Natural Language Processing CZ4045

Group Report (G20C)

Ben Trovato*
G.K.M. Tobin*
trovato@corporation.com
webmaster@marysville-ohio.com
Institute for Clarity in Documentation
Dublin, Ohio

Lars Thørväld
The Thørväld Group
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
San Antonio, Texas
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

Our task covered data processing on a dataset provided by the review platform *yelp*. We had to analyze the data descriptively and we had to focus on the Adjectives in the reports. Therefore we had to compare different methods on how the reviews can be represented by adjectives, which also became our application model. In our application model we were able to find specific properties of the business reviewed in the data.

CCS CONCEPTS

• **Natural Language Processing → Group Assignment**.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Ben Trovato, G.K.M. Tobin, Lars Thørväld, Valerie Béranger, Aparna Patel, Huifen Chan, Charles Palmer, John Smith, and Julius P. Kumquat. 2018. Natural Language Processing CZ4045: Group Report (G20C). In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

*Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

1 DATASET ANALYSIS

Writing Style

For the sentence segmentation we used the library spacy. Each category is displayed in the graph below. I

Table 1: Average length of the sentences in the different star categories

1 Star	2 Star	3 Star	4 Star	5 Star
30.5	25.9	24.0	25	24

Sentence Segmentation

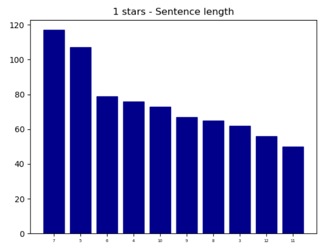
Tokenization and Stemming

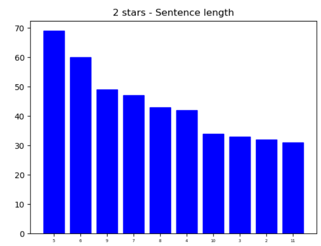
POS Tagging

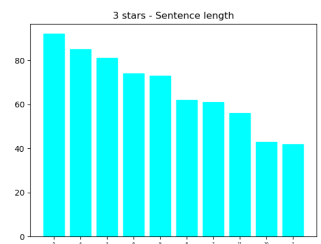
Most Frequent Adjectives for each Rating

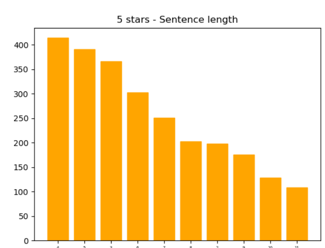
2 DEVELOPMENT OF A NOUN - ADJECTIVE PAIR SUMMARIZER

For the extraction of the Noun-Adjectives, we used a dependency-grammar approach. In dependency grammar, adjectives are labeled as 'amod'(sources) for sentences like "… bad service …" while in the case of Adjectives in sentences like "… the service at … is bad …", the adjectives are labeled as 'acomp'. With this labels we were able to identify the adjectives. For the nouns we searched the dependency tree of the parent









node of the adjective for the 'nsubj' relation. All of these relations were being stored in an array. After another check of the POS-Tag of the collected 'nsubj' relations, we were able to write the ADJ-NOUN Tuples to the final array. Before the adding we lemmatized the words so we can have more accurate number of the used words in the final. We decided to do so since our corpus is to small to distinct between different tenses and therefore increasing the accuracy for the summary of the used adjectives and nouns. The lemmatization was done after the Adjective and Noun were identified For the research we looked at the top twenty appearing pairs. In the results we were able to see some characteritics of the reviewed business. It was possible for example, to make an assumption about the business of the following review: "' Results oLb3-eXUFtCFJl2DuBhcvA [(('front', 'desk'), 20), (('free', 'wifi'), 5), (('clean', 'room'), 5), (('next', 'day'), 5), (('next', 'door'), 4), (('light', 'sleeper'), 4), (('rental', 'car'), 3), (('friendly', 'staff'), 3), (('comfortable', 'bed'), 3), (('great', 'breakfast'), 3), (('hot', 'food'), 3), (('free', 'breakfast'), 3), (('continental', 'breakfast'), 3), (('clean', 'rooms'), 3), (('new', 'room'), 3), (('next', 'morning'), 3), (('complimentary', 'breakfast'), 3), (('free', 'shuttle'), 3), (('first', 'night'), 3), (('first', 'day'), 2)] "' Here we were able to assume, that the review describes an hotel. Some other reviews also indicated the sector of the business. "' Results 2xrpo-LXV9uGIwpvy0dwUw [(('clean', 'car'), 4), (('other', 'locations'), 3), (('great', 'job'), 3), (('basic', 'wash'), 2), (('terrible', 'wash'), 2), (('poor', 'job'), 2), (('high', 'pressure'), 2), (('terrible', 'job'), 2), (('helpful', 'guy'), 2), (('horrible', 'service'), 2), (('bad', 'service'), 2), (('worth', 'place'), 2), (('different', 'options'), 2), (('only', 'place'), 2), (('friendly', 'staff'), 2), (('terrible', 'service'), 2), (('horrible', 'smell'), 2), (('happy', 'camper'), 2), (('classic', 'wash'), 2), (('synthetic', 'change'), 2)] "' It was also possible to identify the kind of food which is served in some restaurants, like mexican, vietnamese or chinese for example. "' Results DcfkRb2bS2c8z21WHaS6A [(('carne', 'asada'), 13), (('mexican', 'food'), 5), (('free', 'chips'), 4), (('great', 'place'), 4), (('authentic', 'food'), 3), (('red', 'sauce'), 3), (('mexican', 'restaurants'), 3), (('good', 'food'), 3), (('friendly', 'staff'), 3), (('best', 'food'), 3), (('little', 'flavor'), 2), (('toasted', 'bread'), 2), (('iced', 'tea'), 2), (('reasonable', 'price'), 2), (('many', 'restaurants'), 2), (('many', 'people'), 2), (('good', 'salsa'), 2), (('great', 'tacos'), 2), (('good', 'taco'), 2), (('favorite', 'place'), 2)] "'

```
Results c1_adyjYG6JEa1PZAXMOBg
[(('south', 'indian'), 14), (('indian', 'food'), 14), (('indian

Results R4EhR8xhONLFqqI6ZnzNWw
[(('good', 'dumpling'), 8), (('good', 'food'), 8), (('korean',
'food'), 7), (('korean', 'dishes'), 7), (('steamed', 'dumplings'
6), (('chinese', 'food'), 5), (('chinese', 'dumplings'),
4), (('chinese', 'cuisine'), 4), (('korean', 'soup'),
4), (('great', 'service'), 4), (('fried', 'pork'), 4),
(('huge', 'fan'), 3), (('cheap', 'food'), 3), (('awesome',
'dumpling'), 3), (('fresh', 'noodle'), 3), (('north',
'korean'), 3), (('other', 'dishes'), 3), (('fried', 'rice'),
3), (('hidden', 'gem'), 3), (('northern', 'chinese'),
3)]
```

cell1	cell2	cell3	cell2	cell3
cell4	cell5	cell6	cell2	cell3
cell7	cell8	cell9	cell2	cell3

With this summarizer we were able to find very specific characteristics of the business. There are alot of useful adjectives to specific offers of the restaurant (e.g.: (('best', 'buffet'), 4), (('quick', 'service'), 2), (('good', 'food'), 5)). With this results we can get extract the most important pairs of the reviews. But still there are some pairs which are not useful to extract the main information of the review like time information which are not useful without their contextes (e.g. : ('single', 'time'), 2),(('first', 'night'), 3), (('first', 'day'), 2)). We assume that the reason for this is the ambigousness of the POS-Tagging. Neither NLTK nor spacy where able to distinguish between the function of a determiner and an adjective. Since these time words can also be used in other contextes as adjectives, an exclusion of these words would not be reasonable. There were also some inaccurancies apperearing like (('tomato', 'sauce'), 3) which are not usefull for the reflection. We can conclude, that we can extract alot of useful knowledge from the reviews using this Adjective-Noun-Summarizer.

3 APPLICATION