**A3. BBC News Articles Workflow**

Dear Client XYZ,

Following are the step I’ve taken to build information extraction pipeline and subsequent topic modeling :

1. Imported the input CSV file (Containing 2225 articles from BBC)
2. [Pre-processed dataset](http://localhost:8888/notebooks/Desktop/US/Spring%202021/Text%20Analytics/BBC%20News%20Articles.ipynb#2.-Pre-processing-Dataset) by removing all unwanted punctuation, fullstops, stopwords
3. Later I’ve devised 3 ways of filtering the dataset and 2 models for topic modelling, following are the brief description of filters and models :
4. [Normal Cleaning](http://localhost:8888/notebooks/Desktop/US/Spring%202021/Text%20Analytics/BBC%20News%20Articles.ipynb#3.-Normal-Cleaning) (lemmatized and tokenized )
5. [Term Frequency Filter](http://localhost:8888/notebooks/Desktop/US/Spring%202021/Text%20Analytics/BBC%20News%20Articles.ipynb#4.-Term-Frequency-Filter) (excluded top 10% and words freq < 5 times)
6. [Part of Speech Filter](http://localhost:8888/notebooks/Desktop/US/Spring%202021/Text%20Analytics/BBC%20News%20Articles.ipynb#5.-Part-of-Speech-Filter) (filtered the word list with nouns)
7. LSI (Latent Semantic Indexing)
8. [LDA](http://localhost:8888/notebooks/Desktop/US/Spring%202021/Text%20Analytics/BBC%20News%20Articles.ipynb#7.-LSI-LDA-2) (Linear discriminant analysis)
9. 2 models \* 3 datasets = 6 results
10. I have found the optimal number of topics each model should generate by calling a technique called coherencemodel
11. I have saved the topics generated by 6 combinations into 6 columns and appended it to the input CSV
12. Below is the snapshot of the output CSV

Table

Description automatically generated with medium confidence

1. **SUMMARY TABLE –** In my opinion LDA Normal works best, I came to this conclusion after going through various texts, the topic relevancy is high with this combination.

|  |  |  |
| --- | --- | --- |
|  | **LSI** | **LDA** |
| **Normal** | Topics : 2  mobile, phone, film, award, best  labour, election, blair, tax, brown  Text  Description automatically generated | Topics : 7  casino, mobile, hsdpa, gambling, band  cabir, argonaut, mock, marvel, election  holmes, turkey, renault, job, johansson  phone, camera, mobile, music, bmw  conte, film, halifax, juninho, price  game, award, blair, bush, film  search, mobile, tax, labour, phone  Text  Description automatically generated |
| **Term\_Freq Filter** | Topics : 8  yukos, music, mobile, software, security  mobile, england, growth, phone, sales  labour, blair, election, people, brown  yukos, england, wales, kenteris, olympic  film, best, awards, england, award  labour, election, blair, brown, party  yukos, russian, gazprom, court, oil  film, mobile, best, economy, growth  Text  Description automatically generated | Topics : 3  blair, labour, party, election, prime  mobile, music, phones, holmes, sales  search, market, dollar, growth, china  Text  Description automatically generated |
| **Part of Speech Filter** | Topics : 1  election, blair, government, party, people  Text  Description automatically generated | Topics : 6  film, sales, tax, party, blair  search, software, attacks, sales, google  dollar, exports, juninho, growth, spyware  blair, ferguson, oil, government, posters  holmes, games, technology, players, game  blair, election, party, music, minister  Text  Description automatically generated |