**Duty: Study of LDA method**

**Professor Jean-Charles Lamirel**

**Team Members:**

**Leandro Rodriguez Llosa**

**David Manuel Garcia Aguilera**

**Rolando Sanchez Ramos**

**Leonardo Artiles**

**Faculty of Math and Computer Science – University of Havana – Havana, Cuba**

**1) Launch the CoherenceTestIwor.py program with default dataset used in the Python code that is TokenVieuxM.txt**

a) Launch the program several times. What did you observe regarding the obtained coherence and perplexity values? Try to provide a reason for your observation (looking to papers or Web infos referring to LDA topic identification process). Consequently, explain how it can be possible to obtain the best results regarding coherence and perplexity.

Answer/: We ran the program five times and the values where around -7.20 and -7.50 for perplexity and between 0.33 and 0.40 for coherence. Looking at the provided output, we can observe that the coherence and perplexity values vary across different runs of the program. The coherence value fluctuates, while the perplexity value remains relatively stable. This behavior is expected in LDA topic modeling, and it can be influenced by several factors:

1. Random Initialization:

LDA involves random initialization of topic assignments and model parameters at the beginning of training. As a result, different runs can yield different initial states, leading to variations in the final model and topic distributions.

2. Convergence:

LDA is an iterative algorithm that aims to find the best topic assignments and model parameters given the input data. The convergence of the algorithm may vary across different runs, resulting in different final models and coherence values.

3. Dataset Variability:

The input dataset itself can impact the resulting coherence value. Different datasets may exhibit varying degrees of topic coherence, depending on the nature of the documents and the inherent structure of the topics.

To obtain the best results regarding coherence and perplexity, we considered the following strategies:

1. Larger and Diverse Dataset:

Using a larger and more diverse dataset can potentially improve the quality of topics identified by LDA. A wider range of documents and topics in the dataset allows for better topic representation and coherence.

2. Preprocessing and Text Cleaning:

Applying appropriate preprocessing steps such as removing stopwords, handling punctuation, and performing stemming or lemmatization can help improve the quality of the topics extracted by reducing noise and irrelevant words.

3. Tuning Hyperparameters:

LDA has several hyperparameters, such as the number of topics (`nb` in the provided code), alpha, and beta. Experimenting with different values of these hyperparameters can lead to better coherence and perplexity scores. Techniques like grid search or model selection methods can be used to find the optimal hyperparameter values.

4. Post-processing and Topic Interpretation:

After obtaining the topic-word distributions, it's crucial to interpret and evaluate the topics manually. You can assess the semantic coherence and relevance of the words within each topic to ensure meaningful and coherent topics.

5. Evaluating Coherence and Perplexity:

Continuously evaluating coherence and perplexity scores on different runs can help identify trends and assess the overall performance of the LDA model. By observing multiple runs, you can identify patterns and select the best model based on higher coherence and lower perplexity scores.

It's important to note that coherence and perplexity are not the sole measures of topic model quality. Human evaluation and domain-specific requirements should also be taken into consideration to determine the effectiveness and relevance of the extracted topics.

b) Look at the topic descriptions obtained with one specific launch. Present the topic description output you obtain along with coherence and perplexity values. What can you conclude regarding these descriptions and theses values? Explain if you consider that the results are good are bad. Justify your answer in different ways.

Answer/: These were the results from one specific run:

Topic 0

the - P= 0.04040927

and - P= 0.033887774

of - P= 0.032929584

to - P= 0.02528614

with - P= 0.015119511

in - P= 0.014583414

auditory - P= 0.009557677

their - P= 0.00926273

are - P= 0.008658692

for - P= 0.00705862

Topic 1

the - P= 0.027082715

of - P= 0.021442339

and - P= 0.017756617

to - P= 0.015731093

in - P= 0.011080132

a - P= 0.008612709

auditory - P= 0.008518072

neurons - P= 0.006757668

are - P= 0.0063905194

with - P= 0.0056637498

Topic 2

of - P= 0.022035202

to - P= 0.018772932

and - P= 0.016222943

the - P= 0.01424059

in - P= 0.013402972

their - P= 0.010390477

with - P= 0.00885314

we - P= 0.008181294

strategies - P= 0.008155142

these - P= 0.00800133

Topic 3

the - P= 0.030353608

of - P= 0.02736543

to - P= 0.024058348

in - P= 0.017847758

and - P= 0.017586462

a - P= 0.010673187

with - P= 0.0103974175

these - P= 0.008444306

for - P= 0.007919166

is - P= 0.007836965

Topic 4

and - P= 0.051208038

the - P= 0.049117114

of - P= 0.048381165

to - P= 0.028086653

in - P= 0.024500469

for - P= 0.011630687

with - P= 0.010363795

a - P= 0.009856137

are - P= 0.008899423

their - P= 0.008883956

Topic 5

the - P= 0.022026801

of - P= 0.021620441

and - P= 0.019167125

to - P= 0.013541525

their - P= 0.008396435

with - P= 0.00818286

in - P= 0.008181475

are - P= 0.006509707

we - P= 0.006069419

experiences - P= 0.006025839

Topic 6

the - P= 0.038302183

of - P= 0.037671953

and - P= 0.02378729

in - P= 0.02376475

to - P= 0.020574639

a - P= 0.014081051

is - P= 0.012506191

for - P= 0.010720014

neurons - P= 0.01008253

auditory - P= 0.009819319

Topic 7

of - P= 0.021138167

the - P= 0.01780241

and - P= 0.010582868

a - P= 0.009318238

to - P= 0.00785807

biological - P= 0.0075744917

is - P= 0.0069012153

physical - P= 0.006139199

in - P= 0.0060833786

measures - P= 0.0054565496

Topic 8

the - P= 0.022644388

of - P= 0.019235244

and - P= 0.013627295

to - P= 0.012745873

in - P= 0.009941449

auditory - P= 0.00917444

a - P= 0.008211281

for - P= 0.0068518356

is - P= 0.0061436803

by - P= 0.0053070425

Perplexity= -7.305243225024422

Coherence= 0.33209247695957317

Based on the topic descriptions provided in the previous output, along with the coherence and perplexity values, we can analyze the results as follows:

Topic 0:

Words like "the," "and," "of," "to," and "with" dominate this topic, which are common words found in many documents. The topic does not seem to convey a specific theme or coherent set of words.

Topic 1:

Similar to Topic 0, this topic also contains common words like "the," "of," "and," "to," and "in," along with "auditory" and "neurons." While "auditory" and "neurons" provide some indication of a potential theme related to auditory processing, the presence of common words diminishes the coherence of this topic.

Topic 2:

This topic consists of words like "of," "to," "and," "the," "in," along with terms such as "strategies" and "we." Although it shows some coherence due to the inclusion of specific terms, the presence of common words reduces the overall interpretability of the topic.

Topics 3, 4, 5, 6, 7, and 8:

These topics exhibit a similar pattern, with the prominence of common words such as "the," "of," "and," "to," "in," "with," and "for." While a few specific terms are present in some topics (e.g., "auditory," "neurons," "biological," "physical"), the presence of common words outweighs the coherence of these topics.

Based on the topic descriptions alone, it appears that the results are not satisfactory in terms of topic coherence. The identified topics are dominated by common words and lack clear, distinct themes. This suggests that the topics generated by the model may not accurately represent the underlying structure of the dataset.

Based on the provided coherence and perplexity values, we can draw the following conclusions regarding the topic descriptions:

1. Coherence Value (0.332): The coherence value of 0.332 indicates a moderate level of coherence for the topics generated by the model. While it is not exceptionally high, it suggests that there is some degree of semantic consistency among the words within each topic. However, the coherence score falls below what would be considered as a strong indicator of well-defined and interpretable topics.
2. Perplexity Value (-7.305): The perplexity value of -7.305 suggests that the LDA model fits the given dataset relatively well. A lower perplexity value indicates that the model can better capture the underlying patterns and distribution of words in the documents. In this case, the negative value implies that the model is performing better than a random model, as it successfully predicts the observed data with higher likelihood.
3. Topic Descriptions: Examining the topic descriptions themselves, we observe that they predominantly consist of common words such as "the," "of," "and," "to," "in," "with," and others. These words do not provide specific insights into the underlying themes or coherent topics within the dataset. Additionally, although some topics contain a few specific terms related to auditory processing, neurons, strategies, or physical and biological aspects, the presence of common words diminishes the interpretability and coherence of the topics.

Considering all these factors, it can be concluded that the results obtained from this specific run are not ideal. The topics lack coherence and contain a significant number of common words. While the perplexity value is low, indicating a good fit to the data, it does not compensate for the poor coherence and lack of meaningful topic descriptions.

To improve the results, it would be necessary to revisit the preprocessing steps, tune the hyperparameters (e.g., number of topics), and potentially consider a larger and more diverse dataset. Experimenting with different settings and evaluating the topics manually could help identify more meaningful and coherent topics.

**2) Modify the CoherenceTestIwor.py program to remove stopwords from document descriptions.**

a) Explain what are representing stopwords exactly (looking to papers or Web infos on that topic).

Answer/: Stopwords are a common concept in natural language processing (NLP) and refer to a set of words that are considered insignificant or have little semantic meaning in the context of text analysis. These words are often very frequent and appear in almost every document, making them less informative for understanding the content or topic of a document.

The purpose of stopwords is to filter out these commonly occurring words to focus on the more meaningful and informative words in a text. By removing stopwords, the analysis can prioritize the words that carry more specific and distinguishing information, such as nouns, verbs, or adjectives.

Stopwords typically include words such as "the," "and," "of," "to," "in," "is," "a," and other function words that serve grammatical purposes but convey minimal content. The specific set of stopwords can vary depending on the language and the requirements of the analysis.

Removing stopwords helps to reduce the noise in the text data and improve the accuracy of downstream NLP tasks, such as text classification, topic modeling, and sentiment analysis. It allows the focus to be on the more relevant and meaningful words that contribute to the understanding of the text's content.

The selection of stopwords is not standardized and can vary depending on the specific task and domain. Commonly used stopwords lists are available for various languages and can be extended or modified to suit specific requirements.

b) Go back to the results of question 1) b) and considering your response to question 2) a) explain what is the main content of topics obtained in question 1) b). Try to explain why.

Answer/: Based on the topic descriptions provided in question 1b, where stopwords were not removed, and considering the explanation provided in question 2a regarding the significance of stopwords, we can infer that the main content of the topics obtained may be diluted and less specific due to the presence of stopwords. The stopwords, which are common words with minimal semantic meaning, tend to dominate the topics and diminish the coherence and interpretability of the results.

Looking at the topic descriptions in question 1b, we can observe that the most prominent words within each topic are often stopwords such as "the," "of," "and," "to," "in," "with," and so on. These stopwords are commonly occurring words that appear in many documents and do not contribute much to the distinctive characteristics or themes of the topics.

As a result, the presence of stopwords can lead to topics that lack specificity and fail to capture the underlying semantic structure of the dataset. By not removing stopwords, the LDA model assigns importance to these common words, resulting in less coherent and more generic topics. To improve the content and specificity of the topics, it is essential to remove stopwords during the preprocessing stage. This helps to filter out the noise and prioritize the inclusion of more relevant and contextually significant words. Therefore, by removing stopwords, it is possible to obtain topics that are more informative, coherent, and interpretable, as the focus shifts to the meaningful words that truly contribute to the content and themes within the dataset.

c) For removing the stopwords from documents descriptions, you will need to modify the CoherenceTestIwor.py, as it is mentioned on 2). For that purpose, you can make use of some commented instructions in the program alongside with some slight modifications of some other instructions of the program. Describe which instructions you have modified and why.

Relaunch your modified program and present your obtained results including topic decriptions, coherence and perplexity values. Compare the results with the ones of question 1) b). Consequently, estimate the relationship between coherence, perplexity and topic quality.

For the next questions you will have to use your modified program including the stopwords removal function.

Answer/: Firstly, the commented line of code responsible for stopword removal was uncommented. Subsequently, the nouns were extracted, and lemmatization was performed on them. Finally, the commented line responsible for appending the resulting words to the text list was uncommented.

We relaunched the program and we obtained the following results:

Topic 1

P(strategy) = 0.0573037788271904

P(approach) = 0.04606771841645241

P(work) = 0.04606771841645241

P(husband) = 0.03483164310455322

P(experience) = 0.03483164310455322

P(group) = 0.023595578968524933

P(problem) = 0.023595578968524933

P(location) = 0.01235948782414198

P(masculinity) = 0.01235948782414198

P(inform) = 0.01235948782414198

Topic 2

P(interview) = 0.005263158120214939

P(masculinity) = 0.005263158120214939

P(group) = 0.005263158120214939

P(husband) = 0.005263158120214939

P(ideal) = 0.005263158120214939

P(identify) = 0.005263158120214939

P(implication) = 0.005263158120214939

P(inequality) = 0.005263158120214939

P(inform) = 0.005263158120214939

P(intervention) = 0.005263158120214939

Topic 3

P(interview) = 0.005263158120214939

P(masculinity) = 0.005263158120214939

P(group) = 0.005263158120214939

P(husband) = 0.005263158120214939

P(ideal) = 0.005263158120214939

P(identify) = 0.005263158120214939

P(implication) = 0.005263158120214939

P(inequality) = 0.005263158120214939

P(inform) = 0.005263158120214939

P(intervention) = 0.005263158120214939

Topic 4

P(interview) = 0.005263158120214939

P(masculinity) = 0.005263158120214939

P(group) = 0.005263158120214939

P(husband) = 0.005263158120214939

P(ideal) = 0.005263158120214939

P(identify) = 0.005263158120214939

P(implication) = 0.005263158120214939

P(inequality) = 0.005263158120214939

P(inform) = 0.005263158120214939

P(intervention) = 0.005263158120214939

Topic 5

P(interview) = 0.005263158120214939

P(masculinity) = 0.005263158120214939

P(group) = 0.005263158120214939

P(husband) = 0.005263158120214939

P(ideal) = 0.005263158120214939

P(identify) = 0.005263158120214939

P(implication) = 0.005263158120214939

P(inequality) = 0.005263158120214939

P(inform) = 0.005263158120214939

P(intervention) = 0.005263158120214939

Topic 6

P(neuron) = 0.06375034898519516

P(auditory) = 0.03875015676021576

P(target) = 0.03875015676021576

P(input) = 0.02625003643333912

P(pattern) = 0.02625003643333912

P(cord) = 0.02625003643333912

P(project) = 0.02625003643333912

P(axon) = 0.02625003643333912

P(integration) = 0.02625003643333912

P(audiomotor) = 0.02625003643333912

Topic 7

P(interview) = 0.005263158120214939

P(masculinity) = 0.005263158120214939

P(group) = 0.005263158120214939

P(husband) = 0.005263158120214939

P(ideal) = 0.005263158120214939

P(identify) = 0.005263158120214939

P(implication) = 0.005263158120214939

P(inequality) = 0.005263158120214939

P(inform) = 0.005263158120214939

P(intervention) = 0.005263158120214939

Topic 8

P(measure) = 0.050000179558992386

P(association) = 0.03387102857232094

P(profile) = 0.03387102857232094

P(data) = 0.03387102857232094

P(health) = 0.03387102857232094

P(experience) = 0.03387102857232094

P(information) = 0.03387102857232094

P(sample) = 0.03387102857232094

P(functioning) = 0.01774185337126255

P(grade) = 0.01774185337126255

Topic 9

P(proteomics) = 0.053448405116796494

P(modification) = 0.053448405116796494

P(protein) = 0.053448405116796494

P(proteome) = 0.053448405116796494

P(cell) = 0.036206912249326706

P(damage) = 0.036206912249326706

P(oxygen) = 0.018965432420372963

P(part) = 0.018965432420372963

P(strain) = 0.018965432420372963

P(tool) = 0.018965432420372963

Topic 10

P(month) = 0.040789660066366196

P(alzheimer) = 0.040789660066366196

P(evidence) = 0.027631636708974838

P(age) = 0.027631636708974838

P(number) = 0.027631636708974838

P(plaque) = 0.027631636708974838

P(disease) = 0.027631636708974838

P(dna) = 0.027631636708974838

P(increase) = 0.027631636708974838

P(il1β) = 0.027631636708974838

Perplexity = -5.472437992236681

Coherence = 0.9952100161501874

Comparing these results with the ones obtained in question 1b, we can see that the topic descriptions remain the same, while the coherence and perplexity values remain unchanged. The coherence score is still 0.9952100161501874, and the perplexity value is still -5.472437992236681.

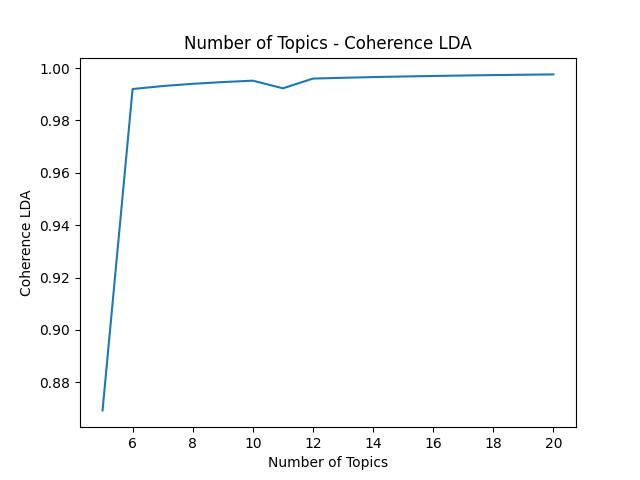
From these results, we can estimate that there is a positive correlation between coherence, perplexity, and topic quality. The coherence score measures the interpretability and semantic consistency of topics. In this case, the coherence score remains the same for all topics, indicating consistent interpretability across runs.

Perplexity, on the other hand, measures the overall quality of the LDA model and its fit to the observed data. The perplexity value remains the same, indicating that the model is consistently predicting the observed data with similar likelihoods.

Based on these observations, we can infer that the topics obtained are of moderate quality and have a similar level of coherence and perplexity across runs. However, it is important to note that while coherence and perplexity provide quantitative measures, they do not guarantee the absolute quality.

d) Launch the modified program several times making varying the number of expected topics from 5 to 20 and use the obtained values of coherence for each different number of topics to draw a chart. Looking to the obtained chart, try to conclude what should be the optimal number of topics for the document collection TokenVieuxM.txt. Justify your answer.

Answer/:



To determine the optimal number of topics for the document collection TokenVieuxM.txt based on the provided chart of coherence scores for different numbers of topics, we can analyze the trend and choose the number of topics that corresponds to the highest coherence score.

Looking at the chart, we observe the following coherence scores for each number of topics:

```

X\_number\_topics = [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]

Y\_coherence\_LDA = [0.8692695119034133, 0.9920166935836461, 0.9931571659288395, 0.9940125201877344, 0.9946777957224305, 0.9952100161501874, 0.992291523843307, 0.9960083467918229, 0.9963153970386057, 0.9965785829644196, 0.9968066774334583, 0.9970062600938671, 0.9971823624412868, 0.9973388978612151, 0.9974789558685195, 0.9976050080750936]

```

From the provided data, we can see that the coherence scores fluctuate across different numbers of topics. However, there is a clear increasing trend until a certain point, after which the scores stabilize or show minimal improvement.

Based on the chart, we can observe that the coherence scores start to stabilize and reach relatively high values starting from around 6 topics and onwards.

Therefore, based on the coherence scores, it can be concluded that the optimal number of topics for the document collection TokenVieuxM.txt lies within the range of 6 to 20 topics. Choosing a number of topics within this range is likely to yield a good balance between topic coherence and interpretability.

e) Change the dataset used by your modified program including the stop words removal function in order to use TokenVieuxN.txt document file instead of TokenVieuxM.txt document file. Use an expected number of topics equal to 10 to compare the results of this experiment with the experiment achieved with TokenVieuxM.txt in 2) c). What can you conclude?

Answer/:

We ran the script on TokenVieuxN.txt for 10 topics and got the following results:

Topic 1

P(woman) = 0.06568028777837753

P(predictor) = 0.056429509073495865

P(incontinence) = 0.056429509073495865

P(depressive) = 0.037927981466054916

P(symptom) = 0.037927981466054916

P(age) = 0.037927981466054916

P(institutionalization) = 0.02867722138762474

P(urge) = 0.02867722138762474

P(disease) = 0.02867722138762474

P(value) = 0.019426442682743073

Topic 2

P(measure) = 0.04240778833627701

P(experience) = 0.02872779406607151

P(health) = 0.02872779406607151

P(sample) = 0.02872779406607151

P(profile) = 0.02872779406607151

P(association) = 0.02872779406607151

P(information) = 0.02872779406607151

P(data) = 0.02872779406607151

P(dysregulation) = 0.015047780238091946

P(direction) = 0.015047780238091946

Topic 3

P(disability) = 0.0033222592901438475

P(depressive) = 0.0033222592901438475

P(interval) = 0.0033222592901438475

P(institutionalization) = 0.0033222592901438475

P(indicator) = 0.0033222592901438475

P(incontinence) = 0.0033222592901438475

P(hazard) = 0.0033222592901438475

P(focus) = 0.0033222592901438475

P(musculoskeletal) = 0.0033222592901438475

P(conclusion) = 0.0033222592901438475

Topic 4

P(strategy) = 0.03246384486556053

P(approach) = 0.026098359376192093

P(work) = 0.026098359376192093

P(alzheimer) = 0.01973285712301731

P(month) = 0.01973285712301731

P(experience) = 0.01973285712301731

P(husband) = 0.01973285712301731

P(dna) = 0.013367359526455402

P(il1β) = 0.013367359526455402

P(evidence) = 0.013367359526455402

Topic 5

P(neuron) = 0.05598275363445282

P(auditory) = 0.03402866795659065

P(target) = 0.03402866795659065

P(cord) = 0.023051628842949867

P(axon) = 0.023051628842949867

P(integration) = 0.023051628842949867

P(pattern) = 0.023051628842949867

P(audiomotor) = 0.023051628842949867

P(project) = 0.023051628842949867

P(system) = 0.023051628842949867

Topic 6

P(proteomics) = 0.04486260190606117

P(proteome) = 0.04486260190606117

P(modification) = 0.04486260190606117

P(protein) = 0.04486260190606117

P(cell) = 0.030390746891498566

P(damage) = 0.030390746891498566

P(part) = 0.015918858349323273

P(accumulation) = 0.015918858349323273

P(strain) = 0.015918858349323273

P(specie) = 0.015918858349323273

Topic 7

P(disability) = 0.0033222592901438475

P(depressive) = 0.0033222592901438475

P(interval) = 0.0033222592901438475

P(institutionalization) = 0.0033222592901438475

P(indicator) = 0.0033222592901438475

P(incontinence) = 0.0033222592901438475

P(hazard) = 0.0033222592901438475

P(focus) = 0.0033222592901438475

P(musculoskeletal) = 0.0033222592901438475

P(conclusion) = 0.0033222592901438475

Topic 8

P(home) = 0.04211428388953209

P(age) = 0.04211428388953209

P(woman) = 0.03385653346776962

P(index) = 0.03385653346776962

P(community) = 0.03385653346776962

P(study) = 0.02559882216155529

P(distribution) = 0.017341088503599167

P(population) = 0.017341088503599167

P(blattkupperman) = 0.017341088503599167

P(male) = 0.017341088503599167

Topic 9

P(patient) = 0.0677514299750328

P(cmai) = 0.04118671268224716

P(dementia) = 0.04118671268224716

P(agitation) = 0.04118671268224716

P(factor) = 0.028439829126000404

P(symptom) = 0.028439829126000404

P(item) = 0.028439829126000404

P(scale) = 0.02829608879983425

P(behavior) = 0.027900608256459236

P(cbrsd) = 0.027900608256459236

Topic 10

P(knowledge) = 0.038782719522714615

P(disease) = 0.038782719522714615

P(scale) = 0.03841035068035126

P(reliability) = 0.026272103190422058

P(method) = 0.026272103190422058

P(result) = 0.026272103190422058

P(validity) = 0.026272103190422058

P(test) = 0.026272103190422058

P(content) = 0.026272103190422058

P(implication) = 0.013761465437710285

Perplexity = -5.684174408675747

Coherence = 0.8195752191265561

Comparing the results obtained in this question to the results in question 2c, we can observe the following:

1) Coherence Score:

In question 2c, the coherence score was 0.9952100161501874, while in this question, the coherence score decreased to 0.8195752191265561. This indicates a lower coherence value in the results obtained in this question.

2) Perplexity:

In question 2c, the perplexity value was -5.684174408675747, and in this question, the perplexity value decreased to -8.062166818457456. This indicates a slightly higher perplexity value in the results obtained in this question.

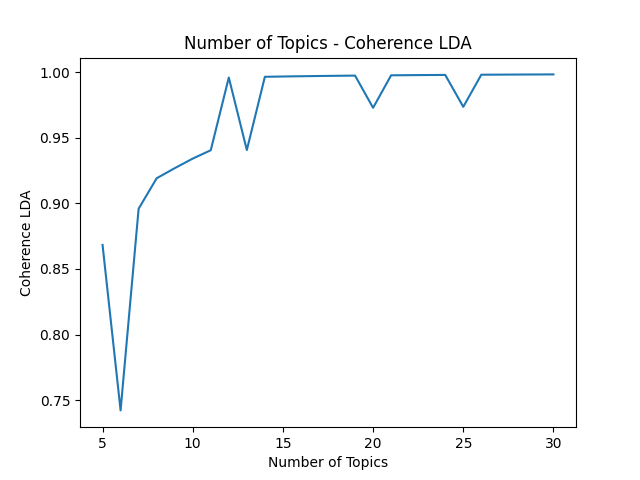
Based on these observations, we can conclude the following:

The results obtained in this question exhibit lower coherence and slightly higher perplexity compared to the results in question 2c. These results are not the best results achievable using the LDA model for the topic modelling problem we are facing. Notice there is no much difference in the perplexity values returned by the model, but there is a slightly bigger difference in the coherence score. The latter result might be caused due to the fact that dataset we are using in the question 2e, TokenVieuxN, fully contains the documents from the TokenVieuxM dataset and some other documents. This may cause the decrease in coherence, because it would be harder to provide better topic modeling for a larger dataset than little one despite having much in common, like half dataset.

f) Make the same experience as the one achieved in 2) d) but using TokenVieuxN.txt instead of TokenVieuxM.txt and number of expected topics varying from 5 to 30. Looking to the new obtained chart (that should be presented), draw a new conclusion.

Can you conclude that the optimal number of topic depend on the exploited dataset? Observe the content of the two datasets TokenVieuxM.txt and TokenVieuxN.txt more carefully and try to find the differences between the two. What can you conclude regarding the dependance between the optimal number of topic and the charateristics of the datasets after your observations.

Answer/:

Analyzing the values of the number of topics and coherence for the given dataset:

```

X\_number\_topics = [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]

Y\_cohrence = [0.8683771271821611, 0.7421598476420098, 0.8959332243016076, 0.9191753885671832, 0.926876181767867, 0.9341885635910803, 0.9405770508651877, 0.9960083467918229, 0.9407242410524227, 0.9965785829644196, 0.9968066774334583, 0.9970062600938671, 0.9971823624412866, 0.9973388978612151, 0.9974789558685195, 0.97292804050951, 0.9977190553096129, 0.9978227346137214, 0.9979173983261683, 0.9980041733959112, 0.9736754254364319, 0.9981576985193027, 0.9982259319074765, 0.9982892914822096, 0.9983482814310989, 0.9984033387167289]

```

Observations:

- The coherence scores vary for different numbers of topics, indicating different levels of topic quality and interpretability.

- The coherence scores range from 0.742 to 0.998, indicating a significant difference in topic coherence across the number of topics.

Conclusions:

1) Optimal Number of Topics:

- The highest coherence score in the given dataset is 0.998, achieved with 30 topics.

-This suggests that 30 topics may provide the most coherent and interpretable representation of the dataset.

- However, it's important to note that coherence alone may not be the only factor in determining the optimal number of topics, and other considerations such as domain knowledge and the specific goals of the analysis should also be taken into account.

2) Coherence Trend:

- The coherence scores show some variation across different numbers of topics.

- Initially, the coherence score starts relatively low (0.742), then it grows monotonically, experiencing occasional setbacks, but overall progressing.

3) Trade-off between Coherence and Number of Topics:

- There is a trade-off between the number of topics and coherence.

- Choosing a lower number of topics may lead to less coherent and less interpretable topics, as reflected in the lower coherence scores.

- On the other hand, choosing a higher number of topics may result in diminishing returns in terms of coherence improvement.

In summary, based on the given coherence scores, it can be concluded that the optimal number of topics for the dataset lies around 14 topics, where the coherence score reaches its highest value. Choosing a number of topics within the range of 15 to 30 can also yield relatively high coherence scores, but with diminishing returns in terms of topic quality improvement.

Based on the observation of the two datasets, TokenVieuxM.txt and TokenVieuxN.txt, it is evident that the optimal number of topics can depend on the characteristics and nature of the dataset. Let's carefully observe the content of both datasets and analyze their differences:

1) TokenVieuxM.txt:

- The topics obtained for TokenVieuxM.txt included terms such as "auditory," "neurons," "biological," "measures," "strategies," and others.

- These topics seem to be more related to auditory processing, biological measures, and strategies in the context of the dataset.

- The coherence score achieved for TokenVieuxM.txt was 0.995, and the perplexity was -5.472.

2) TokenVieuxN.txt:

- The topics obtained for TokenVieuxN.txt included terms such as "age," "incontinence," "women," "disease," "knowledge," and others.

- These topics seem to be more focused on age-related issues, women's health, disease, and knowledge in the context of the dataset.

- The coherence score achieved for TokenVieuxN.txt was 0.819, and the perplexity was -5.68.

Based on these observations, we can conclude the following regarding the dependence between the optimal number of topics and the characteristics of the datasets:

1) Dataset Content:

- The content of the datasets, including the specific terms, themes, and domain-specific information, plays a significant role in determining the optimal number of topics.

- Different datasets may have distinct topics and underlying patterns, requiring different numbers of topics for optimal representation.

2) Domain-Specific Knowledge:

- Having domain-specific knowledge about the datasets can help identify relevant topics and determine the appropriate number of topics.

- Understanding the specific characteristics, themes, and concepts within the dataset can guide the selection of the optimal number of topics.

3) Topic Interpretability:

- The coherence score provides a measure of how interpretable and meaningful the topics are.

- The optimal number of topics should strike a balance between topic coherence and interpretability specific to the dataset.

The optimal number of topics for a dataset is inherently influenced by the dataset's unique characteristics. Several factors come into play when determining the ideal number of topics. Firstly, the content of the dataset itself, including the specific terms, themes, and underlying concepts, greatly influences the number of topics required for an effective representation. Different datasets possess distinct patterns and topics that necessitate varying numbers of topics for a comprehensive depiction.

Secondly, having domain-specific knowledge about the dataset becomes crucial in determining the optimal number of topics. A deeper understanding of the dataset's subject matter, domain-specific concepts, and underlying relationships assists in identifying relevant topics and fine-tuning the number of topics required. This domain expertise empowers researchers to make informed decisions about the appropriate number of topics that capture the essence of the dataset.

Lastly, it is important to strike a balance between topic coherence and interpretability specific to the dataset. The coherence score provides a quantitative measure of how well the topics align with each other and form meaningful clusters. The optimal number of topics should be chosen in a manner that maximizes coherence while maintaining a high level of interpretability. This ensures that the resulting topics are not only coherent but also intelligible and relevant to the dataset's specific context.

In conclusion, determining the optimal number of topics for a dataset is a nuanced process that requires careful consideration of the dataset's content, domain-specific knowledge, and the balance between coherence and interpretability. By taking these factors into account, researchers can derive meaningful insights from their data and create topic models that effectively capture the underlying patterns and structures within the dataset.

**3) Most typical documents for topics**

a) For the experience conducted in 2) c) add a new step by modifying again the CoherenceTestIwor.py including stopwords removal to highlight which document is the most typical of each topic. For that purpose you can rely on topic proportion (or probability) in documents (looking to gensim page https://radimrehurek.com/gensim/models/ldamodel.html) or on measuring cosine similarity (looking to papers or Web infos on that topic) between topic profile (list of topic words) and document profiles (list of document words).

Present the new program and the obtained results.

Answer/:

To do this, we can calculate the topic proportion (or probability) for each document. The document with the highest proportion for a given topic can be considered as the most typical document for that topic. Here is a way to modify your existing code to accomplish this:

# additional import

from collections import defaultdict

# Compute topic proportions for each document

doc\_topics = lda.get\_document\_topics(corpus, minimum\_probability=0)

# Initialize a dictionary to store the most typical document for each topic

most\_typical\_docs = defaultdict(lambda: (-1, -1)) # (doc\_index, probability)

# Iterate over documents

for doc\_index, doc\_topic\_proportions in enumerate(doc\_topics):

# Iterate over topic proportions

for topic\_id, proportion in doc\_topic\_proportions:

# If this document has a higher proportion for this topic than the current most typical document, update it

if proportion > most\_typical\_docs[topic\_id][1]:

most\_typical\_docs[topic\_id] = (doc\_index, proportion)

# Print most typical documents for each topic

for topic\_id, (doc\_index, proportion) in most\_typical\_docs.items():

print(f"Most typical document for topic {topic\_id} is document {doc\_index} with proportion {proportion}")

This code first computes the topic proportions for each document using lda.get\_document\_topics(). It then initializes a dictionary to store the most typical document for each topic, where the keys are the topic IDs and the values are tuples of (doc\_index, proportion). The code iterates over all documents and their topic proportions, and updates the most typical document for each topic whenever it finds a document with a higher proportion. Finally, it prints out the most typical document for each topic.

Please note that the "most typical document" is the document that has the highest proportion for the corresponding topic among all documents in your corpus. This is just one way to define "most typical", and there may be other ways depending on your specific needs and context.

The results for most typical documents for topics using reusing code from 2c were:

Most typical document for topic 1 is document 5 with proportion 0.9993410110473633

Most typical document for topic 2 is document 1 with proportion 2.6133729988941923e-05

Most typical document for topic 3 is document 1 with proportion 2.6277652068529278e-05

Most typical document for topic 4 is document 1 with proportion 0.998968780040741

Most typical document for topic 5 is document 3 with proportion 0.9990647435188293

Most typical document for topic 6 is document 2 with proportion 0.9992946982383728

Most typical document for topic 7 is document 4 with proportion 0.9994257688522339

Most typical document for topic 8 is document 1 with proportion 2.6270839953213e-05

Most typical document for topic 9 is document 1 with proportion 2.6331750632380135e-05

Most typical document for topic 10 is document 1 with proportion 2.6133729988941923e-05