Machine Learning for NLP - 1

Exercise 6

Exercise Tasks:

- Use count vectoriser and Latent Dirichlet Allocation (LDA) to find common topics and trends Computer Science research among three periods
- Use Combined Topic Modelling to do the same.

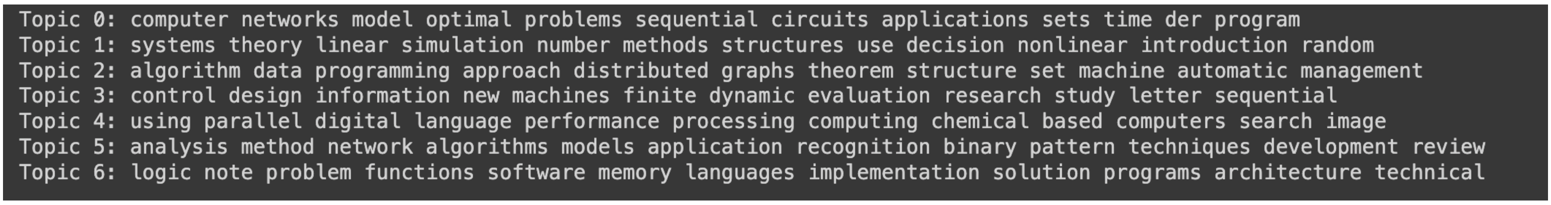
Preprocessing:

Our dataset comprises titles of research papers on which we perform LDA. Since the results of the LDA vary greatly depending on the preprocessing of the text we perform a number of experiments to determine the ideal set of hyper parameters. We choose them based on the coherence and distinctness of the topics. Specifically we do the following:

A. Remove all punctuation and non alphanumerics values.
B. Lowercase the text.
C. Try different N-gram ranges (1, 2) and (1, 3)
D. Try not removing the stop words
E. Change max_df i.e threshold of term frequency in document
F. Change maximum number of features

Finally we also experiment with different number of topics specifically 5, 7 and 10. Based on a qualitative analysis of the topics generated we choose 7 topics.

LDA Topic and Trend analysis:



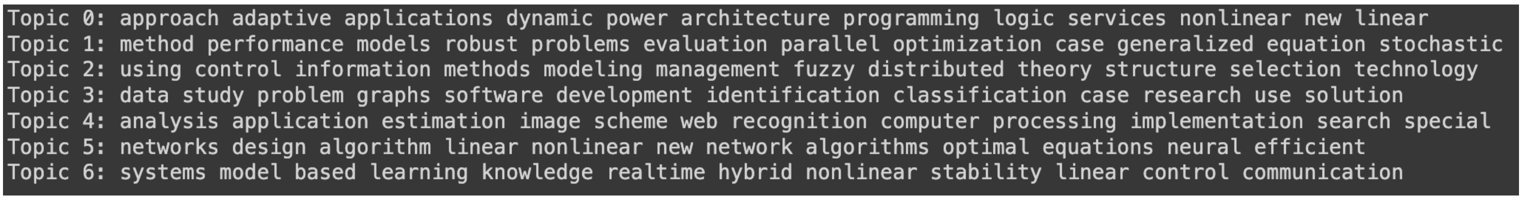
```
Topic 0: computer networks model optimal problems sequential circuits applications sets time der program
Topic 1: systems theory linear simulation number methods structures use decision nonlinear introduction random
Topic 2: algorithm data programming approach distributed graphs theorem structure set machine automatic management
Topic 3: control design information new machines finite dynamic evaluation research study letter sequential
Topic 4: using parallel digital language performance processing computing chemical based computers search image
Topic 5: analysis method network algorithms models application recognition binary pattern techniques development review
Topic 6: logic note problem functions software memory languages implementation solution programs architecture technical
```

Topics before 1990

Based on these results we claim our topics are:

1. Computer networks
2. Linear and non linear methods
3. Distributed algorithms and graph theorems

4. Design control and finite machines
5. Parallel processing and digital computing
6. Pattern recognition techniques
7. Logic implementation

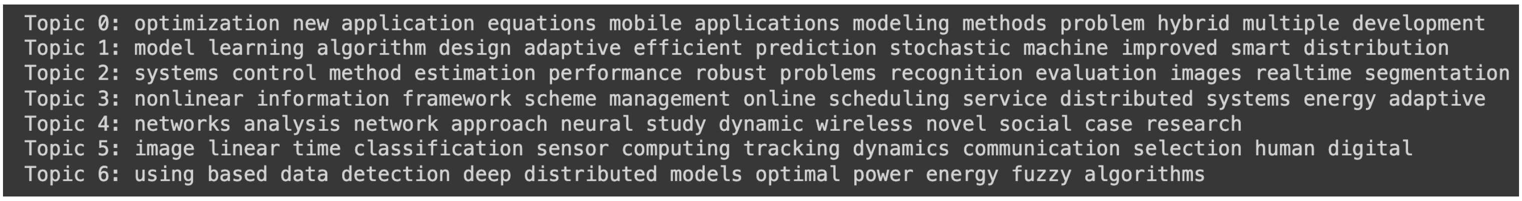


```
Topic 0: approach adaptive applications dynamic power architecture programming logic services nonlinear new linear
Topic 1: method performance models robust problems evaluation parallel optimization case generalized equation stochastic
Topic 2: using control information methods modeling management fuzzy distributed theory structure selection technology
Topic 3: data study problem graphs software development identification classification case research use solution
Topic 4: analysis application estimation image scheme web recognition computer processing implementation search special
Topic 5: networks design algorithm linear nonlinear new network algorithms optimal equations neural efficient
Topic 6: systems model based learning knowledge realtime hybrid nonlinear stability linear control communication
```

Topics from 1990 to 2010

Based on these results we claim our topics are:

1. Non linear and Linear approaches
2. Generalised and robust optimisation methods
3. Information theory and management
4. Identification and classification (using data and graphs possibly)
5. Image recognition and web search
6. Neural network algorithms
7. Realtime knowledge learning systems



```
Topic 0: optimization new application equations mobile applications modeling methods problem hybrid multiple development
Topic 1: model learning algorithm design adaptive efficient prediction stochastic machine improved smart distribution
Topic 2: systems control method estimation performance robust problems recognition evaluation images realtime segmentation
Topic 3: nonlinear information framework scheme management online scheduling service distributed systems energy adaptive
Topic 4: networks analysis network approach neural study dynamic wireless novel social case research
Topic 5: image linear time classification sensor computing tracking dynamics communication selection human digital
Topic 6: using based data detection deep distributed models optimal power energy fuzzy algorithms
```

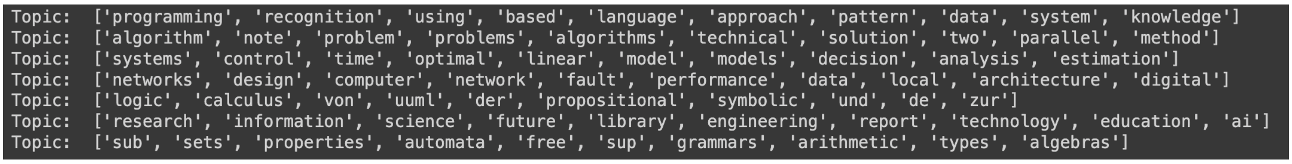
Topics from 2010 onwards

Based on these results we claim our topics are:

1. New mobile applications
2. Machine Learning and smart algorithms
3. Realtime image recognition and segmentation
4. Distributed systems
5. Neural networks
6. Image classification and human tracking
7. Data based models (perhaps Deep Learning)

We can see that there is a clear trend towards more and more Artificial Intelligence with more and more topics focusing on fields such as Image classification, optimisation and so on. There has been relatively stable interest in distributed systems and computer networks, while focus on digital computing, circuits, logic and finite state machines have reduced.

CTM Topic and Trend analysis:

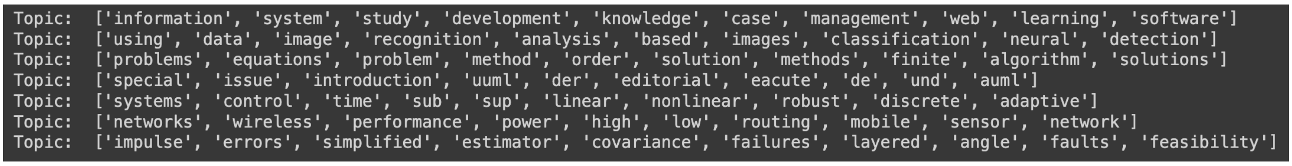


```
Topic:  ['programming', 'recognition', 'using', 'based', 'language', 'approach', 'pattern', 'data', 'system', 'knowledge']
Topic:  ['algorithm', 'note', 'problem', 'problems', 'algorithms', 'technical', 'solution', 'two', 'parallel', 'method']
Topic:  ['systems', 'control', 'time', 'optimal', 'linear', 'model', 'models', 'decision', 'analysis', 'estimation']
Topic:  ['networks', 'design', 'computer', 'network', 'fault', 'performance', 'data', 'local', 'architecture', 'digital']
Topic:  ['logic', 'calculus', 'von', 'uuml', 'der', 'propositional', 'symbolic', 'und', 'de', 'zur']
Topic:  ['research', 'information', 'science', 'future', 'library', 'engineering', 'report', 'technology', 'education', 'ai']
Topic:  ['sub', 'sets', 'properties', 'automata', 'free', 'sup', 'grammars', 'arithmetic', 'types', 'algebras']
```

Topics until 1990

The following topics can be inferred:

1. Pattern recognition using data
2. Parellel algorithms
3. Optimal decision model
4. Computer networks
5. Propositional logic and calculus
6. Science and engineering reports
7. Math in Computer Science (such as algebra, arithmetic and automata)

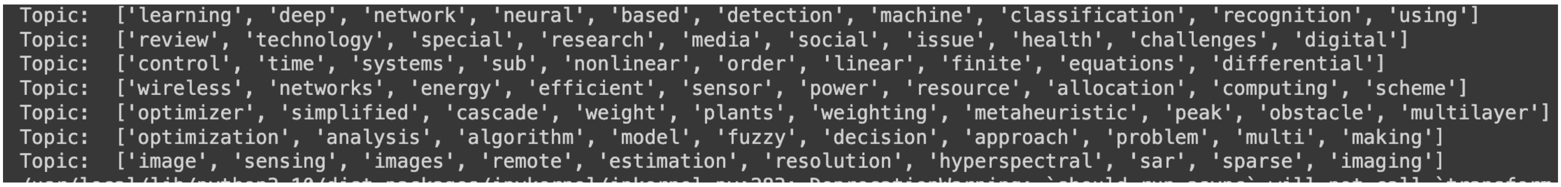


```
Topic:  ['information', 'system', 'study', 'development', 'knowledge', 'case', 'management', 'web', 'learning', 'software']
Topic:  ['using', 'data', 'image', 'recognition', 'analysis', 'based', 'images', 'classification', 'neural', 'detection']
Topic:  ['problems', 'equations', 'problem', 'method', 'order', 'solution', 'methods', 'finite', 'algorithm', 'solutions']
Topic:  ['special', 'issue', 'introduction', 'uuml', 'der', 'editorial', 'eacute', 'de', 'und', 'auml']
Topic:  ['systems', 'control', 'time', 'sub', 'sup', 'linear', 'nonlinear', 'robust', 'discrete', 'adaptive']
Topic:  ['networks', 'wireless', 'performance', 'power', 'high', 'low', 'routing', 'mobile', 'sensor', 'network']
Topic:  ['impulse', 'errors', 'simplified', 'estimator', 'covariance', 'failures', 'layered', 'angle', 'faults', 'feasibility']
```

Topics from 1990 to 2010

The following topics can be inferred:

1. Information system
2. Image recognition, classification and detection using neural networks.
3. Finite order methods
4. No clear topic
5. Robust and adaptive discrete control systems
6. Wireless mobile networks
7. No clear topics



```
Topic:  ['learning', 'deep', 'network', 'neural', 'based', 'detection', 'machine', 'classification', 'recognition', 'using']
Topic:  ['review', 'technology', 'special', 'research', 'media', 'social', 'issue', 'health', 'challenges', 'digital']
Topic:  ['control', 'time', 'systems', 'sub', 'nonlinear', 'order', 'linear', 'finite', 'equations', 'differential']
Topic:  ['wireless', 'networks', 'energy', 'efficient', 'sensor', 'power', 'resource', 'allocation', 'computing', 'scheme']
Topic:  ['optimizer', 'simplified', 'cascade', 'weight', 'plants', 'weighting', 'metaheuristic', 'peak', 'obstacle', 'multilayer']
Topic:  ['optimization', 'analysis', 'algorithm', 'model', 'fuzzy', 'decision', 'approach', 'problem', 'multi', 'making']
Topic:  ['image', 'sensing', 'images', 'remote', 'estimation', 'resolution', 'hyperspectral', 'sar', 'sparse', 'imaging']
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform
```

Topics from 2010 onwards

The following topics can be inferred:

1. Deep Learning
2. Health issues relating to social media
3. Finite order methods and differential equations
4. Energy efficient sensors
5. No clear topics
6. Model optimisation and fuzzy algorithms
7. Remote imaging sensing

There is still a trend towards Artificial Intelligence especially images and deep learning. There has been relatively stable interest in distributed systems and computer networks, while focus on digital computing, circuits, logic and finite state machines have reduced.

Comparison and Conclusion

In general Combined Topic Modelling seems to perform much better with topics more coherent and distinctive. It also covers topics that remain obscure using LDA. However in general there is a lot of overlap and due to the much greater computational demands of CTM, there is a case to be made for using LDA.