A Pizza My Mind

Willie Boag
MIT
wboag@mit.edu

Alan Larochelle
UMass Lowell
alan_larochelle@student.uml.edu

Abstract

Alfredo's Italian Kitchen. Brothers Pizza. Papa Gino's. ABC Pizza. What do these all have in common? They are pizza stores. For whatever reason, the phrase "pizza store" seems to inspire such vitrol among my so-called "friends". "That's not a real phrase" they always snidely remark. Oh yeah? Well I'm the one with a 3.8-billion-word dump of English Wikipedia! I decided to run some experiments to see just how "not a thing" my very normal phrase is. How frequently did a few common pizza phrases occur? My experiments suggest that although synonymous phrases might be modestly more popular, many of the acceptable phrases are all within the same order of magnitude. In addition, a popular internet search engine does an arguably better job with understanding "pizza store' than it does "pizza place". In other words, my friends are jerks.

1. Introduction

Natural language is a very interesting thing. Using some nearly arbitrary combination of letters, we are able to form words, phrases, and sentences. These inventions are able to convey concepts, and what makes language so amazing is that it is creative! We can describe things we've never seen before, such as a shark flying through the air on a jetpack. In fact, that might be the first time anyone has ever uttered the sentence "a shark flying through the air on a jetpack" , yet you all knew what it meant anyway! That's awesome!

Figure 1. A classic pizza store.

One thing that makes natural language so difficult (even for people and *especially* for computers) is its ambiguity. Ambiguity means that I can say a sentence and you don't know what I mean. I might give the classic example of "I saw a man on a hill with a telescope." And then I ask you what I saw. Well, maybe I saw a man standing on a hill while he holds a telescope, or maybe I used a telescope to see a man standing on a hill, or maybe it was I standing on the hill in the first place. We're clear on what ambiguity means, right? [1]

Along similar lines as ambiguity, there are certainly multiple ways to express the same concept in a given language. As an example, consider the two sentences:

1. Dane Cook is so funny.
2. Dane Cook is hilarious.

While it would be fair to say that neither sentence is true, the important takeaway is that these sentences mean the exact same thing. The words "funny" and "hilarious" are synonymous; you could replace one with the other and almost always have the same meaning. This is, of course, an oversimplification. For one, some words might have many senses (the classic example being "bank" as both a financial institution and a river bank) and only one of the senses is a synonym with another word. But more importantly, there are many words and phrases that are approximately synonyms, but don't mean *exactly* the same thing. An example of these near synonyms would be "near" and "neighboring" While these two words essentially convey the same closeness in proximity, they have nuanced differences - one might say that some of France's "neighboring"

[1] that was a hilarious joke

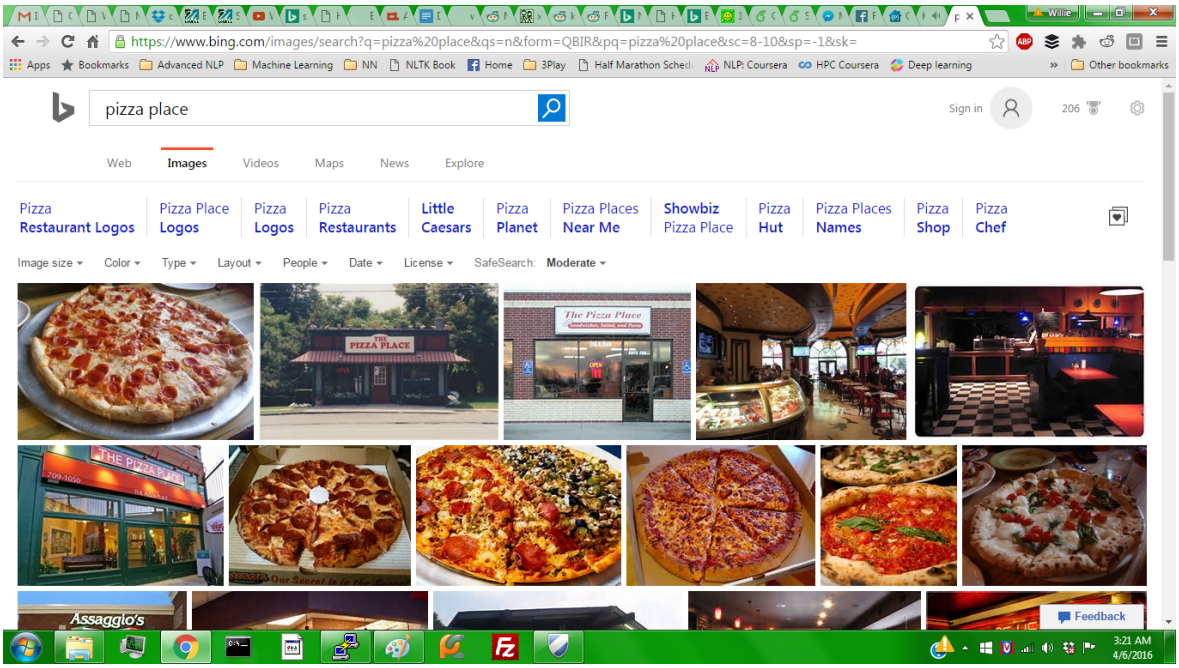

Figure 2. Bing Image results for the query "pizza place"

counties include Germany and Belguim, whereas the Netherlands are simply "near" France. Nuances and such.

2. Problem Formulation

So how does all of this fit together? Well you see, my friends are apparently very judgmental people. They're the kind of people who, when you call something a *pizza store* will laugh snarkily at you and insist that such a phrase does not exist. "We'll see who's laughing after I prove you wrong", I say to myself, quickly developing a hypothesis.

Though my friends mistakenly believe that it is inappropriate to say *pizza store*, they are content to accept the terms: *pizzeria*, *pizza place*, and *pizza joint*. Consequently, I hypothesized that *pizza store* is used frequently enough to be considered "a thing".

3. Experiments

Fortunately for me, the Natural Language Processing community is very concerned with the statistical analysis of language, especially English. There are many datasets available which I can use to definitively answer "How frequently do English speakers actually use this phrase compared to that one?" I decided to use a recent dump of the English Wikipedia [2]. Not only is this dataset quite large and representative of a wide variety of styles and domains, such as training word2vec models [1]!

After downloading the dataset, I cleaned it up (tokenized the text and removed the markup language) using a perl script for preprocessing made available at the bottom of Matt Mahoney's page [3]. This brought 53G xml document down to a 21G tokenized text document. This final text document contained 3,870,207,414 total tokens.

To analyze the text file, I wrote some very simple scripts using the python programming language [2]. Most notably, I was curious how frequently a few various pizza-themed phrases appeared in the data. We can see the results in Figure 1. We can see the frequencies that numerous reasonable phrases occur as well as a few nonsense phrases occur. These nonsense phrases *pizza land* and *pizza park* were used as baselines to demonstrate that there truly are phrases that simply "aren't a thing". It just so happens that *pizza store* is not one of those nonsense phrases.

4. Results

4.1. Quantitative

When you're comparing against nearly 4 billion word dataset, the differences between 70 words and 280 words is not all that meaningful. Is *pizza store* a

[2] http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[3] http://mattmahoney.net/dc/textdata.html

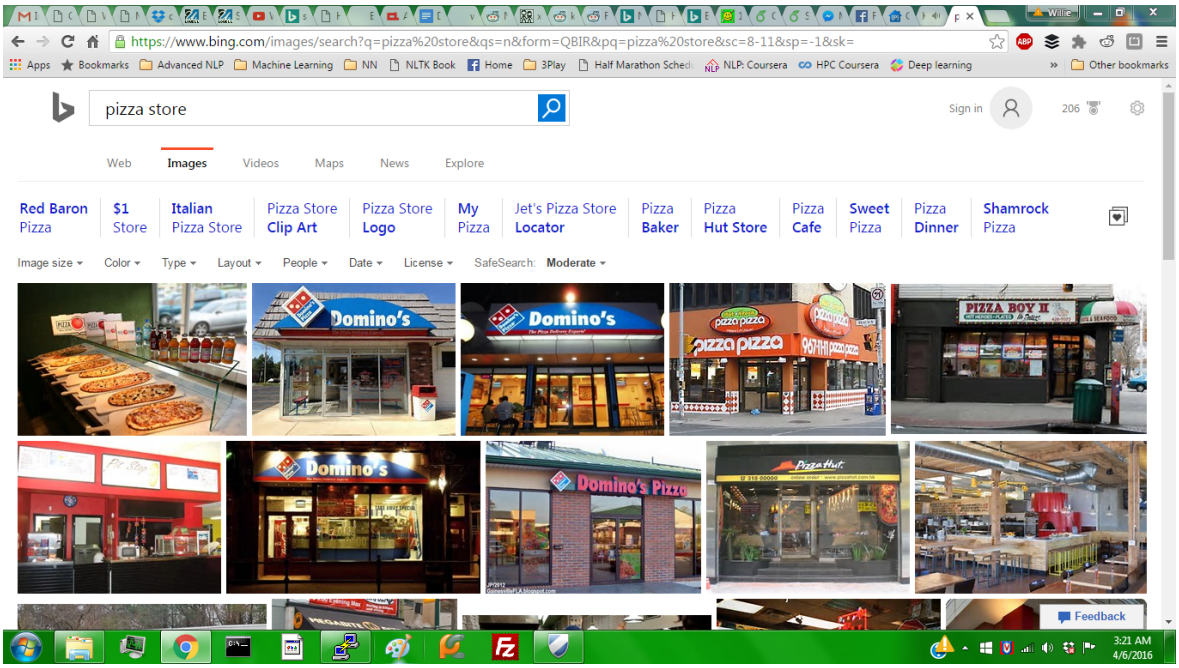

Figure 3. Bing Image results for the query "pizza store"

little less popular? Sure. But it's the same order of magnitude.

The data suggests that *pizzeria* is the most popular term for describing the concept, however it should be noted that 188 instances were preceded by the tokenized word "s" (from the possessive "'s"), indicating that many of these uses were proper noun names of stores. Even still, if only 1100-1400 occurrences were the generic "pizzeria", it would still be, by far, the most used phrase for describing the concept.

I suspect that perhaps "pizzeria" is used so frequently on Wikipedia because it is both concise and specific, thus reducing ambiguity. One of Wikipedia's very appealing features is its ability to disambiguate confusing queries, so it would not be surprising to see such a descriptive word used so frequently throughout the dataset. It should be noted that I, as a native English speaker, have never used the word "pizzeria" as my first choice to describe a pizza store. Sure there are awesome stores that call themselves a pizzeria, and I certainly follow suit, but by default the term "pizzeria" seems so cumbersome for me to say in conversation.

4.2. Qualitative

I also performed a little qualitative analysis to measure *pizza store* vs *pizza place*. I searched the queries "pizza store" and "pizza place" in the popular internet search engine Bing Images [4]. The images can be seen in Figure 3 and 2. We can see that *pizza store* is actually a superior phrase because it emphasizes the *store*. On the other hand, *pizza place* seems to be more concerned with the pizza, itself, and does not convey the correct concept.

[4]https://www.bing.com/explore/rewards?PUBL=REFERAFRIEND&CREA=RAW&rrid=_46938fa1-61e5-3d37-d710-9b94ef606482

phrase frequencies	
PHRASE	OCCURENCES
pizzeria	1657
pizza place	287
pizza shop	208
pizza joint	68
pizza store	59
pizza land	9
pizza park	2

Table 1. How frequently various phrases appeared in the English Wikipedia sump.

5. Acknowledgments

I would also like to thank my parents for their useful discussion with me as they were trying to sleep at 3:30 am. Their input was very valuable to me.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *In Proceedings of NIPS*, 2013.

[2] Python-Software-Foundation. *Python Language Reference, version 2.7*. Available at http://www.python.org.