introduction

September 24, 2019

1 Twitter Topic Modelling

1.1 Introduction

Due to the increasing volume of digitized text, the automatic classification and organization of written information becomes more and more important. The widely used tools to address this problem are topic models. These models originated from heuristic approaches, like singular value decomposition (SVD) in latent semantic indexing (LSI). The evolution of topic modelling continued in generative probabilistic models, two key developments are probabilistic LSI (pLSI) [#] and its continuation: latent Dirichlet allocation (LDA) [#]. The latter was established as the state-of-the-art method, and was widely used for classification, recommendation, and even political analysis [#].

```
<left>
    <b>Fig. 0.</b> This figure shows the connection between the probabilistic model of topics a
</left>
```

However, in a recent (2018) paper [#] the authors argue that network based community finding methods (for example hierarchical Stochastic Block Model - hSBM) are superior to the LDA model. They showed, that their approach achieves superior results on real life data (Wikipedia, New York Times, etc.), and even on synthetic data constructed from the generative process of LDA.

(The figures below are copies from the original paper [#])

1.2 Plan of action

My project is to understand the new topic modelling approach introduced in [#] and compare it to LDA on tweets. The following tasks should be done during the semester:

```
1. Download tweets of some popular topic from Twitter, e.g. #Brexit, or #meetoo
(or tweets around a given location,  e.g. New York, London)
    a) Download (I am using tweepy, see tweet_collector.html)
    b) Clean the data (stopwords, usernames, plural forms)

2. Understand the two approaches discussed in the article
(which is quite hard in itself, because of the complex statistical models introduced by the au

3. Find suitable libraries (e.g. networkx), which implement the necessary algorithms
(e.g. SVD, hSBM, other community finders)
```

```
4. Test the methods on the dataset, and assess whether the results are meaningful. If not, fin
said to be hard to apply topic modelling on tweets, due to their shortness, and the nature of
```

[ ]: