Topical Extraction on the Book of Psalms

Timothy Lee (xql2001)

Introduction

- Book of Psalms – an anthology of 150 prayers/worship songs/poems from biblical Israel (Bullock, 2001)
- Several authors and likely editorial & compilation phases (Bullock, 2001)
  - Potential stitching/separation of psalms (e.g. 42 and 43)

Psalms – Types & Themes

- Arguably classifiable into 'types' with subject matter/form – the form-critical approach (Weiser, 1962)
  - Main ones are Hymns (focused on praise), Laments, and Thanksgiving
- Arguable themes (Weiser, 1962)
  - Of Praise (Hymns) – Nature/Rule of God, People's act of "proclaiming knowledge of God"
  - Of Lamentation – Judgment, Calamity, Supplication, Penitence, Saving Mercy of God
  - Of Thanksgiving – Mix of both
- However, this is messy
  - Stitching of psalms together (Bullock, 2001)
  - "The form-critical approach is, however, not sufficient by itself to explore the nature of the poetry of the psalms…the mixing of different types is to be found even in the earliest poetry of Israel" (Weiser, 1962)

Latent Dirichlet Allocation (LDA)

- Blei, Ng & Jordan (2003)
- Topic Modelling technique
- In NLP, LDA typically takes a set of text documents and outputs topics
- Topics
    - Encompass the themes of the whole set of data
    - Formed by weightings of individual word "features"
    - Each document is modelled as a weighted mix of these topics
- Works on a form of Bayesian inference
- Topic Coherence (Röder, Both & Hinneburg, 2015)
    - Measure of interpretability of topics from a topic model
    - Can be used to evaluate topic models – a good model should present interpretable, coherent, topics

Research Question and Hypothesis

- Is topical extraction using Latent Drichlet Allocation (LDA) able to identify distinct themes in the psalms?
    - Hypothesis: Clear themes corresponding to/captured by topics with high coherence scores for LDA model

- Is topical extraction using Latent Drichlet Allocation (LDA) able to reflect blending of types and the influence of editorship?
    - Hypothesis: Messy topic dominance - Mixed dominance between topics in many psalms

The Data

- Book of Psalms World English Bible (WEB) Translation
  - Free public domain translation of the Bible, and is an updated version of the American Standard Version (ASV).
  - Uses more modern English – hence works better with NLTK tools for lemmatisation and stemming
  - Tried ASV, but Word Net Lemmatiser and Porter Stemmer unable to detect ‘th’ suffix

Method and Implementation

Apply LDA to Psalms	Full Text – Book of Psalms Document – Individual Psalms 1 to 150
Text extraction	1. Copy + Paste full text to text file 2. Read as string object in Python
Pre-processing	1. Remove authors and supplementary information 2. Split into individual psalms using re.split 3. Remove stopwords 4. Lemmatisation
Model training	1. Tested 30 LDA models with 1-30 topics 2. Took model compromising coherence with n_topics
Model analysis and visualisation	1. Identify topic themes 2. Find dominant topics and its relative dominance for each psalm

Summary of Findings – Topic Coherence

- Poor topic coherence of the model at around 0.3 consistent for n_topics = 1 to 30
  - For n_topics = 5:
  - Can derive themes from looking at word weightings in topic
  - Largely coherent with different themes of praise form, as mentioned by Weiser (1962)
    - God’s nature (being proclaimed/praised)
    - The people praising
    - God’s actions/rule
  - Low salience of words with links to lamentation forms
    - Despite high incidence of lamenting forms of psalms (about 1/3)
    - Believable given that praise is also involved in lamentation forms
  - However, hard to trust inferences on topics given low coherence

Summary of Findings – Topic Coherence

- High overlap between words in different topics
  - Biggest culprits – Yahweh and God were taken out, because their outsize frequency is obvious
  - Still the high overlap suggests high word frequency – similarity/interconnectedness in themes in praise form the psalms
  - Makes sense given the hymn form: Call to Praise Yahweh, then Praise of Nature/Rule of God (Weiser, 1962)

Summary of Findings – Topic Dominance

- High topic dominance in most of the individual psalms
  - only 27/150 had dominant topic weightage <= 0.75, which should suggest limited influence of editing, stitching, and mixed forms across the book of Psalms
  - Confounded by high overlap between topics – editing, stitching, and mixed forms may have manifested in the topics instead, explaining low coherence

Future Refinements and Directions

- Refine current methodology
  - Feature Engineering - Add stopwords, Use Ngrams
  - Increase n_topics beyond 30
  - Try LDA MALLET and other Topical Extraction techniques
- Analysis based on grouping similar topics
- Apply supervised learning to classification of psalms using form-critical theory classifications as expert labels
  - Apply to psalms outside Book of Psalms and other Near-East writings
  - Investigating model output and process may yield new insights on structure of writings from Ancient Near-East

Thoughts

- Liked Best – Worked with an interesting dataset, data wrangling is satisfying
- Liked Least – Feature Engineering and explanation – hard to draw insights with rudimentary understanding of both the methods and topic

References

Bullock, C. H. (2001). *Encountering the Book of Psalms: A Literary and Theological Introduction*. Grand Rapids, MI: Baker Academic.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (pp. 399–408). New York, New York, USA: ACM Press. https://doi.org/10.1145/2684822.2685324

Weiser, A. (1962). *The Psalms: A Commentary*. (G. E. Wright, J. Bright, J. Barr, & P. Ackroyd, Eds.) (Fifth Revi). Philadelphia: Westminster John Knox Press.