ML4NLP Exercise 6 - Topic Modeling

Data Preprocessing

In our data preprocessing approach, we implemented lowercasing and punctuation removal. While we considered employing stemming to reduce duplication (such as consolidating "algorithm" and "algorithms"), we ultimately decided against it. This decision was based on the observation that stemming can sometimes produce results that are less intuitive or meaningful. For example, stemming would transform "optimal" into "optim," which could detract from the clarity and interpretability of the data. Therefore, we opted not to use stemming in our preprocessing steps.

Part 1: Topic Modeling using LDA

For each period, assign a name to each generated topic based on the topic's top words. List all topic names in your report. If a topic is incoherent to the degree that no common theme is detectable, you can just mark it as incoherent (i.e., no need to name a topic that does not exist).
The result of the topic modeling in each period is shown in below tables:

Table 1: LDA Result before 1990

Topic	Content	Name
1	note problem functions optimal method technical linear decision solution problems algorithm using	algorithm
2	control new implementation digital optimal linear approach design theory using systems problems	incoherent
3	software processing applications finite research parallel digital computer data design theory information	software
4	analysis application languages performance algorithms networks theory computer data design digital linear	incoherent
5	programming simulation linear digital computer problems language languages approach parallel using networks	programming
6	design algorithm data networks information approach digital performance using parallel computer linear	data analysis
7	computer using theory linear problems algorithms parallel models digital performance decision design	incoherent
8	language recognition sets time pattern solution linear using problems parallel problem approach	incoherent
9	logic distributed programs parallel networks using computer functions approach design algorithms theory	software development
10	systems model network linear performance computer digital decision design information models theory	computer system

Table 2: LDA Result from 1990 to 2009

Topic	Content	Name
1	algorithm new linear problem algorithms optimal robust equations efficient detection optimization multiple	algorithm
2	networks approach nonlinear network models problems neural wireless mobile evaluation scheduling robust	neural network
3	systems based distributed nonlinear linear robust control approach optimal adaptive evaluation detection	incoherent
4	control analysis methods software development computing robust nonlinear optimal linear adaptive problems	incoherent
5	applications scheme web power efficient wireless mobile new robust control networks evaluation	wireless network
6	model performance image time graphs parallel digital evaluation algorithms robust scheduling optimal	image processing
7	using method dynamic simulation equations nonlinear detection models multiple problems efficient new	incoherent
8	adaptive application estimation learning modeling fuzzy theory recognition robust nonlinear control approach	adaptive learning
9	design information management evaluation approach robust development systems network new optimal mobile	information management
10	data study programming models approach linear multiple evaluation analysis using problems algorithms	data analysis

Table 3: LDA Result 2010 onwards

Topic	Content	Name
1	detection design linear mobile time recognition computing stochastic images segmentation optimal deep	image recognition
2	control learning efficient machine deep optimal nonlinear tracking adaptive distributed energy fuzzy	deep learning
3	networks model algorithm wireless novel sensor energy based improved distributed optimal neural	wireless networks
4	analysis optimization study power distributed hybrid multiple case feature smart performance energy	information management
5	information application framework applications prediction methods management social scheduling cloud computing energy	incoherent
6	data approach image nonlinear new classification online tracking improved research review based	incoherent
7	systems evaluation sensing nonlinear linear performance stochastic distributed fuzzy optimal control tracking	optimisation
8	method estimation robust scheme problem problems communication nonlinear optimal based new improved	mathematical analysis
9	network adaptive neural dynamic modeling sensor selection deep nonlinear wireless feature based	neural network
10	using based deep performance models optimal energy fuzzy machine algorithms classification improved	machine learning

- **Do the topics make sense to you? Are they coherent? Do you observe trends? Discuss in 4-6 sentences.** Some of the topics do not make much sense, and look more like combinations of irrelevant keywords. For example, in the topic 6 in Table 3, the keywords ”image”, ”classification” and ”data” and ”nonlinear” are

all related to image recognition and machine learning, but ”based”, ”review”, ”online”, ”tracking”, ”new” are not much relevant.
Some of the topics are coherent, and the keywords are related to each other. For example, in the topic 1 in Table 3, the keywords ”detection”, ”linear”, ”recognition”, ”computing”, ”stochastic”, ”images”, ”segmentation”, ”optimal” and ”deep” are all related to image recognition and machine learning.
Based on the information from the three tables, it's clear that the topics prior to 1990 were predominantly focused on computer systems and software development. However, from 1990 onwards, the fields of machine learning and deep learning have gained increasing popularity.

Part 2: Topic Modeling using Combined Topic Models (CTMs)

- **Again: Assign a name to each topic based on the topic's top words (for each period). List all topic names in your report.**
  The result of the topic modeling in each period is shown in below tables:

Table 4: CTM Result before 1990

Topic	Content	Name
1	logic theory theorem symbolic set association logics modal meeting proof calculus propositional	logical theory
2	information computer system data management retrieval systems review science database chemical decision	information system
3	recognition using analysis digital image pattern method approach application processing detection images	image recognition
4	networks network simulation von der performance de zur und communication des local	incoherent
5	computers introduction research future operations chess editor letter report technology guest history	incoherent
6	language design languages programming software implementation program hardware machine memory environment development	software engineering
7	control systems optimal model time linear stochastic adaptive estimation analysis nonlinear distributed	system optimization
8	magnetic surfaces degrees risk sets focus conversion forecasting enumerable generalization geometric compact	physics
9	algorithm problem note problems technical linear solution programming equations scheduling algorithms integer	algorithm
10	graphs automata trees binary machines finite sequential algorithms number random complexity grammars	theory of computation

Table 5: CTM Result from 1990 to 2009

Topic	Content	Name
1	systems control linear robust nonlinear stability feedback adaptive optimal uncertain discretetime output	incoherent
2	number graphs complete trees classes automata groups note graph theorem sets complexity	theory of computation
3	system design distributed software decision process framework applications support development language simulation	system
4	networks wireless mobile network routing access performance sensor protocol dynamic efficient service	wireless networks
5	information web review research technology electronic paper book case online internet use	electronic resources
6	using image based recognition classification images detection neural segmentation feature face algorithm	image recognition
7	problems problem method equations solutions methods solution solving numerical optimization boundary differential	mathematical analysis
8	analysis data models study model molecular functional dynamics structure comparison brain fmri	medical data anaylsis
9	power frequency channels cmos channel estimation low signal circuit high noise modulation	signal processing
10	underwater editorial terminal und der section interacting optimisation von fuumlr vehicles guest	incoherent

Table 6: CTM Result 2010 onwards

Topic	Content	Name
1	optimization multiobjective algorithm system hybrid problem power swarm scheduling particle electric planning	incoherent
2	image images sparse segmentation fusion reconstruction feature based sensing remote color detection	image recognition
3	number selfadaptive complexity spaces minimum graphs degree note weight metric bound multi	graph theory
4	molecular magnetic field connectivity thermal temperature resonance measurements functional changes radiation simulations	physics
5	nonlinear systems linear class equations control differential equation boundary solutions stability fractional	mathematical analysis
6	computing cloud internet smart applications things special security iot secure issue edge	IT security
7	learning deep neural network machine classification recognition convolutional detection prediction using graph	deep learning
8	analysis fuzzy decision data model series models time making regression using application	fuzzy regression
9	online social technology media knowledge review information case research digital use perspective	degital technology
10	wireless networks sensor allocation channel channels cognitive radio performance relay communication massive	wireless networks

- **Bianchi et al. 2021 claim that their approach produces more coherent topics than previous methods. Let's test this claim by comparing the coherence of the topics produced by CTM with the topics produced by**

LDA. Describe your observations in 3-4 sentences.
In our analysis, we've observed that the topics generated by CTM are notably more coherent compared to those produced by LDA. This distinction is evident in our results, where LDA shows a higher degree of overlapping keywords across various topics, leading to more instances of incoherence and greater ambiguity in the thematic interpretation. In contrast, the results from CTM display less overlap between topics, which simplifies the process of identifying a common theme for each topic, thereby enhancing the overall clarity and interpretability of the model.

- **Do the two models generate similar topics? Can you discover the same temporal trends (if there are any)? Discuss in 5-6 sentences.**
  Both models exhibit similar trends in topic generation for the period before 1990, with keywords related to algorithm systems and software engineering being prominent in both.
  However, CTM produces a more diverse range of topics compared to those generated by LDA. We can find topics like theory of computation, graph theory, IT security, etc. from the results produced by CTM.
  Despite this diversity, the overall trends in both models align: topics primarily focus on computer systems prior to 1990, shifting towards a growing emphasis on machine learning after the 90s.
  For CTM, we also discovered that the topics became more diverse after 1990.