I'm going to talk about my proposal for doing topic modeling on twitter data.

When talking about social issues, one might think that most of the people is up to date. Talking with friends, colleagues and other people refreshes our understanding of current situations. However, it is not the case for issues that happened in the past.

What is the investigated phenomenon?

„You can neither remember nor forget what you do not understand".

Memory is vulnerable, it can be easily distorted to fit beliefs that are more comfortable than accurate. Individuals does not appear to be willing to cause destruction (or they are incapable). This is not the case for groups of individuals. The dangers of self-deception about past events can be magnified in the social media. On the other hand, social media can be used to gather information about past events, thus things can be put back into the correct past narrative frame. So, we need some method to differentiate between the two.

What I intend to do my research on, is the tweets about a rally (Unite the Right rally) in the United States, which took place in Charlottesville in 2017. The rally occurred when controversy emerged by the removal of confederate statues. There were clashes between armed protesters and counter-protesters, and the rally turned violent. The clashes culminated in the event of a white supremacist ramming his car into the crowd of the counter-protesters killing one person and injuring many then fleeing the scene.

I got interested what I'm going to do, because I got interested at looking at motivation for social conflict. My research's goal is to shed some light onto the narrative structure through tweets surrounding the Unite the Right rally and the series of events after it. It can be achieved objectively by using unsupervised learning algorithms.

People use Twitter (the famous social media platform) to share their thoughts with other people in the form of short posts, tweets. This shortness causes that the tweets are brief, dense and not so complex. Usually one or two sentences. The low complexity means, that each tweet is focused on mostly one thing. The high information density means, that we can get results on relatively small data.

What data should be still collected?

I'm going to use the Twitter API to retrieve tweets with the #Charlotessville hashtags or the tweets mentioning "Charlottesville". I've already applied for permissions at Twitter and my developer account is activated. According to the documentation I'm able to gather 500k tweets /month with a standard product with a stricter limit of 300 request/15 min. I plan to download about 50k tweets, which should take about ~50 hours with some idle time.

Does the data need to be cleaned and if yes how?

Firstly, we should get rid of retweets, as they are not a document in and of themselves. Secondly, we need to filter for the location (US) and the time interval (august 10-20. 2017.)

The steps I'm going to take for each report

I'd like to make the following progress:

For the first report I should be done with acquiring the data from the Twitter API. Other than that I'm going to find a way to configure graph-tools (which is a network analysis library). It might be problematic, as I use windows, but it works on linux only. As a solution, I might use docker or windows subsytems for linux.

For the second report I'm going try to get a grasp on the LSA, pLSA and LDA algorithms and apply them in order to reach some intermediate results, and to have them as a baseline for the final part.

Finally, as for the third report, I'm going to use the hSBM community detection method for the same task, compare with previous results and draw the final conclusions about the social media activity surrounding the series of events in Charlottesville.