FROM DISTANT TO CLOSE READING: EVALUATING THE USE OF LDA TOPIC MODELING

Christoph Aurnhammer c.j.aurnhammer@uvt.nl
Iris Cuppen iriscuppen@gmail.com
Inge van de Ven i.g.m.vdven@uvt.nl
Menno van Zaanen m.m.vanzaanen@uvt.nl

Tilburg University

Introduction

- Huge amounts of available texts requires rethinking reading approaches
  - Not all texts can be read in extreme detail
  - Can we identify the most interesting texts in large collections?
  - Can we identify overall trends found in texts in large collections?
- There are two seemingly conflicting approaches:

  Close reading: detailed (manual) analysis of texts: shows details

  Distant reading: automatic analysis of texts: shows general trends
- Can we develop approaches that allow for a combination of close and distant reading?

Idea

- Clustering texts will allow us to perform close reading on an interesting selection of texts
- How can we cluster documents from a close reading perspective?
- Can we develop a distant reading approach that mimics the close reading approach?

Dataset

- Reddit discussion thread: 461 comments
- "Should the Democrats nominate a celebrity in 2020? What would be the pros and cons?"

Close reading

- Manual, hypothesis-free close reading of all comments results in assignment into 16 classes

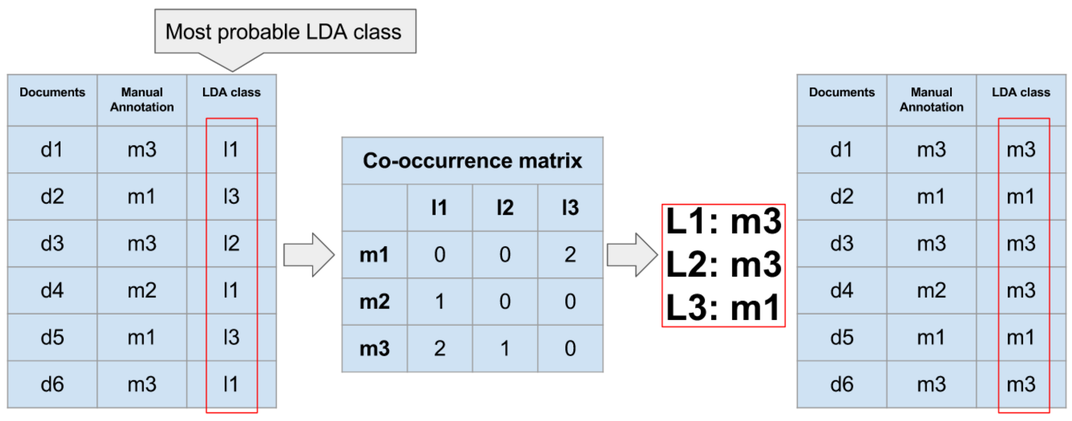
	Class support	Category
1.	24	Classic leadership characteristics: experience, competence, seriousness
2.	21	Celebrity characteristics: charm, likeability, humor
3.	26	Bad leadership characteristics: incompetence, foolishness, populism
4.	50	Media, branding, etc.
5.	45	Entertainment Industry, celebrities
6.	20	Analyzing the system and deciding which candidates could fit this system
7.	133	Red and blue, Republicans vs. Democrats
8.	8	Target groups, minorities, religion
9.	20	Misogyny, gender
10.	18	Ability of the public to vote
11.	12	Fact-checking, truth-finding
12.	2	Conspiracy thinking
13.	23	Jokes
14.	2	Responses to other comments
15.	39	Purely emotional comments
16.	18	Deleted or removed comment

Distant reading

- LDA (Latent Dirichlet Allocation) topic modeling (Blei et al., 2003)
- Varying (predefined) number of topics: 1–461 (= number of documents)

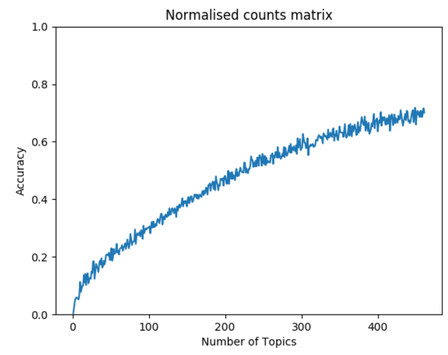
Distant reading classes → close reading classes

- Identify co-occurrence counts (and normalize)
- Select best close reading class for each distant reading class
- Apply mapping and calculate accuracy



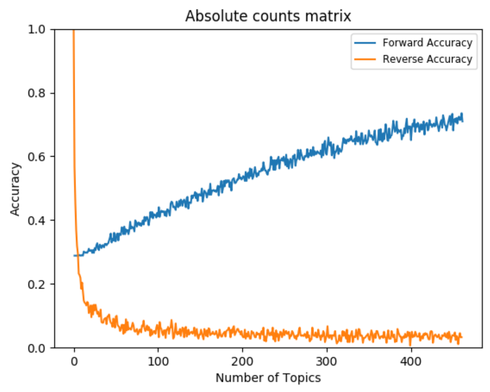
Finding the optimal number of LDA classes

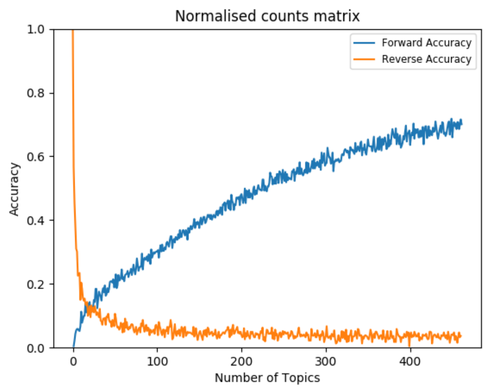
- One LDA class for all comments does not distinguish
- One LDA class for each comment does not generalize (cluster)



- Increasing number of LDA classes increases accuracy
- Increasing number of LDA classes reduces generalization (at some point)
- Which point has highest distinctiveness and generalization combined?
- Incorporate reverse overlap: (Also) map close reading classes to distant reading classes

Results





- Optimal number of classes is at the intersection of forward and backward accuracy
- Baseline: Chance assignment over 16 classes

	Accuracy	Optimal # of topics
Baseline (random assignment)	6.03%	-
Forward + Reverse (absolute counts matrix)	29.17%	6
Forward + Reverse (normalized counts matrix)	13.01%	23
Manual analysis		16

Conclusion

- Manual, hypothesis-free close reading provides additional and extra-textual information
  - Behavior in Reddit thread
  - Allows for further investigation (e.g., order of types of comments)
- Automatic, LDA-based distant reading is efficient
- Finding optimal number of classes is difficult without knowledge of goal
- The consistency of close and distant reading can be assessed with a hand-labeled sample.

Future work

- Exactly which questions do we want to answer using close reading?
  - In how far can distant reading help when dealing with large document collections?
- How consistent are identified classes and the LDA system over topics?
- Given the manual annotations, can labeled LDA perform better?
- What are the limitations of the overlap measure?

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3:933–1022.