 'side', 'taking', 'first', 'ive', 'mg', 'peri od', 'pill', 'take', 'started', 'get', 'pain', 'lik e', 'feel'] |
| ['period', 'pill', 'birth', 'month', 'c ontrol', 'bleeding', 'acne', 'weight', 'cramp', 'ive ', 'sex', 'spotting', 'gain', 'swing', 'mood', 'got'  , 'heavy', 'gained', 'light', 'pregnant'] | ['day', 'mg', 'pain', 'take', 'medication', 'effect  ', 'sleep', 'anxiety', 'side', 'medicine', 'work',  'night', 'taking', 'hour', 'feel', 'dose', 'drug',  'life', 'prescribed', 'doctor'] |
| ['pain', 'day', 'period', 'hour', 'back  ', 'cramp', 'took', 'relief', 'got', 'bleeding', 'st omach', 'pill', 'insertion', 'severe', 'knee', 'take  ', 'cramping', 'painful', 'surgery', 'chronic'] | ['day', 'period', 'pill', 'cramp', 'took', 'bleedin g', 'got', 'next', 'first', 'heavy', 'hour', 'spott ing', 'th', 'later', 'light', 'cramping', 'two', 'e very', 'plan', 'insertion'] |
| ['effect', 'side', 'pain', 'weight', 'medicine', 'gr eat', 'medication', 'control', 'gain', 'year', 'birt h', 'month', 'acne', 'severe', 'worked', 'ive', 'mig raine', 'lb', 'loss', 'experienced'] | ['pain', 'year', 'back', 'get', 'take', 'doctor', ' time', 'would', 'work', 'hour', 'got', 'life', 'rel ief', 'severe', 'month', 'never', 'cramp', 'help',  'period', 'like'] |
| ['day', 'effect', 'side', 'pill', 'week', 'took', 'h our', 'nausea', 'lb', 'headache', 'first', 'taking', 'started', 'infection', 'next', 'lost', 'water', 'la ter', 'stomach', 'eat'] | ['im', 'mg', 'feel', 'ive', 'anxiety', 'year', 'lik e', 'started', 'life', 'sleep', 'week', 'depression ', 'time', 'night', 'dont', 'taking', 'panic', 'att ack', 'get', 'better'] |
| ['lb', 'weight', 'pain', 'lost', 'week'  , 'im', 'pound', 'started', 'gained', 'eat', 'gain', 'month', 'lose', 'loss', 'back', 'ive', 'eating', 'e xercise', 'appetite', 'diet'] | ['month', 'mg', 'year', 'started', 'first', 'week',  'ago', 'taking', 'anxiety', 'day', 'doctor', 'went'  , 'period', 'back', 'two', 'stopped', 'depression',  'life', 'dose', 'every'] |
| ['work', 'acne', 'skin', 'face', 'using', 'product',  'medicine', 'use', 'really', 'well', 'clear', 'used'  , 'dry', 'great', 'im', 'cream', 'lb', 'get', 'good'  , 'infection'] | ['week', 'pain', 'month', 'first', 'started', 'im',  'feel', 'mg', 'felt', 'back', 'lb', 'like', 'two', 'still', 'better', 'lost', 'feeling', 'went', 'naus ea', 'second'] |
| ['work', 'pill', 'weight', 'great', 'ta ke', 'sleep', 'lb', 'gain', 'medicine', 'well', 'goo d', 'control', 'lost', 'gained', 'birth', 'hour', 'p ound', 'works', 'period', 'dont'] | ['pill', 'take', 'time', 'like', 'took', 'hour', 'w ould', 'get', 'work', 'feel', 'taking', 'night', 's leep', 'felt', 'one', 'dont', 'didnt', 'make', 'per iod', 'anxiety'] |
| ['mg', 'month', 'great', 'year', 'work', 'acne', 'st arted', 'worked', 'well', 'day', 'taking', 'period', 'skin', 'works', 'two', 'ago', 'using', 'back', 'wee k', 'lb'] | ['week', 'year', 'time', 'first', 'got', 'ago', 'tw o', 'old', 'would', 'period', 'one', 'went', 'took' , 'last', 'life', 'felt', 'using', 'never', 'skin',  'later'] |
| ['pill', 'acne', 'taking', 'control', 'mg', 'birth', 'skin', 'sleep', 'night', 'take', 'pain', 'face', 'h elp', 'also', 'hour', 'clear', 'made', 'weight', 'mo od', 'gain'] | ['month', 'like', 'time', 'feel', 'get', 'would', ' day', 'first', 'skin', 'medication', 'felt', 'life' , 'better', 'bad', 'work', 'go', 'got', 'acne', 'us e', 'using'] |

In terms of performance and grouping of similar words we can see that LSA with BOW performed well compared to TFIDF, higher the value of Explained variance, better the performance of the model( in the above graph we can see the explained variance is high for bow than tfidf based lda) . For instance we can see the highlighted row from the above where BOW grouped well and repetitions are less compared to TFIDF.

LDA Model:

For training LDA model, test dataset is used since it is small compared to Training dataset and due to System Constraint. Still it is test dataset it should be sampled well from the Training Dataset.

2.1)Number of components for LDA can be found same as LSA model and we can make use of loglikelihood value to figure it out. After trial and error, Taking number of components as 8. 2.2)

Performance Table:

|  |  |  |
| --- | --- | --- |
|  | Bag Of words | TFIDF |
| Alpha (Document topic Prior) | 0.1 | 0.01 |
| Beta (Topic word Prior) | 0.1 | 0.5 |
| Log Likelihood score | -3484350.7764869346 | -549217.6295709757 |

2.3)

Dirichlet Parameter:

The alpha controls the mixture of topics for any given document, whereas the beta hyperparameter controls the distribution of words per topic. The lesser the alpha makes the documents less number of topics and higher the beta the topics will have more words.

With this in mind, we should find alpha which will be low and beta with high value. Typically both these values lies within range 1(but not necessarily). From above data we can see that LDA with tfidf has less alpha value, high beta and Likelihood value compared to LDA with BOW, hence TFIDF performed better with LDA than BOW.

|  |  |
| --- | --- |
| BOW | TF-IDF |
| ['week', 'day', 'started', 'effect', 'side', 'weight', 'lb'  , 'month', 'first', 'im', 'lost', 'taking', 'feel', 'year',  'time', 'pound', 'ive', 'eat', 'took', 'like'] | ['chantix', 'smoking', 'smoked', 'quit', 'smoke', 'cigarett e', 'smoker', 'cialis', 'pack', 'year', 'free', 'nicotine', 'erection', 'viagra', 'stendra', 'dream', 'dysfunction', 'z yban', 'erectile', 'day'] |
| ['pain', 'year', 'day', 'skin', 'back', 'month', 'take', 'i ve', 'time', 'side', 'taking', 'work', 'week', 'use', 'doct or', 'medicine', 'effect', 'im', 'one', 'amp'] | ['day', 'mg', 'pain', 'effect', 'year', 'side', 'taking', ' take', 'medication', 'work', 'week', 'medicine', 'im', 'fee l', 'time', 'month', 'started', 'like', 'anxiety', 'ive'] |
| ['day', 'hour', 'took', 'time', 'take', 'like', 'night', 'f irst', 'work', 'get', 'felt', 'im', 'feel', 'one', 'pill', 'went', 'sleep', 'would', 'pain', 'taking'] | ['hour', 'taste', 'day', 'burning', 'water', 'took', 'night  ', 'prep', 'itching', 'infection', 'like', 'minute', 'first ', 'time', 'sleep', 'product', 'work', 'good', 'pm', 'drink '] |
| ['day', 'week', 'month', 'im', 'year', 'period', 'got', 'fi rst', 'bleeding', 'pain', 'like', 'time', 'would', 'inserti on', 'get', 'ive', 'since', 'cramp', 'mirena', 'doctor'] | ['period', 'month', 'day', 'pill', 'bleeding', 'week', 'got  ', 'im', 'ive', 'cramp', 'year', 'birth', 'sex', 'control',  'get', 'first', 'time', 'insertion', 'mirena', 'spotting'] |
| ['mg', 'taking', 'day', 'feel', 'year', 'take', 'like', 'sl eep', 'effect', 'medication', 'medicine', 'started', 'side' , 'would', 'life', 'night', 'get', 'time', 'doctor', 'feeli ng'] | ['excellent', 'lotrel', 'cellcept', 'azor', 'product', 'exe lon', 'pressure', 'prograf', 'namenda', 'controls', 'menopu r', 'pmt', 'fantastic', 'hiccup', 'united', 'states', 'ever thing', 'result', 'factive', 'neutrasal'] |
| ['pain', 'work', 'im', 'take', 'year', 'get', 'great', 'mon th', 'dont', 'back', 'day', 'medication', 'ive', 'time', 's ide', 'effect', 'would', 'doctor', 'injection', 'patch'] | ['works', 'great', 'well', 'ditropan', 'relieves', 'goes', 'efects', 'detrol', 'psuedocholonesterase', 'exhibit', 'sup ervised', 'effective', 'spasticity', 'iop', 'subjectively', 'fianc', 'masterfully', 'wonderfulonly', 'fewer', 'superior  '] |
| ['period', 'month', 'pill', 'ive', 'control', 'birth', 'im' , 'weight', 'day', 'year', 'first', 'get', 'week', 'acne',  'mood', 'time', 'side', 'effect', 'taking', 'started'] | ['period', 'pill', 'month', 'birth', 'control', 'ive', 'wei ght', 'acne', 'mood', 'im', 'gain', 'first', 'swing', 'year ', 'get', 'week', 'sex', 'taking', 'gained', 'started'] |
| ['mg', 'effect', 'year', 'side', 'anxiety', 'medication', ' day', 'take', 'taking', 'month', 'life', 'depression', 'wor k', 'medicine', 'time', 'ive', 'doctor', 'im', 'help', 'dru g'] | ['viral', 'load', 'cd', 'cholesterol', 'undetectable', 'sid e', 'effect', 'crestor', 'ldl', 'viagra', 'hep', 'month', ' year', 'harvoni', 'work', 'count', 'expensive', 'works', 'd rug', 'great'] |

From the above result itself it shows TF-IDF performed well along with the lda model compared to BOW.

2.4)

With the help of gridsearchcv function from sklearn we were able to find the Best optimal value of Alpha and Theta. While performing this we will give log likelihood as the performance metrics, which will help us in fetching the best model with high likelihood. With TF-IDF and alpha with 0.01 and beta with 0.5 gives high likelihood model. This in turns satisfies the relation with Alpha and beta.

3)

2) Group the drugs from the resultant model.

group the documents(Reviews) based on the max similarity for a topic by fitting the data to the model with optimal params that we found. After grouping the documents fetch the relevant drug from the data.

Sample:

Topic\_Zero Drugs:

{'Alfuzosin', 'Rabeprazole', 'Bupropion', 'Atacand', 'Belsomra', 'Efaviren

z / emtricitabine / tenofovir', 'Travatan', 'Ezetimibe / simvastatin', 'Suvorexant', 'Tilmanoc ept', 'Clocortolone', 'Testosterone', 'Chantix', 'Gabapentin', 'Disulfiram', 'Lymphoseek', 'Cl onazepam', 'Copper', 'Humira', 'Toujeo Solostar', 'Aciphex', 'Promethazine', 'Dexedrine', 'Pen tazocine', 'Suboxone', 'Varenicline', 'Phenergan', 'Antabuse', 'Zoledronic acid', 'Jentadueto' , 'Zanaflex', 'Lorazepam', 'Metronidazole', 'Xtampza ER', 'Propranolol', 'Atripla', 'Venlafaxi ne', 'Buprenorphine / naloxone', 'Reclast', 'WP Thyroid'}