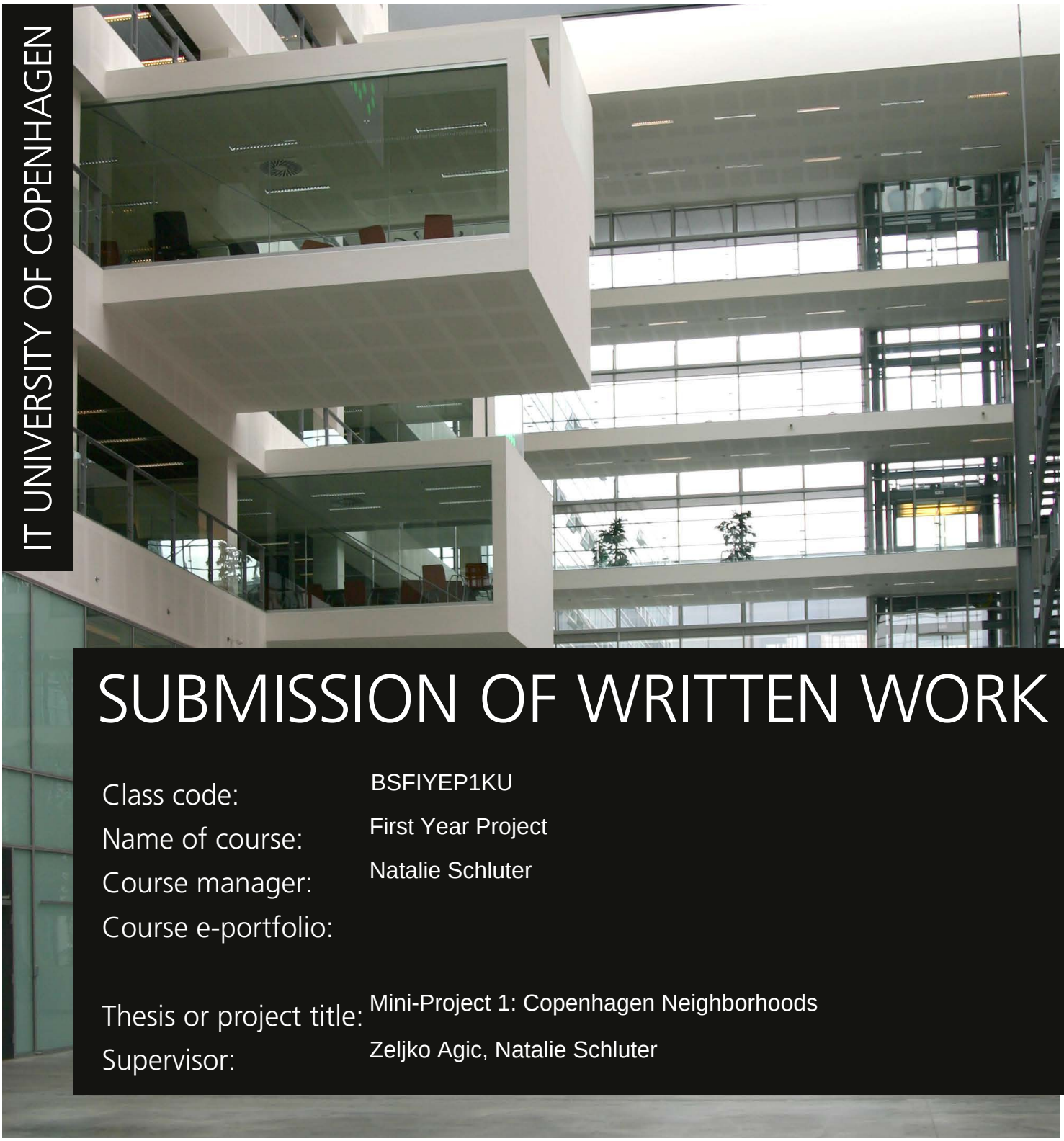
IT UNIVERSITY OF COPENHAGEN



SUBMISSION OF WRITTEN WORK

Class code:	BSFIYEP1KU
Name of course:	First Year Project
Course manager:	Natalie Schluter
Course e-portfolio:	
Thesis or project title:	Mini-Project 1: Copenhagen Neighborhoods
Supervisor:	Zeljko Agic, Natalie Schluter

	Full Name:	Birthdate (dd/mm-yyyy):	E-mail:
1.	Robert Bayer	08/06-1998	roba@itu.dk
2.	Joachim Moe Osterhammel	07/12-1995	joos@itu.dk
3.	Steffan Maksimovic	13/01-1999	stma@itu.dk
4.	Oskar Martin	14/03-1995	osma@itu.dk
5.	Bence Zoltan Balazs	18/12-1996	beba@itu.dk
6.	Katerina Rique Anastasia Østergaard	17/03-1998	kaoe@itu.dk
7.			@itu.dk
8.			@itu.dk

1 Introduction

Living in Copenhagen differs vastly depending on the neighbourhood. Such neighbourhoods give rise to different interactions and social cultures.
In this report, we will analyze how people interact with different neighbourhoods in the city of Copenhagen. The main objective is to discover the likes, dislikes and typical days of a Copenhagener in the following neighbourhoods: Amager, Nørrebro, Vesterbro, Østerbro, Indre by, Valby and Torvehallerne.

2 Methodology

The analysis of this report rely on four over-all methodologies within the domain of NLP. (1) term-frequency, which is the foundation of the (2) term-frequency, inverse document frequency. Furthermore, an analysis of sentence structures and overall topics will be analyzed via (3) n-grams & (4) Latent Dirichlet Allocation (LDA).
To calculate the TF-IDF, we take product of the TF and IDF.

$$TF_{t,d} = \frac{t_d}{\sum_{d=1}^{D} w_d} \qquad IDF_t = \log \frac{N}{n_t} \tag{1}$$

The n-gram method was used to real-estate listings, which predicts the $n't$ word given a sequence of $n-1$ words. LDA generates overall topics and populates these with most frequently used words belonging to that topic. The model is as follows:

$$\prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, Z_{d,n})) \tag{2}$$

3 Data

For the foundation of our analysis we chose captions of Instagram posts that contain hashtags with the names of the districts, real-estate postings from the specific districts and finally minutes notes from meetings[1] held by local committees in the districts.

The Instagram data contained captions in several languages, which were separated and only English and Danish were used. The $243,878$ captions were distributed differently throughout different district, with the size ranging from $3,816$ to $83,611$ captions.

The real-estate listings were scraped from Home[2] using a custom scraper, scraping the listings through ZIP codes corresponding to the districts, adding up to $\sim$1500 listings.

The meeting minutes were scraped manually, where we scraped data from all the meetings since January 1st 2018, to get the most recent, unresolved, problems and plans

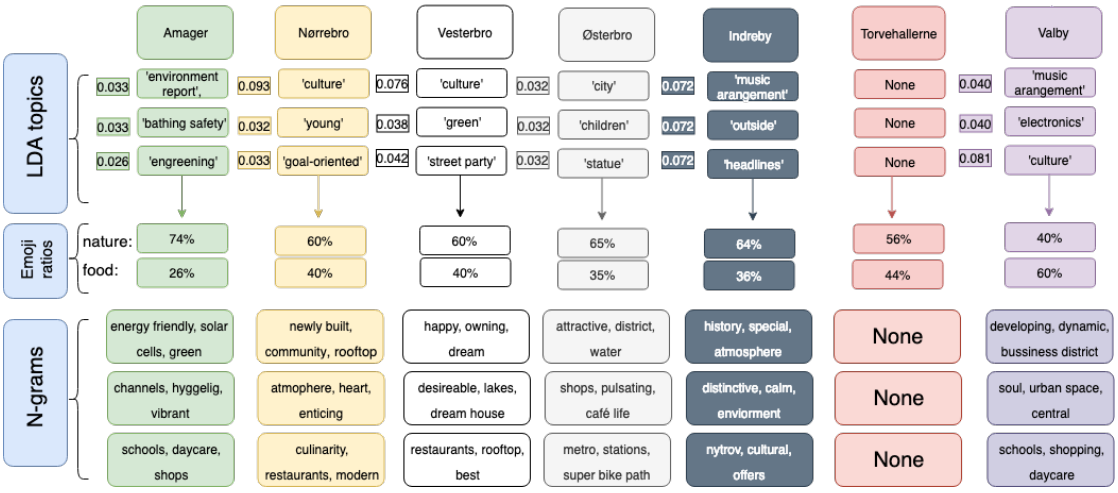
[1] https://www.kk.dk/artikel/lokaludvalg
[2] https://home.dk/

in the districts. Each district has a meeting every month with between 12 - 20 lines about each meeting, which results in 168 to 280 lines of data from each local committee. There are two local committees for Amager and none for Torvehallerne as this is not a separate district.

4 Results

Through our analysis, we found matches between the topics of minutes, Instagram emoji usage and n-grams from real-estate listings. The local committee in Amager overall sums to environmental matters, which also reflects in the Instagram captions where the most frequent emojis belong to the nature category. We also see an overall theme for Copenhagen is being a green, family friendly city which is shown in all of the LDA topics as well as the real-estate n-grams.
The differences between the neighbourhoods are illustrated by Vesterbro, Nørrebro and Indreby all sharing topics such as street parties, music arrangements and youth, whilst Amager and Østerbro are more focused on family matters such as daycare institutions, calm environments and children. The following figure (1) sum up all of our results.

Figure 1: Results of LDA, emoji frequency and n-grams



5 Interpretation

For Nørrebro, we found that the positive sides of the neighbourhood primarily revolves around the social and culinary aspects which implies that this part of the city is more focused on going out than for example Amager where family life and the nature plays a more important role. The downside of the Nørrebro is the *"ghetto"* aspect highlighted

in political minutes which might imply higher crime rates, which can bother the local residents. For Østerbro, the positives side are that it is an exclusive neighbourhood with green areas and access to the newly made bay.

For both Vesterbro and Indre By, there's a big overlap in up and down sides of the areas. Both highlight cultural life, such as musical events – with Vesterbro emphasizing Distortion and Indre By being bothered by noise pollution, which is a big downside of living in the city centre. The real-estate listings reflects the typical days of the local residents, with focus on transportation for Østerbro and daycare as example.

6 Error Analysis

Due to the demographics of Instagram, these captions mainly reflect a younger segment of Copenhageners. Inherently both the Instagram captions and real-estate listings are positive. This results in few negatives being highlighted in our data. Another problem arose due to language differences between Danish and English text data. Another source of bias is that we filtered out irrelevant findings resulting in increasing thresholds for both n-grams, LDA and TF-IDF. Lastly, a larger data set for Torvehallerne is needed for a meaningful analysis of this area, but was not possible due to time constraints.

7 Concluding remarks and future work

As stated, different neighbourhoods give rise to different interactions within them. Our results all point to `nature` and `culinary culture` being overall positives for all of Copenhagen. The more negative sides are inherent to the neighbourhoods and less general, such as `ghettos` for Nørrebro and `noise pollution` for Indre By and Vesterbro. For further analysis an analysis of contextual meanings of emojis could prove beneficial – this would reveal deeper meanings behind emoji usage.