CoLi – Final Project – Diacritics Restoration

Herbert Ullrich (**2576412**)

March 15, 2019

1 Introduction

1.1 Project files

- **movies-pp.txt** - the test data, as unzipped from Piazza
- **movies-50.txt** - first 50 sentences of the test data, included with submission, used for debugging
- **lda.py** - a runnable python module to execute LDA over the given test data
- **topics.txt** - text output of a single run of **lda.py** describing the 20 modelled topics

2 Implementation

As for time reasons, implementation of the LDA itself was skipped. Instead, we are at least including a short script **lda.py** which solves the given challenge using the Python package gensim by Radim Řehůřek [1]

3 Results

We observe that the most of our model topics include dominant words like *film*, *movie*, *like*,... Let us therefore try to omit the words and topics we consider uninteresting. Full output of a run of our program over the full testing set can be found in the file **topics.txt**.

4 Conclusions

The topics modelled by our LDA model sometimes indeed represent thematically coherent semantic fields: different movies, as shown in topics 4, 6, 9 or e.g. a movie series, topic 2.

The main limit was in our case the fixed context of movies – words like *movie*, *film*,... made a lot of appearances across the topics and were misguiding the algorithm trying to distinguish smaller thematic nuances.

4.1 Improvements

In our opinion, (even a very dummy) lemmatization of the corpus before its evaluation with LDA would improve the accuracy dramatically. Melting down the semantically equivalent words (e.g. the synonyms or inflected forms) to some computer-distinguishable unique representative would make the semantically important words occur more frequently (and would reduce bias by changing the speaker while preserving the topic).

References

[1] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.