02582 Computational Data Analysis Case Study

Authors:
Anshul Chauhan s192164
Md Alamgir Kabir s193074
Ghassen Lassoued s196609

Content

- Data preprocessing
- Analysis method
- Results
- Conclusion

Data preprocessing

The data provided contains transcripts of parliamentary speeches from the sittings in the Chamber of the Danish Parliament from year 2009 to 2017.

The corpus consists of xml files, by using *python* script we extracted four features from the dataset (MeetingId, Sagtype, Tale and Navn) and made CSV files. During the data extraction process the script failed to read about 10-15 speeches (Tale) due to unicode formatting error. The resulted CSV files contains 854 unique meetings with 378387 speeches.

The Approach:

We are trying to gather all the speeches that were made by politicians and extract the unique words, all the words with their frequency and probability of being in a particular document should be our features in rows and all the documents will be in columns. As for the base case, We are planning to use NMF and LDA approach to cluster the words into the optimal number of clusters k (for every document) and then compare the approaches.

Furthermore, In order to find this k, we are calculating coherence score for LDA and comparing both the approaches for these optimal values of clusters.

Note that, for different documents, we have different optimal values as the analysis is data dependent.

Speaker Ghassen

Analysis method NMF and LDA

We have chosen unsupervised method for analysis to find the latent features in the corpus.

To achieve the goal we have chosen Non-negative Matrix factorization (NMF) because first of all, this algorithm works better with text mining. In text mining the columns of W matrix of NMF contains set of words for different topics and H matrix contains the weights for the topics. Therefore by passing a set of documents, NMF identifies topics and simultaneously classifies the documents among these different topics.

The main idea of Latent Dirichlet Allocation (LDA) is that each document can be described by a distribution of topics and each topic can be described by a distribution of words.

Algorithm:

- Go through each document and randomly assign each word in the document to one of ***k*** topics
- For each document ***d***, go through each word ***w****:*
  - **p(topic *t* | document *d*)**: the proportion of words in document *d* that are assigned to topic *t*
  - **p(word *w*| topic *t*)**: the proportion of assignments to topic *t* over all documents that come from this word *w*
  - Update the probability for the word *w* belonging to topic *t*, as

    p(word w with topic t) = p(topic t | document d) * p(word w | topic t)

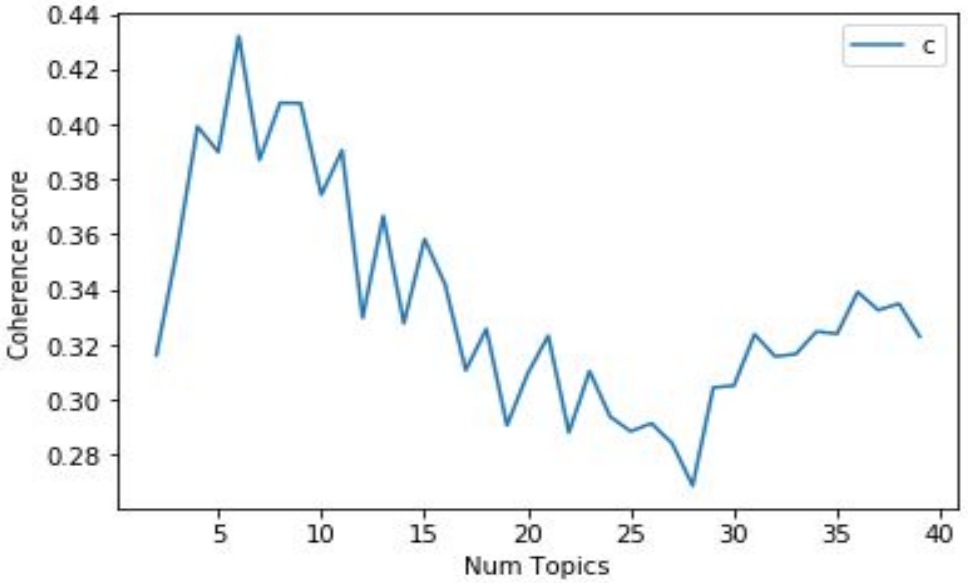
Speaker Ghassen

Optimal number of topics in LDA

The approach to finding the optimal number of topics is to build many LDA models with different values of number of topics (k) and pick the one that gives the highest coherence score. The number of unique Sagtype is taken into consideration to choose the range of k i.e this range contains this number of unique Sagtype.

For instance, the data from year 2009 is taken

Number of unique Sagtype= 9



Optimal k= 6

Speaker Ghassen

Result Comparison Between LDA and NMF for 20091 Dataset

NMF

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06
0	organer	mio	så	status	væsentligste	hurtig
1	ordningen	tages	tages	sten	spørgsmål	bekymret
2	forringelse	lo	finansielle	stemt	stammer	krise
3	så	dennis	ordningen	dennis	opretholde	fremsætte
4	endelig	jeg	sjælland	ind	krisecenter	personen
5	heraf	forbundet	justering	stabilitet	dele	canada
6	flyttes	sund	spørgsmål	skærpende	seniorførtidspension	tages
7	dagen	afspejler	berørte	svarede	stabilitet	legitimt
8	dele	metode	undersøgelse	afsætte	mad	billede
9	organisationerne	jakobsen	socialdemokrater	rolle	opfyldt	bakket

LDA

A	B	C	D	E	F	G
	0	1	2	3	4	5
0	('hr.', 0.08054142)	('Det', 0.02310077)	('kr.', 0.024515452)	('fru', 0.03168209)	('Det', 0.014165726)	('Det', 0.010567865)
1	('Så', 0.043171898)	('Jeg', 0.01478280	('mia.', 0.019717628)	('stemte:', 0.021506164)	('Vi', 0.0060842345)	('Danmark', 0.00608948
2	('Hr.', 0.034589518)	('det,', 0.01461665	('offentlige', 0.00966366	('Fru', 0.020202199)	('altså', 0.006049500	('nye', 0.0052761664)
3	('kort', 0.034260202)	('–', 0.010948039)	('regeringen', 0.008966	('Ministeren.', 0.016793	('det,', 0.005905672)	('derfor', 0.0047603506
4	('ordføreren.', 0.02182014)	('om,', 0.00694441	('Det', 0.008914496)	('nr.', 0.016456354)	('for,', 0.0054335766)	('regeringen', 0.004575
5	('bemærkning.', 0.021022	('sige,', 0.0067993	('kr.,', 0.007360642)	('imod', 0.015067891)	('på,', 0.0052774716)	('Vi', 0.0045664487)
6	('fru', 0.01950443)	('Dansk', 0.006272	('mio.', 0.007185776)	('Der', 0.011476173)	('–', 0.0050222725)	('danske', 0.003961649
7	('Tak', 0.018370908)	('på,', 0.00591462	('penge', 0.00644305)	('ændringsforslag', 0.01	('om,', 0.0045222454)	('bl.a.', 0.0036867305)
8	('korte', 0.011989564)	('Men', 0.00577042	('økonomiske', 0.00604	('Hr.', 0.010237717)	('Jeg', 0.004414215)	('mellem', 0.003448668
9	('bemærkning', 0.0118219	('synes', 0.005542	('pct.', 0.0057537598)	('Line', 0.008857835)	('børn', 0.00436949)	('række', 0.003402859)

Result Comparison Between 20091 and 20101 Dataset

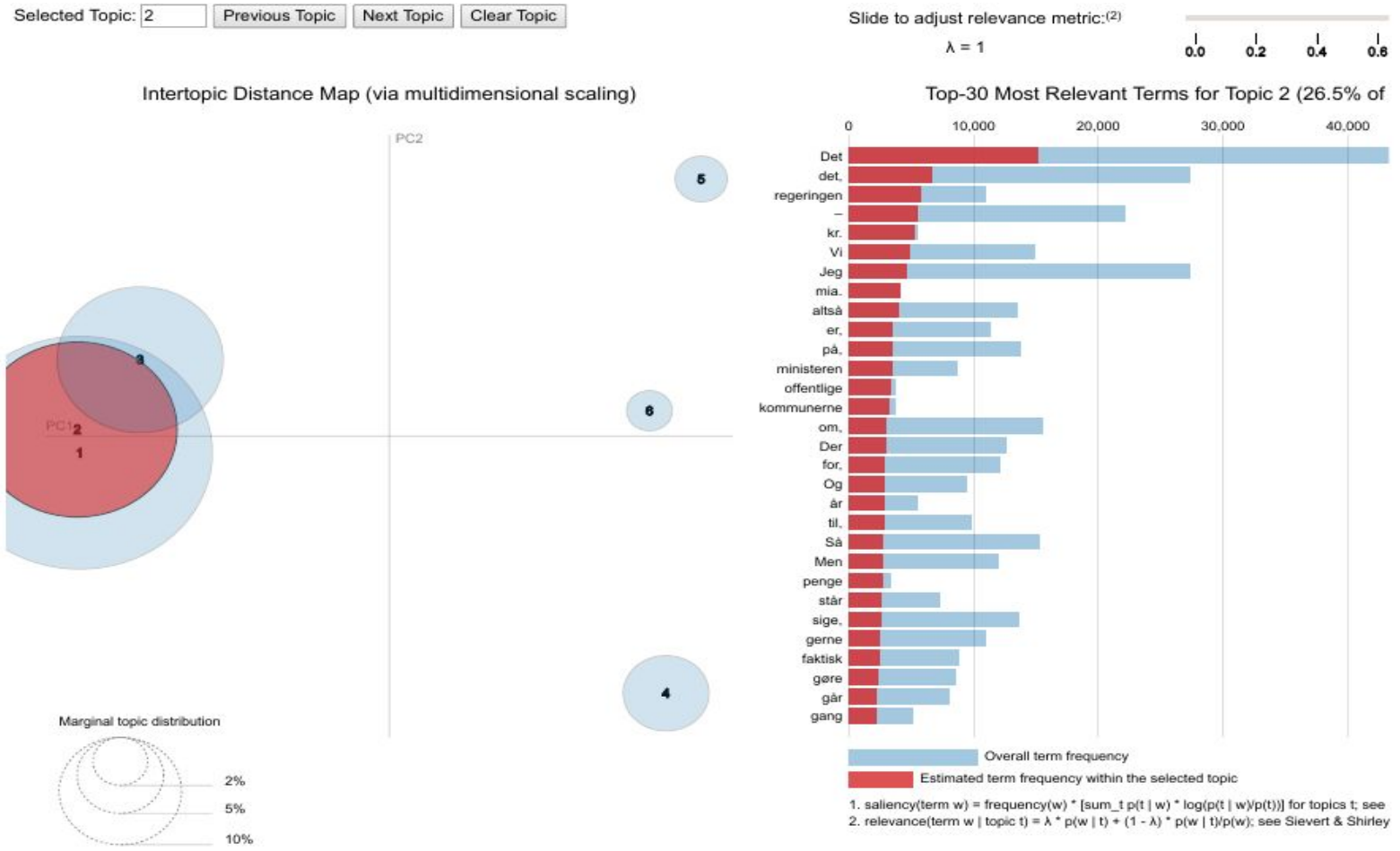
20091

	0	1	2	3	4	5
0	('hr.', 0.08054142)	('Det', 0.02310077)	('kr.', 0.024515452)	('fru', 0.03168209)	('Det', 0.014165726)	('Det', 0.010567865)
1	('Så', 0.043171898)	('Jeg', 0.01478280	('mia.', 0.019717628)	('stemte:', 0.021506164)	('Vi', 0.0060842345)	('Danmark', 0.0060894
2	('Hr.', 0.034589518)	('det,', 0.01461665	('offentlige', 0.00966366	('Fru', 0.020202199)	('altså', 0.006049500	('nye', 0.0052761664)
3	('kort', 0.034260202)	('–', 0.010948039)	('regeringen', 0.008966	('Ministeren.', 0.016793	('det,', 0.005905672)	('derfor', 0.0047603506
4	('ordføreren.', 0.02182014)	('om,', 0.00694441	('Det', 0.008914496)	('nr.', 0.016456354)	('for,', 0.0054335766)	('regeringen', 0.004575
5	('bemærkning.', 0.0210223	('sige,', 0.0067993	('kr.,', 0.007360642)	('imod', 0.015067891)	('på,', 0.0052774716)	('Vi', 0.0045664487)
6	('fru', 0.01950443)	('Dansk', 0.006272	('mio.', 0.007185776)	('Der', 0.011476173)	('–', 0.0050222725)	('danske', 0.003961649
7	('Tak', 0.018370908)	('på,', 0.00591462	('penge', 0.00644305)	('ændringsforslag', 0.01	('om,', 0.0045222454)	('bl.a.', 0.0036867305)
8	('korte', 0.011989564)	('Men', 0.00577042	('økonomiske', 0.00604	('Hr.', 0.010237717)	('Jeg', 0.004414215)	('mellem', 0.003448668
9	('bemærkning', 0.0118219	('synes', 0.005542	('pct.', 0.0057537598)	('Line', 0.008857835)	('børn', 0.00436949)	('række', 0.003402859)

20101

	0	1	2	3	4	5
0	('ministeren', 0.013523	('Det', 0.022434467)	('hr.', 0.07381847)	('Det', 0.011476103)	('Tak', 0.033156715)	('Sophie', 0.012174774)
1	('Det', 0.013195513)	('det,', 0.012686162)	('kort', 0.02829237)	('danske', 0.00698447	('Så', 0.028819239)	('Hæstorp', 0.009729966)
2	('det,', 0.010636545)	('Jeg', 0.011212312)	('Hr.', 0.021114072)	('Danmark', 0.006638	('fru', 0.027325032)	('Møller', 0.0078461245)
3	('Jeg', 0.01037084)	('–', 0.009091881)	('fru', 0.016672745)	('Ministeren.', 0.00503	('hr.', 0.027176626)	('Flemming', 0.0077422247)
4	('–', 0.01000666)	('om,', 0.0068316013	('bemærkning.', 0.0	('–', 0.0040270193)	('ordføreren.', 0.019	('Færøerne', 0.0073418533)
5	('Ordføreren.', 0.008690	('sige,', 0.00670813)	('stemte', 0.013465	('regeringen', 0.00366	('Hr.', 0.017078847)	('Statsministeren.', 0.0071385144)
6	('kr.', 0.0071882447)	('på,', 0.0063619222	('Der', 0.010772334	('Vi', 0.0036257636)	('ordfører', 0.015624	('Fru', 0.007052431)
7	('om,', 0.005501918)	('synes', 0.00612991	('Så', 0.010614502)	('derfor', 0.003534266	('Der', 0.013504196)	('Andersen.', 0.0061408565)
8	('ordføreren', 0.004988	('for,', 0.0058367867)	('imod', 0.01039662	('år', 0.0032413986)	('Fru', 0.012660165)	('færøske', 0.0055554863)
9	('kommunerne', 0.0049	('Vi', 0.0054290714)	('Per', 0.008912313	('altså', 0.0030992348	('går', 0.009814737)	('Hr.', 0.0051148497)

Another method of displaying same topics is done using pyLDAvis library but giving them in different order



Analysis over both approaches

Cross Validation:

The major problem faced with NMF is the validation part. Since it is unsupervised method, one can not perform naively the crossvalidation. Probably due to the large amount of data, and to the fact that crossvalidation is not very compatible with unsupervised learning, the program that was used for crossvalidation for NMF did not give results.

From here came the idea of performing the other method for NLP which is LDA(Latent Dirichlet allocation) to compare the results of both methods using the optimal number of topics provided by LDA.

Unexpected results for 20141 dataset:

1) Got k = 2 as optimal number of clusters
2) Very close coherance values

Speaker Anshul