Urban Perception Extraction From Texts Shared on Social Media: Framework and Applications

Candidate: Frances Albert Santos
Supervisor: Prof. Leandro Aparecido Villas
Co-supervisor: Prof. Thiago Henrique Silva (UTFPR)

Campinas,
July 30, 2021

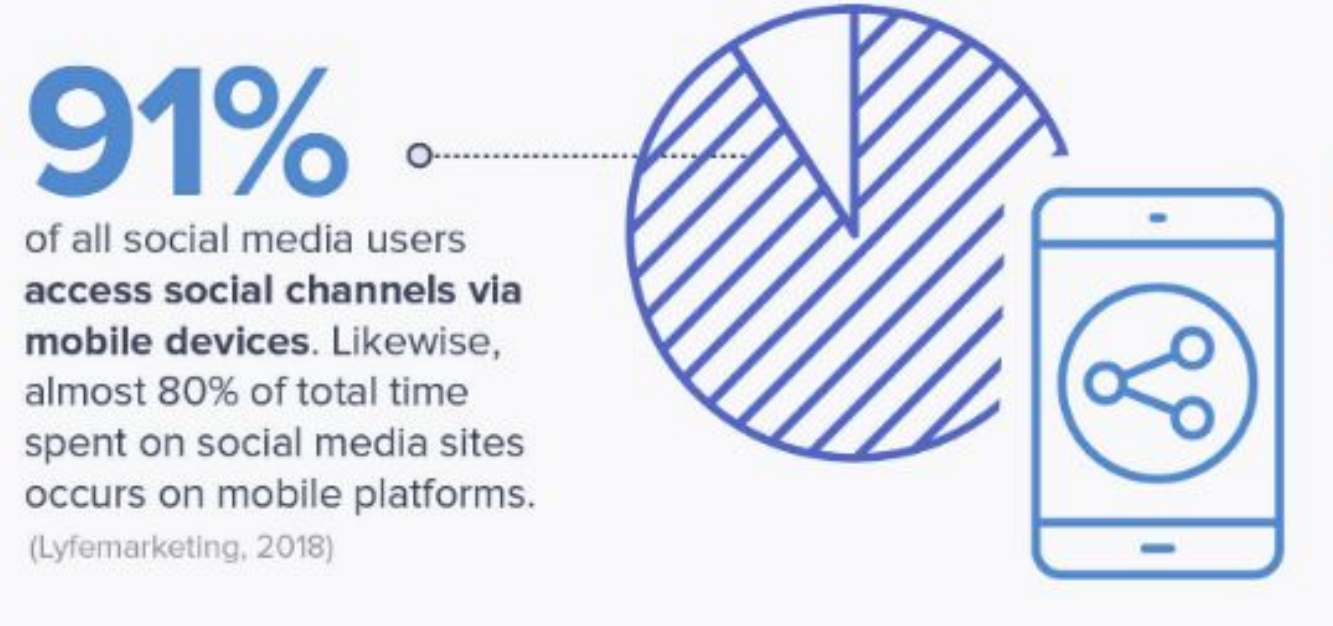
Agenda

- Motivation
- Goal
- Our Framework
- Results
- DEMO
- Final Remarks

Motivation

- **Why social media?**

Social media is an important part of many people's lives



https://www.oberlo.com/blog/social-media-marketing-statistics

The power and influence of social media

comes a rich
containing
nation about
g on a global
e

Motivation

- **Why texts?**

Looks like quite similar, but ...

Motivation

Pedestrian in critical condition after being hit by truck in Vanier

Woman in her 80s rushed to hospital trauma centre, paramedics say

CBC News · Posted: Nov 03, 2017 11:02 AM ET | Last Updated: November 3, 2017

Traffic accident
Nov 03 2017, 10:45 a.m.

Murder
Nov 27 2017, 09:30 p.m.

Shot fired during holdup at Vanier cannabis dispensary

SHAAMINI YOGARETNAM

More from Shaamini Yogaretnam

Published on: February 18, 2018 | Last Updated: February 18, 2018 12:08 AM EST

Shot Fired
Fev 18 2018, 12:08 p.m.

Motivation


Daddy_Jazz 1.47
stuffy nose feelin miserable :-(
Hershey_Madness 0:93
UptownEssence 1.35
afrosemary 1.27
STRUCK_BYglam 0.63
NooSoul 1.04
95KOVG 0.86
MrsBGerard 0.91
MELLIZA04 1.47

Goals

“Extract useful urban perceptions using public social media content to help better understanding urban areas and leverage new services and applications.”

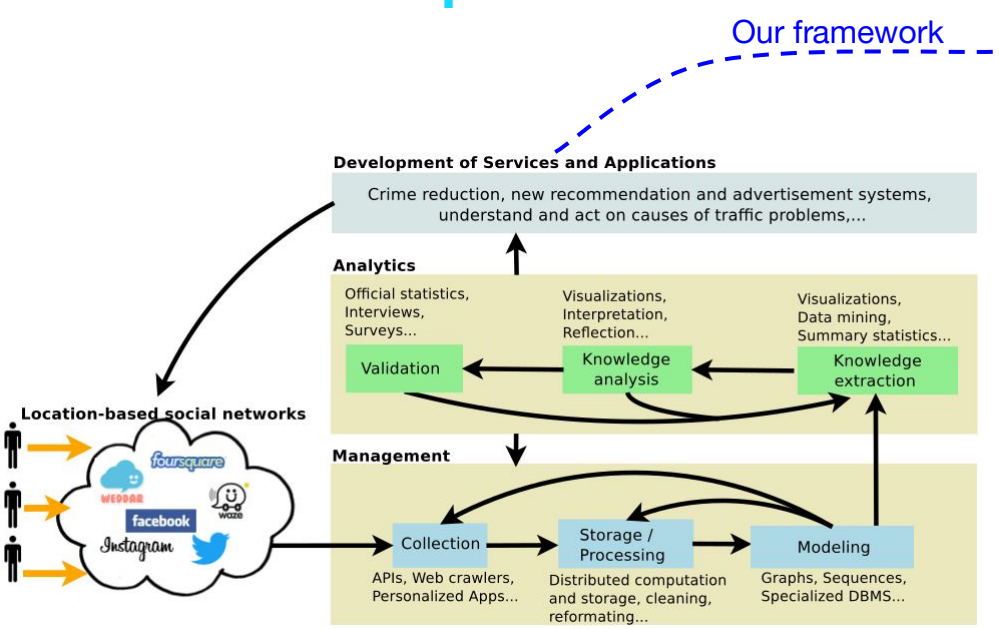
- How can we extract urban perceptions from free texts?
- What is the level of agreement among extracted perceptions with respect to a “ground truth”?
- How urban perception can be exploited to leverage new services and applications?
- How can we combine multiple layers of perception to obtain aggregated knowledge regarding urban areas to enhance the decision-making process for services and applications?

Related work

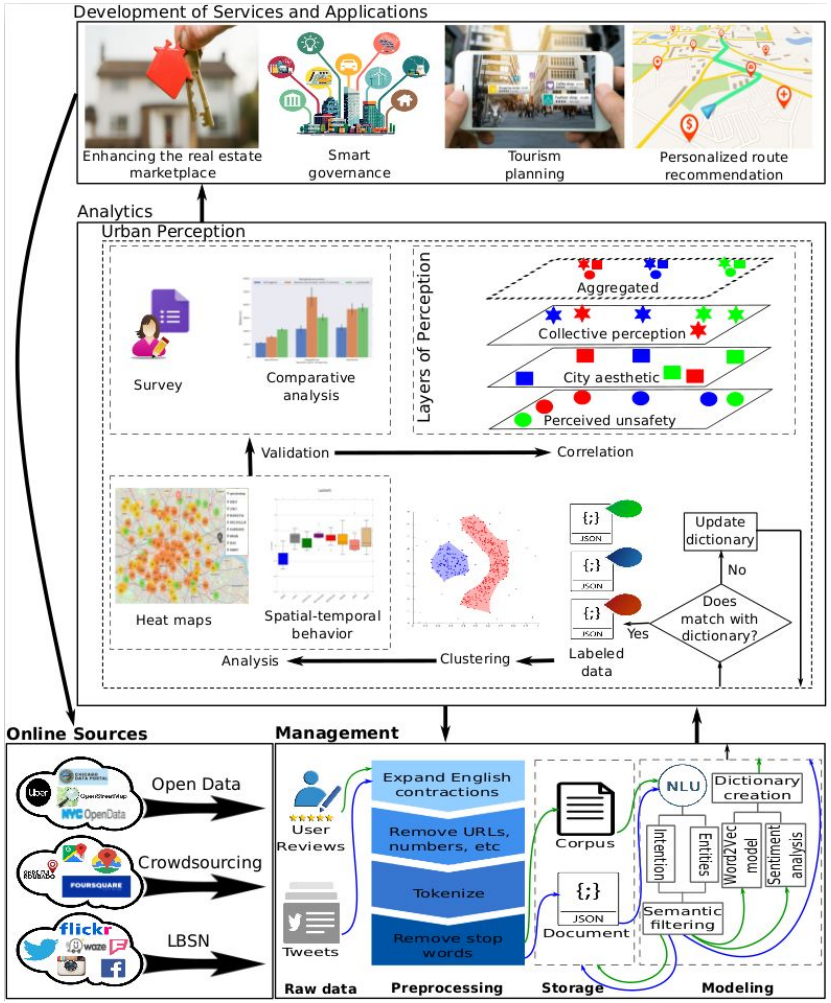
Authors	Perception	Source(s)	Automatic?	Scalable?	Generic?
Quercia et al. 2015	Smell	Sensory walks (texts) and LBSN (tags, hashtags and texts)	No	No	No
Hsu et al. 2019	Smell	Crowdsourcing (texts)	No	Yes*	No
Aiello et al. 2016	Sound	Public project and repository (texts) and LBSN (texts)	No	Yes	No
Quercia et al. 2014	Visual	Crowdsourcing (votes)	No	Yes*	No
Dubey et al. 2016	Visual	Crowdsourcing (votes)	No	Yes*	No
Leng et al. 2019	Taste	Crowdsourcing (texts) and LBSN (texts)	No	Yes	No
Jang and Kim 2019	*Identity*	LBSN (hashtags)	No	Yes	No
Redi et al. 2018	*Spirit*	LBSN (photos and tags)	No	Yes	No
Our proposal	***Collective***	**Crowdsourcing (texts) and LBSN (texts)**	**Yes**	**Yes**	**Yes**

How can we extract urban perceptions from free texts?

Urban Perception Framework

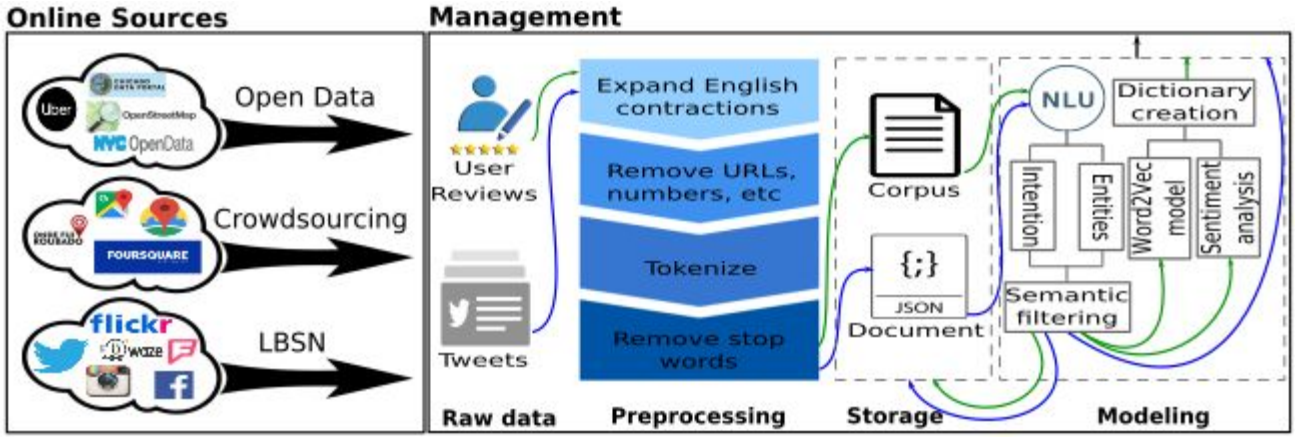


A urban computing framework with LBSN data proposed by Silva et al., 2018.



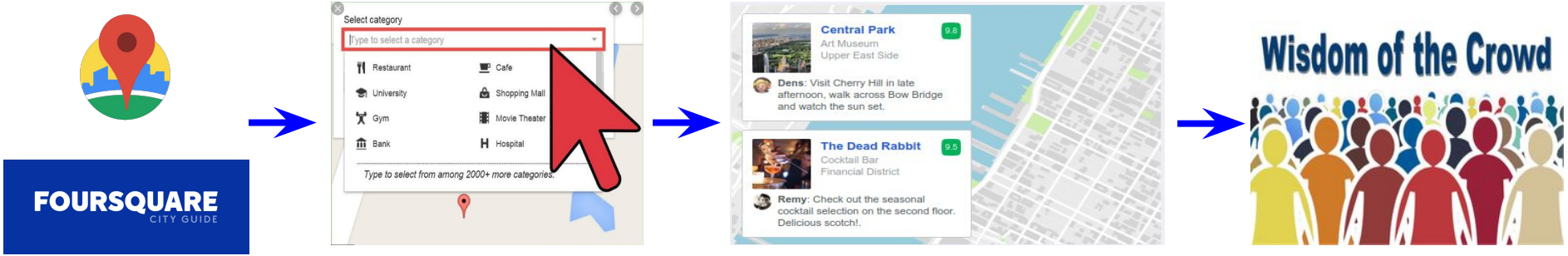
T. H. Silva, A. C. Viana, F. Benevenuto, L. Villas, J. Salles, A. Loureiro, and D. Quercia, "Urban computing leveraging location-based social network data: a survey", ACM Computing Surveys (CSUR), vol. 52, no. 1, p. 17, 2019.

Management Layer



Raw Data: Tweets

Raw Data: User Reviews



Preprocessing/Storage

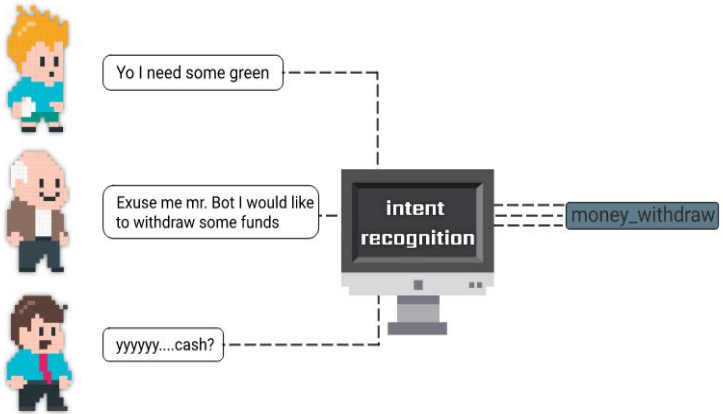

```
{
        "quote_count": 0,
        "contributors": null,
        "truncated": false,
        "text": "Want to work in #Chicago, IL? View our latest opening:
https://t.co/s3PH44ldKj #Education #Job #Jobs #Hiring #CareerArc",
        "is_quote_status": false,
        "in_reply_to_status_id": null,
        "reply_count": 0,
        "id": 923998983780941824,
        "favorite_count": 0,
        "source": "<a href=\"http://www.tweetmyjobs.com\"
rel=\"nofollow\">TweetMyJOBS</a>",
        "retweeted": false,
        "coordinates": {
                "type": "Point",
                "coordinates": [
                        -87.6297982,
                        41.8781136
                ]
        },
        "timestamp_ms": "1509133499461",
        "entities": {**},
        "in_reply_to_screen_name": null,
        "id_str": "923998983780941824",
        "retweet_count": 0,
        "in_reply_to_user_id": null,
        "favorited": false,
        "user": {**},
        "geo": {**},
        "in_reply_to_user_id_str": null,
        "possibly_sensitive": false,
        "lang": "en",
        "created_at": "Fri Oct 27 19:44:59 +0000 2017",
        "filter_level": "low",
        "in_reply_to_status_id_str": null,
        "place": {**}
}
```



```
{
   "t923998983780941824": {
      "data": {
         "text": "Want to work in #Chicago, IL?
View our latest opening:
https://t.co/s3PH44ldKj #Education #Job
#Jobs #Hiring #CareerArc",
         "timestamp": "1509133499.461",
         "geolocation": {
            "latitude": "41.8781136",
            "longitude": "-87.6297982"
         },
         "source": "twitter",
         "location": "chicago",
         "language": "en_US"
      }
   }
}
```

Modeling: Natural Language Understanding (NLU)

- Intent Recognition



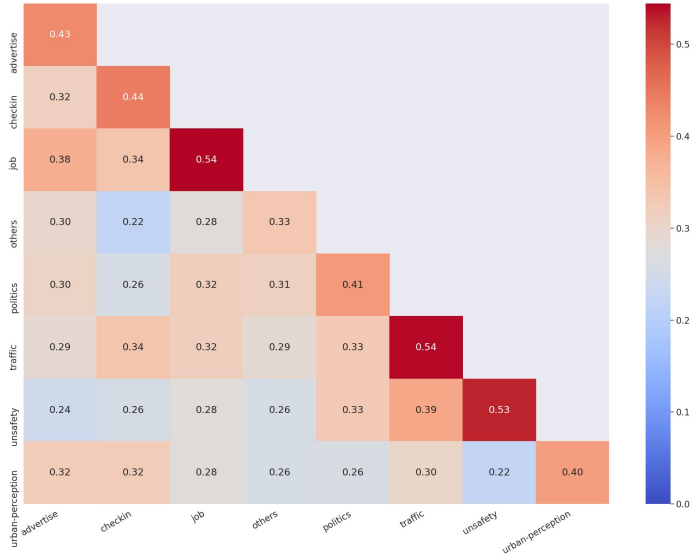
- Named Entity Recognition (NER)

When Sebastian Thrun PERSON started at Google ORG in 2007 DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.

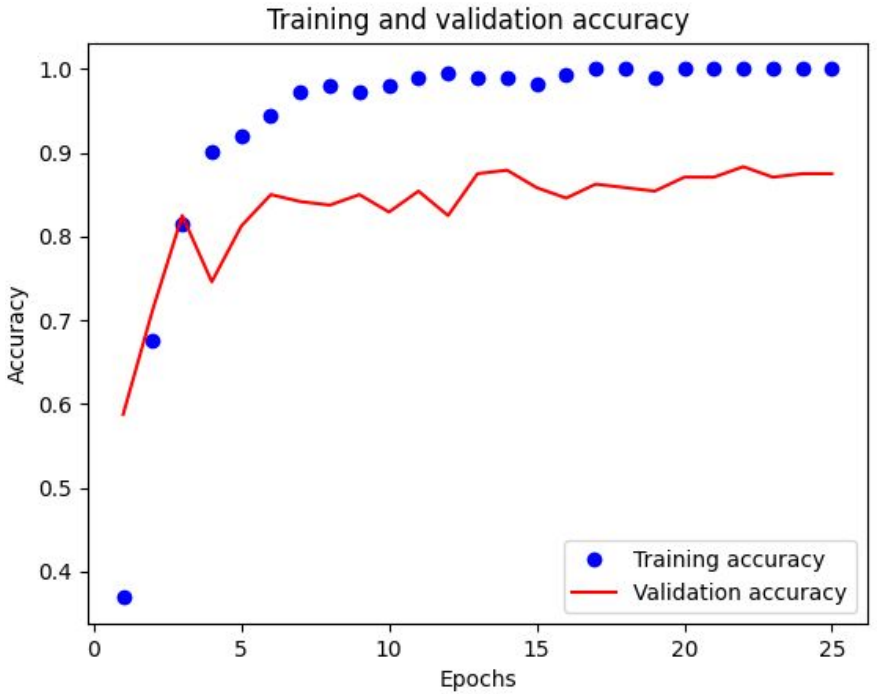
A little less than a decade later DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

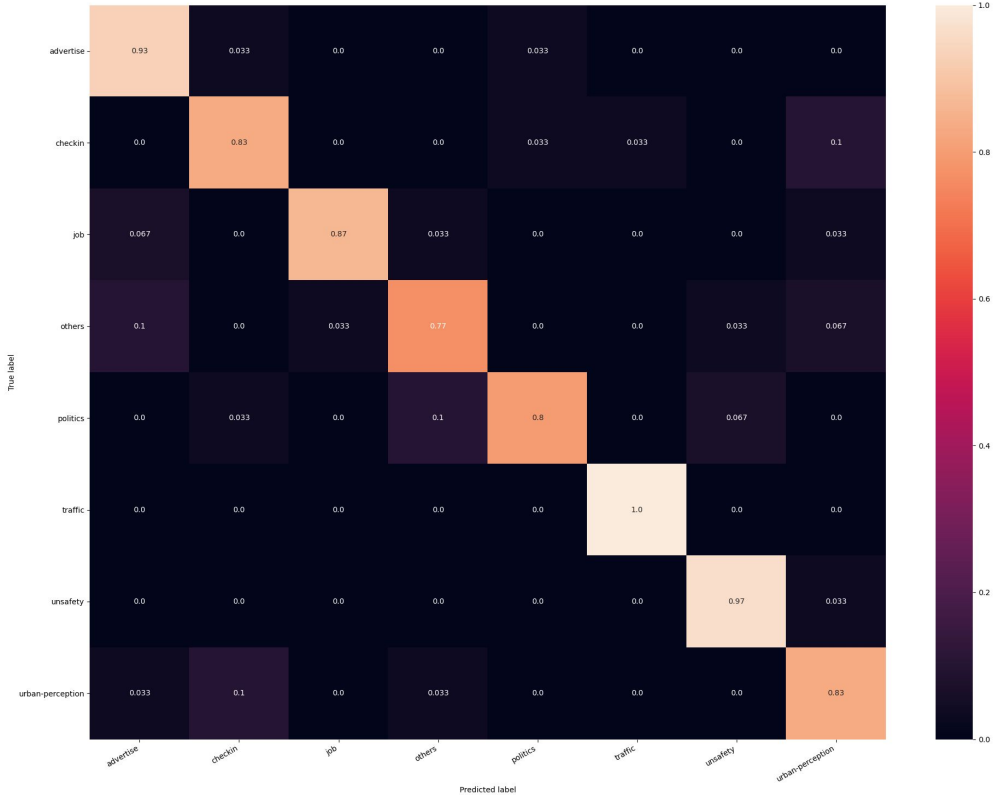
NLU: Intent Recognition

- Define common topics in social media data
  - Advertise, check-in, job, others, politics, traffic, unsafety, urban perception
- Manually label a data set to train a supervised model
  - 1,200 samples; 150 for each intent; 120 for training and 30 for testing
- Create a RNN model, training and evaluate it
  - BiLSTM + LSTM
  - Max sequence length: 100; Vocabulary size: 30,522 (BERT)
  - Batch size: 32; Number of hidden layers: 300; Epochs: 25



NLU: Intent Recognition

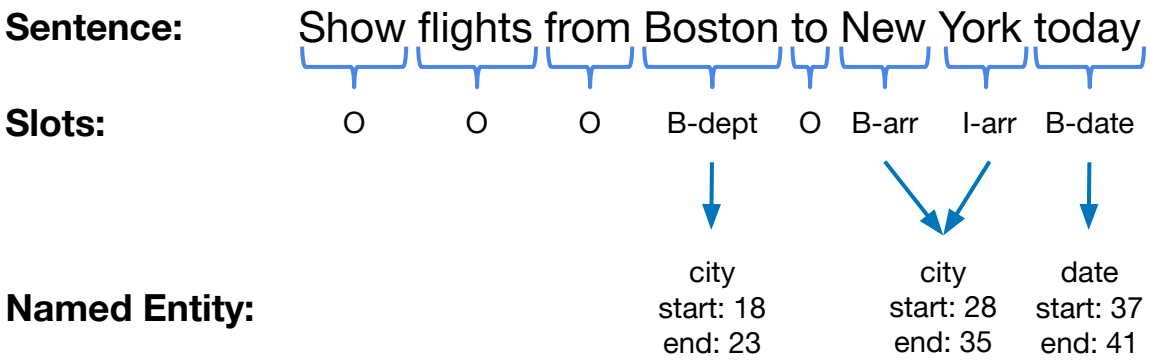




NLU: Intent Recognition

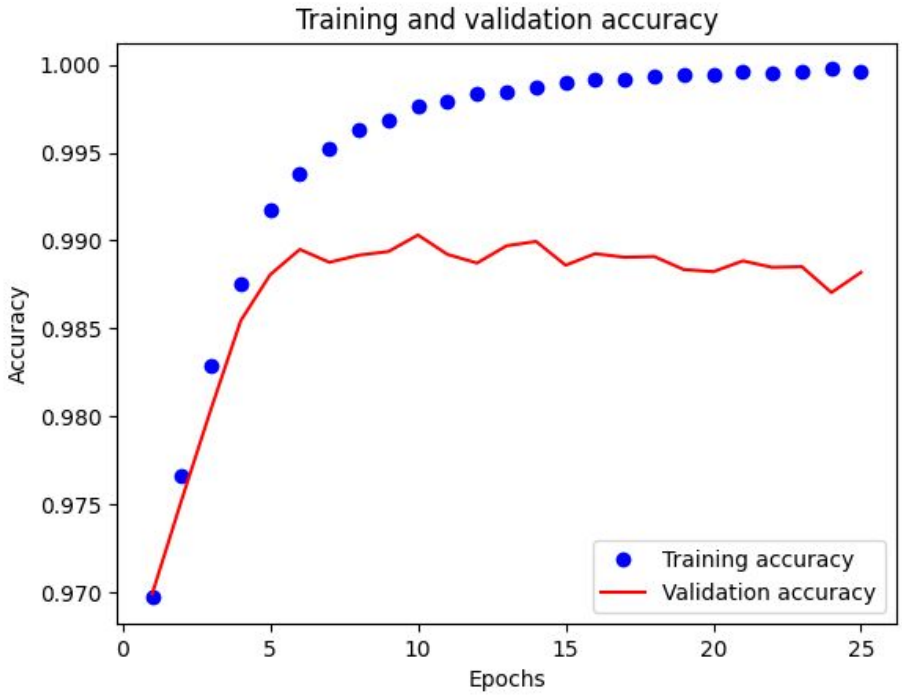
Intents	Precision	Recall	F1-score
advertise	0.82	0.93	0.87
checkin	0.83	0.83	0.83
job	0.96	0.86	0.91
others	0.82	0.76	0.79
politics	0.92	0.80	0.85
traffic	0.96	1.0	0.98
unsafety	0.90	0.96	0.93
urban-perception	0.78	0.83	0.80

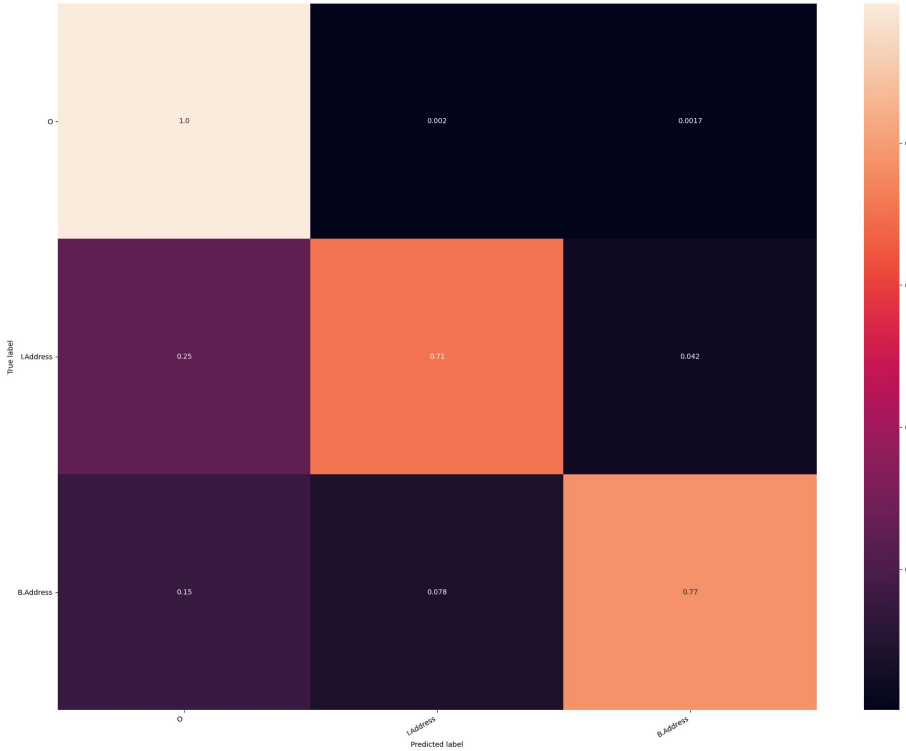
NLU: NER



- Manually label a data set to train a supervised model to predict address entities (POI, City, State, Country, Neighborhood, Avenue/Road/Street)
  - 1,200 samples
- Create a RNN model, training and evaluate it
  - BiLSTM
  - Max sequence length: 100; Vocabulary size: 30,522 (BERT)
  - Batch size: 32; Number of hidden layers: 300; Epochs: 25

NLU: NER


Training and validation accuracy
Accuracy
Epochs
Training accuracy
Validation accuracy
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0
5
10
15
20
25


O
I.Address
B.Address
True label
Predicted label
1.0
0.002
0.0017
0.25
0.71
0.042
0.15
0.078
0.77

NLU: NER

- Several tweets*, mainly those related to traffic and unsafety, are posted by news agencies
  - **Problem 1**: Using geolocation present in tweet metadata drive us to think there are many issues (about traffic/unsafety) in surrounding area of these places.
  - **Problem 2:** Only a small part of collected tweets are geotagged.
- Usually, there is an address (just a part or full address) in the tweet text
  - **Hypothesis**: If there is an address in the text, we can use NER to get it and applying a geocoding algorithm to obtain the most probably geolocation
- Issues
  - Famous NER libraries, like Facebook Duckling[1], do not support address recognition
  - Slot filling model created by us didn't show a good performance
  - Geocoding might to represent a high cost (execution time and monetary)
- Last shot
  - Test spaCy's NER model

[1]https://github.com/facebook/duckling

NLU


```
{
   "t923998983780941824": {
      "data": {
         "text": "Want to work in #Chicago, IL? View our latest opening:
https://t.co/s3PH44ldKj #Education #Job #Jobs #Hiring #CareerArc",
         "timestamp": "1509133499.461",
         "geolocation": {
            "latitude": "41.8781136",
            "longitude": "-87.6297982"
         },
         "source": "twitter",
         "location": "chicago",
         "language": "en_US"
      }
   }
}
```

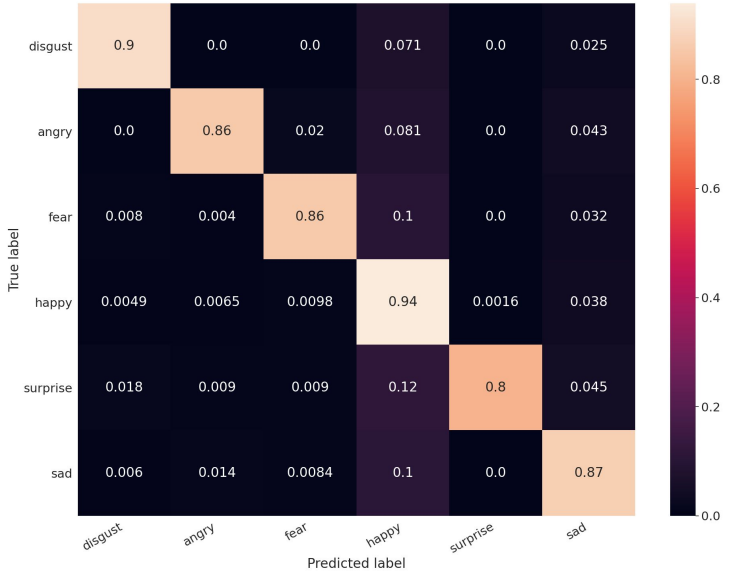


```
{
   "t923998983780941824": {
      "data": {
         "text": "Want to work in #Chicago, IL? View our latest opening:
https://t.co/s3PH44ldKj #Education #Job #Jobs #Hiring #CareerArc",
         "timestamp": "1509133499.461",
         "geolocation": {
            "latitude": "41.8781136",
            "longitude": "-87.6297982"
         },
         "source": "twitter",
         "location": "chicago",
         "language": "en_US"
      }
   }
}
```

Modeling: Emotion Analysis

- Considering the following basic emotions: anger, disgust, fear, happiness, sadness, and surprise; used to create a Twitter Emotion Data Set available on Kaggle[1]

- We trained a supervised model (similar to Intent Recognition model), to predict the emotion of tweets



[1]https://www.kaggle.com/shainy/twitter-emotion-analysis

Modeling: Dictionary Creation

- Create Semantic Graph
  - Part-of-Speech (PoS) Tagging by using spaCy (Penn Treebank Project[1])
  - Get all adjectives: JJ, JJS, JJR
  - Double check with NLTK WordNet
  - Set the Graph vertices, where each vertex represents an adjective (on the lemma form)
  - Set the Graph edges, linking all vertices, where the weight is defined by:

$$\text{score}_{\text{sim}} = \alpha \text{ x } \text{word}_{\text{sim}} + (1 - \alpha) \text{ x } \text{sentiment}_{\text{sim}}$$

considering $\alpha = 0.8$

- Find Communities
  - Define a threshold to eliminate “lightweight” edges as follows:

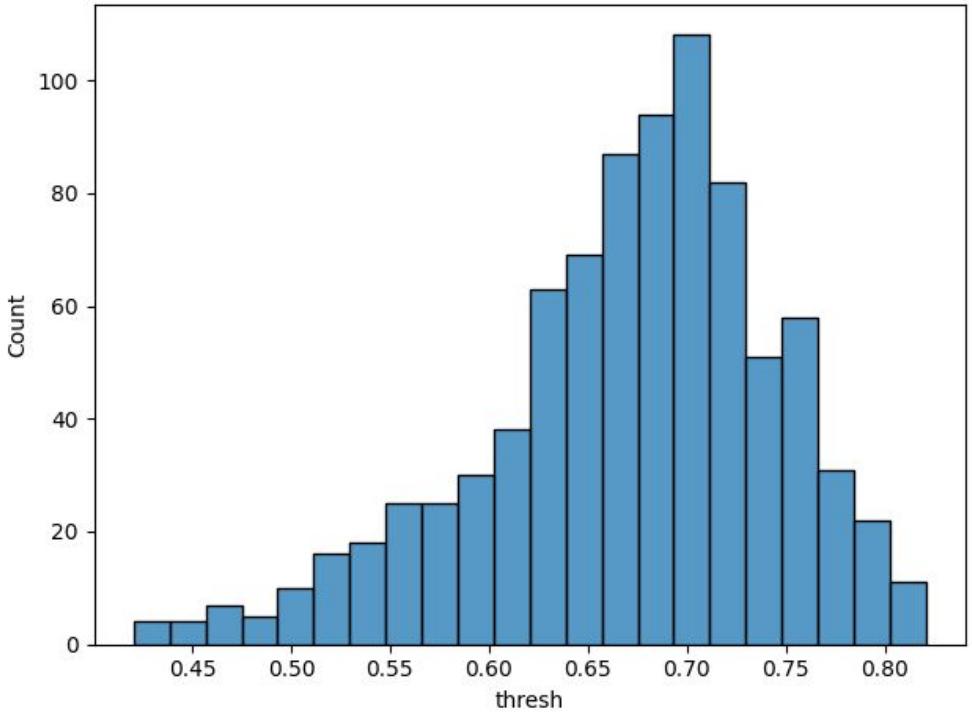
$$\text{thresh}_{u} = \overline{X}_{u} + \beta \text{ x } \sigma_{u}$$

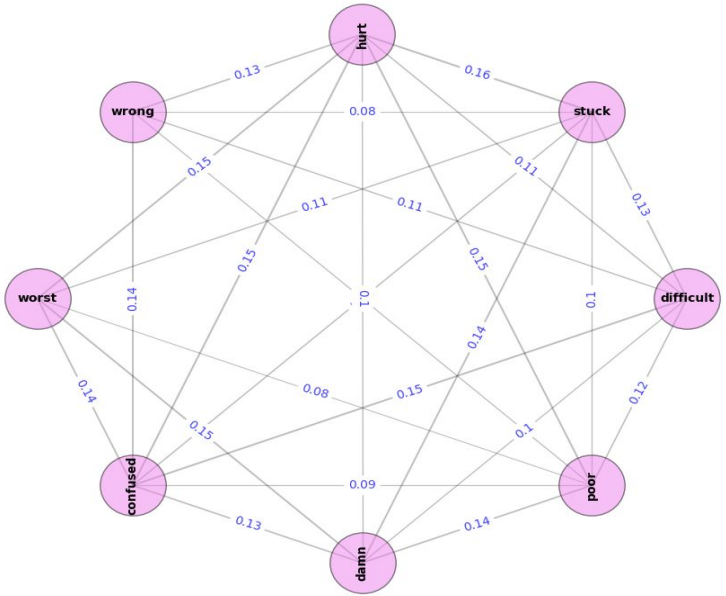
considering $\beta = 2$

  - Find communities by using k-clique community algorithm, k=5

[1]https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Threshold Distribution Analysis





Dictionary: Urban Perception (26 communities)

solar&nuclear: "solar", "cosmic", "subatomic", ...

good&beautiful: "surprised", "embarrassed", "cute", "honest", "frustrating", "painful", ...

simple&honest: "immaculate", "honest", "sane", "intact", "simple"

excited&exciting: "excited", "ambitious", "energetic", "exciting", "eager"

inappropriate&atrocious: "incomprehensible", "inappropriate", "atrocious", "unrealistic", ...

cold&cloudy: "frosty", "cloudy", "cold", "frosted", "chilling", "chilly"

musical&acoustic: "choral", "instrumental", "vocal", "acoustic", "musical", "jazzy", "philharmonic"

big&high: "loud", "big", "mighty", "crowded", "massive", "huge", ...

outdoor&outside: "grassy", "outside", "outdoor", "wooded", "leafy"

patient&medical: "patient", "geriatric", "pediatric", "medical", "clinical"

Dictionary: Traffic (21 communities)

great&amazing: "excellent", "outstanding", "fantastic", "terrific", "fabulous", "brilliant"

bad&wrong: "insane", "inappropriate", "broken", "fatal", "tragic", "unpleasant", "unacceptable", "disappointing", chaotic", ...

big&high: "heavy", "crowded", "loud", "huge", "massive", "intense", ...

crazy&weird: "insane", "strange", "odd", "unconventional", "bizarre", "surprising", "confusing"

normal&peaceful: "relaxed", "calm", "peaceful", "normal", "mild"

weeklong&nightly: "seasonal", "hundredth", "weeklong", "nightly"

third&fourth: "oriental", "legislative", "eastbound", "interstate", "westbound", "crosstown", "fifth", "sixth", "east"

slow&normal: "unchanged", "clueless", "boring", "powerless", "partial", "conventional", "limited", ...

happy&cute: "friendly", "pleasant", "attractive", "generous", "pretty"

strange&surprising: "astonishing", "unique", "mysterious", "unexpected"

Dictionary: Unsafety (28 communities)

desperate&nervous: "desperate", "restless", "nervous", "impatient", "anxious"

dangerous&violent: "offensive", "confrontational", "aggressive", "violent", "dangerous"

simple&honest: "immaculate", "honest", "sane", "intact", "simple"

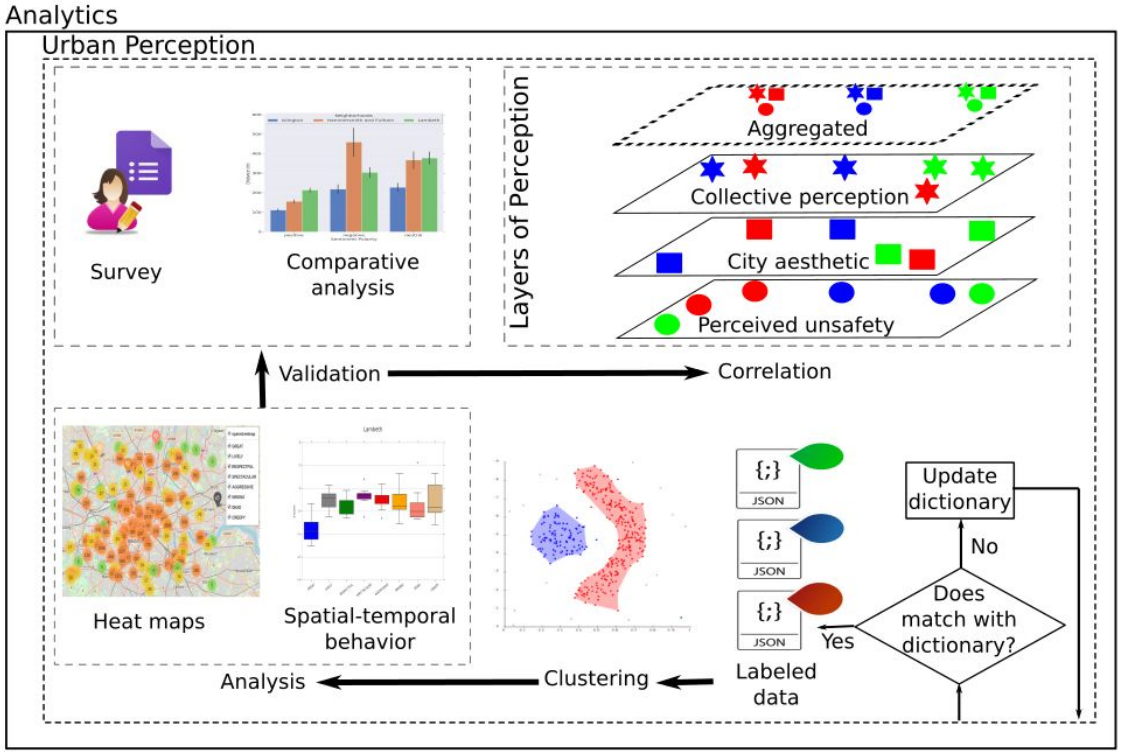
unrealistic&inconvenient: "incomprehensible", "disconcerting", "unrealistic", "atrocious", "inconvenient"

weird&odd: "bizarre", "disconcerting", "weird", "confusing", "odd"

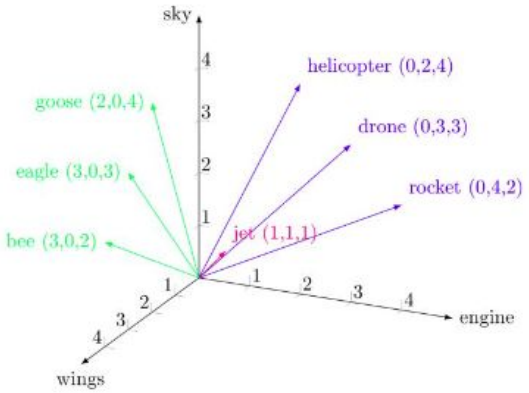
crazy&weird: "mysterious", "bizarre", "strange", "freaky", "insane", ...

bad&stupid: "unethical", "irresponsible", "embarrassed", "dirty", "moronic", "cynical", "horrified", "unjust", "hopeless", "unpleasant", "uncomfortable", "oppressive", "fatal", "disappointing", "pathetic", "dangerous", "tragic", "ill", "nasty", "awful", "abusive", "abused", "controversial", "hostile", "sad", "frustrating", "violent", "disturbed", "exasperated", "grim", "painful", "bizarre", "freaky", "irritable", "adverse", "inconsiderate", "insensitive", "appalling", "incompetent", "hateful", "insane", "evil", "alarming", "bad", "disgusting", "depressing", "destructive", "impatient", "hurt", "annoying", "lunatic", "annoyed", "negative", "rotten", "distraught", "harsh", ...

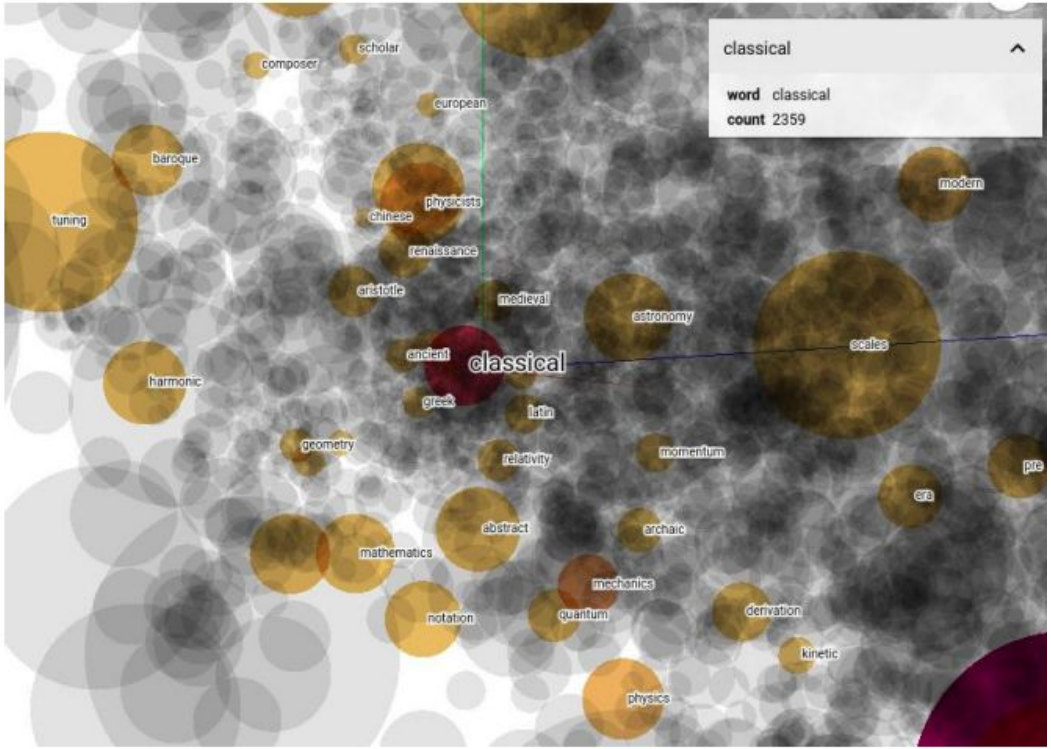
Analytics Layer



Semantic Similarity



Word Embeddings



Words similar to “classical”

Semantic Tagging: NLU + Emotion + Dictionary

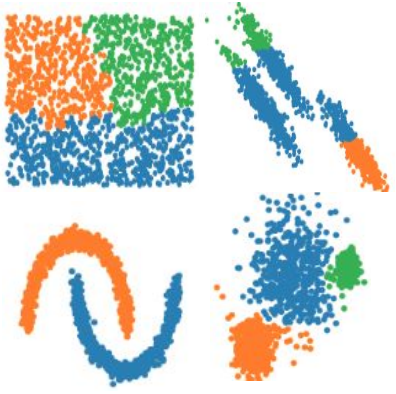

```
"t924018874386997248": {
        "data": {
                "text": "Strange things are afoot in Wicker Park tomorrow night \ud83d\udc7b I've been given the green light to\u2026 https://t.co/ElSS2R8Uau",
                ...
        },
        "nlu": {
                "0": {
                        "sentence": "strange things are afoot in wicker park tomorrow night i ve been given the green light to",
                        "intent": {
                                "value": "urban-perception",
                                "probability": "0.7431086"
                        },
                        "entities": [{
                                "address": "wicker park",
                                "start": 7,
                                "end": 9
                        }]
                }
        },
        "dictionary": {
                "emotion": {
                        "value": "sad",
                        "probability": "0.7396196"
                },
                "semantic_similarity": {
                        "id": "7",
                        "top2_words": "insane&absurd",
                        "adjectives": ["strange afoot green"],
                        "score": "0.6491906046867371"
                }
        }
}
```

Clustering

- We splitted data according to:
  - **Perception category**: Urban perception, unsafety, and traffic
  - **Semantic similarity** with the Dictionary's communities

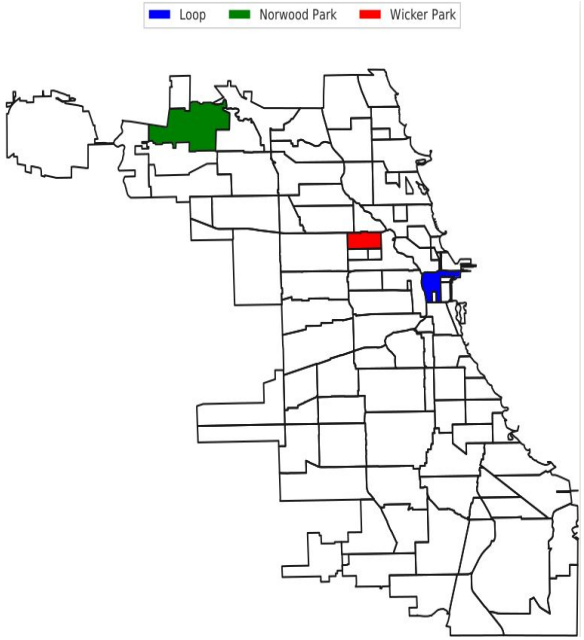
- Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBScan)
  - We should conduct future analyses to determine the suitable clustering algorithm for our framework

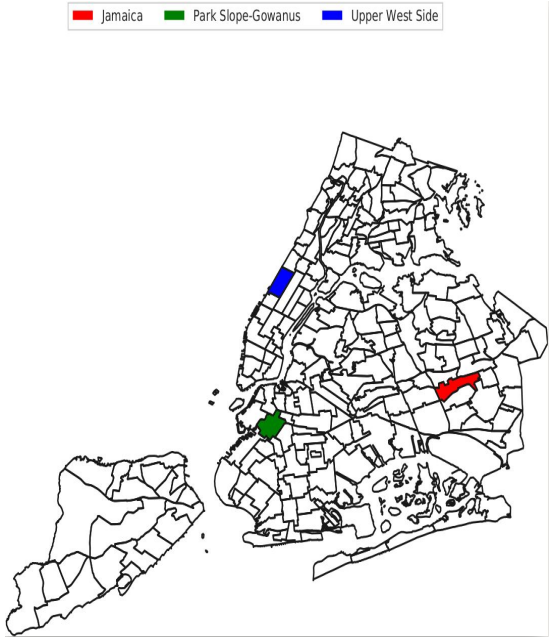
- Why clustering?
  - Reduce the relevance of individual perceptions and highlight the collective ones
  - Mitigate noises
  - Find the most remarkable characteristics of urban spaces

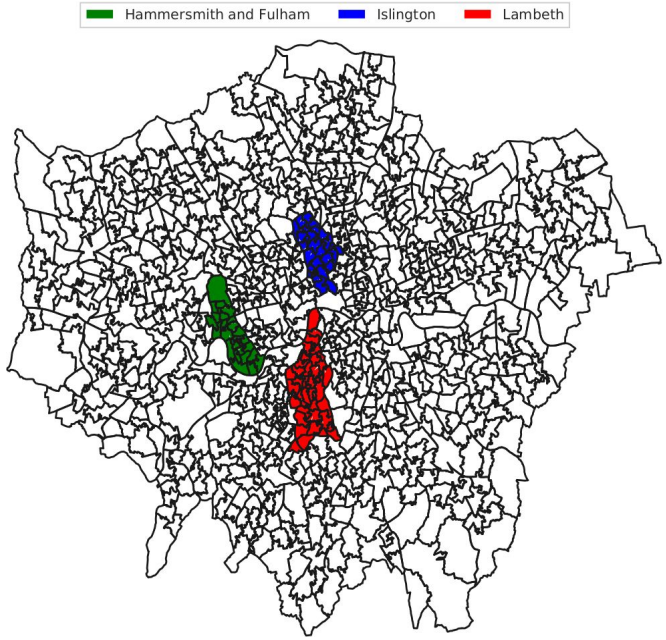


Analysis

Evaluated areas of Chicago, NYC and London:







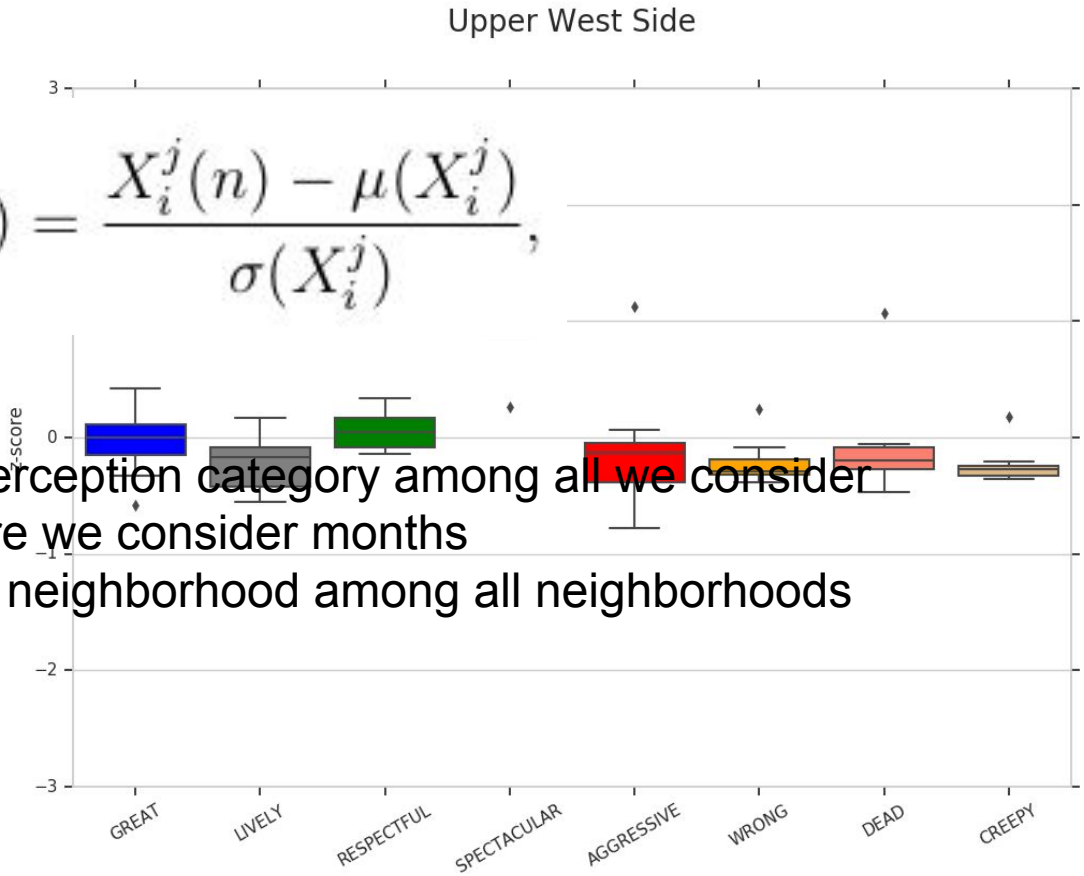
Urban Perception – NYC's neighborhoods

$$\text{z-score}_i^j(n) = \frac{X_i^j(n) - \mu(X_i^j)}{\sigma(X_i^j)},$$

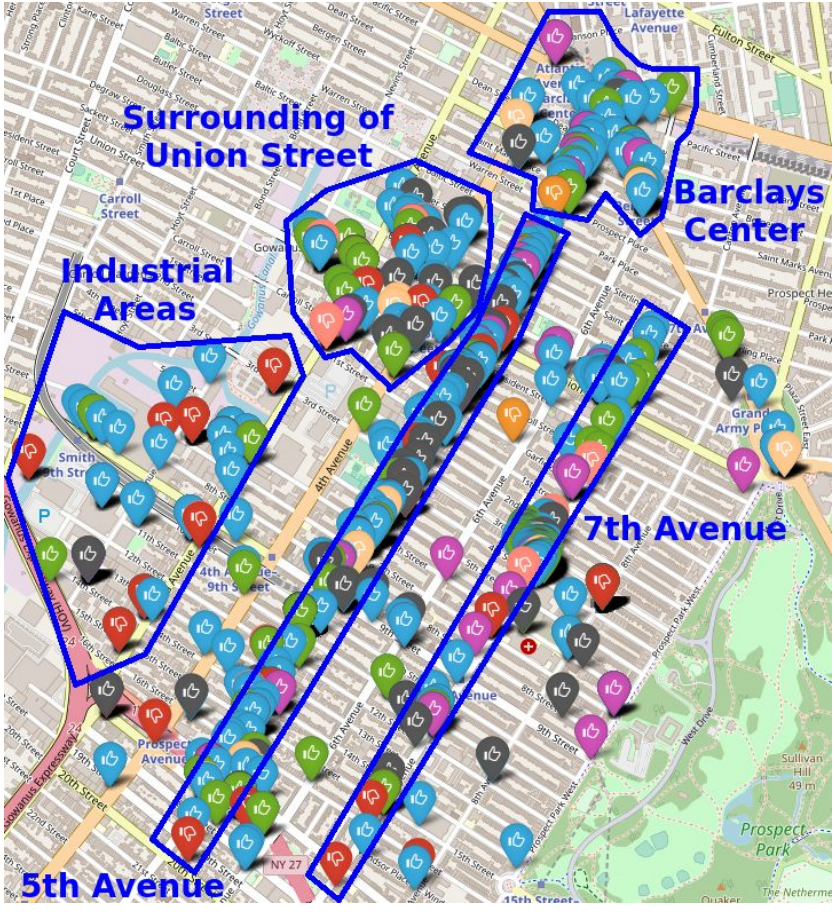
i denotes one specific perception category among all we consider

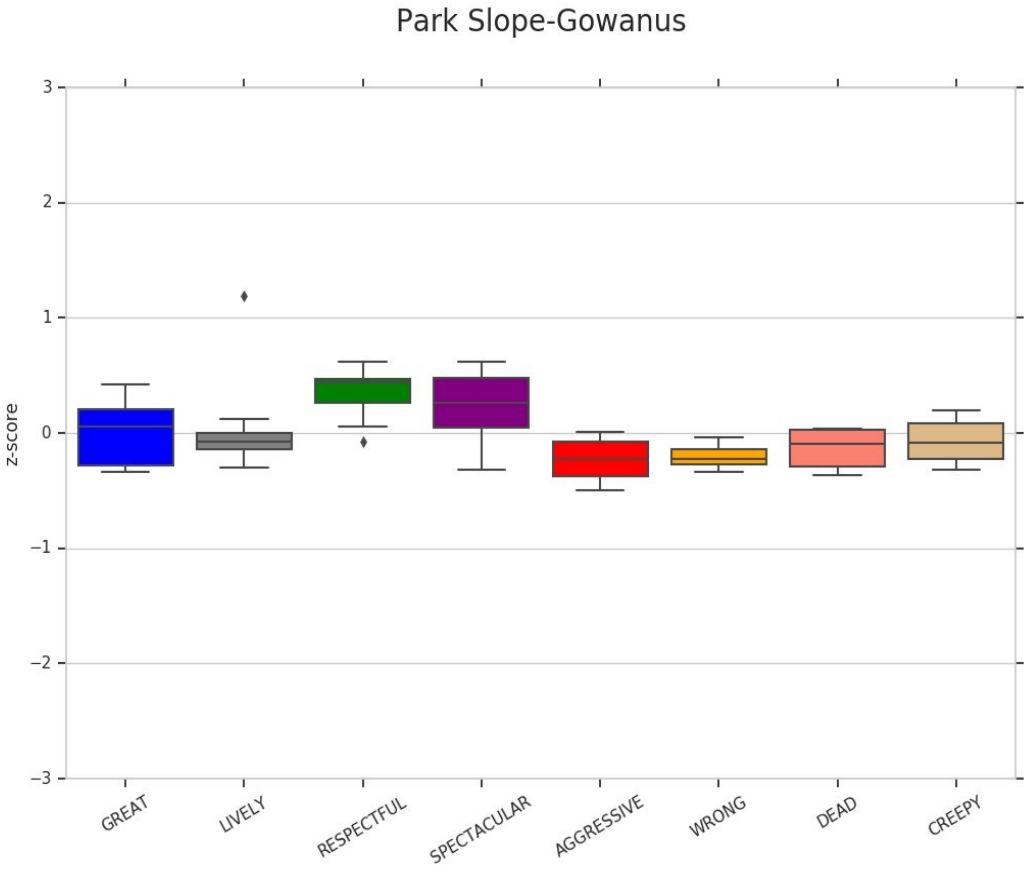
j denotes the period, here we consider months

n denotes a specific city neighborhood among all neighborhoods

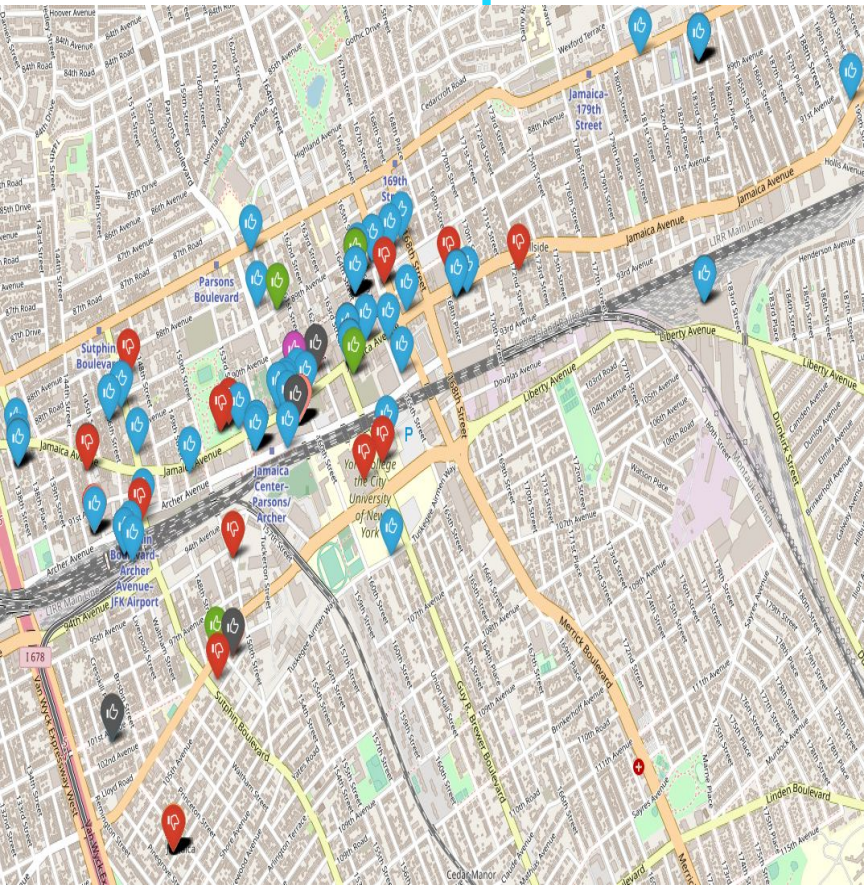


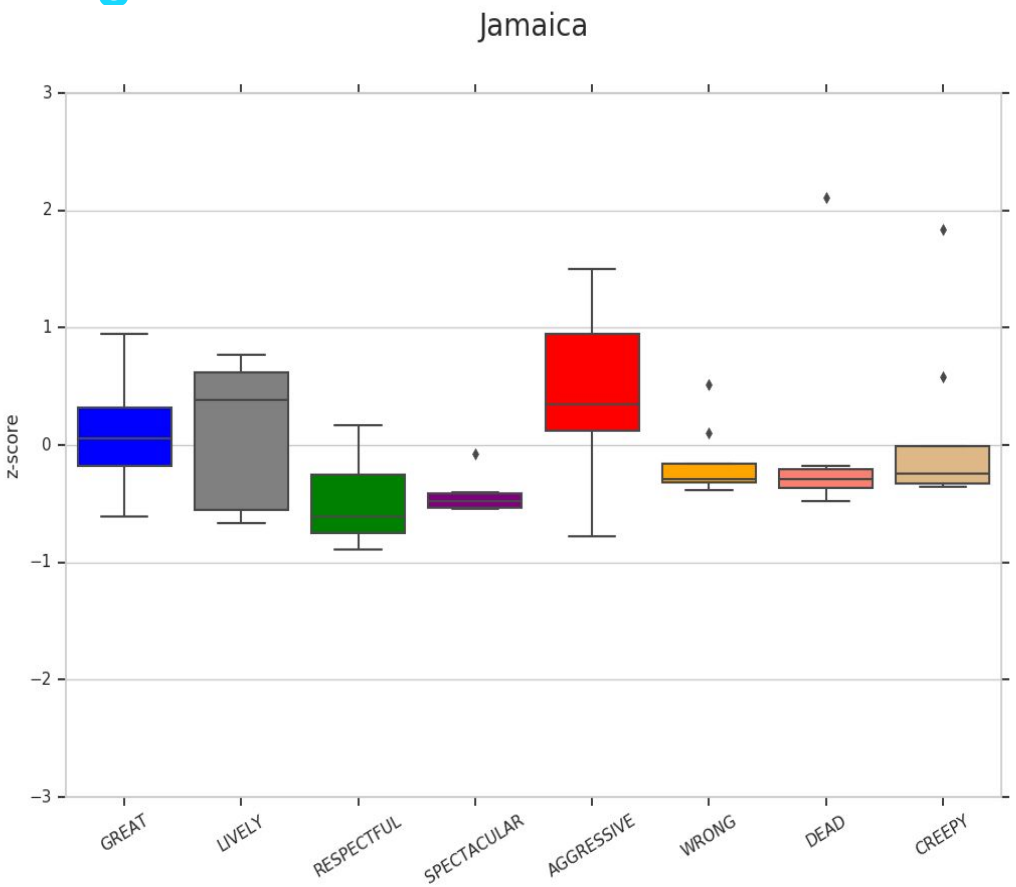
Urban Perception – NYC's neighborhoods





Urban Perception – NYC's neighborhoods





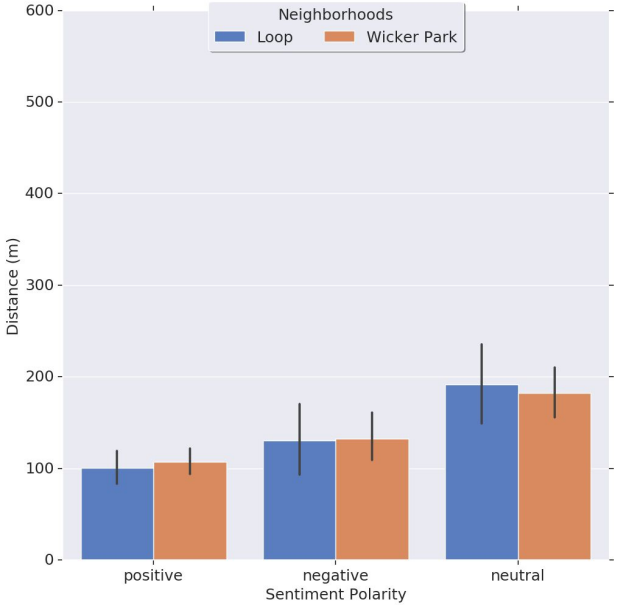
What is the level of agreement among extracted perceptions with respect to a "ground truth"?

Consistency of the Extracted Urban Perception

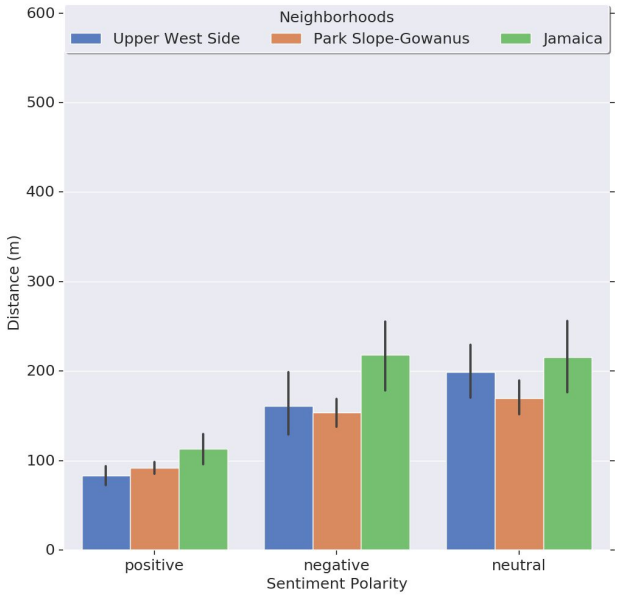
- **Project PlacePulse 2.0 (http://pulse.media.mit.edu/data)**

  - Collects from volunteers their perception based on features present in urban outdoor images of **56 cities**
  - More than **1,2M comparisons** between pair of images
  - Considered perceptions: **safety, wealthy, beautiful, lively, boring and depressing**
  - Since there is no direct mapping between our perception classes and the ones considered by Pulse Place, we aggregate them according to their sentiment polarity

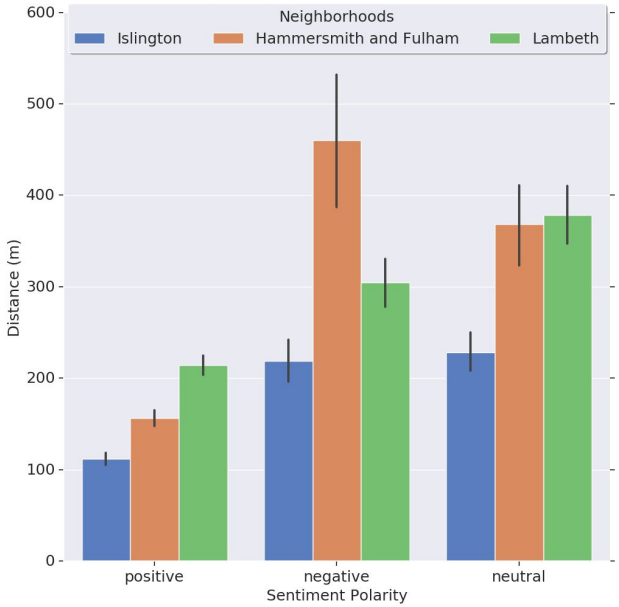
Spatial similarity between perception points from both approaches



Chicago

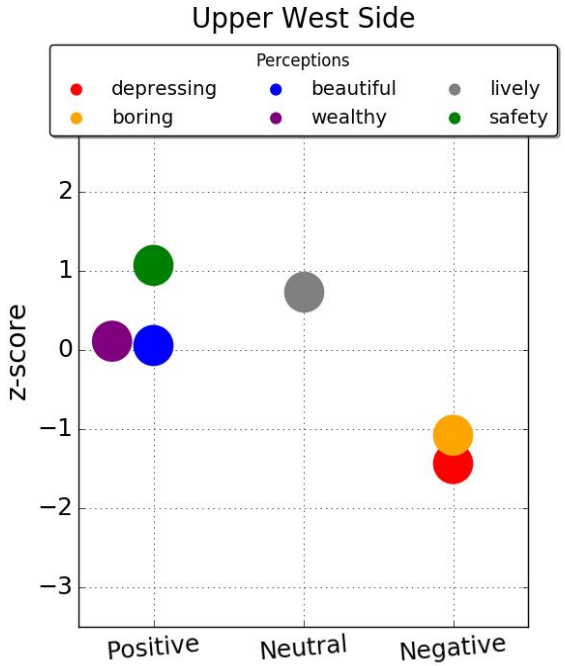


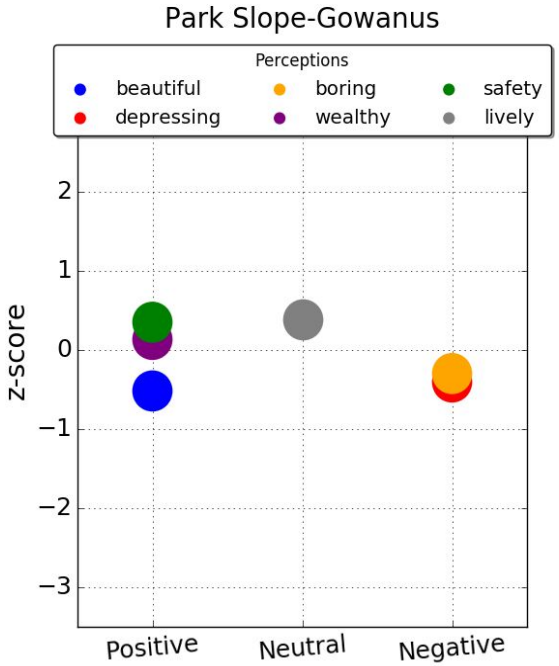
NYC

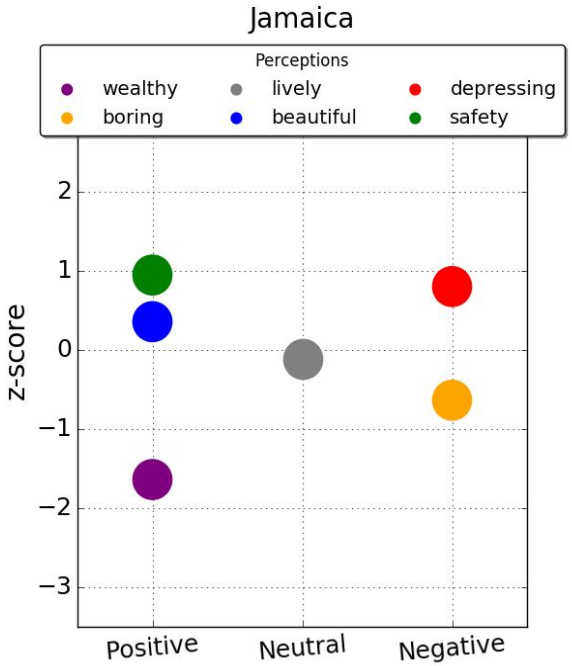


London

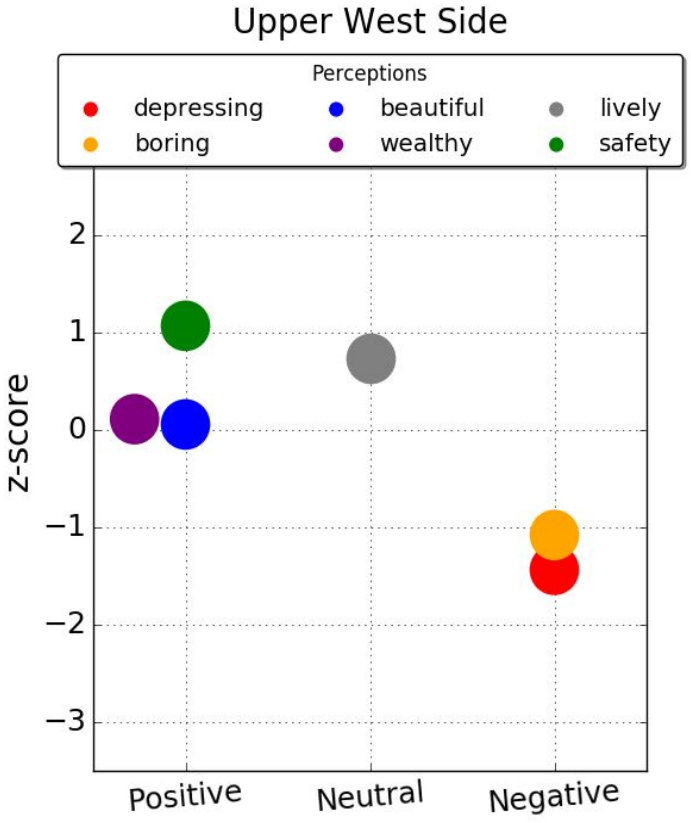
The perception strength of NYC's neighborhoods according to Place Pulse

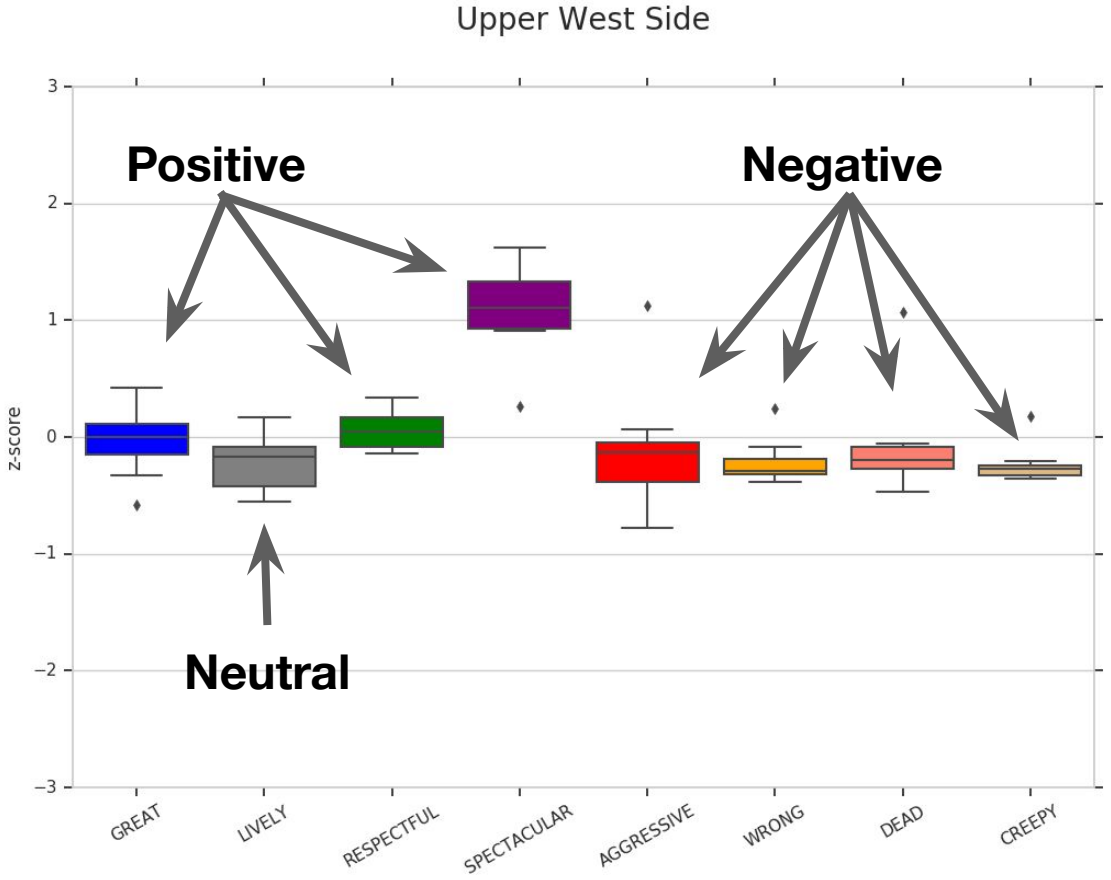




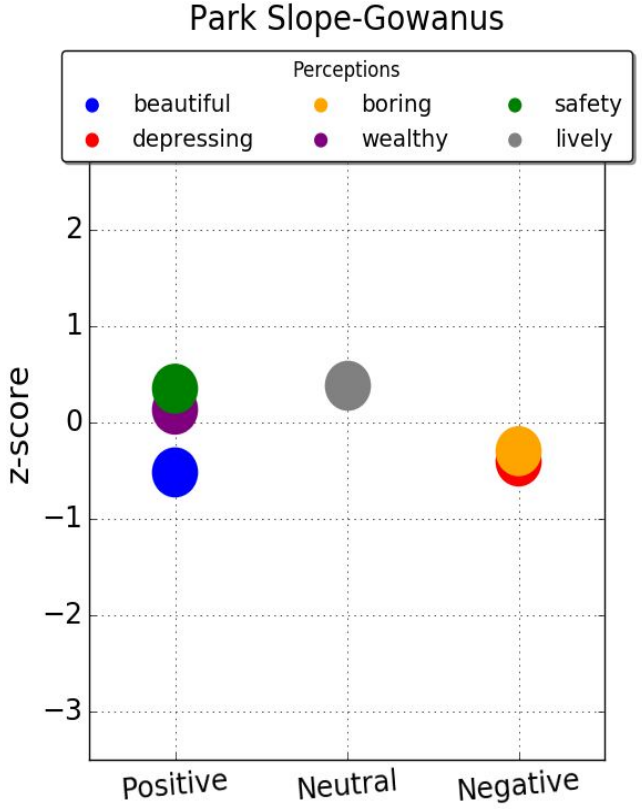


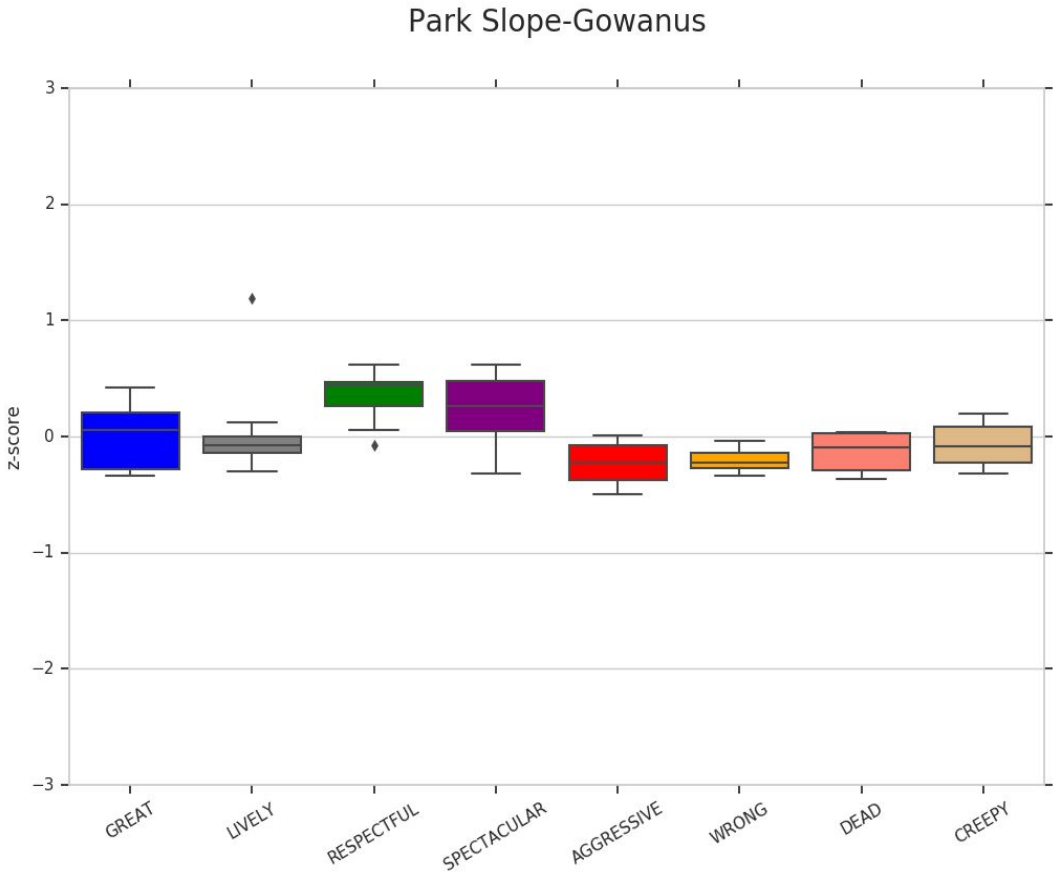
Comparative analysis - NYC



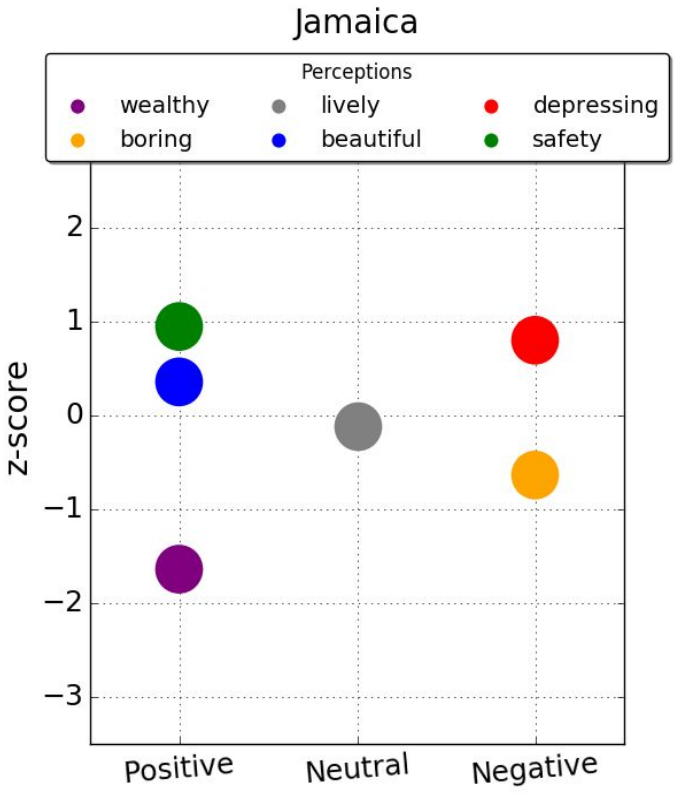


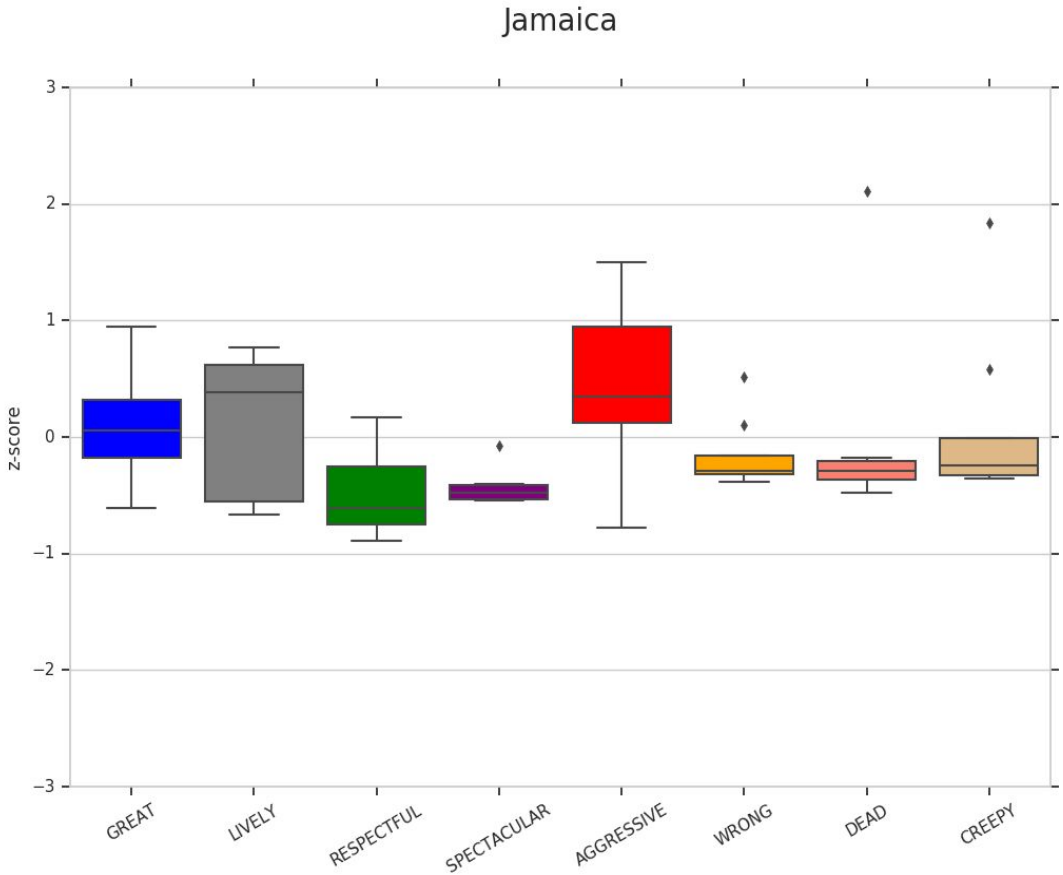
Comparative analysis - NYC





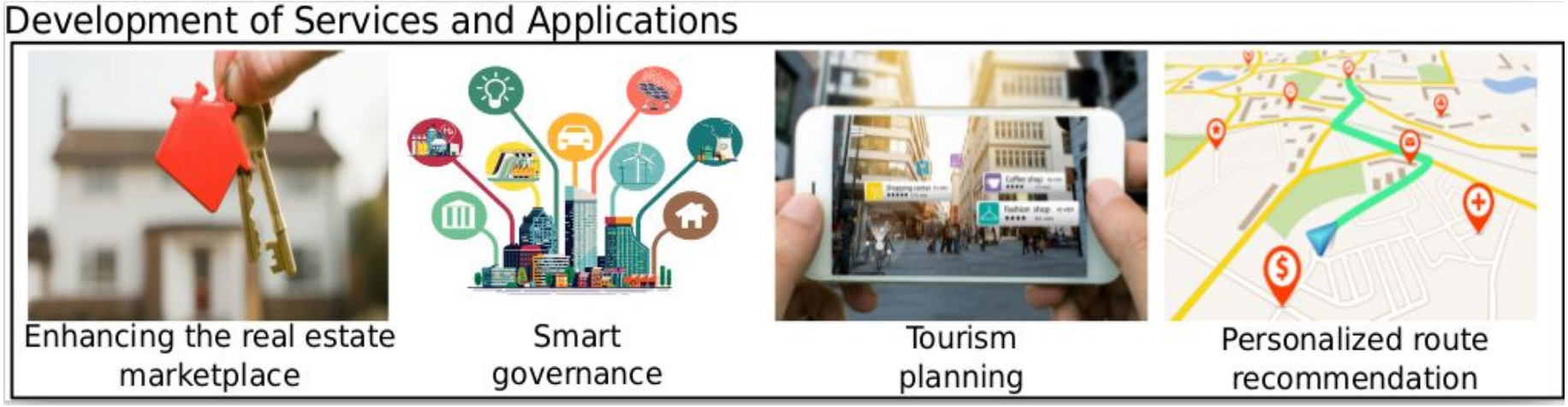
Comparative analysis - NYC





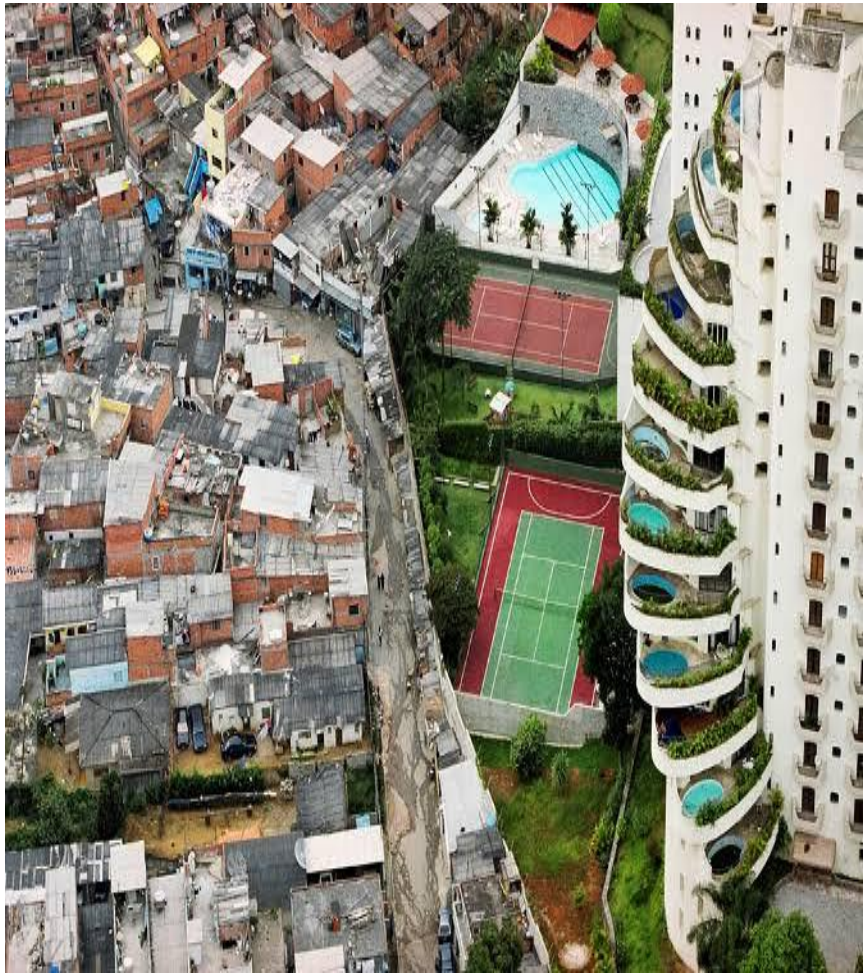
How urban perception can be exploited to leverage new services and applications?

Development of Services and Applications Layer



Enhancing the Real Estate Marketplace

Precisão	★★★★☆	Localização	★★★★½
Comunicação	★★★★½	Check-In	★★★★½
Limpeza	★★★★☆	Valor	★★★★☆



DEMO

Final remarks

Conclusion

Q1

- We propose a framework to support the urban perception extraction from free-texts shared on social media, which is generic, automatic and scalable.
- We demonstrate our framework considering real data to extract the collective urban perception.

Q2

- We conducted a comparative analysis based on Place Pulse 2.0 data, where we could observe that both results yield a very similar level of agreement.

Final remarks

Conclusion

Q3

- Develop a DEMO to support the real estate marketplace.

Final remarks

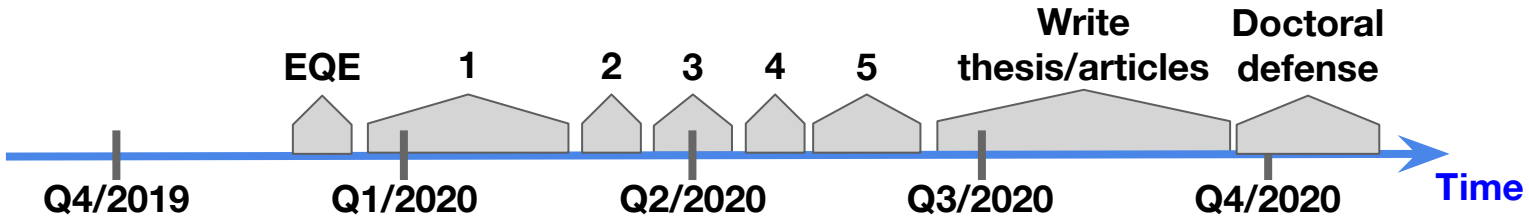
Next steps

Q3

1. Finalize the NLU component and, subsequently, the component to update the dictionary.
2. Incorporate these components into our approach.
3. Extract multiple perceptions, e.g., perceived unsafety and traffic conditions, regarding urban areas.

Q4

4. Develop a personalized route recommendation system that will operate as a service in Intelligent Transportation Systems (ITS) field.
5. Develop an application to support the real estate marketplace.



Final remarks

Publications directly related to proposal

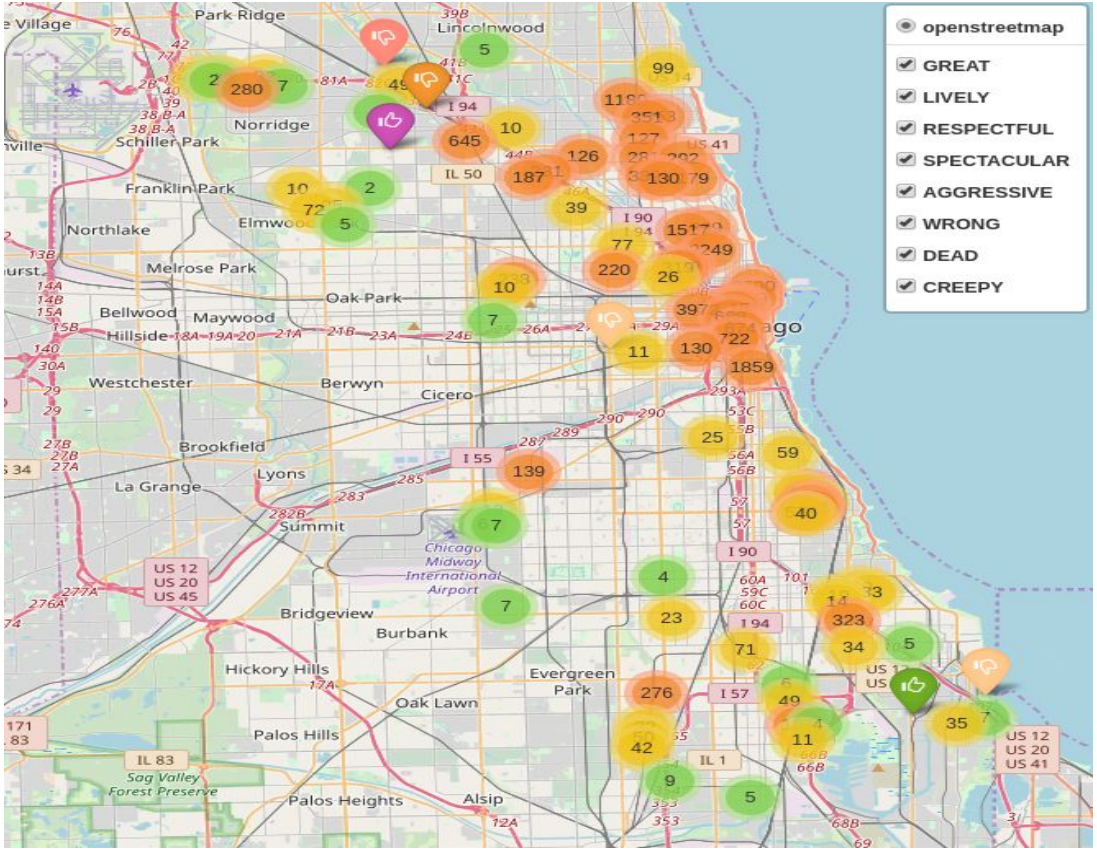
- Springer SNAM 2020: **Automatic extraction of urban outdoor perception from geolocated free texts.**
- SBC Minicursos SBRC 2019: **Computação Urbana da Teoria à Prática: Fundamentos, Aplicações e Desafios.**
- IEEE/WIC/ACM WI 2018: **Uncovering the Perception of Urban Outdoor Areas Expressed in Social Media.**
- SBC SBRC 2018: **Identificação da Reputação de Áreas Urbanas Externas com Dados de Mídias Sociais.**
- IEEE ICC 2018: **Context-aware Vehicle Route Recommendation Platform: Exploring Open and Crowdsourced Data.**
- SBC WGRS 2017 (Best paper award): **Rotas Veiculares Cientes de Contexto: Arcabouço e Análise Usando Dados Oficiais e Sensoriados por Usuários sobre Crimes.**

Final remarks

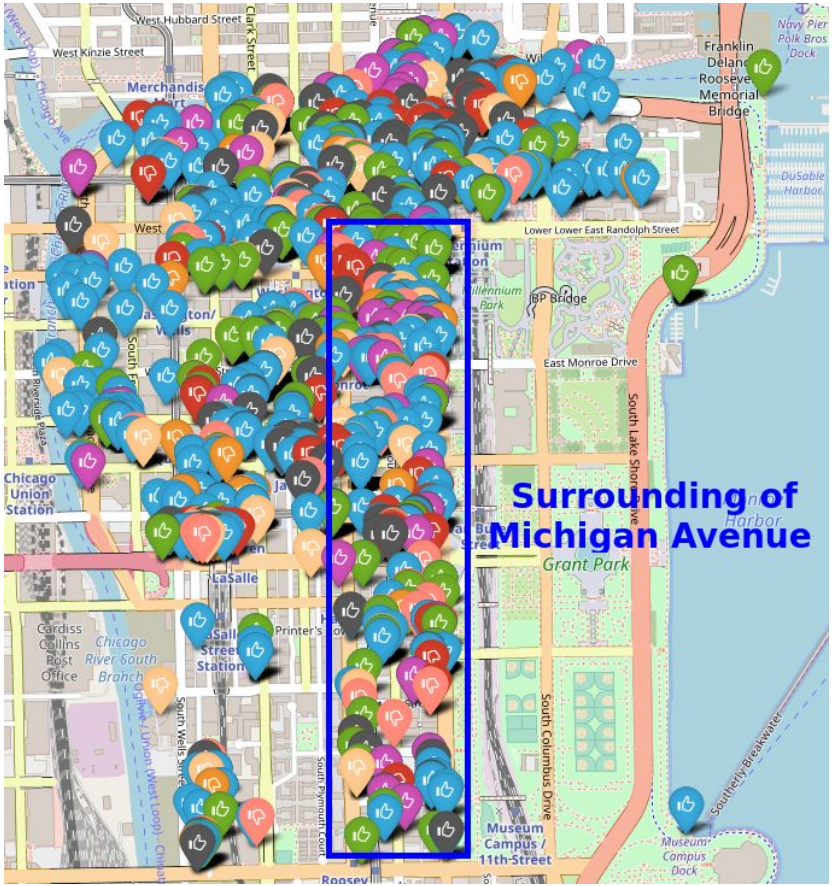
Other publications

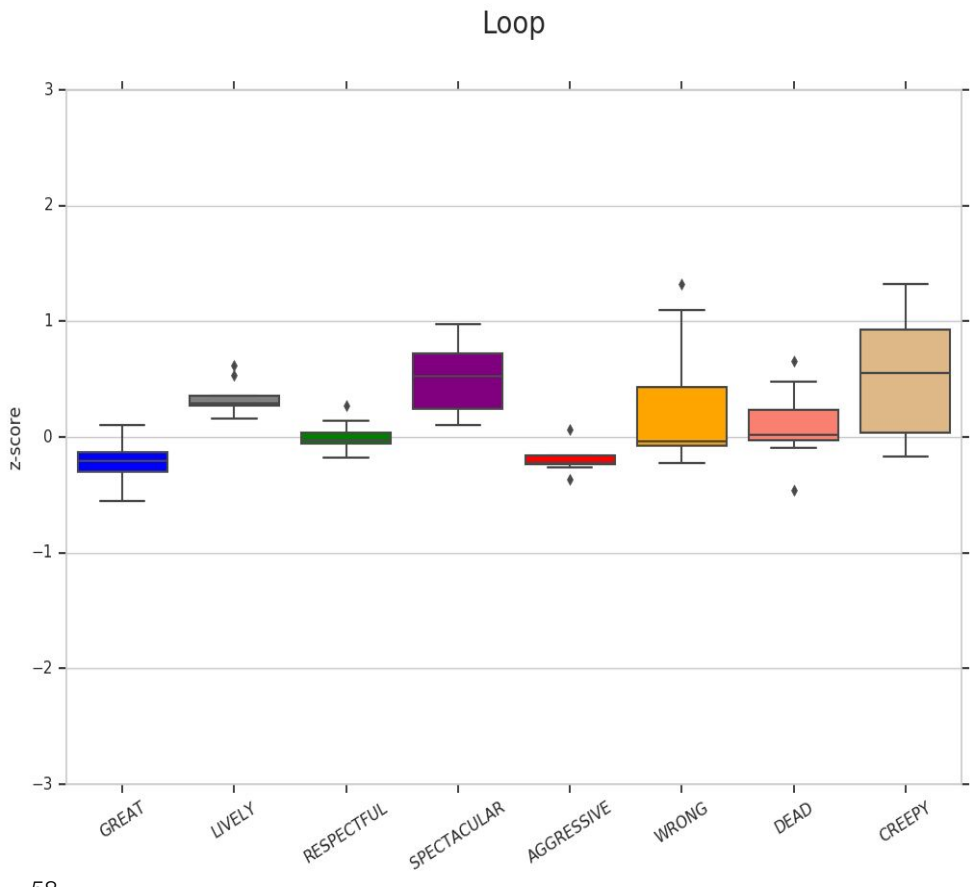
- IEEE ICC 2017: **Towards a sustainable people-centric sensing.**
- IEEE VTC-Fall 2016: **A roadside unit-based localization scheme to improve positioning for vehicular networks.**

Perception Maps - Chicago

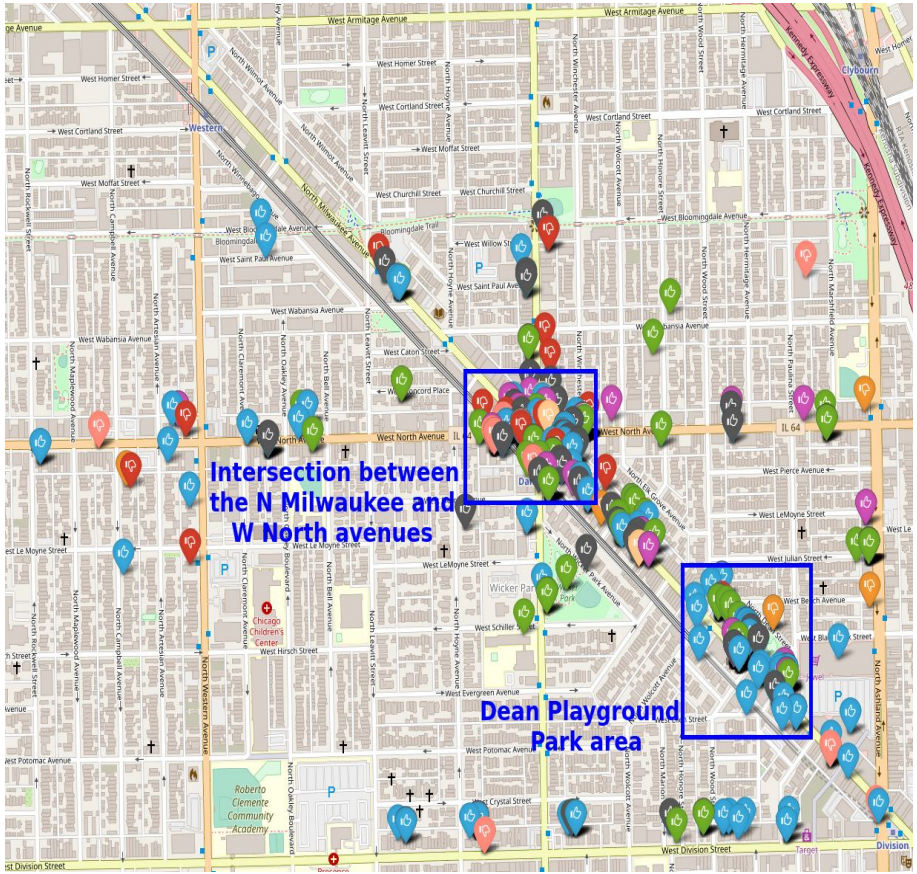


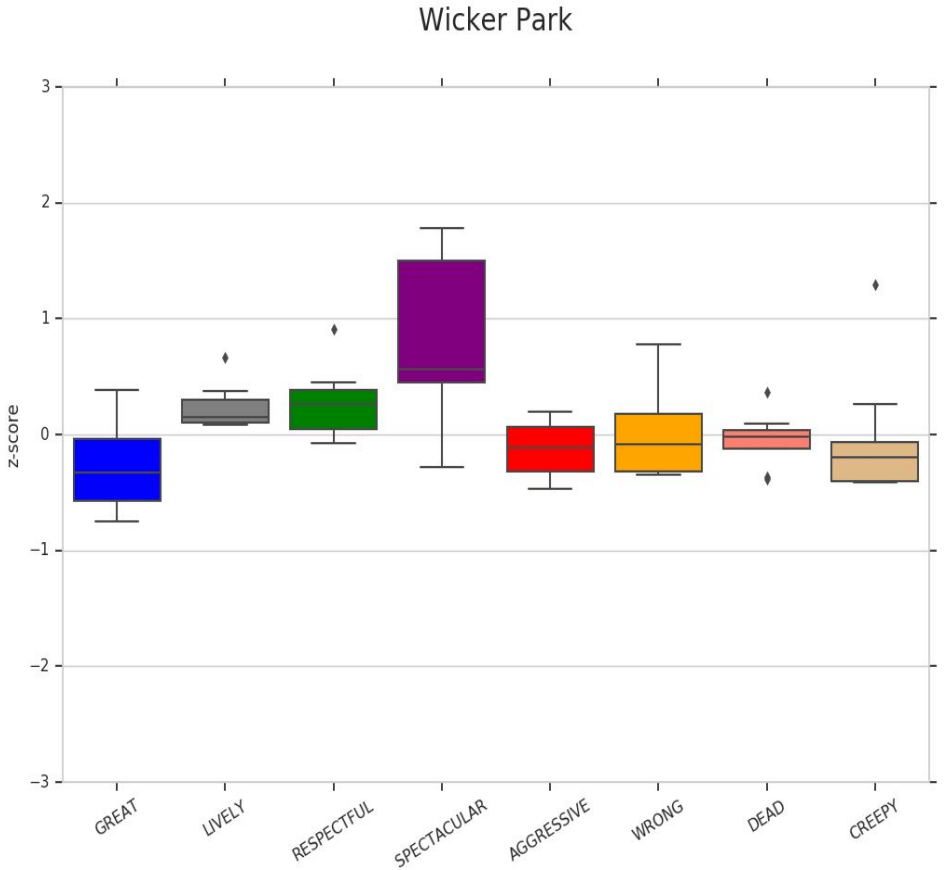
Urban Perception – Chicago's neighborhoods



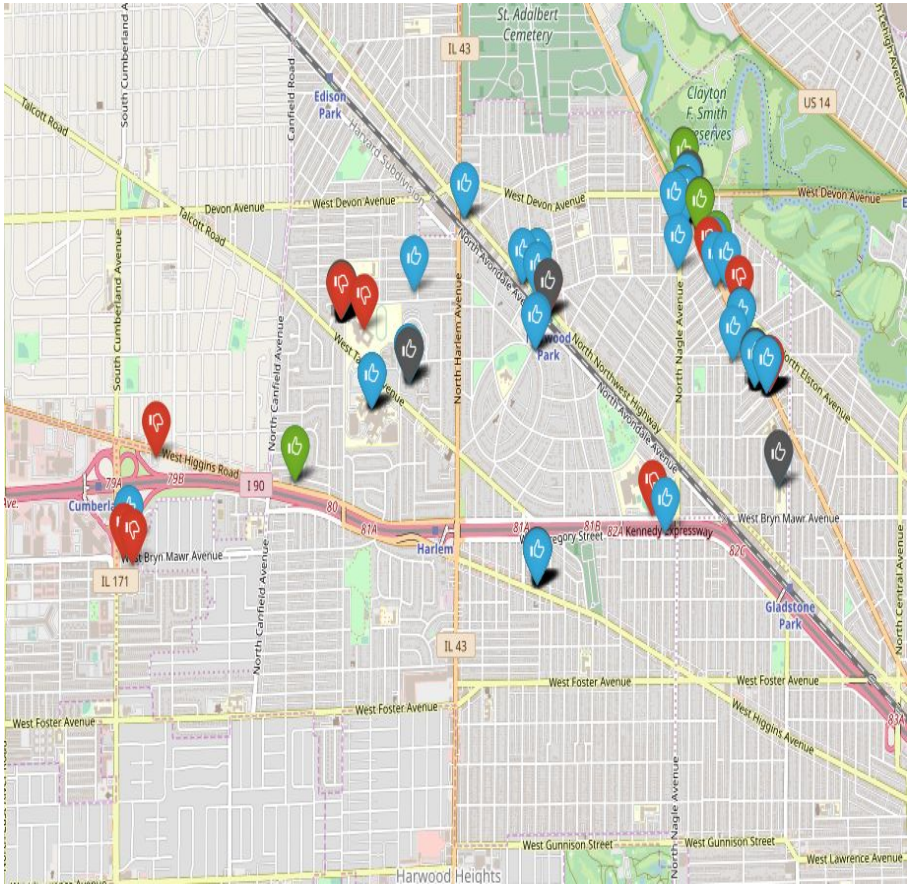


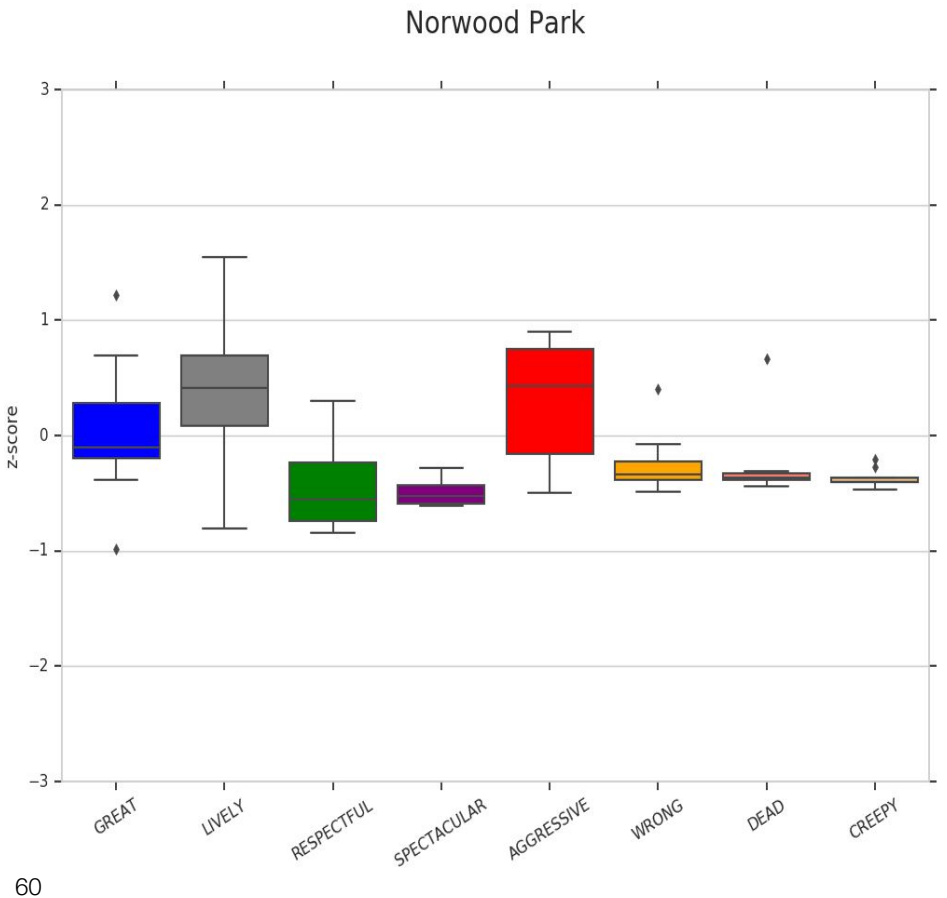
Urban Perception – Chicago's neighborhoods



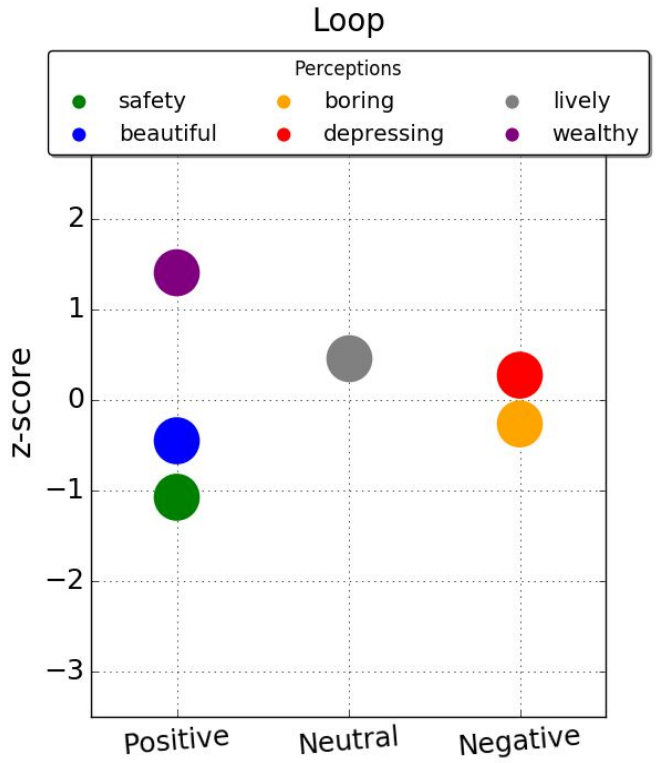


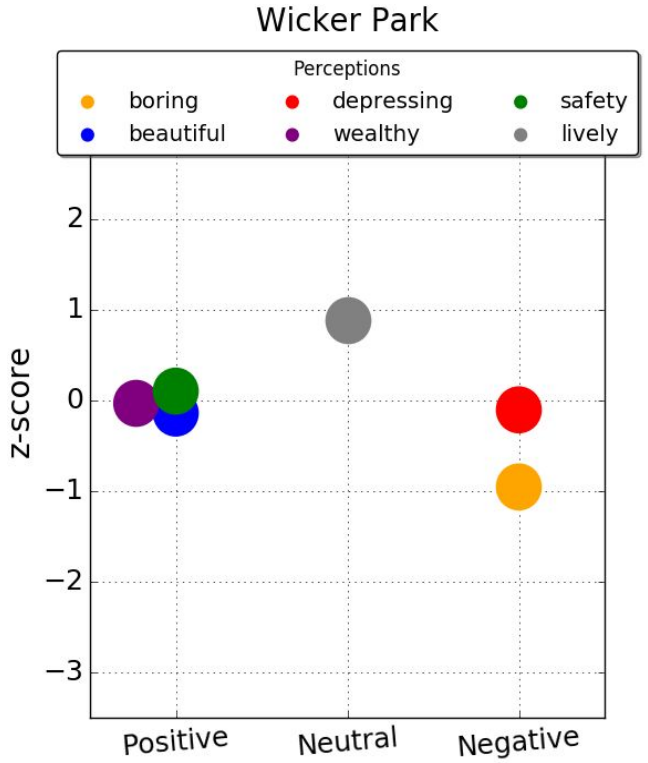
Urban Perception – Chicago's neighborhoods



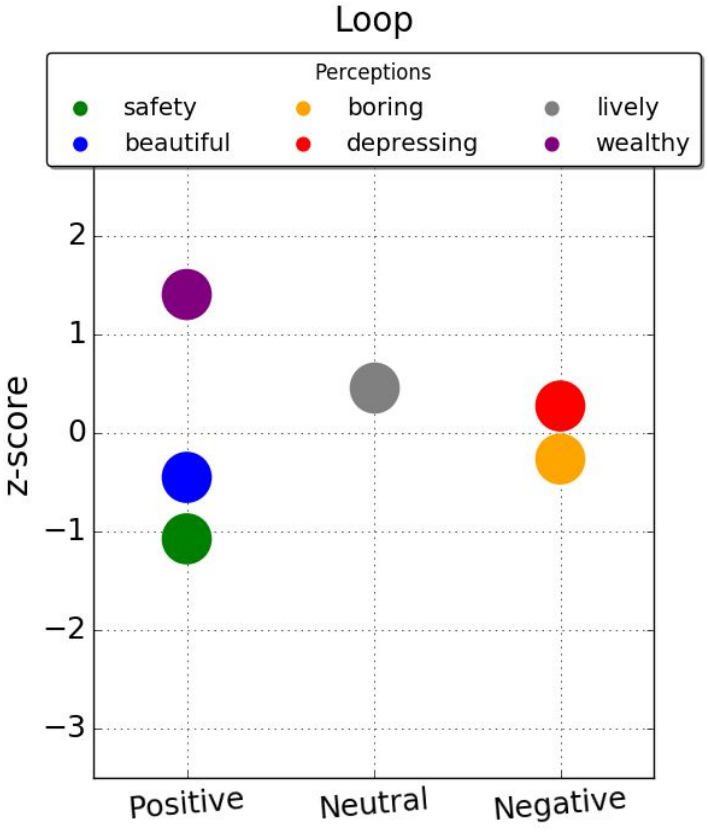


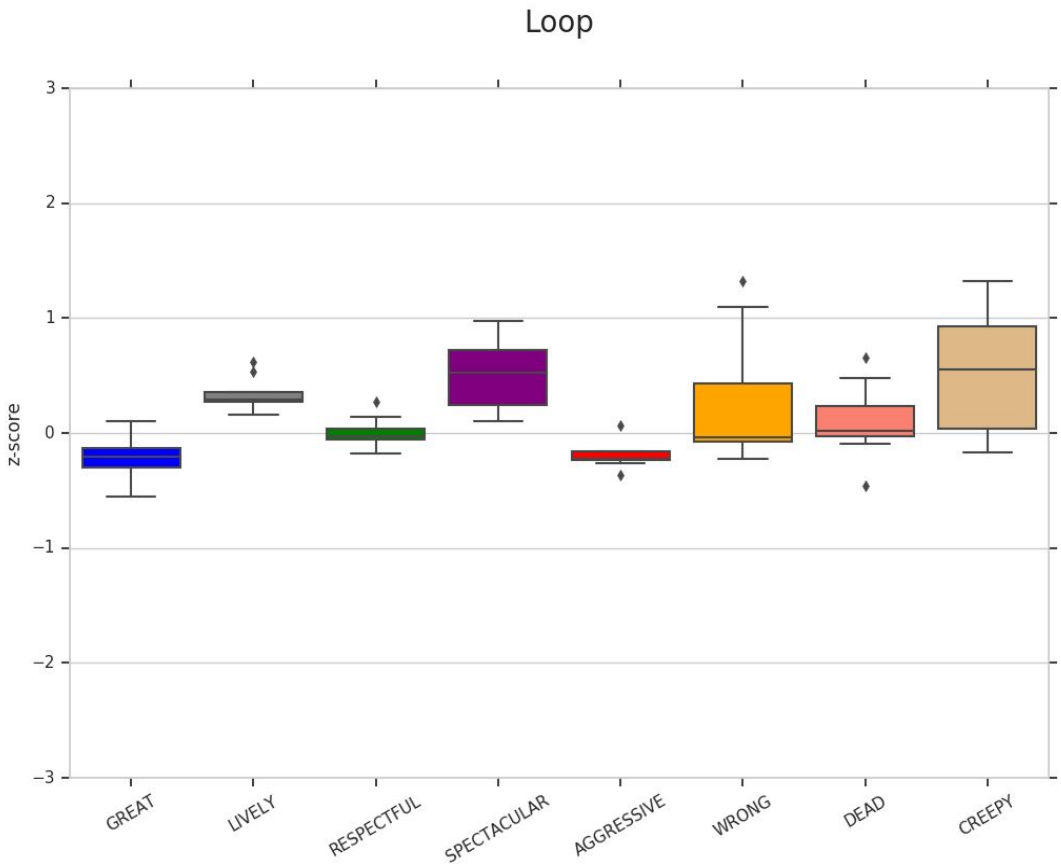
The perception strength of Chicago's neighborhoods according to Place Pulse



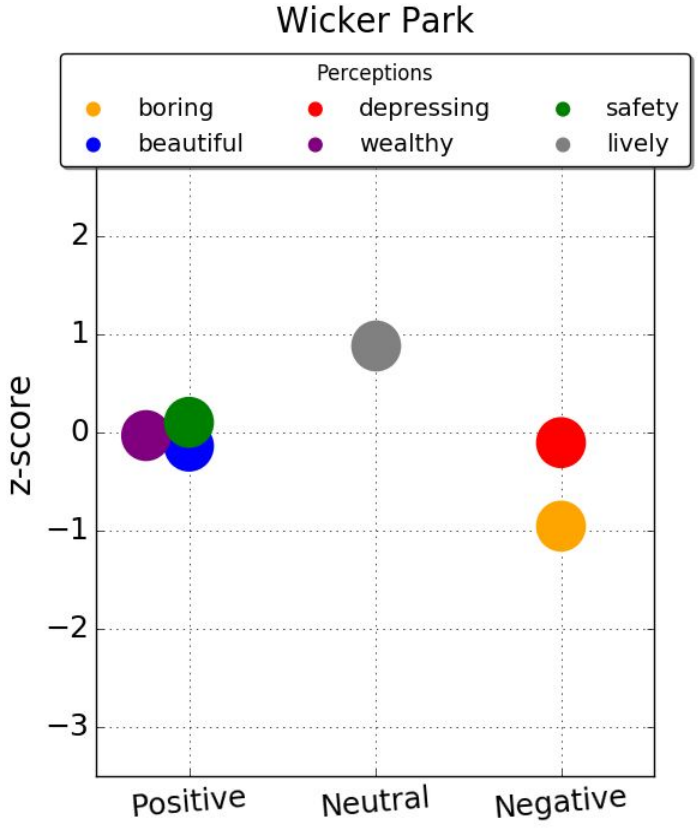


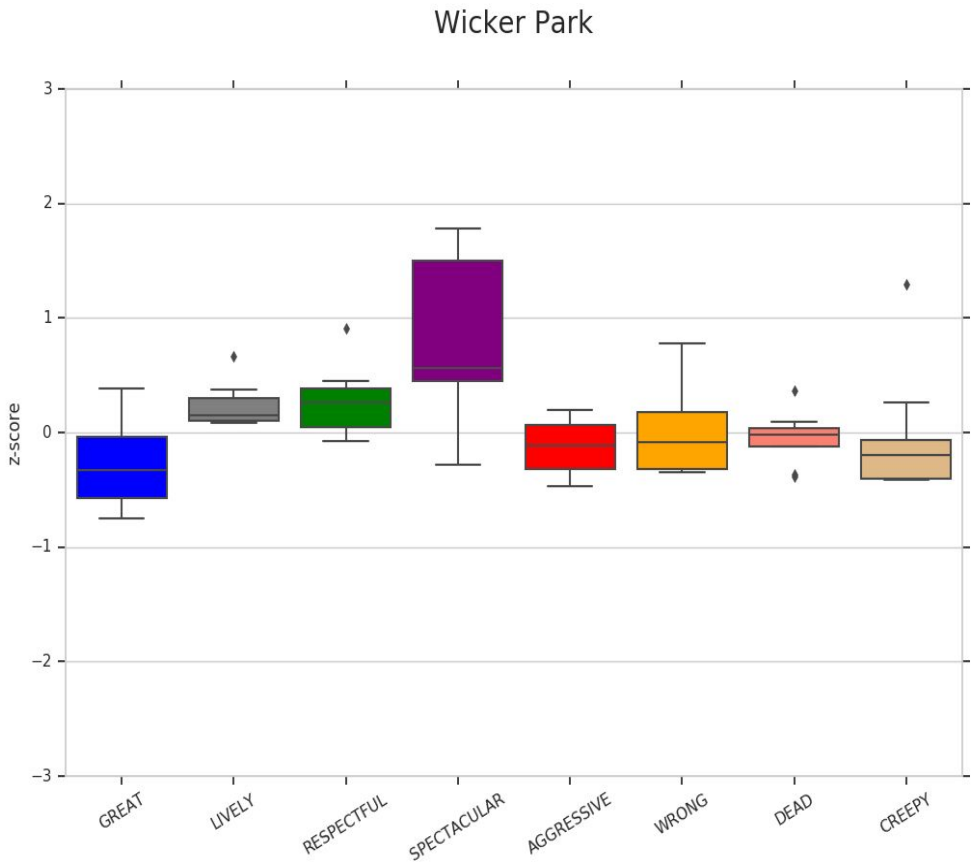
Comparative analysis - Chicago



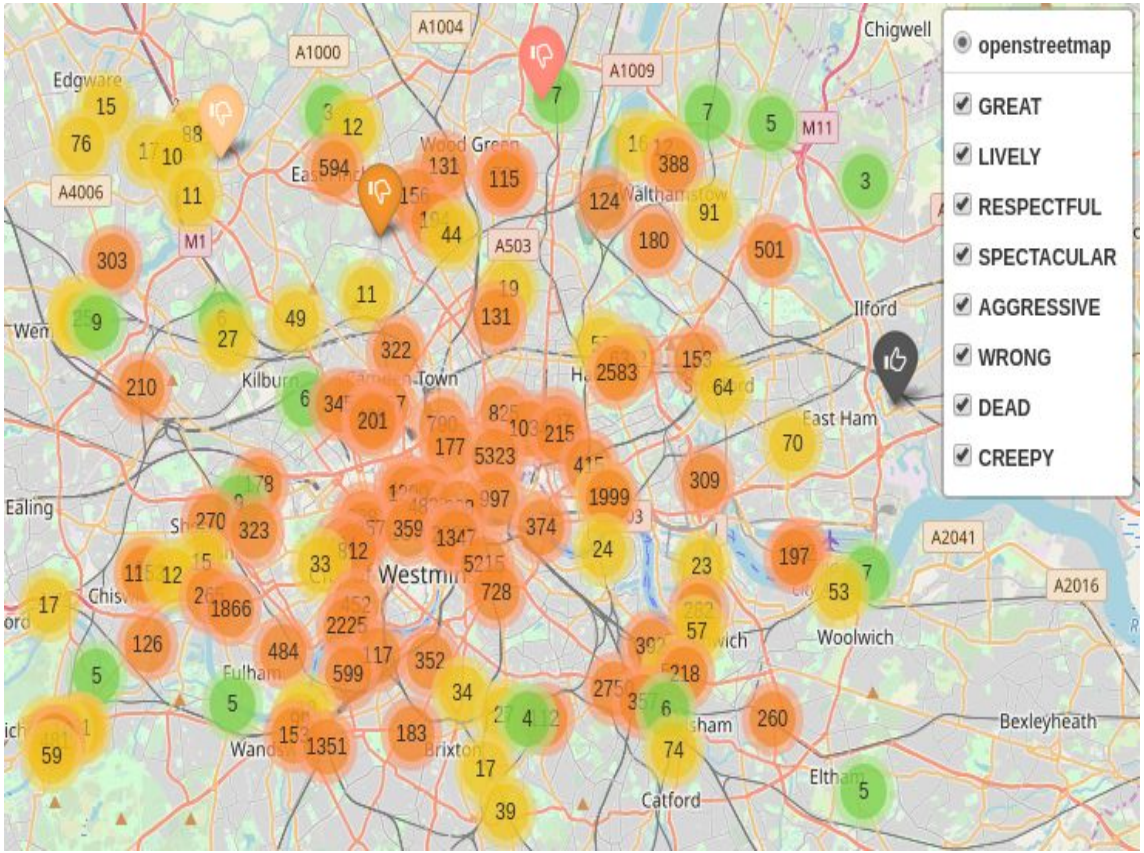


Comparative analysis - Chicago

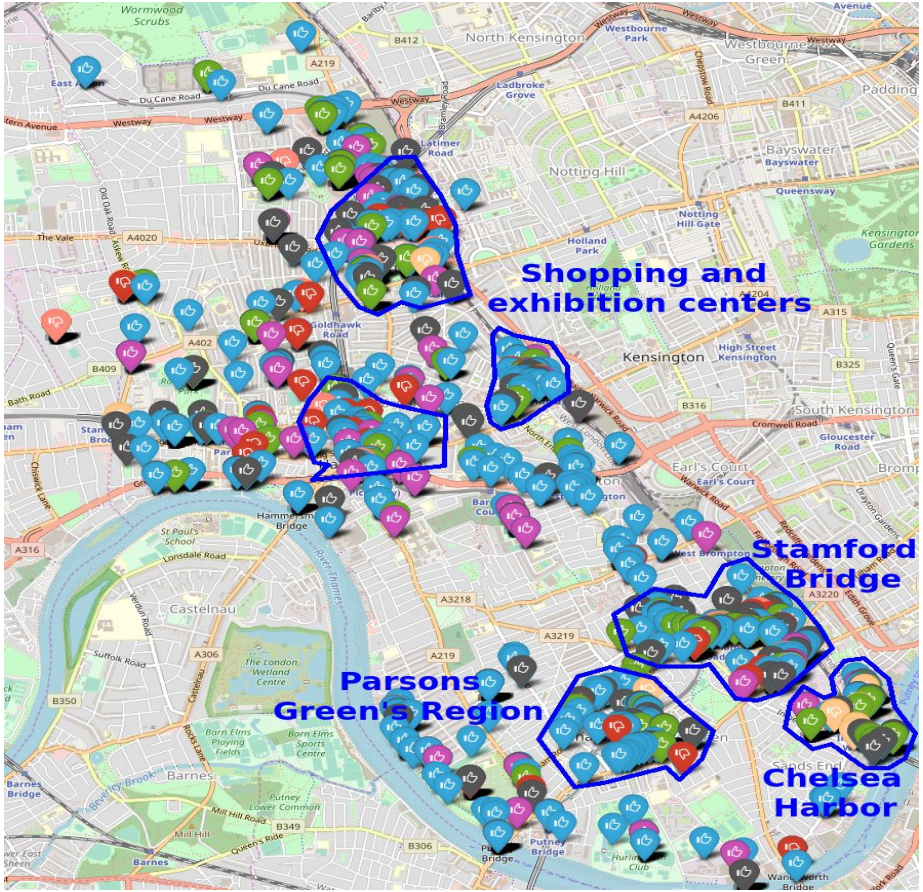


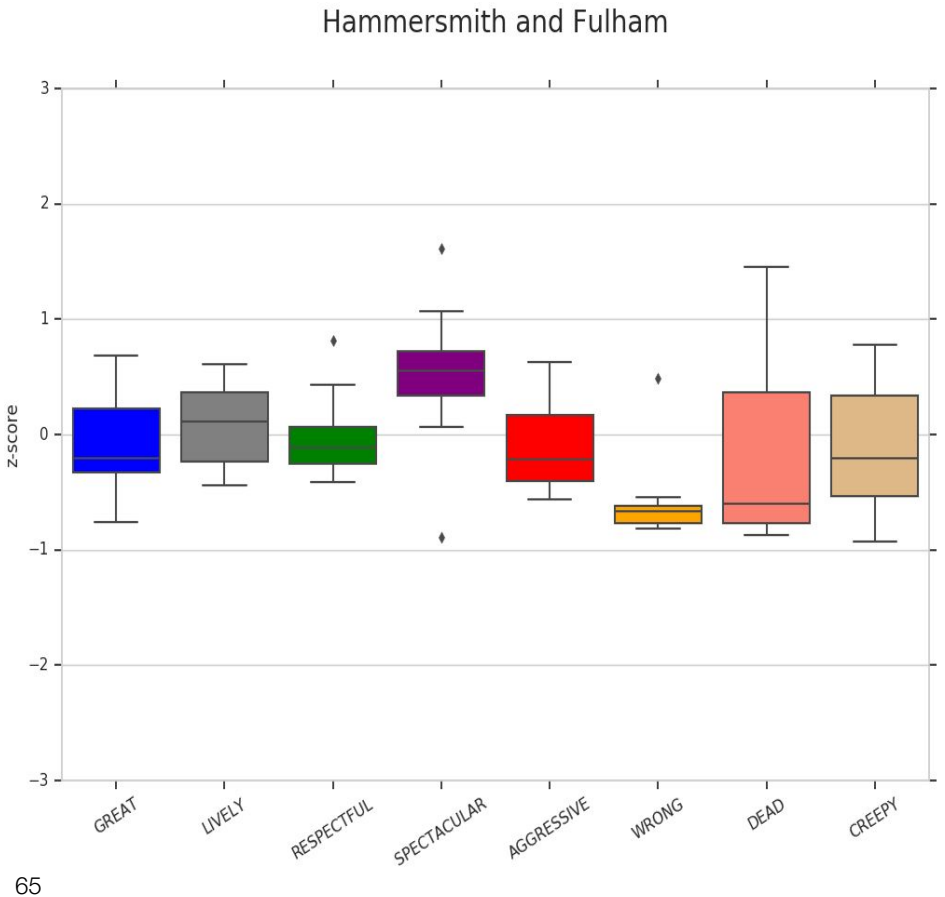


Perception Maps - London

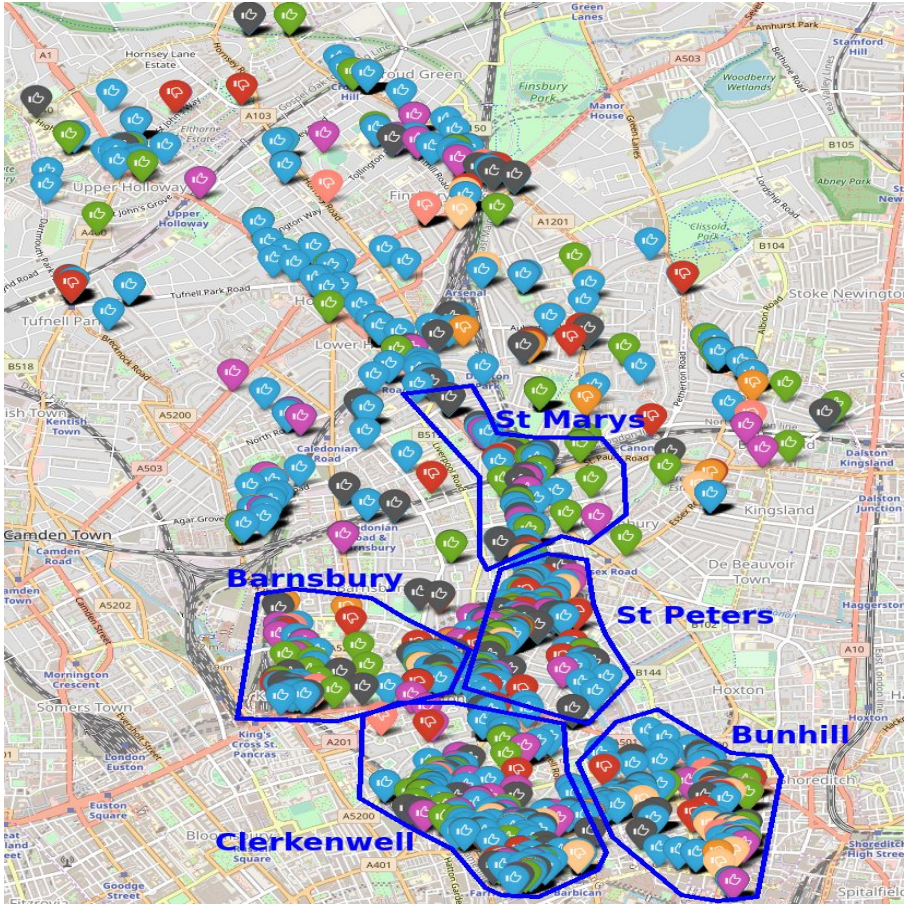


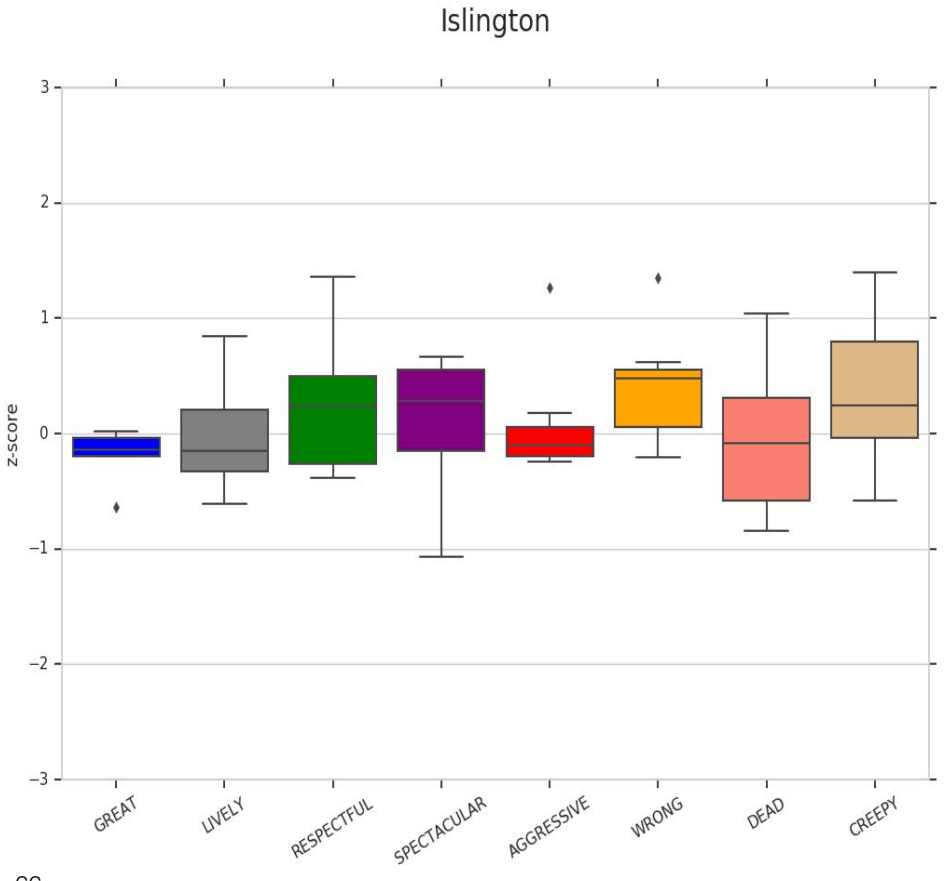
Urban Perception – London's neighborhoods



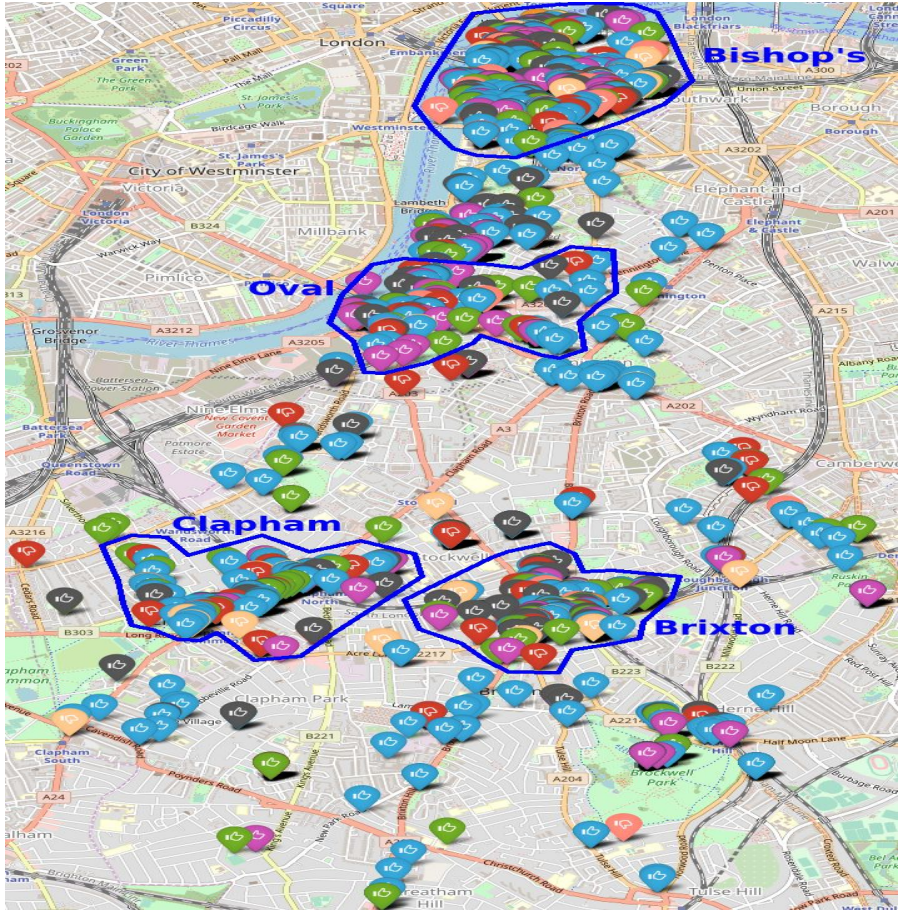


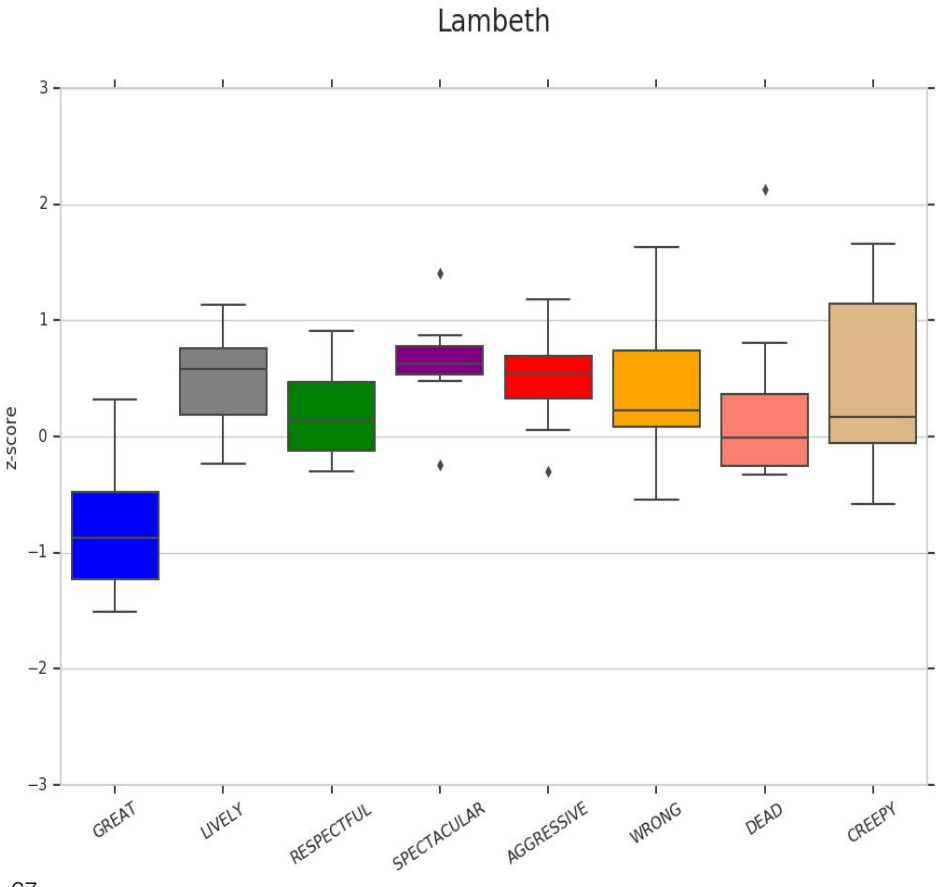
Urban Perception – London's neighborhoods



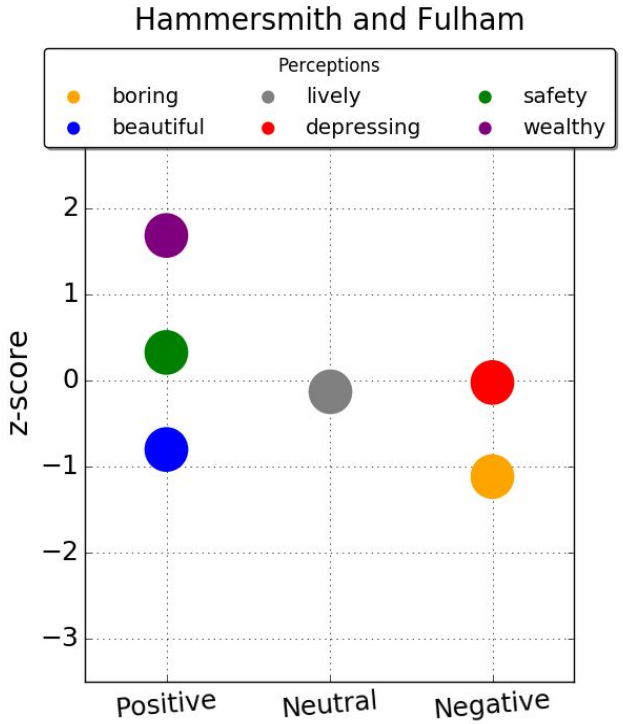


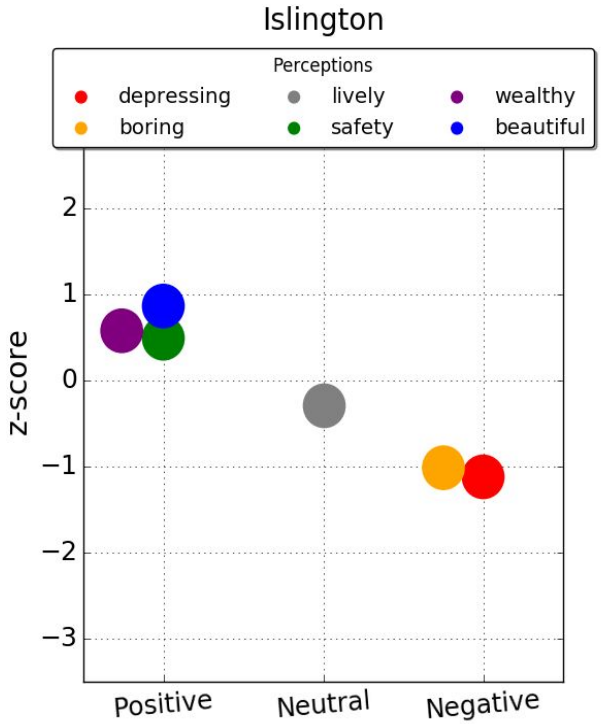
Urban Perception – London's neighborhoods

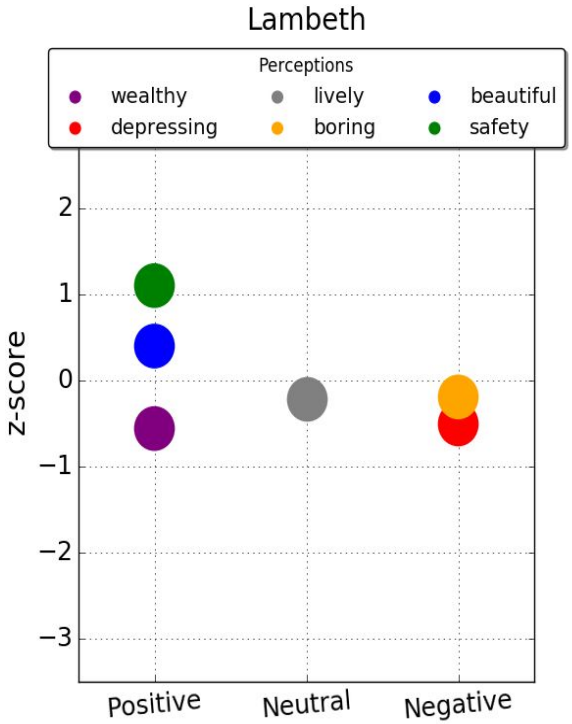




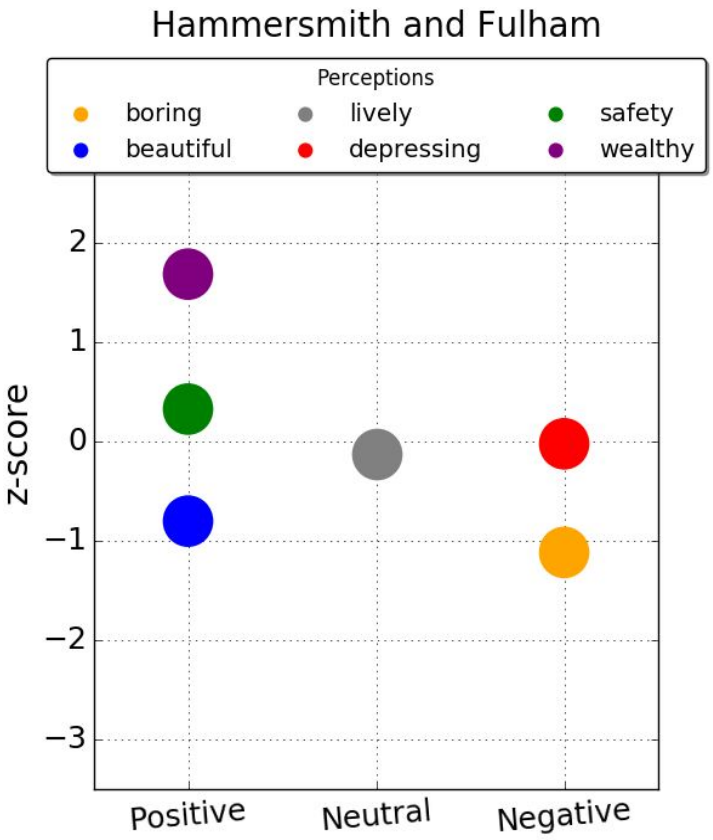
The perception strength of London's neighborhoods according to Place Pulse

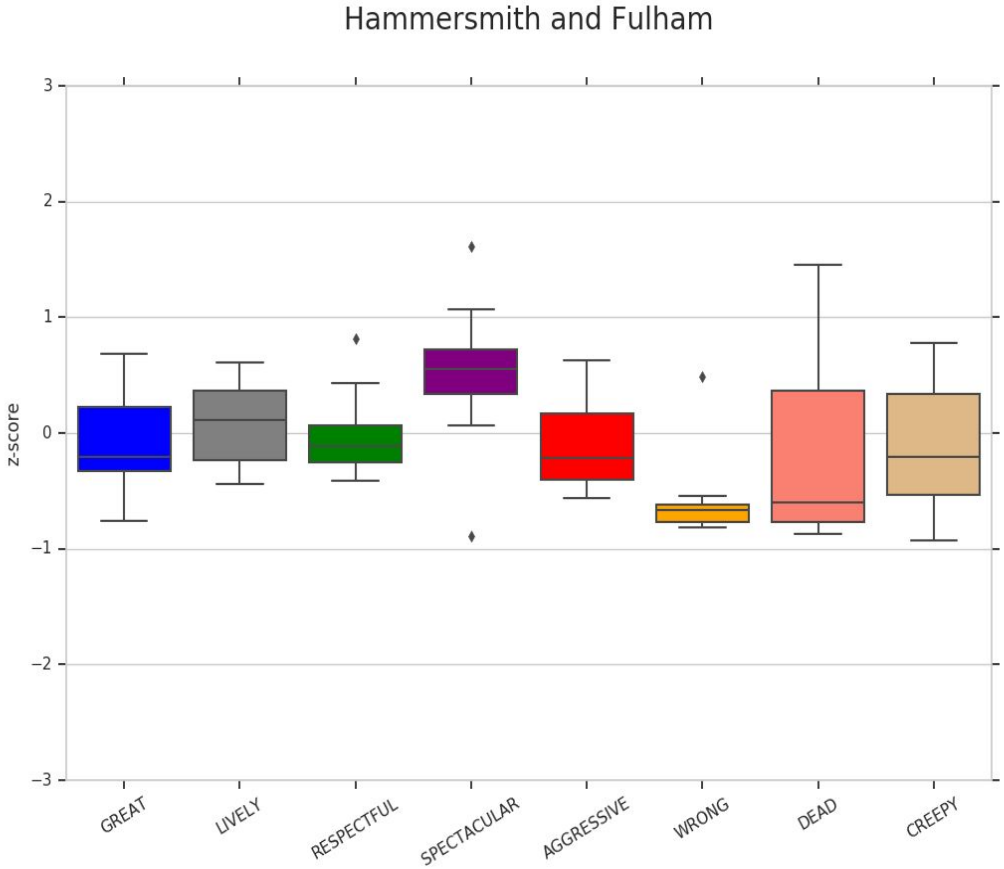




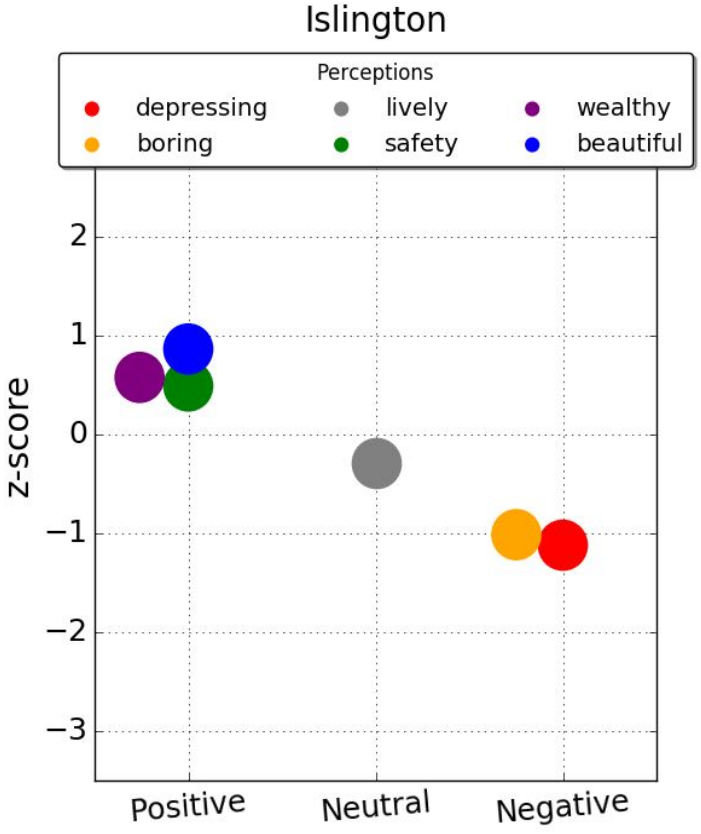


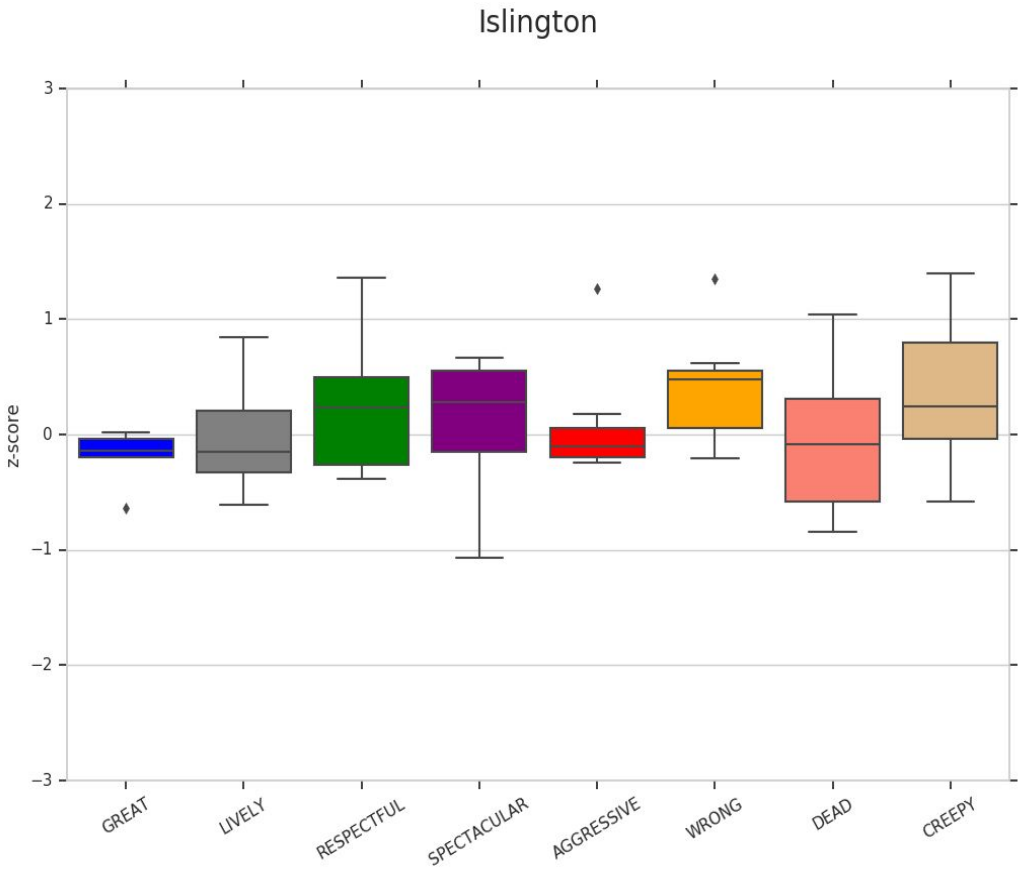
Comparative analysis - London



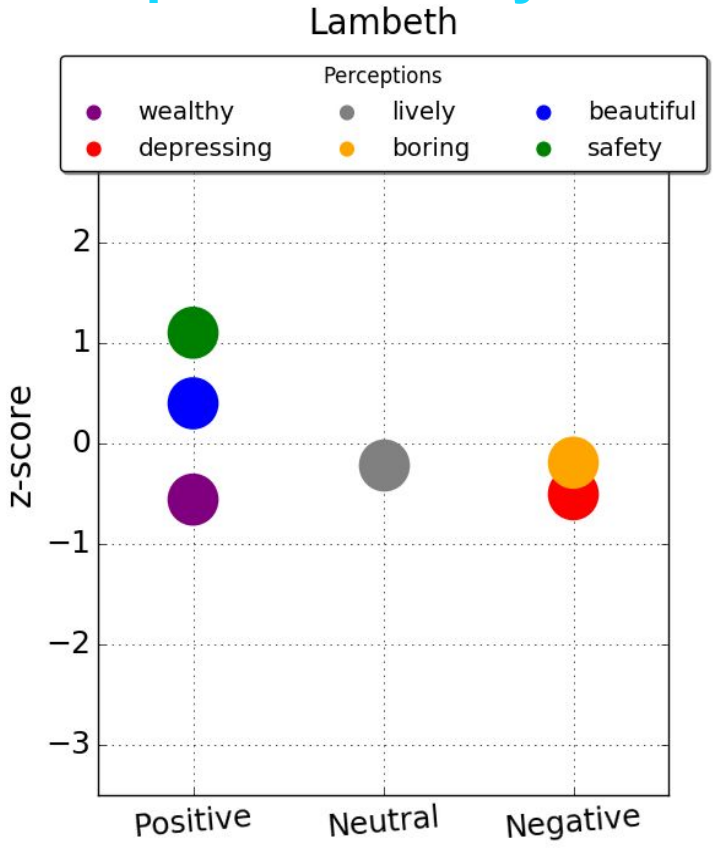


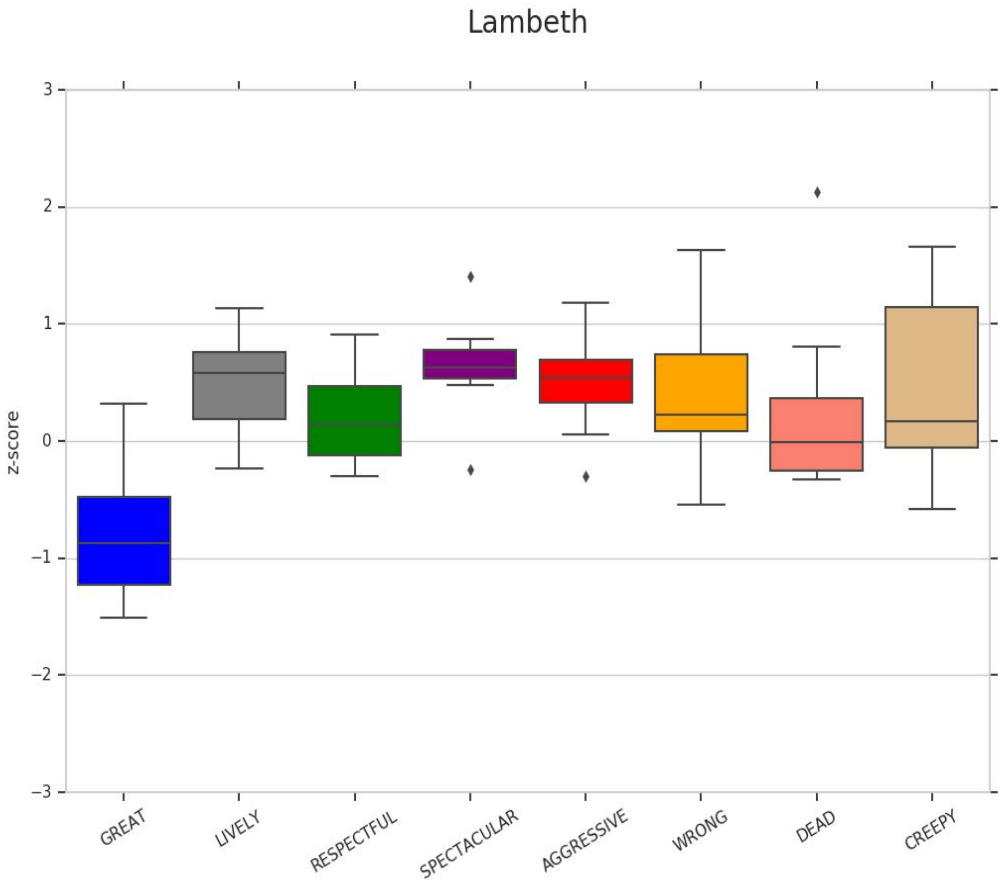
Comparative analysis - London





Comparative analysis - London

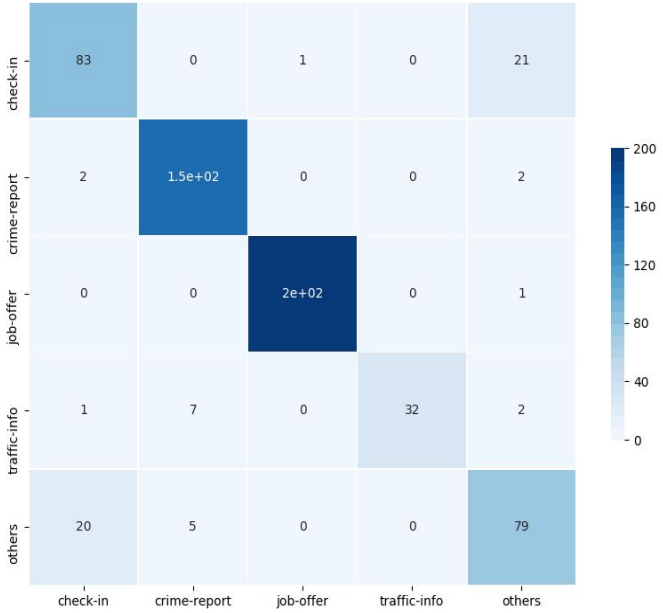


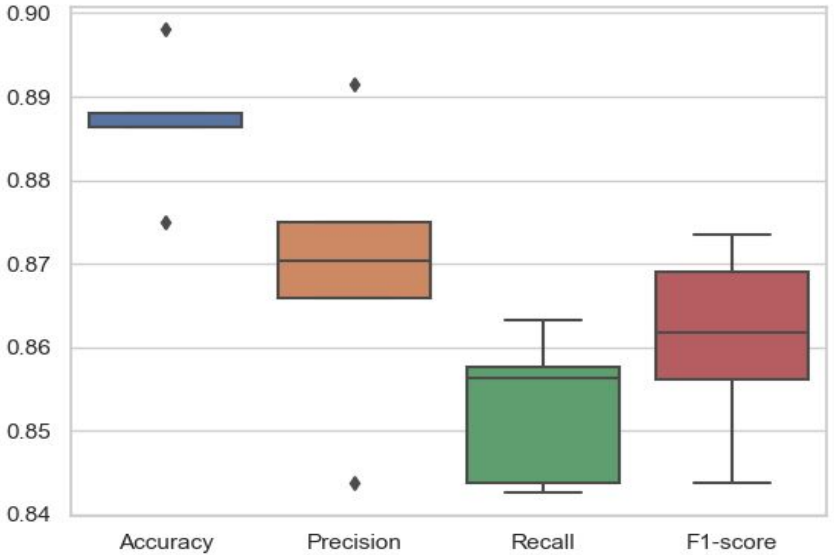


Management Layer

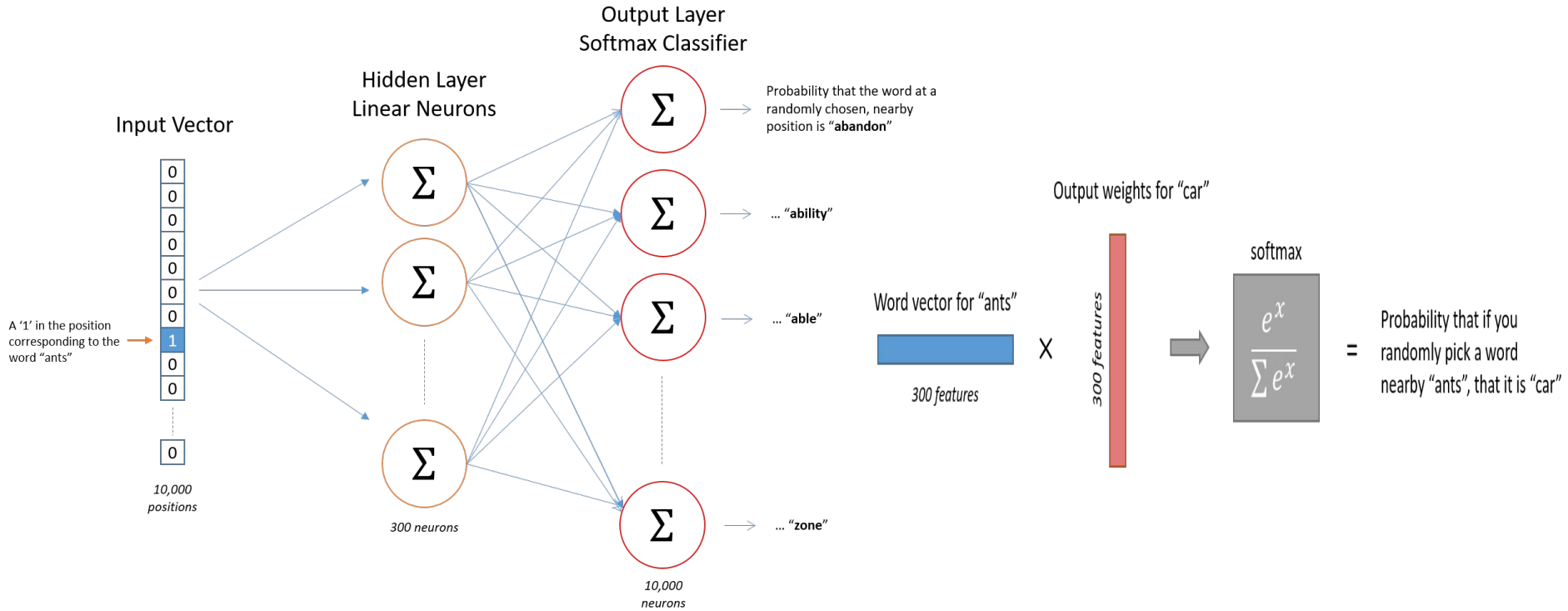
RASA NLU

check-in	*crime-report*	*job-offer*	*traffic-info*	*others*
521	785	1000	208	525





Word2Vec


Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
10,000 positions
300 neurons
10,000 neurons
Output weights for "car"
Word vector for "ants"
300 features
X
300 features
softmax
Probability that if you randomly pick a word nearby "ants", that it is "car"

Sentiment Analysis

VADER Sentiment Analysis

“VADER is smart, handsome and funny.”

{‘neg’: 0.0, ‘neu’: 0.254, ‘pos’: 0.746, **‘compound’: 0.8316**}

“Today SUX!”

{‘neg’: 0.779, ‘neu’: 0.221, ‘pos’: 0.0, **‘compound’: -0.5461**}

Perception Maps - NYC