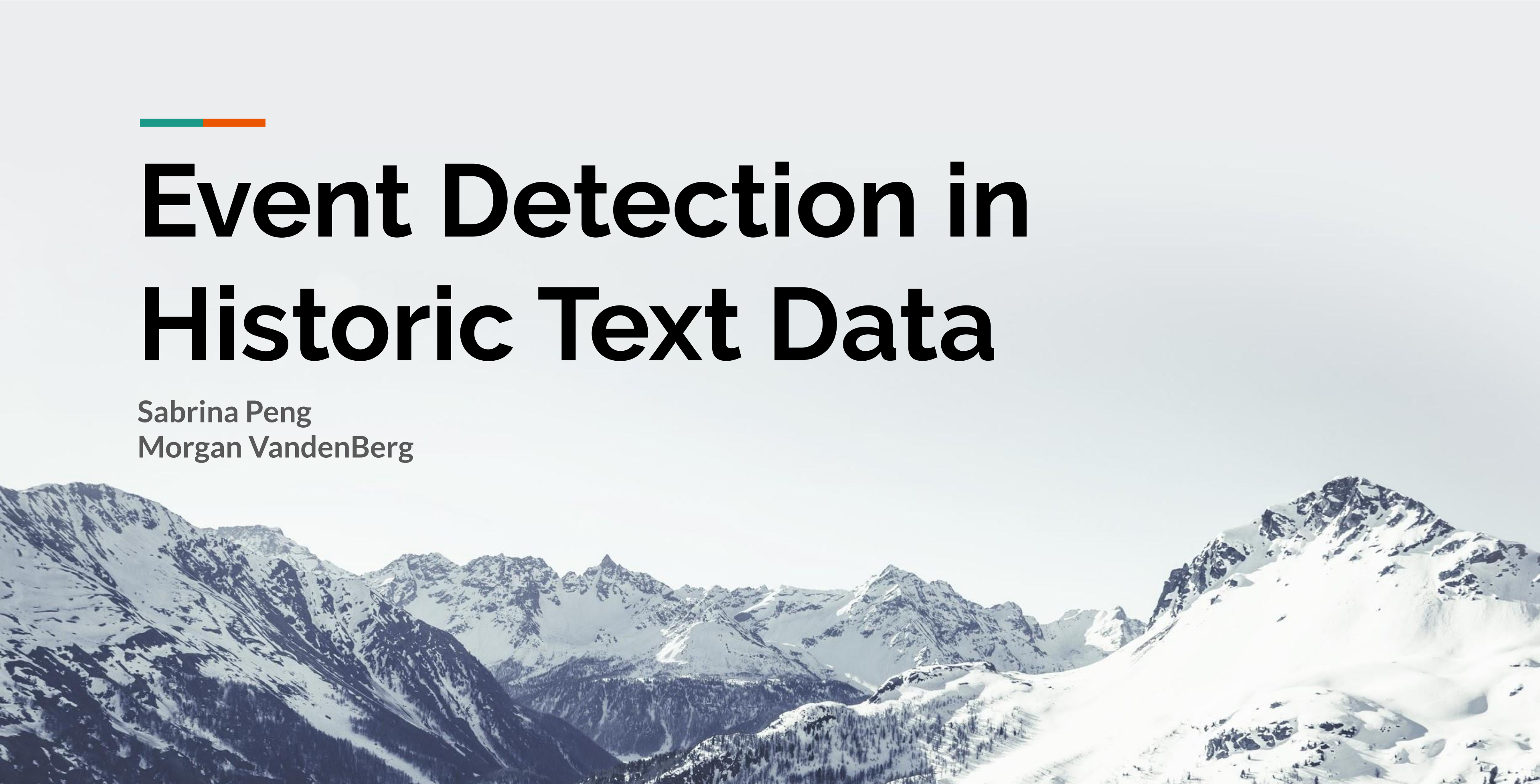


Event Detection in Historic Text Data

Sabrina Peng
Morgan VandenBerg

Problem Overview

Using natural language processing techniques on historic text data can be valuable in determining what causes significant historical events. Specifically, issues, ideas, and sentiments can suddenly become viral and become triggers for influential events. This project uses time series mining, topic modeling, and unsupervised machine learning techniques on British parliamentary documents to detect what viral events occurred in Britain during the 19th century.

Experimental Setup

1. Stratified subsampling by speech date required due to volume of data

2. Remove stop words, given list of domain-specific terms

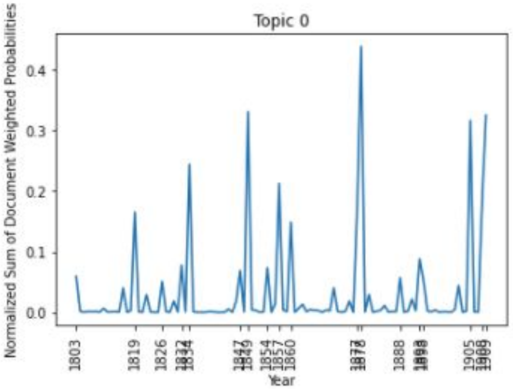
3. Filtering of texts by parts of speech, lemmatization

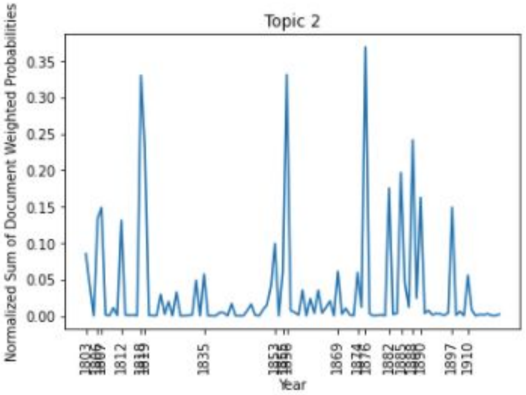
4. Evaluation metric: consult with domain experts to inspect results

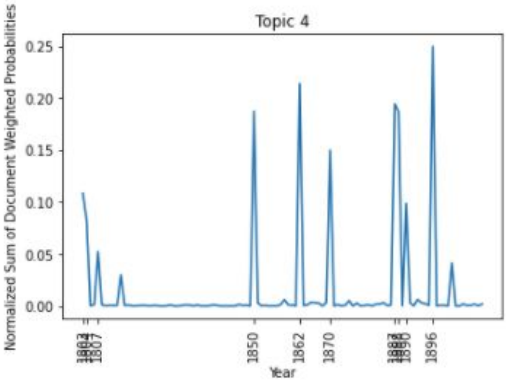
Experiment: Basic Topic Modelling

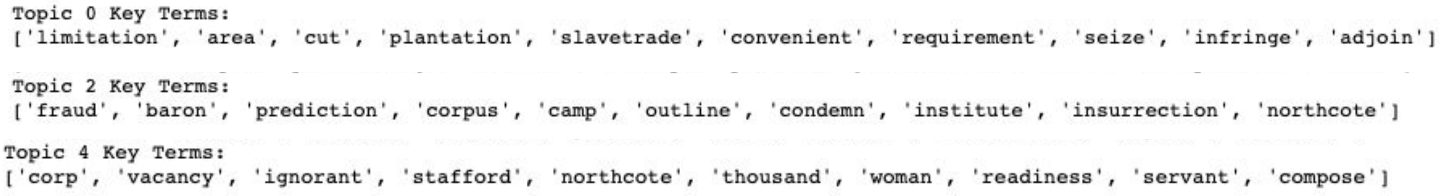
- Use LDA to derive topics from speeches in sample
- Create better representations of topics by removing common terms from resulting word-probability pairs
- Plot sum of probability weights of speeches for topics over time

Experiment: Topic Modeling, n=50



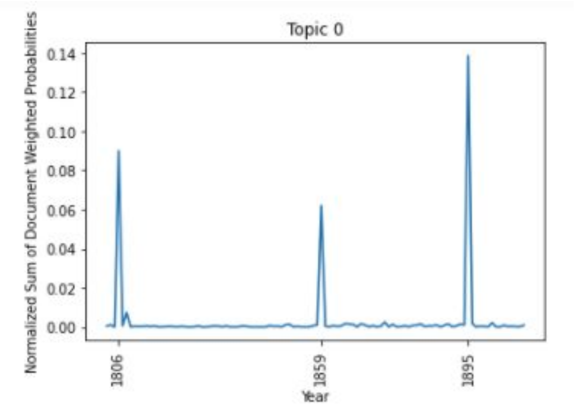


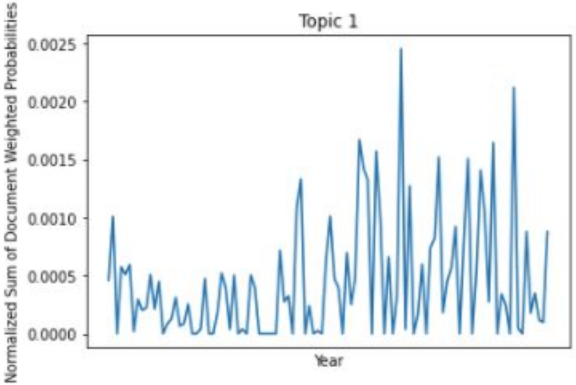


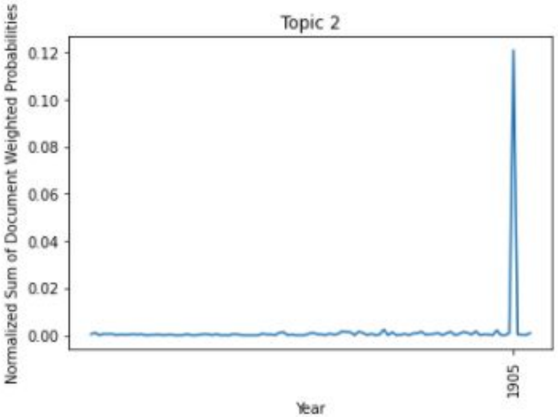


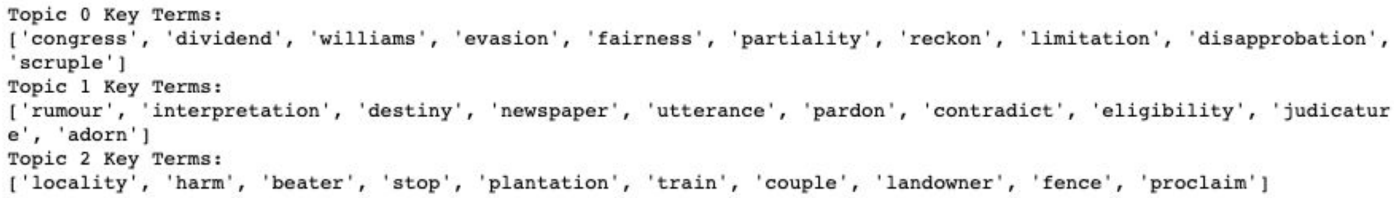
```
Topic 0 Key Terms:
['limitation', 'area', 'cut', 'plantation', 'slavetrade', 'convenient', 'requirement', 'seize', 'infringe', 'adjoin']

Topic 2 Key Terms:
['fraud', 'baron', 'prediction', 'corpus', 'camp', 'outline', 'condemn', 'institute', 'insurrection', 'northcote']

Topic 4 Key Terms:
['corp', 'vacancy', 'ignorant', 'stafford', 'northcote', 'thousand', 'woman', 'readiness', 'servant', 'compose']
```

Experiment: Basic Topic Modeling, n=100





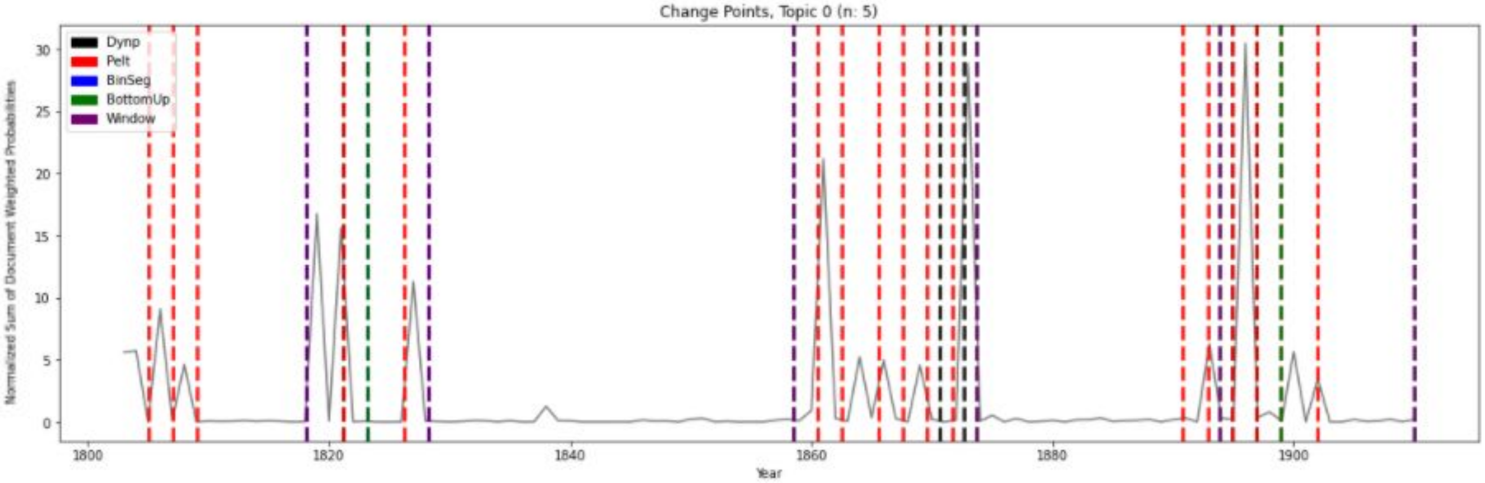




```
Topic 0 Key Terms:
['congress', 'dividend', 'williams', 'evasion', 'fairness', 'partiality', 'reckon', 'limitation', 'disapprobation',
'scruple']
Topic 1 Key Terms:
['rumour', 'interpretation', 'destiny', 'newspaper', 'utterance', 'pardon', 'contradict', 'eligibility', 'judicatur
e', 'adorn']
Topic 2 Key Terms:
['locality', 'harm', 'beater', 'stop', 'plantation', 'train', 'couple', 'landowner', 'fence', 'proclaim']
```

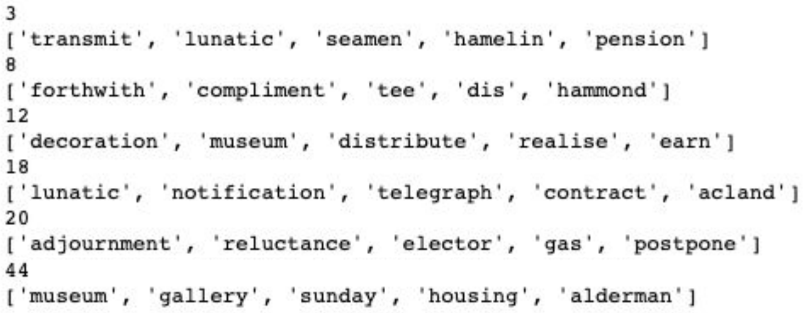
Experiment: Change Point Detection

- Generate time series out of sums of document weighted probabilities over time period for each topic
- Conduct change point detection using 5 different algorithms to find breakpoints

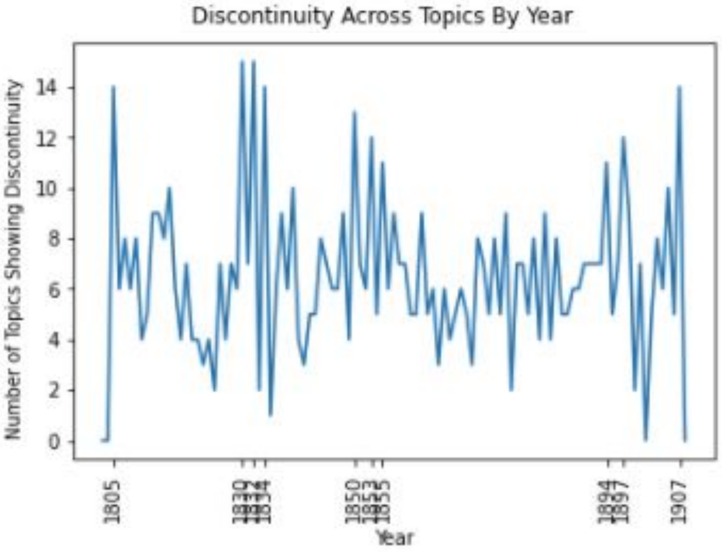


Experiment: Discontinuities Across Topics

- Sum discontinuities (breakpoints) across all generated topics
- Examine if there exists discontinuity for same topics in different years
  - 1825 and 1827



```
3
['transmit', 'lunatic', 'seamen', 'hamelin', 'pension']
8
['forthwith', 'compliment', 'tee', 'dis', 'hammond']
12
['decoration', 'museum', 'distribute', 'realise', 'earn']
18
['lunatic', 'notification', 'telegraph', 'contract', 'acland']
20
['adjournment', 'reluctance', 'elector', 'gas', 'postpone']
44
['museum', 'gallery', 'sunday', 'housing', 'alderman']
```

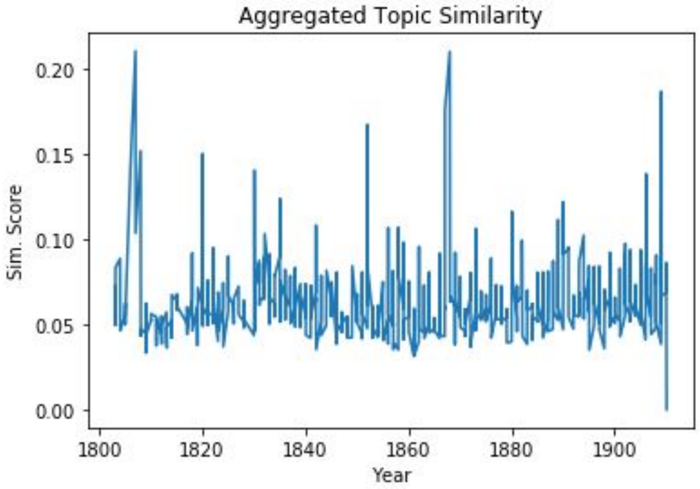


Alternative Experimental Methods

- Aggregate data by month and year
- Perform filtering by parts of speech, stopwords
- Build LDA topic models for each subset

Experiment: Topic Dissimilarity Over Time

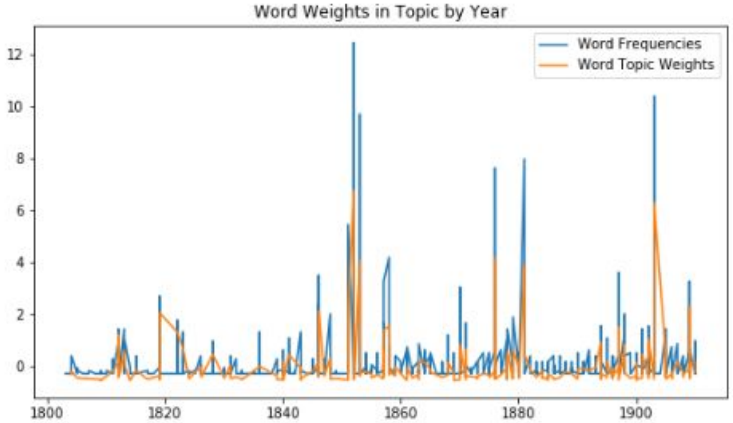
- Calculate year-after-year topic set similarities
  - Jaccard
  - Word-by-word embeddings
  - Aggregated embeddings

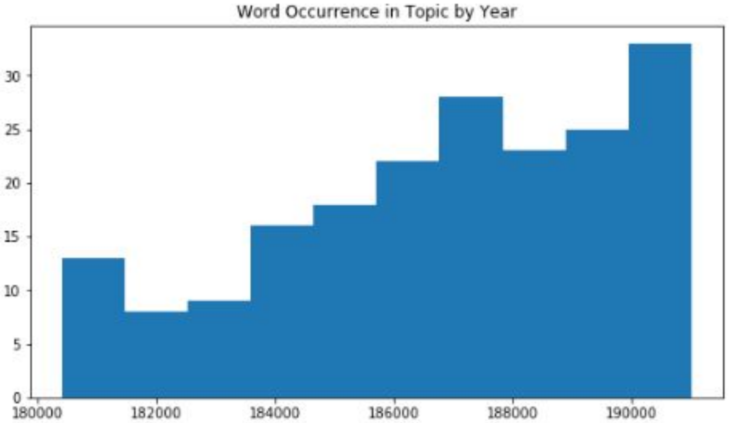


Experiment: Weights by Year

- Look at change of word weights
  - Take weights from topic set models
- Should mirror word frequency
  - Discrepancies could suggest unexpected significance

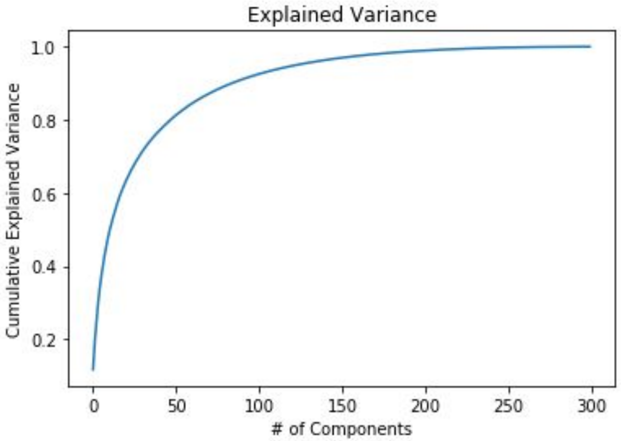
```
Use of 'India' in Topic Models
```



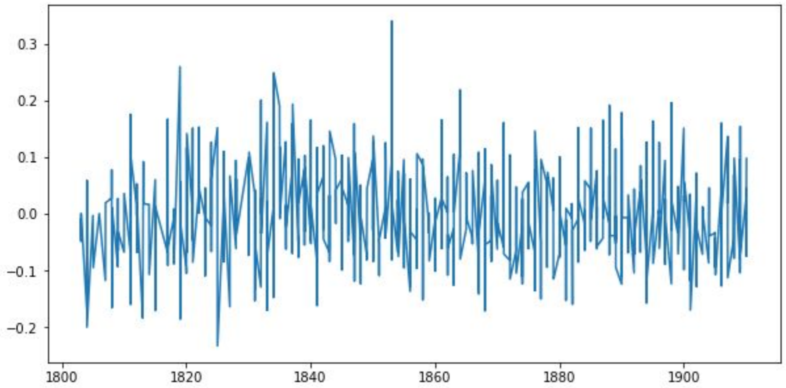


Experiment: PCA Element Change-Point

- PCA reduction on topic embedding elements
  - Reduce dimension of embedding vectors
- Advantage: unsupervised “word” detection
- Change point on reduced vectors

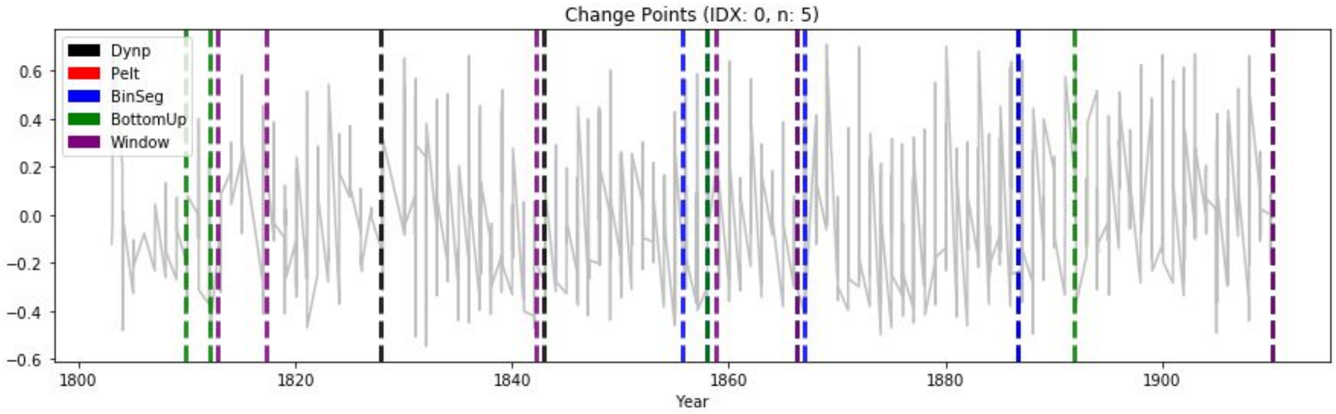


MOVEMENTS IDX: 21



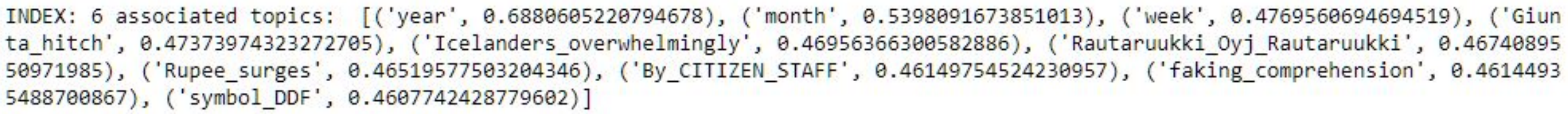
PCA Element Change Point (cont'd)

- Graph element over time
    - Run different change-point algorithms
- Reconstruct vector to get represented terms
    - Simple vector similarity

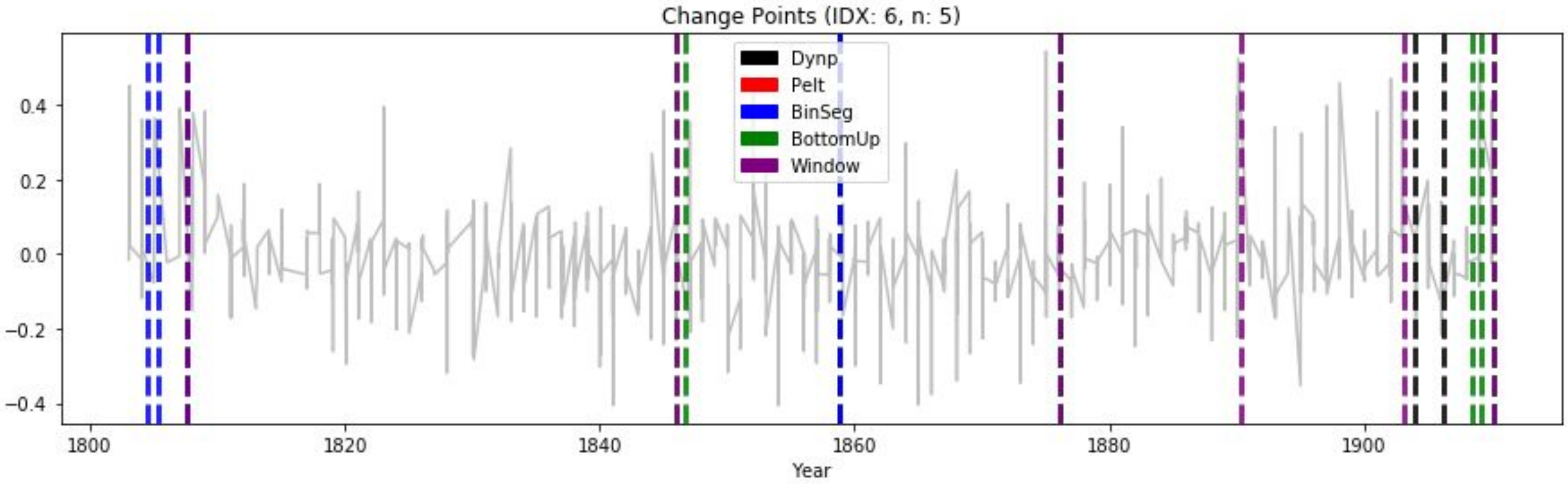


```
('lord', 0.6021789312362671), ('house', 0.5635101795196533), ('ship', 0.5448867082595825)
```

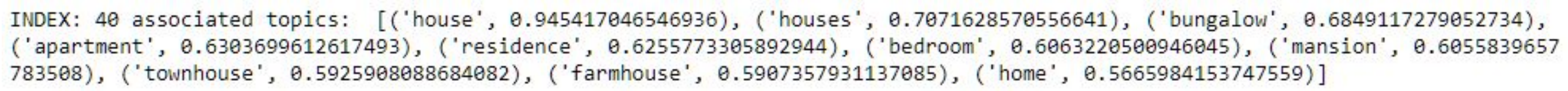
PCA Element Change Point (cont'd)



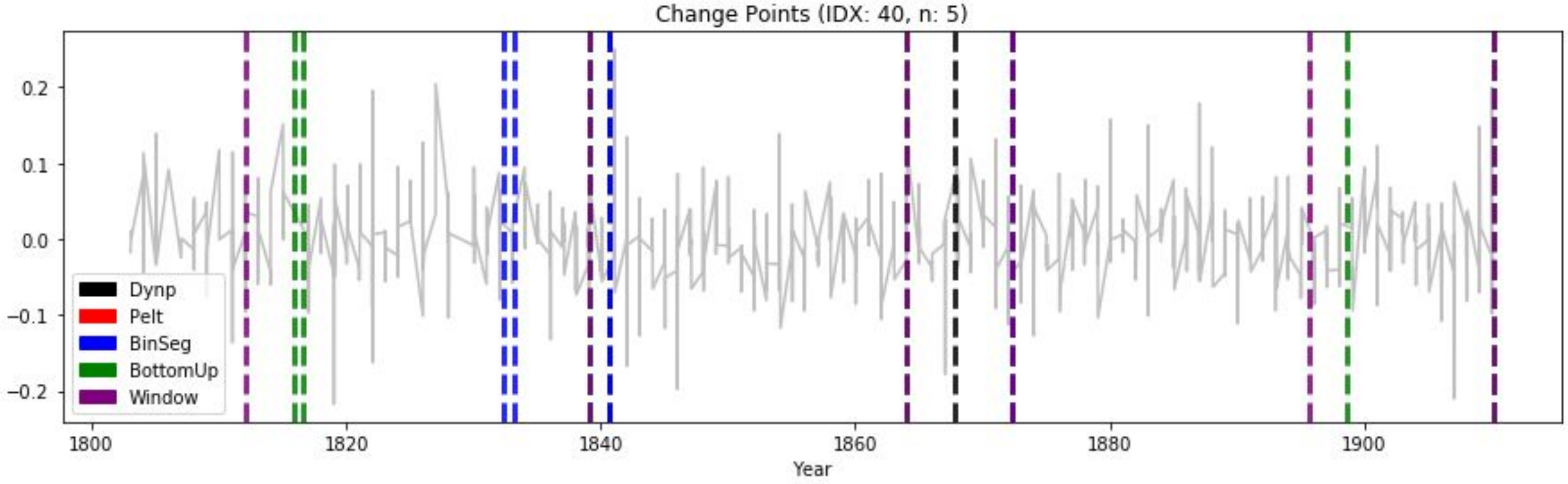
```
INDEX: 6 associated topics:  [('year', 0.6880605220794678), ('month', 0.5398091673851013), ('week', 0.4769560694694519), ('Giun
ta_hitch', 0.47373974323272705), ('Icelanders_overwhelmingly', 0.46956366300582886), ('Rautaruukki_Oyj_Rautaruukki', 0.46740895
50971985), ('Rupee_surges', 0.46519577503204346), ('By_CITIZEN_STAFF', 0.46149754524230957), ('faking_comprehension', 0.4614493
5488700867), ('symbol_DDF', 0.4607742428779602)]
```



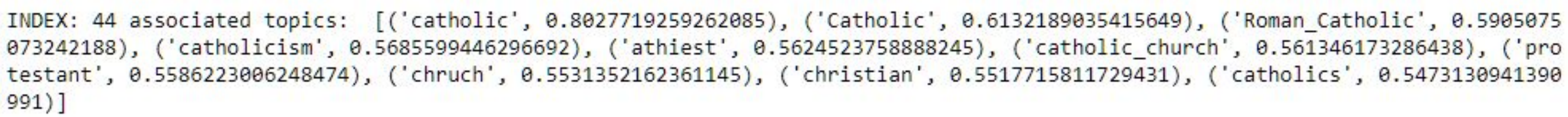
PCA Element Change Point (cont'd)



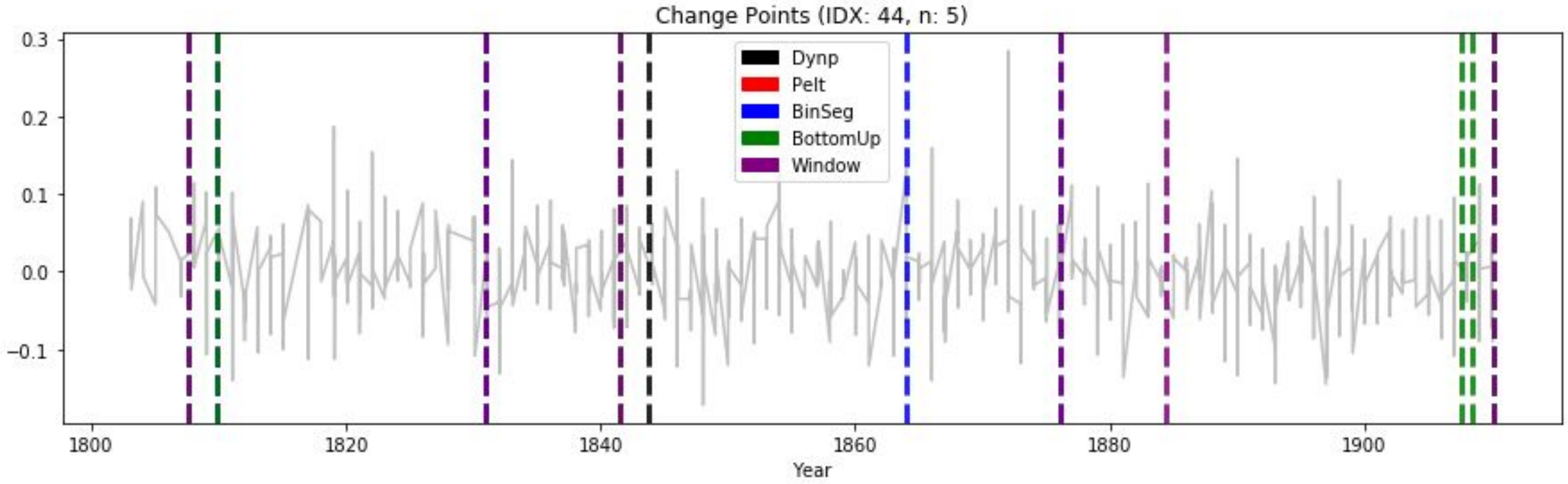
```
INDEX: 40 associated topics:  [('house', 0.945417046546936), ('houses', 0.7071628570556641), ('bungalow', 0.6849117279052734),
('apartment', 0.6303699612617493), ('residence', 0.6255773305892944), ('bedroom', 0.6063220500946045), ('mansion', 0.6055839657
783508), ('townhouse', 0.5925908088684082), ('farmhouse', 0.5907357931137085), ('home', 0.5665984153747559)]
```



PCA Element Change Point (cont'd)



```
INDEX: 44 associated topics:  [('catholic', 0.8027719259262085), ('Catholic', 0.6132189035415649), ('Roman_Catholic', 0.5905075
073242188), ('catholicism', 0.5685599446296692), ('athiest', 0.5624523758888245), ('catholic_church', 0.561346173286438), ('pro
testant', 0.5586223006248474), ('chruch', 0.5531352162361145), ('christian', 0.5517715811729431), ('catholics', 0.5473130941390
991)]
```



Future Work

- Larger sample sizes, hyperparameter tuning
- Transfer learning with word embeddings
  - Different historical vernacular and vocabulary
- Embedding-aware topic models
  - DMM, LFTM techniques
- Further consultation with domain experts
  - Explore the detected items, verify technique effectiveness

Questions?