TTIC 31220 Final Project:

Topic Modelling of British Parliamentary Debates

Xiaoyu Sun

Zhi Rong Tan

March 18, 2019

Motivation & Objectives

- Data: Transcripts of British parliamentary debates from 1919 to 2019
- Question: How have public issues of UK evolved?

We investigate two decades: 1990 - 1999 (era 1) vs. 2009-2018 (era 2)

- Compare different topic modelling techniques: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP)
- Compare fitted models for two eras, in terms of topic types, topic proportions, word choices, etc - understand how topics have changed in parliament based on the document weightage

Data Scraping

- Original format: XML
- Tool: ElementTree in Python
- Extract major headings and speeches, saving as TXT

debates2019-02-28a.xml 01-Mar-2019 06:16 558K
debates2019-03-04a.xml 05-Mar-2019 06:18 570K
debates2019-03-05a.xml 06-Mar-2019 06:17 596K
debates2019-03-06a.xml 08-Mar-2019 06:16 262K
debates2019-03-06b.xml 08-Mar-2019 06:16 752K
debates2019-03-07a.xml 08-Mar-2019 06:16 523K
debates2019-03-11a.xml 12-Mar-2019 06:15 637K
debates2019-03-12a.xml 14-Mar-2019 06:16 105K
debates2019-03-12b.xml 14-Mar-2019 06:16 681K
debates2019-03-13a.xml 15-Mar-2019 06:15 172K
debates2019-03-13b.xml 15-Mar-2019 06:15 898K
debates2019-03-14a.xml 16-Mar-2019 06:15 103K
debates2019-03-14b.xml 16-Mar-2019 06:15 833K
debates2019-03-15a.xml 16-Mar-2019 06:15 338K


```
−<publicwhip scraperversion="a" latest="yes">
  <major-heading id="uk.org.publicwhip/debate/2019-03-15a.663.0" nospeaker="true" colnum="663" time="" url=""> SPEAKER'S STATEMENT: NEW ZEALAND TERROR ATTACKS </major-heading>
 −<speech id="uk.org.publicwhip/debate/2019-03-15a.663.1" speakername="John Bercow" person_id="uk.org.publicwhip/person/10040" colnum="663" time="" url="">
  −<p pid="a663.1/1">
      In respectful memory of the 49 people who horrendously lost their lives in the terrorist attack in Christchurch, New Zealand, and of the apparently dozens who were injured in the attack on the two mosques, as well as in solidarity with the people of New Zealand and Muslims around the world, I humbly suggest to the House—I know that both sides of the House are on the same page as me in this regard—that we hold one minute's silence at 11 am. I think that some colleagues will want to say something about this matter now, before we get on to today's business, sitting in private or any of that. I therefore call Minister Ben Wallace.
    </p>
  </speech>
 −<speech id="uk.org.publicwhip/debate/2019-03-15a.663.2" speakername="Ben Wallace" person_id="uk.org.publicwhip/person/11668" colnum="663" time="" url="">
  −<p pid="a663.2/1">
      Let me say to the House on behalf of the Government that we send our sincere condolences to the victims and people of New Zealand for their loss, and that they have our offer of any assistance required to deal with this repugnant attack. The UK stands shoulder to shoulder with New Zealand against terrorism, and we will not falter in our commitment to uphold the values of tolerance, religious freedom and democracy that we both hold so dear.
    </p>
  −<p pid="a663.2/2">
      Later today, the Home Secretary and I will be speaking to police counter-terrorism leaders and the security services to discuss what further measures we can take to protect our mosques and communities from any threat here in the United Kingdom. No one should be in any doubt that our police and security services treat all threats the same and all terrorists the same. No matter what community, religion or background they come from, a terrorist is a terrorist, and we shall deal with them exactly the same.
    </p>
  </speech>
```

Data Processing

13,206 documents for era 1; 9,231 documents for era 2

- Tokenization: Split each document into a list of words
- Remove stopwords (e.g. ‘a’, ‘the’, ‘in’) and words shorter than 3 characters
- Lemmatization: Convert words to first person, present tense forms
- Stemming: Convert to root forms

- Create a dictionary of all the unique words
- Remove words whose appearance < 5, or > 50%
- Calculate the **tf-idf** matrix (tf-idf: term frequency - inverse document frequency)

Topic Modelling - Methods

Explored 3 methods:

1. **Latent Semantic Analysis**
   - Linear method using SVD on document-term matrix
   - 1 hyperparameter - number of topics

2. **Latent Dirichlet Allocation**
   - Learnt in class!
   - 2 hyperparameters - α (prior parameter of Dirichlet distribution from which topics are drawn), number of topics

Topic Modelling - Methods

3. Hierarchical Dirichlet Process

- Bayesian nonparametric method - another layer being added to the generative model to produce the number of topics
- Method uses a Dirichlet process for each group/topic of data, with the Dirichlet processes for all groups sharing a base distribution which is itself drawn from a Dirichlet process
- Model generates topic associated with the n-th word in the j-th document, then generate the word from the topic
- Hyperparameter - α, H

$$G_0 \sim DP(\alpha_0, H) \qquad \theta_{jn} \sim G_j$$

$$G_j \sim DP(\alpha, G_0) \text{ for each } j \qquad w_{jn} \sim multi(\theta_{jn})$$

Topic Modelling - Steps

For each era:

1. Split dataset into training set and validation set

For each method:

1. Tune hyperparameter(s) by evaluating trained model on the validation set and selecting hyperparameter that gives highest coherence score on validation corpus
2. Train final model on the entire set (training + validation)
3. Get final coherence score

Topic Modelling - Coherence Score

Evaluate the model using the average coherence score of the topic

$$coherence(V) = \sum_{(v_i, v_j) \in V} score(v_i, v_j)$$

$$coherence(model) = \frac{1}{T} \sum_{t} coherence(V_t)$$

Used UCI metric, which defines each word pair's score as the pointwise mutual information between the 2 words, over the sum of all words in the topic

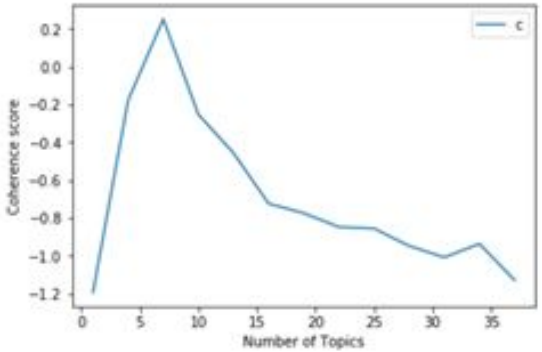
$$score(v_i, v_j) = log \frac{p(v_i,\ v_j)}{p(v_i)p(v_j)}$$

Extrinsic metric - measures coherence of model on external corpus

Latent Semantic Analysis

Era 1: #topics = 7. Coherence Score: -0.0829

Era 2: #topics = 6. Coherence Score: -0.823



Method Evaluation:

- Simple - only 1 hyperparameter to tune
- Some topics are not clearly interpretable

Topic	Words
1	-0.386*"amend" + -0.317*"claus" + -0.315*"insert" + -0.211*"page" + -0.195*"subsect" + -0.148*"ireland" + -0.143*"line" + -0.142*"section" + 0.133*"tax" + -0.131*"lord"
2	-0.374*"school" + 0.274*"tax" + 0.266*"pension" + -0.258*"educ" + -0.208*"health" + -0.186*"hospit" + -0.160*"patient" + -0.156*"teacher" + -0.149*"ireland" + -0.143*"nhs"

Hierarchical Dirichlet Process

Era 1: Coherence Score: -9.9

Era 2: Coherence Score: -10.1

5.	0.000*courthous + 0.000*nomenclatur + 0.000*benidorm + 0.000*wallenberg + 0.000*creepi + 0.000*room + 0.000*hmip + 0.000*beat + 0.000*curmudgeon
10.	0.000*skylin + 0.000*goug + 0.000*portland + 0.000*unfurnish + 0.000*achill + 0.000*charterhouse

Method Evaluation:

- Each era has ~ 50+ topics
- BUT… only 2-3 topics are interpretable
- Most likely to get good results only with very involved tuning of > 2 parameters, and increasing no. of iterations or convergence specifications

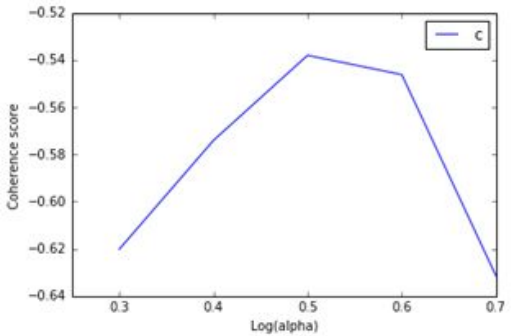
Latent Dirichlet Allocation

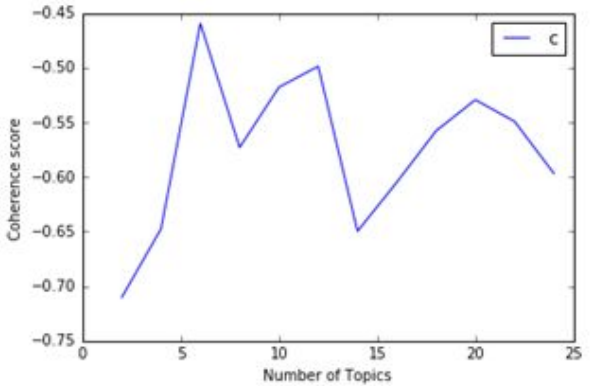
Exists tradeoff between speed/convenience and topic model performance

Among the 3 methods - LDA seems to have the best tradeoff

Era 1: Coherence Score: -0.0591, alpha=0.5

Era 2: Coherence Score: -0.5067, alpha=0.5





Final Topic Model for Era 1

LDA

(T=7)

Topic 0: FINANCE

Words: 0.009*"session" + 0.006*"deposit" + 0.003*"clerk" + 0.003*"petit" + 0.003*"fee" + 0.003*"agent" + 0.003*"thereto" + 0.003*"lord" + 0.003*"deem" + 0.003*"refund" + 0.003*"descript" + 0.003*"chargeabl" + 0.002*"privat" + 0.002*"amend" + 0.002*"relief" + 0.002*"tax" + 0.002*"acquisit" + 0.002*"journal" + 0.002*"suspend" + 0.002*"proceed"

Topic 1: AGRICULTURE & ENVIRONMENT

Words: 0.006*"petit" + 0.006*"farmer" + 0.006*"road" + 0.005*"beef" + 0.004*"signifi" + 0.004*"agricultur" + 0.004*"food" + 0.004*"petition" + 0.004*"traffic" + 0.004*"anim" + 0.004*"farm" + 0.004*"expedi" + 0.003*"payabl" + 0.003*"transport" + 0.003*"authoris" + 0.003*"attribut" + 0.003*"payment" + 0.003*"environment" + 0.003*"queen" + 0.003*"fisheri"

Topic 2: HEALTH + LAW ENFORCEMENT

Words: 0.003*"health" + 0.003*"madam" + 0.003*"polic" + 0.003*"ireland" + 0.002*"northern" + 0.002*"court" + 0.002*"amend" + 0.002*"hospit" + 0.002*"prime" + 0.002*"prison" + 0.002*"pension" + 0.002*"leader" + 0.002*"defenc" + 0.002*"nhs" + 0.002*"claus" + 0.002*"european" + 0.002*"foreign" + 0.002*"patient" + 0.002*"crime" + 0.002*"lord"

Topic 3: FOREIGN POLICY

Words: 0.002*"kosovo" + 0.002*"iraq" + 0.002*"saddam" + 0.002*"bosnia" + 0.002*"nato" + 0.002*"hussein" + 0.001*"militari" + 0.001*"troop" + 0.001*"iraqi" + 0.001*"serb" + 0.001*"amend" + 0.001*"refuge"

Topic 4: EDUCATION & WELFARE POLICY

Words: 0.004*"tax" + 0.004*"school" + 0.003*"educ" + 0.003*"industri" + 0.002*"labour" + 0.002*"invest" + 0.002*"wale" + 0.002*"compani" + 0.002*"unemploy" + 0.002*"scottish" + 0.002*"scotland" + 0.002*"budget" + 0.002*"employ" + 0.002*"prime" + 0.002*"london" + 0.002*"conserv" + 0.002*"chancellor" + 0.002*"pension" + 0.002*"fund" + 0.002*"sector"

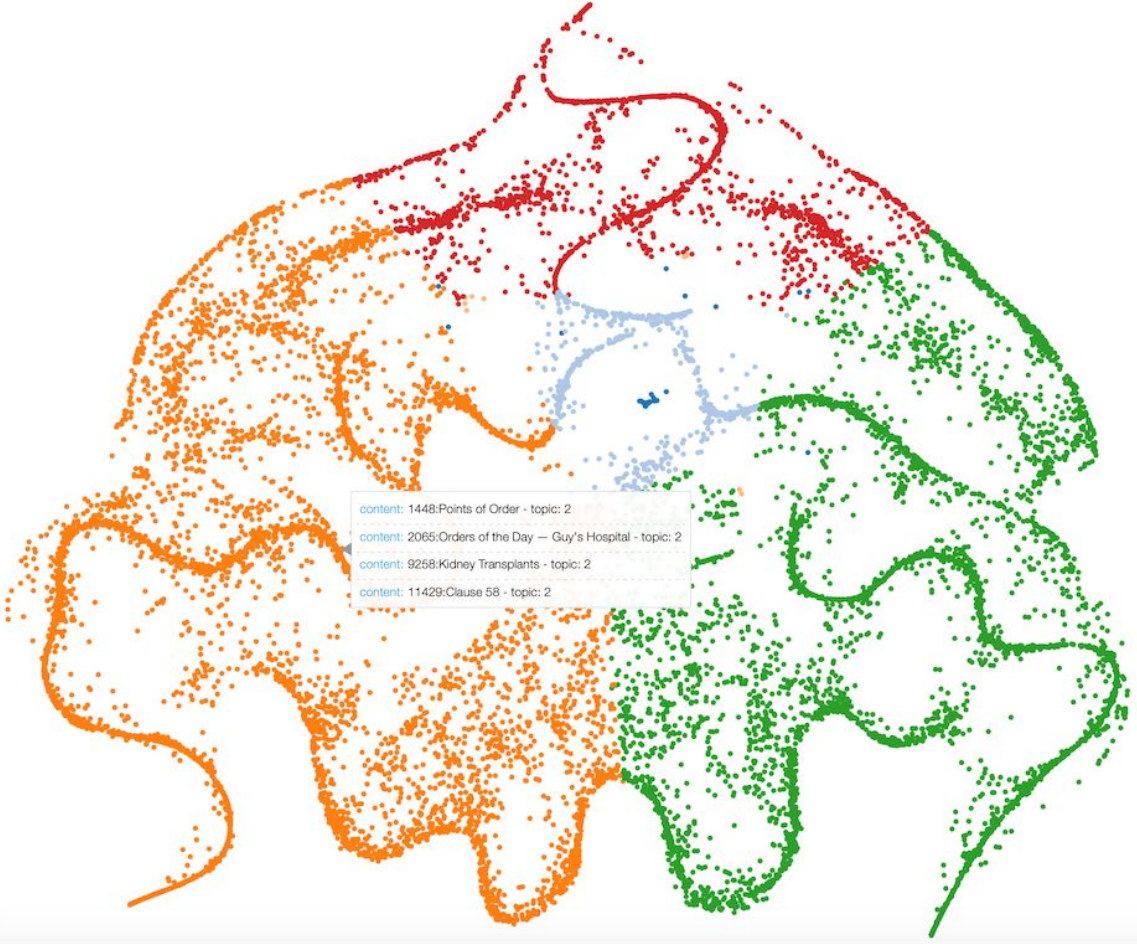
Topic 5: RELIGION

Words: 0.013*"church" + 0.006*"commission" + 0.006*"majesti" + 0.005*"gracious" + 0.005*"humbl" + 0.004*"clergi" + 0.002*"sovereign" + 0.002*"address" + 0.002*"loyal" + 0.002*"resum" + 0.002*"bishop" + 0.002*"dioces" + 0.002*"synod" + 0.002*"parish" + 0.002*"assembl" + 0.002*"stipend" + 0.001*"dome" + 0.001*"ireland" + 0.001*"cathedr" + 0.001*"adjourn"

Topic 6: ADMINISTRATIVE

Words: 0.006*"insert" + 0.006*"claus" + 0.006*"sir" + + 0.004*"subsect" + 0.004*"page" + 0.004*"motion" + 0.004*"andrew" + 0.004*"section" + 0.004*"paragraph" + 0.003*"schedul" + 0.003*"proceed"

Era 1 2-D Visualization



Final Topic Model for Era 2

LDA (T=6)

Topic 0: EDUCATION + HEALTHCARE

Words: 0.002*"school" + 0.002*"petition" + 0.001*"educ" + 0.001*"student" + 0.001*"proceed" + 0.001*"pupil" + 0.001*"children" + 0.001*"amend" + 0.001*"park" + 0.001*"social" + 0.001*"cut" + 0.001*"young" + 0.001*"nhs" + 0.001*"hospit" + 0.001*"nurs"

Topic 1: SOCIAL & WELFARE POLICY

Words: 0.002*"prison" + 0.002*"tax" + 0.002*"women" + 0.001*"disabl" + 0.001*"brexit" + 0.001*"children" + 0.001*"vote" + 0.001*"pension" + 0.001*"educ" + 0.001*"payment" + 0.001*"employ" + + 0.001*"authoris"

Topic 2: ECONOMY & INFRASTRUCTURE

Words: 0.002*"brexit" + 0.002*"nhs" + + 0.002*"vote" + 0.002*"trade" + 0.002*"union" + 0.001*"european" + 0.001*"tax" + 0.001*"invest" + 0.001*"rail" + 0.001*"industri" + 0.001*"bank" + 0.001*"crime" + 0.001*"leader" + + 0.001*"sector" + 0.001*"economi" + 0.001*"billion"

Topic 3: BREXIT / BORDER ISSUES

Words: 0.005*"ireland" + 0.004*"northern" + 0.002*"church" + 0.002*"brexit" + 0.002*"european" + 0.002*"petition" + 0.002*"border" + 0.002*"union" + 0.002*"leasehold" + 0.001*"proceed" + 0.001*"tax" + 0.001*"polic" + 0.001*"agreement" + 0.001*"amend" + 0.001*"clad" + 0.001*"custom"

Topic 4: ADMINISTRATIVE

Words: 0.009*"proceed" + 0.005*"draft" + 0.004*"amend" + 0.004*"approv" + 0.004*"lay" + 0.003*"conclus" + 0.003*"commenc" + 0.003*"lord" + 0.003*"conclud" + 0.002*"deleg" + 0.002*"regul" + 0.002*"prison" + 0.002*"grand" + 0.002*"petit" + 0.002*"claus" + 0.002*"committ"

Topic 5: FOREIGN POLICY + DEFENCE

Words: 0.002*"russian" + 0.002*"church" + 0.002*"nato" + 0.002*"petit" + 0.001*"cancer" + 0.001*"arm" + 0.001*"defenc" + 0.001*"proceed" + 0.001*"saudi" + 0.001*"yemen" + 0.001*"syria" + 0.001*"foreign" + 0.001*"weapon" + 0.001*"humanitarian"

Era 2 2-D Visualization

Analysis

- Did not expect number of topics to be small (6 or 7)
- Some topics contain > 1 issue (e.g. Defence + Foreign Policy, *or* Healthcare + Education, *or* Economy + Infrastructure)
  - Further tuning needed? :(
  - More likely that in that era, these 2 issues are highly integrated together, but these integration differ across eras
  - Different ways of integration cause problems in comparison

Still,we tried...

Document Comparison *(Incomplete)*

Topic	Era 1 Proportion	Era 2 Proportion
Administrative	30%	20%
Finance / Economy	25%	20%
Agriculture / Environment	9%	N/A**
Social Policy (Health, Education, Welfare)	25%	40%
Religion	1%	N/A
Foreign Policy / Defence	10%	~10%
Border Issue	N/A**	~10%

Extensions

From Unigram / Bag-of-words Model to Bigram/Trigram model where word sequence matters

More complex methods - such as LDA2Vec, Markov Hidden Topic Model

Compare speeches from other countries!


THANK YOU!