STATISTICAL LANGUAGE MODEL AND PERPLEXITY USING THE YELP DATASET

Kekoa Riggin
kekoar@uw.edu
Source code: https://github.com/unclenachoduh/yelp

1 INTRODUCTION

The sheer size of each Yelp Dataset allows us an interesting look into the type of communication that goes on online. Independent of Yelp's Dataset Challenge, researchers are constantly observing how humans act, what they consume, and what they say online.

The Yelp Dataset is particularly interesting because it features data from clustered geographic regions. This geographic spread provides the opportunity to objectively compare the uses of language by region, and possibly define the ways in which different people and cultures talk online.

The goal of this project is to use the Yelp Dataset as training data for a statistical language model. The data will be separated by region, allowing for the creation of a language model for each. A sample of each region will be set aside to be used as testing data, which will be compared to each language model to produce a perplexity score, indicating the variation in the language use.

2 METHODOLOGY

In order to build the language models, the relevant data needed to be extracted from the dataset, formatted, separated into training and testing data, then utilized. Each is described as follows:

2.1 DATA CREATION

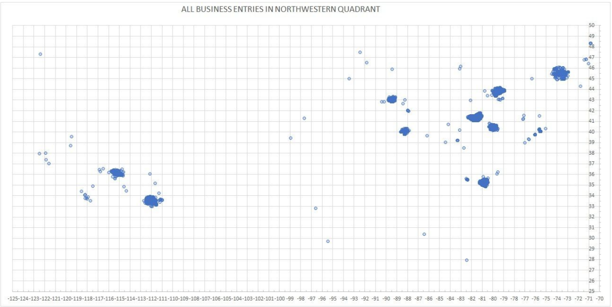
The Yelp Dataset comes in a well-formatted JSON file. However, the data has been spread out over several files and must be searched through in order to pair various pieces of metadata to the data that will be used to build the language model. For the case of comparing reviews to regional language models, the reviews must be paired to the geographic region of the businesses. Additionally, the raw review text needs to be formatted in order to suffice the architecture of the language model.

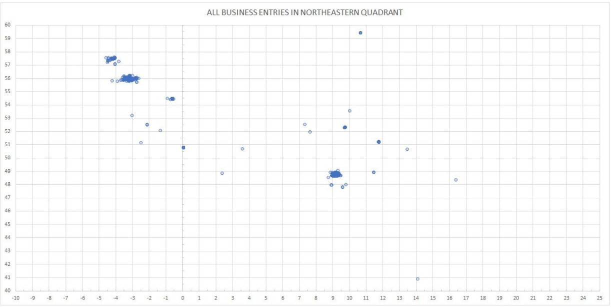
2.1.1 BUSINESSES

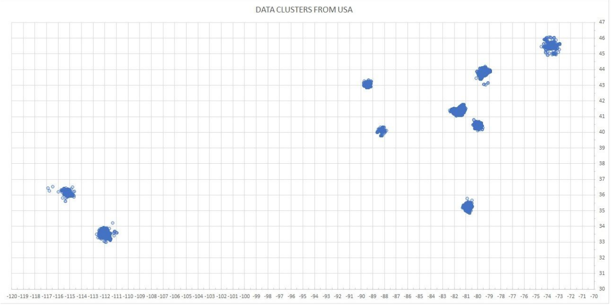
In order to identify the geographic regions of the businesses, the latitude and longitude coordinates were extracted from the business JSON file. These coordinates were saved to a simple spreadsheet and used to generate a plot chart of the businesses.

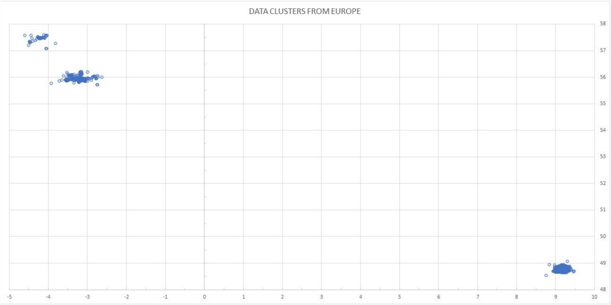
The spread of the business locations revealed two important things: 1) the major clusters that were the metropolitan areas of the dataset and 2) the number and spread of the outliers to those areas.

In an attempt to verify the quality of the data, regions were added to the extraction script in order to filter out businesses that did not belong to the major metropolitan clusters.









The metropolitan areas were labeled as follows, based on the city most well known to the researcher near the cluster: Buenos Aires, Argentina; Champaign, Illinois; Charlotte, North Carolina; Cleveland, Ohio; Edinburgh, United Kingdom; Inverness, United Kingdom; Las Vegas, Nevada; Madison, Wisconsin; Montreal, Canada; Phoenix, Arizona; Pittsburgh, Pennsylvania; Stuttgart, Germany; and Toronto, Canada.

The business IDs from the business JSON file were written to a separate file so that they could be used to identify the geographic location of each review.

The plot charts above so the spread of the data points prior to regional selection (top) and after regional selection (bottom) for both the USA (left) and Europe (right).

2.1.2 REVIEW MATCHING

The review JSON file did not contain geographic data, but did contain the business ID for which the review was written. Thus, each review could be matched to the business location using the file from the business extraction.

The sheer size of the review JSON file presented a problem. It measured 4.7 million lines long, and, to the authoring of this report, the complete review file has not been completely processed.

Instead, the file has been put to read as a background process over the course of several days, writing an output file at every 250k lines. Roughly, 2/3rds of the total reviews was able to be extracted, but even then, the processing time for other phases was far too long.

With the exception of a few metropolitan areas, 200k reviews were used as a sample for the project. This sampling is deemed sufficient by the researcher.

Of the 200k reviews (fewer for some areas), 180k were used as training data and 20k as testing data (or 90% and 10%) for each region.

2.1.3 TOKENIZATION

Once the reviews had been extracted and organized by geographic region, they also needed to be tokenized before they could be used for the language model.

Tokenization separates the functional pieces of a sentence so that they can be analyzed as working parts in a sequence.

In order to maintain the integrity of the language of Yelp reviews, no special characters were removed from the data, with the exception of regular expressions such as `\n` and `\/`. No other changes were made to the reviews.

2.2 LANGUAGE MODEL

The language model is a graph where the nodes are the words of the review and the arcs are the transitions between words. Each arc is assigned a weight, which is the probability of the next word given the ngram series of words that come before it. The language model for this project is a 3-gram model. The probability of the next word is defined by `P(word | ngram)`.

For this language model, unknown words are not given a probability and are subtracted from the total word count to eliminate their effect on the perplexity score.

2.3 PERPLEXITY

The perplexity score indicates how well the test data fits the language model. The perplexity score is the logprob of the path through the model.

For each word in the test data, the probability is calculated using the function

$$\begin{aligned} P(w_i \mid w_{i-2}\, w_{i-1}) \;=\; & \lambda_3\, P_3\,(w_i \mid w_{i-2}\, w_{i-1}) \\ & + \lambda_2\, P_2(w_i \mid w_{i-1}) \\ & + \lambda_1\, P_1(w_i) \end{aligned}$$

where w_i is the current word, w_{i-1} is the previous word, and w_{o-2} is the word prior to the previous word. P_1 represents the probability of the word given the number of words in the language model, P_2 represents the probability of the word given the previous word, and P_3 represents the probability of the word given the previous two words.

Each lambda in the function is used to weight the probabilities. For this project, the lambda values were as follows: L1 0.2, L2 0.5, L3 0.3. In a previous project, these values produced results with the lowest perplexity scores.

The log of `P` is summed for all words in the test data. Then the negative sum is divided by the sum of the total word count and sentence count minus the number of unknown words. The log of this quotient is the perplexity score.

A high perplexity score indicates that the test data does not fit the model while a low perplexity score indicates a better fit. A score of 1 indicates an ideal fit.

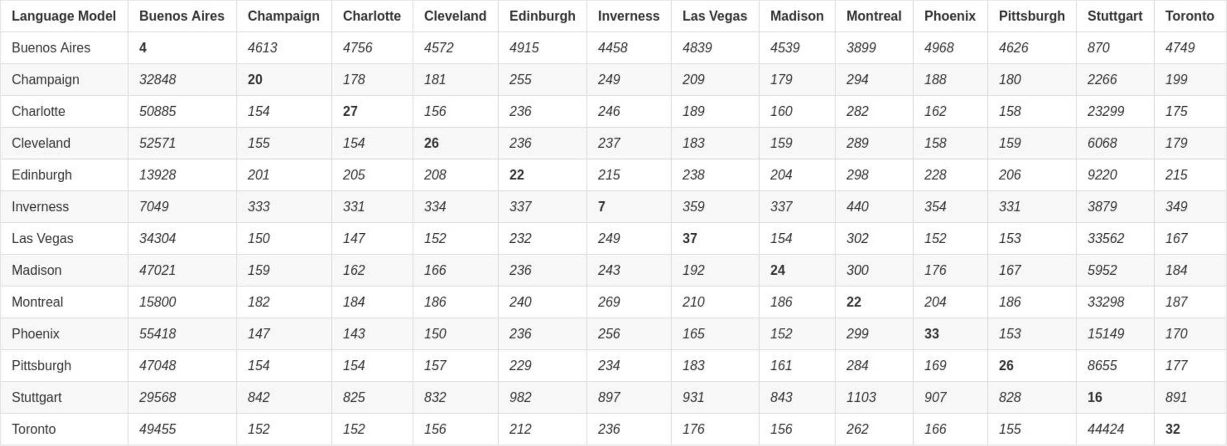
3 RESULTS

The perplexity scores showed some interesting things about the language of Yelp.

First, all testing data, when compared to the language model of its own origin, resulted very low. This indicates that reviews from a metropolitan area tend to follow the same linguistic features. This didn’t come to much of a surprise, but there were no regions that showed much variation from this trend.

Secondly, no region resulted in a false positive for any language model. This did not come as much of a surprise for regions like Buenos Aires or Stuttgart, where the official language is unique to the rest of the regions. However, it was surprising that none of the US, UK, or Canadian cities came as matches. In fact, many of them showed significantly high perplexity scores.

The UK cities provide an interesting example. They showed very similar



Language Model	Buenos Aires	Champaign	Charlotte	Cleveland	Edinburgh	Inverness	Las Vegas	Madison	Montreal	Phoenix	Pittsburgh	Stuttgart	Toronto
Buenos Aires	**4**	4613	4756	4572	4915	4458	4839	4539	3899	4968	4626	870	4749
Champaign	32848	**20**	178	181	255	249	209	179	294	188	180	2266	199
Charlotte	50885	154	**27**	156	236	246	189	160	282	162	158	23299	175
Cleveland	52571	155	154	**26**	236	237	183	159	289	158	159	6068	179
Edinburgh	13928	201	205	208	**22**	215	238	204	298	228	206	9220	215
Inverness	7049	333	331	334	337	**7**	359	337	440	354	331	3879	349
Las Vegas	34304	150	147	152	232	249	**37**	154	302	152	153	33562	167
Madison	47021	159	162	166	236	243	192	**24**	300	176	167	5952	184
Montreal	15800	182	184	186	240	269	210	186	**22**	204	186	33298	187
Phoenix	55418	147	143	150	236	256	165	152	299	**33**	153	15149	170
Pittsburgh	47048	154	154	157	229	234	183	161	284	169	**26**	8655	177
Stuttgart	29568	842	825	832	982	897	931	843	1103	907	828	**16**	891
Toronto	49455	152	152	156	212	236	176	156	262	166	155	44424	**32**

perplexity scores when compared against language models from other cities, but even higher perplexity scores when compared to each other. The same is true of several US cities.

Even cities that fall in close geographic proximity or what could be perceived as similar culture resulted in considerably high perplexity scores. The scores when comparing East Coast Us cities to West Coast cities did not vary much from the scores when comparing East to East and West to West.

4 CONCLUSION

It's not a secret that language use online is unique to language use in conversation or in print, especially on social networks. However, the differences between regional online language use is not as apparent. In conversation, the differences in communication may not even be noticeable.

This 3-gram language model reveals that, at least in the world of Yelp, regional language does, in fact, have variants. Even if the human eye doesn't see them, they are hiding in plain sight.

The use of a language model to identify types of language could have practical application and should be furthered with future studies. For example, if a language model could be trained on verified reviews, perhaps fake reviews or internet trolls could be identified if their review text doesn't fit the model of the verified reviews. In a further study, it might be possible to identify reviews written by tourists as opposed to locals, if a verified dataset could be created.

REGARDS

I am extremely grateful for the opportunity to work with the data from the Yelp Dataset Challenge from Round 10. I would like to thank Yelp for providing the dataset and establishing this opportunity for young researchers.

Although I performed the entirety of the project alone, I would like to thank Josh Matthias for consulting me on tokenization and those affiliated with the Computational Linguistics program at the University of Washington. Without the education I am currently receiving at that university, I could not have completed this project.