Lemmatization for Topic Modeling

1 Introduction

In order to determine if lemmatization has a beneficial impact on topic modeling for English documents, we ran some experiments comparing the effect of both stemming and lemmatization on a standard data set.

2 Experiments

To make the experiments as simple as possible, we adapted existing example code from the scikit-learn Python library[1] . The example code applies both Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) to a corpus of documents taken from the well-known "20 Newsgroups" dataset[2] , which is included as one of the built-in datasets in scikit-learn. Once the two models have been fitted, the code then outputs the most frequent words for each topic.

We modified the code to run the same tests on three copies of the full training set: one without any pre-processing, one with stemming, and one with lemmatization. We also added code to compute the held-out perplexity for each of the LDA models, using the dataset's pre-defined test set.

3 Results

To evaluate the results, we looked at three different aspects:

- Readability of the top word lists for each topic
- Manual comparison of how well the resulting topics matched the original newsgroups
- Held-out perplexity for the LDA models (provided by the scikit-learn library)

The full set of results can be found in the appendix at the end of this document. Below is a summary of our findings.

[1] http://scikit-learn.org/stable/auto_examples/applications/topics_extraction_with_nmf_lda.html

[2] http://qwone.com/~jason/20Newsgroups/

4 Readability of Topic Terms

In terms of readability, the advantages of lemmatization are evident. Consider, for example, the following terms were extracted for three of the topics:

- **LDA (stemmed):** gun state law right govern peopl weapon crime ani control
- **LDA (non-lemmatized):** gun guns law laws control police rate crime state firearms
- **LDA (lemmatized):** gun law government right state case weapon control people

- **NMF (stemmed):** car engin driver mile speed owner buy price model look
- **NMF (non-lemmatized):** car cars engine speed price good new driver bought looks
- **NMF (lemmatized):** car engine buy price mile speed driver sell owner model

Both stemming and lemmatization improve readability by eliminating semantic duplicates such as gun and guns or law and laws. However, stemming produces a number of stems that do not correspond to real words, such as people and ani, which significantly decreases the readability of the frequent word lists, and disqualifies stemming from being a viable option. Given this, we will omit the stemming results from the other aspects of the evaluation.

Lemmatization also improves readability in another way, by correctly dealing with contractions which are otherwise split by the tokenizers, resulting in incomplete words such as don, ll, re and ve:

- **LDA (non-lemmatized):** does think people just don believe point time case say
- **LDA (lemmatized):** say think know people just come tell like thing good

- **NMF (non-lemmatized):** don think time good did really say make way want
- **NMF (lemmatized):** good think just like time make year thing little look

Conclusion: Lemmatization improves the readability of topic terms (user experience) by reducing duplicate topic terms and by correctly dealing with contractions.

5 Topic Quality

The original dataset is divided into 20 newsgroups (discussion groups from Usenet, the legacy worldwide distributed discussion system), differentiated by topic (religion, computer graphics, hockey, etc.). In order to evaluate how well the resulting topics match the original newsgroups, we consider three different measurements:

- The number of topics that can be easily matched to a newsgroup
  - o NMF (non-lemmatized): 14 topics matching 13 newsgroups
  - o NMF (lemmatized): 18 topics matching 16 newsgroups
  - o LDA (non-lemmatized): 14 topics matching 13 newsgroups
  - o LDA (lemmatized): 14 topics matching 13 newsgroups

- The number of topics that can be easily matched to a newsgroup
  - o NMF (non-lemmatized): 14 topics matching 13 newsgroups
  - o NMF (lemmatized): 18 topics matching 16 newsgroups
  - o LDA (non-lemmatized): 14 topics matching 13 newsgroups
  - o LDA (lemmatized): 14 topics matching 13 newsgroups

- The number of newsgroups that cannot be matched to a topic
  - o NMF (non-lemmatized): 7 newsgroups without a matching topic
  - o NMF (lemmatized): 4 newsgroups without a matching topic
  - o LDA (non-lemmatized): 7 newsgroups without a matching topic
  - o LDA (lemmatized): 7 newsgroups without a matching topic

- The number of topics that cannot be matched to a newsgroup
  - o NMF (non-lemmatized): 6 topics without matching newsgroup
  - o NMF (lemmatized): 2 topics without matching newsgroup
  - o LDA (non-lemmatized): 6 topics without matching newsgroup
  - o LDA (lemmatized): 6 topics without matching newsgroup

Conclusion: While lemmatization on English does not seem to affect the LDA model, it clearly improves the NMF results, with the lemmatized NMF model outputting the best results overall.

6 Conclusions

In our initial experiments, the effects of using lemmatization for English topic modeling, compared to not using lemmatization, showed strong improvements for NMF models and provided comparable topic quality for the LDA models. The held-out perplexity of the LDA model was also improved by applying lemmatization. In regard to the readability of topic terms, lemmatization provides significant improvements over not lemmatizing. Since there are benefits to applying lemmatization without any negatives, our conclusion is that lemmatization should be used for English.