AIML – Project Report

Extractive Text Summarization Using Topic Modelling

Abstract:

With the advancement of technology, more and more data is available in digital form. Among which, most of the data (approx. 85%) is in unstructured textual form. Text, so it has become essential to develop better techniques and algorithms to extract useful and interesting information from this large amount of textual data.

Extractive text summarization, a vital natural language processing technique, plays a pivotal role in distilling the essence of extensive textual content by meticulously identifying and selecting the most important sentences or phrases. This process allows for the creation of succinct summaries that encapsulate the key information within the original text. In the realm of extractive summarization, Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet process (HDP), powerful topic modelling algorithms, emerges as a valuable tool for unveiling latent themes within a corpus of text, thereby aiding in the extraction of salient content. In this project, we focus on combining the power of extractive text summarization and LDA, elucidating how this combination can enhance the process of text summarization.

Dataset : Eur-Lex [eurlex · Datasets at Hugging Face] :
EUR-Lex contains a wide range of legal documents, including primary legislation (such as treaties and directives) and secondary legislation (such as regulations and decisions). These documents cover various domains, including economic, environmental, social, and political aspects of EU law.

Project Literature Survey

Paper 1: The Similarity Measure Based on LDA for Automatic Summarization Tiedan Zhu, Kan Lia

Link : Sci-Hub | The Similarity Measure Based on LDA for Automatic Summarization. Procedia Engineering, 29, 2944–2949 | 10.1016/j.proeng.2012.01.419

The paper presents two algorithms for automatic summarization:

LMMR (Latent MMR): A modification of the Maximal Marginal Relevance (MMR) algorithm that uses LDA-Sim for sentence selection.

LSD (LDA Sentence Descending): An algorithm that removes less important sentences one by one until the summary reaches the desired length, with importance judged using LDA-Sim.

The paper conducted experiments on DUC (Document Understanding Conference) datasets from multiple years. LMMR and LSD both performed well, with LSD showing the best results among the tested methods.

Paper 2: LATENT DIRICHLET LEARNING FOR DOCUMENT SUMMARIZATION Ying-Lang Chang and Jen-Tzung Chien

Link : Sci-Hub | Latent Dirichlet learning for document summarization. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing | 10.1109/ICASSP.2009.4959927

The authors propose a hierarchical representation of words, sentences, and documents using LDA. They introduce SLDA, a sentence-based extension of LDA, designed for document summarization. SLDA is described as a sentence mixture model with a mixture of Dirichlet themes, used for generating latent topics in observed words.

The paper discusses the Bayesian variational inference scheme used to estimate the parameters of SLDA.

The paper reports experimental results comparing VSM (vector space model), traditional LDA, and SLDA. SLDA consistently outperforms the other methods in terms of precision, recall, and F-measure.

Paper 3: Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering Erwin Yudi Hidayata, Fahri Firdausillaha, Khafiizh Hastutia, Ika Novita Dewia, Azharib

Link : (PDF) Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering (researchgate.net)

The paper discusses the application of Latent Dirichlet Allocation (LDA) in automatic text summarization to improve the accuracy of document clustering. The research involves using a dataset of 398 public blog articles obtained using a Python scrapy crawler and scraper. The main steps of the research include preprocessing, automatic document compression using feature methods, automatic document compression using LDA, word weighting, and clustering algorithms.

The paper reports that automatic document summarization with LDA achieved a 72% accuracy rate in LDA 40%, compared to traditional k-means clustering, which only reached 66%.

The paper also includes a literature review on web crawling, text mining, clustering methods, and various techniques related to text summarization and document clustering.

Paper 4: Legal Document Summarization using Latent Dirichlet Allocation

Link : http://www.ijcst.org/Volume3/Issue7/p23_3_7.pdf

The paper presents a method for automatically summarizing legal judgments using Latent Dirichlet Allocation (LDA) topic modeling. The authors address the challenge of handling the vast amount of legal documents by extracting essential information and condensing it into concise summaries.The method entails segmenting documents into seven separate categories and creating summaries based on each of these topics. On a dataset of civil case judgements, the authors ran experiments comparing the summaries produced by the system and those produced by legal professionals. In terms of precision, recall, and F-measure, the results demonstrate encouraging performance. The suggested approach may greatly help legal professionals swiftly extract vital information from legal papers. The authors propose that future research might concentrate on improving the summary procedure.

Paper 5: Multi-document Text Summarization Using Topic Model

Link : https://link.springer.com/chapter/10.1007/978-3-642-39712-

The research paper focuses on a novel approach for multi-document text summarization, which combines topic modeling and fuzzy logic. The proposed method involves extracting relevant topic words from source documents, using them as elements of fuzzy sets. Additionally, each sentence in the source document contributes to generating a fuzzy relevance rule, indicating the importance of each sentence. A fuzzy inference system then produces the final summarization.The system design comprises pre-processing stages like sentence boundary detection, stop-word removal, and stemming. The sentences are then scored based on various features, including TF-ISF score, title score, length score, and position score. A fuzzy system is employed to make decisions about sentence importance.In conclusion, the research paper introduces an innovative method for multi-document summarization, leveraging topic modeling and fuzzy logic. The approach shows promise in automating the summarization process, but there is potential for enhancements, such as exploring more advanced fuzzy set systems and applying the approach to specific domains like medical documents.

Methodology

1. Preprocessing:

- Tokenization: The text is divided into smaller units like words, phrases, or sentences.
- Stopword Removal: Common words (e.g., 'and', 'the', 'is') that hold little semantic value are eliminated.
- Bigramarization: Combining bigrams from the tokens to make the topic modelling more efficient.
- Normalization: Words are converted to their base or root form (lemmatization or stemming) to reduce variant forms.

2. Topic Modelling with LDA or HDP:

Latent Dirichlet Allocation (LDA): It's a probabilistic model where each document is assumed to be a mixture of a set of topics, and each topic is a mixture of words. LDA aims to discover these topics and their word distributions.

Training the LDA Model: The algorithm is applied to the pre-processed text corpus to learn the topics present in the documents.

Topic Distribution: Each document is then represented as a distribution over the discovered topics.

Identifying Important Topics: Relevant topics containing key information are identified based on their prevalence and significance.

Hierarchical Dirichlet process (HDP): It is a non-parametric Bayesian Approach, that infers the number of topics automatically from the data. An extension of LDA that allows for an infinite number of topics, making it particularly useful for scenarios where the number of topics is unknown.

Modelling Latent Structure: It assumes a hierarchy of topics where each document can have an unbounded number of topics.

3. Summary Generation:

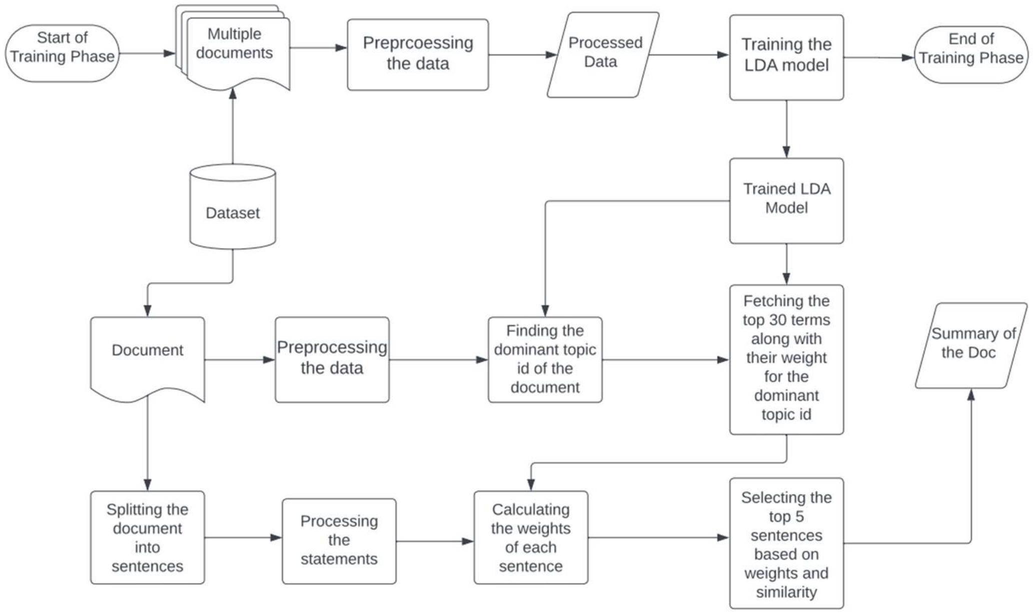
Dominant Topic: Gathering dominant topic for a document to summarize.

Top terms: Gathering top terms and their weights with respect to the dominant topic id from the model.

Sentence Splitting: Sentences are spitted and is stored in an array for further processing.

Sentence Processing: The Sentences which passes minimum length and maximum length criteria are then tokenized and then processed into bag of words format.

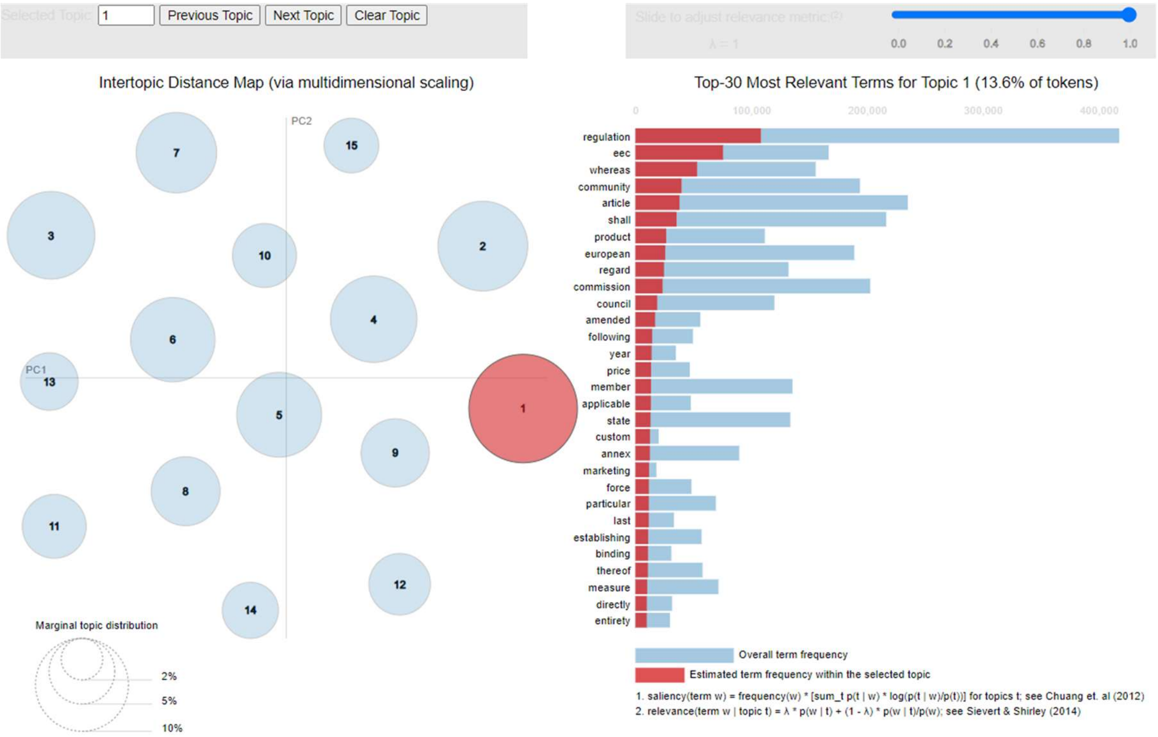
Sentence Selection: Weights for each sentence is calculated with the help of weights of the terms gathered from the dominant topic id and top k sentences are selected based on their weights and the similarity between the already chosen sentences.


Start of Training Phase
Multiple documents
Preprcoessing the data
Processed Data
Training the LDA model
End of Training Phase
Dataset
Trained LDA Model
Document
Preprocessing the data
Finding the dominant topic id of the document
Fetching the top 30 terms along with their weight for the dominant topic id
Summary of the Doc
Splitting the document into sentences
Processing the statements
Calculating the weights of each sentence
Selecting the top 5 sentences based on weights and similarity

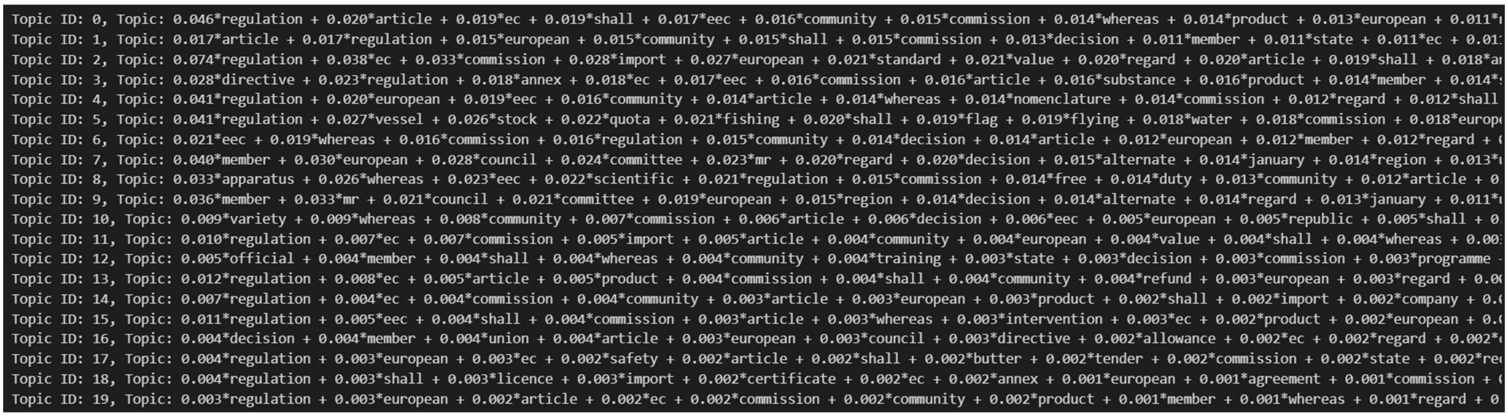
Block Diagram – Extractive Text Summarization using Topic Modelling

Results

We trained our LDA models for 15 topics and the HDP model generated 20 topics based on the input dataset which consisted of 45000 legal documents. Both the models generated the dominant topic id with respect to their inference generated by them during the training respectively.

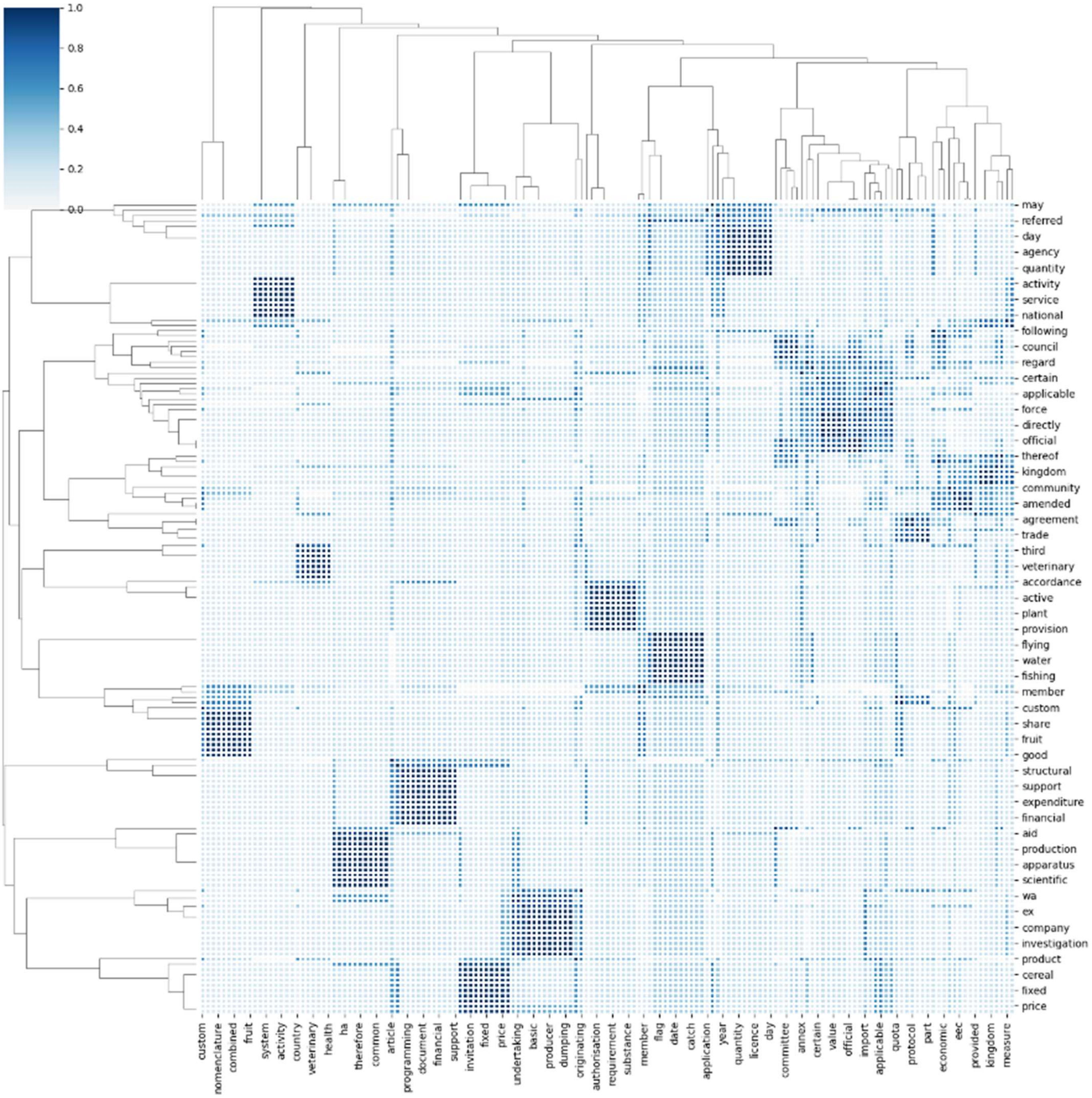


15 Topics Generated by LDA



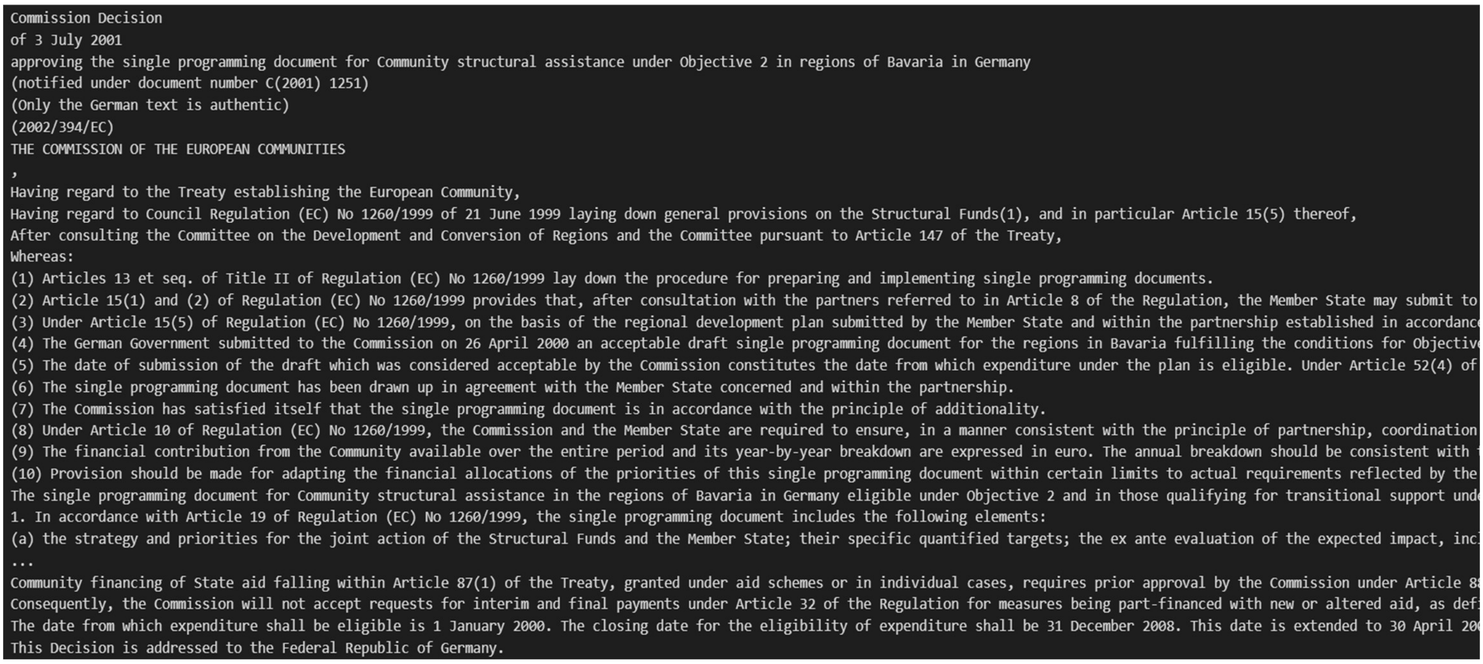
```
Topic ID: 0, Topic: 0.046*regulation + 0.020*article + 0.019*ec + 0.019*shall + 0.017*eec + 0.016*community + 0.015*commission + 0.014*whereas + 0.014*product + 0.013*european + 0.011*
Topic ID: 1, Topic: 0.017*article + 0.017*regulation + 0.015*european + 0.015*community + 0.015*shall + 0.015*commission + 0.013*decision + 0.011*member + 0.011*state + 0.011*ec + 0.01
Topic ID: 2, Topic: 0.074*regulation + 0.038*ec + 0.033*commission + 0.028*import + 0.027*european + 0.021*standard + 0.021*value + 0.020*regard + 0.020*article + 0.019*shall + 0.018*a
Topic ID: 3, Topic: 0.028*directive + 0.023*regulation + 0.018*annex + 0.018*ec + 0.017*eec + 0.016*commission + 0.016*article + 0.016*substance + 0.016*product + 0.014*member + 0.014*
Topic ID: 4, Topic: 0.041*regulation + 0.020*european + 0.019*eec + 0.016*community + 0.014*article + 0.014*whereas + 0.014*nomenclature + 0.014*commission + 0.012*regard + 0.012*shall
Topic ID: 5, Topic: 0.041*regulation + 0.027*vessel + 0.026*stock + 0.022*quota + 0.021*fishing + 0.020*shall + 0.019*flag + 0.019*flying + 0.018*water + 0.018*commission + 0.018*europ
Topic ID: 6, Topic: 0.021*eec + 0.019*whereas + 0.016*commission + 0.016*regulation + 0.015*community + 0.014*decision + 0.014*article + 0.012*european + 0.012*member + 0.012*regard +
Topic ID: 7, Topic: 0.040*member + 0.030*european + 0.028*council + 0.024*committee + 0.023*mr + 0.020*regard + 0.020*decision + 0.015*alternate + 0.014*january + 0.014*region + 0.013*
Topic ID: 8, Topic: 0.033*apparatus + 0.026*whereas + 0.023*eec + 0.022*scientific + 0.021*regulation + 0.015*commission + 0.014*free + 0.014*duty + 0.013*community + 0.012*article + 0
Topic ID: 9, Topic: 0.036*member + 0.033*mr + 0.021*council + 0.021*committee + 0.019*european + 0.015*region + 0.014*decision + 0.014*alternate + 0.014*regard + 0.013*january + 0.011*
Topic ID: 10, Topic: 0.009*variety + 0.009*whereas + 0.008*community + 0.007*commission + 0.006*article + 0.006*decision + 0.006*eec + 0.005*european + 0.005*republic + 0.005*shall + 0
Topic ID: 11, Topic: 0.010*regulation + 0.007*ec + 0.007*commission + 0.005*import + 0.005*article + 0.004*community + 0.004*european + 0.004*value + 0.004*shall + 0.004*whereas + 0.00
Topic ID: 12, Topic: 0.005*official + 0.004*member + 0.004*shall + 0.004*whereas + 0.004*community + 0.004*training + 0.003*state + 0.003*decision + 0.003*commission + 0.003*programme
Topic ID: 13, Topic: 0.012*regulation + 0.008*ec + 0.005*article + 0.005*product + 0.004*commission + 0.004*shall + 0.004*community + 0.004*refund + 0.003*european + 0.003*regard + 0.0
Topic ID: 14, Topic: 0.007*regulation + 0.004*ec + 0.004*commission + 0.004*community + 0.003*article + 0.003*european + 0.003*product + 0.002*shall + 0.002*import + 0.002*company + 0.
Topic ID: 15, Topic: 0.011*regulation + 0.005*eec + 0.004*shall + 0.004*commission + 0.003*article + 0.003*whereas + 0.003*intervention + 0.003*ec + 0.002*product + 0.002*european + 0.
Topic ID: 16, Topic: 0.004*decision + 0.004*member + 0.004*union + 0.004*article + 0.003*european + 0.003*council + 0.003*directive + 0.002*allowance + 0.002*ec + 0.002*regard + 0.002*
Topic ID: 17, Topic: 0.004*regulation + 0.003*european + 0.003*ec + 0.002*safety + 0.002*article + 0.002*shall + 0.002*butter + 0.002*tender + 0.002*commission + 0.002*state + 0.002*re
Topic ID: 18, Topic: 0.004*regulation + 0.003*shall + 0.003*licence + 0.003*import + 0.002*certificate + 0.002*ec + 0.002*annex + 0.001*european + 0.001*agreement + 0.001*commission +
Topic ID: 19, Topic: 0.003*regulation + 0.003*european + 0.002*article + 0.002*ec + 0.002*commission + 0.002*community + 0.002*product + 0.001*member + 0.001*whereas + 0.001*regard + 0
```

20 Topics Generated by HDP



Clustered Heatmap - Extractive Text Summarization using Topic Modelling

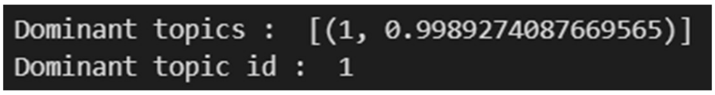
Sample document:



```
Commission Decision
of 3 July 2001
approving the single programming document for Community structural assistance under Objective 2 in regions of Bavaria in Germany
(notified under document number C(2001) 1251)
(Only the German text is authentic)
(2002/394/EC)
THE COMMISSION OF THE EUROPEAN COMMUNITIES
,
Having regard to the Treaty establishing the European Community,
Having regard to Council Regulation (EC) No 1260/1999 of 21 June 1999 laying down general provisions on the Structural Funds(1), and in particular Article 15(5) thereof,
After consulting the Committee on the Development and Conversion of Regions and the Committee pursuant to Article 147 of the Treaty,
Whereas:
(1) Articles 13 et seq. of Title II of Regulation (EC) No 1260/1999 lay down the procedure for preparing and implementing single programming documents.
(2) Article 15(1) and (2) of Regulation (EC) No 1260/1999 provides that, after consultation with the partners referred to in Article 8 of the Regulation, the Member State may submit to
(3) Under Article 15(5) of Regulation (EC) No 1260/1999, on the basis of the regional development plan submitted by the Member State and within the partnership established in accordance
(4) The German Government submitted to the Commission on 26 April 2000 an acceptable draft single programming document for the regions in Bavaria fulfilling the conditions for Objective
(5) The date of submission of the draft which was considered acceptable by the Commission constitutes the date from which expenditure under the plan is eligible. Under Article 52(4) of
(6) The single programming document has been drawn up in agreement with the Member State concerned and within the partnership.
(7) The Commission has satisfied itself that the single programming document is in accordance with the principle of additionality.
(8) Under Article 10 of Regulation (EC) No 1260/1999, the Commission and the Member State are required to ensure, in a manner consistent with the principle of partnership, coordination
(9) The financial contribution from the Community available over the entire period and its year-by-year breakdown are expressed in euro. The annual breakdown should be consistent with t
(10) Provision should be made for adapting the financial allocations of the priorities of this single programming document within certain limits to actual requirements reflected by the
The single programming document for Community structural assistance in the regions of Bavaria in Germany eligible under Objective 2 and in those qualifying for transitional support unde
1. In accordance with Article 19 of Regulation (EC) No 1260/1999, the single programming document includes the following elements:
(a) the strategy and priorities for the joint action of the Structural Funds and the Member State; their specific quantified targets; the ex ante evaluation of the expected impact, incl
...
Community financing of State aid falling within Article 87(1) of the Treaty, granted under aid schemes or in individual cases, requires prior approval by the Commission under Article 8
Consequently, the Commission will not accept requests for interim and final payments under Article 32 of the Regulation for measures being part-financed with new or altered aid, as defi
The date from which expenditure shall be eligible is 1 January 2000. The closing date for the eligibility of expenditure shall be 31 December 2008. This date is extended to 30 April 20
This Decision is addressed to the Federal Republic of Germany.
```

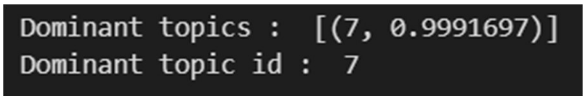
Dominant Topic Id generated by the models:

HDP



```
Dominant topics :  [(1, 0.9989274087669565)]
Dominant topic id :  1
```

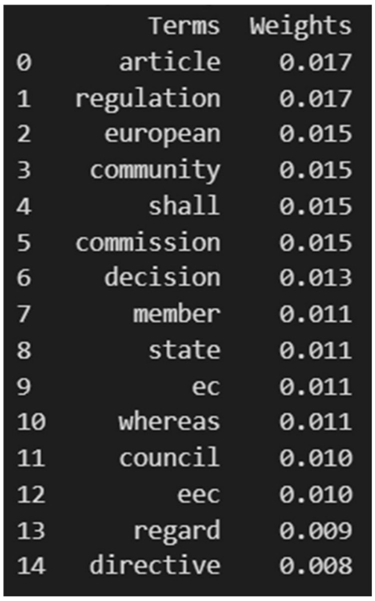
LDA



```
Dominant topics :  [(7, 0.9991697)]
Dominant topic id :  7
```

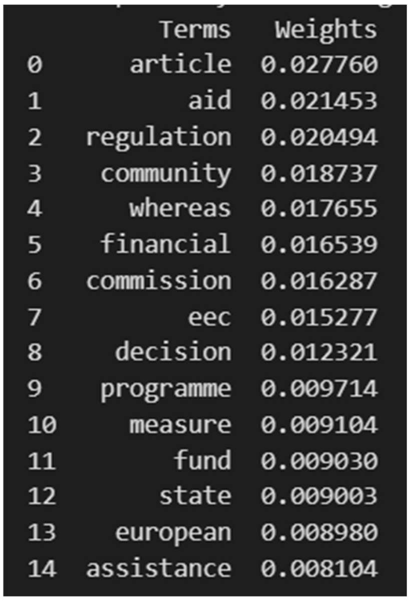
Top 15 terms for the topics generated by the models:

HDP



```
          Terms  Weights
0       article    0.017
1    regulation    0.017
2      european    0.015
3     community    0.015
4         shall    0.015
5    commission    0.015
6      decision    0.013
7        member    0.011
8         state    0.011
9            ec    0.011
10      whereas    0.011
11      council    0.010
12          eec    0.010
13       regard    0.009
14    directive    0.008
```

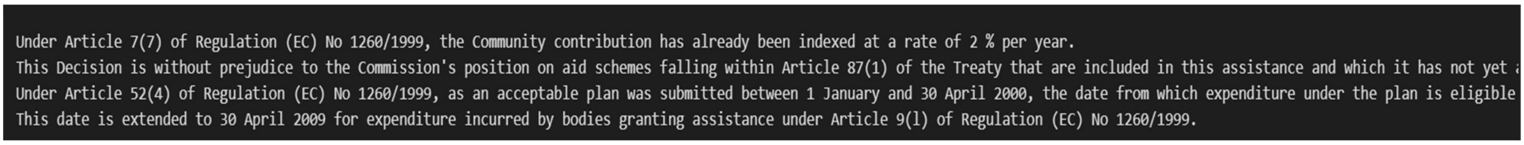
LDA



```
          Terms   Weights
0       article  0.027760
1           aid  0.021453
2    regulation  0.020494
3     community  0.018737
4       whereas  0.017655
5     financial  0.016539
6    commission  0.016287
7           eec  0.015277
8      decision  0.012321
9     programme  0.009714
10      measure  0.009104
11         fund  0.009030
12        state  0.009003
13     european  0.008980
14   assistance  0.008104
```

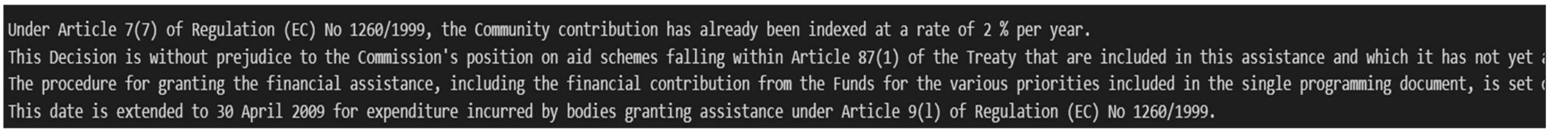
Summary of the given document:

HDP:



```
Under Article 7(7) of Regulation (EC) No 1260/1999, the Community contribution has already been indexed at a rate of 2 % per year.
This Decision is without prejudice to the Commission's position on aid schemes falling within Article 87(1) of the Treaty that are included in this assistance and which it has not yet a
Under Article 52(4) of Regulation (EC) No 1260/1999, as an acceptable plan was submitted between 1 January and 30 April 2000, the date from which expenditure under the plan is eligible
This date is extended to 30 April 2009 for expenditure incurred by bodies granting assistance under Article 9(l) of Regulation (EC) No 1260/1999.
```

LDA:



```
Under Article 7(7) of Regulation (EC) No 1260/1999, the Community contribution has already been indexed at a rate of 2 % per year.
This Decision is without prejudice to the Commission's position on aid schemes falling within Article 87(1) of the Treaty that are included in this assistance and which it has not yet a
The procedure for granting the financial assistance, including the financial contribution from the Funds for the various priorities included in the single programming document, is set o
This date is extended to 30 April 2009 for expenditure incurred by bodies granting assistance under Article 9(l) of Regulation (EC) No 1260/1999.
```

Conclusion

In conclusion, employing topic modeling techniques like Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) in extractive text summarization has showcased considerable promise and efficacy. These models have offered a structured approach to distil large volumes of text into coherent and informative summaries by identifying underlying topics and selecting salient sentences representing these topics. The flexibility of HDP, especially in handling an unknown number of topics, has enhanced the scalability and adaptability of summarization systems.