Do Tourist Review Differently on Yelp?

Web scraping to train an NLP Classifier

Tourism: Some local statistics

- New York City's travel and tourism sector saw 60.5 million visitors in 2016 and reaching $43 billion in spending.
  - Over the period from 2010 to 2016, gains in visitation averaged 4.1% per year rising from 48.8 million.

Source: NYC & Company | New York City Travel + Tourism Trend Report

Hypothesis

- **Hypothesis**: Tourist to a region generally have different expectations, preferences and satisfaction thresholds than a local from the same region. This should be reflective is the speech behavior and utterance patters reflected a corpus of restaurant reviews.

Outline

- In what follows, I will present a corpus analysis of reviews scrapped from yelp.com in order to confirm this hypothesis.

- Then I will see how well traditional logistical regression can perform on classifing yelp reviews as being written by a tourist.

Fields

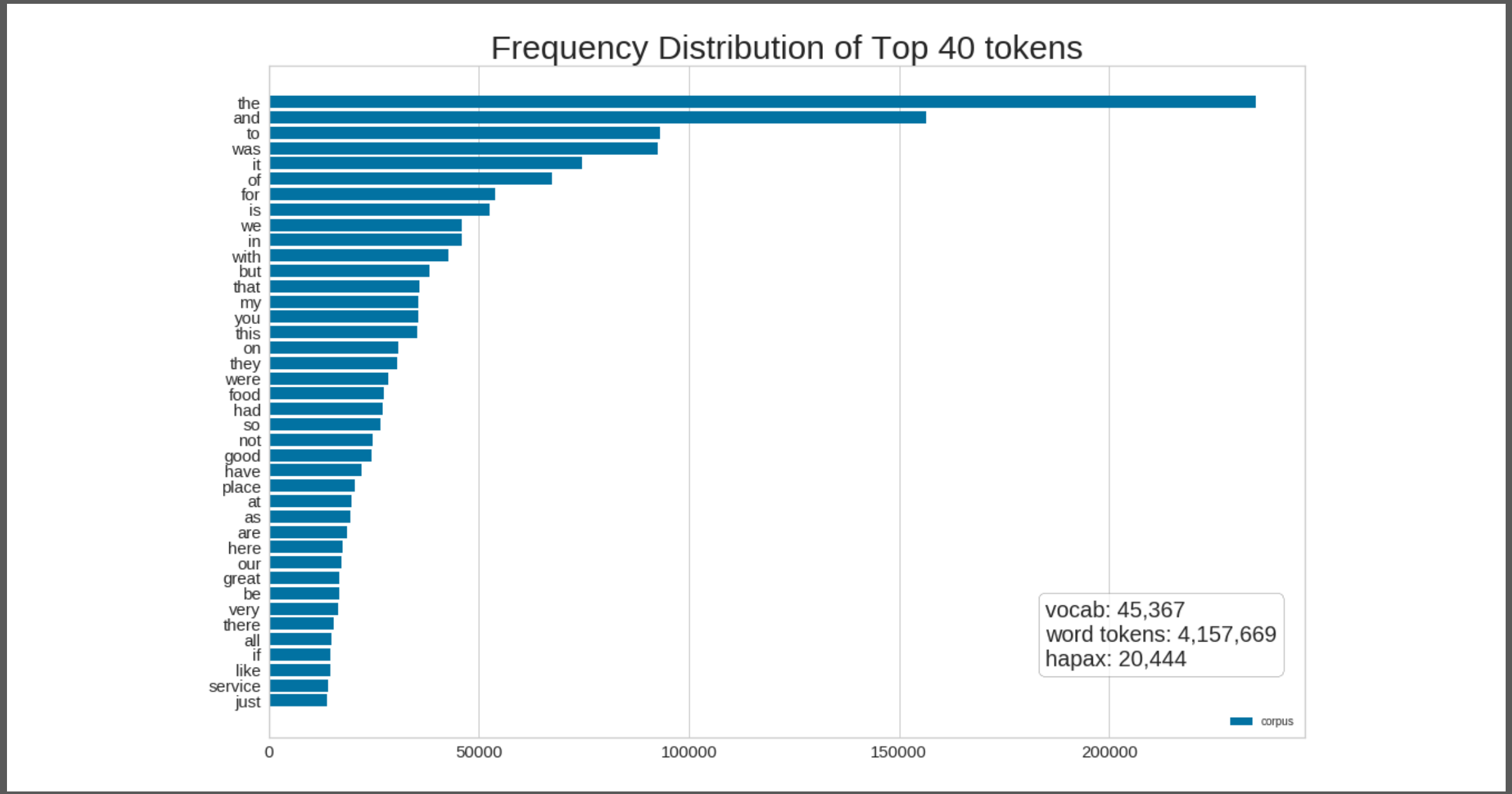
- Business:
  - City
  - Name
  - Average Star Rating
  - State
  - Zip code

- Review:
  - User Location
  - Cool
  - Funny
  - Useful
  - Date
  - Star Rating
  - Reviewer ID

Defining a Corpus

- In order not to overrepresent cities with more reviews I took a sample of the total number of scrapped reviews.
  - 6,859 reviews from each area
- Then I made sure to equally represent local and tourist reviews
- Leaving the final corpus around 32,000 reviews.

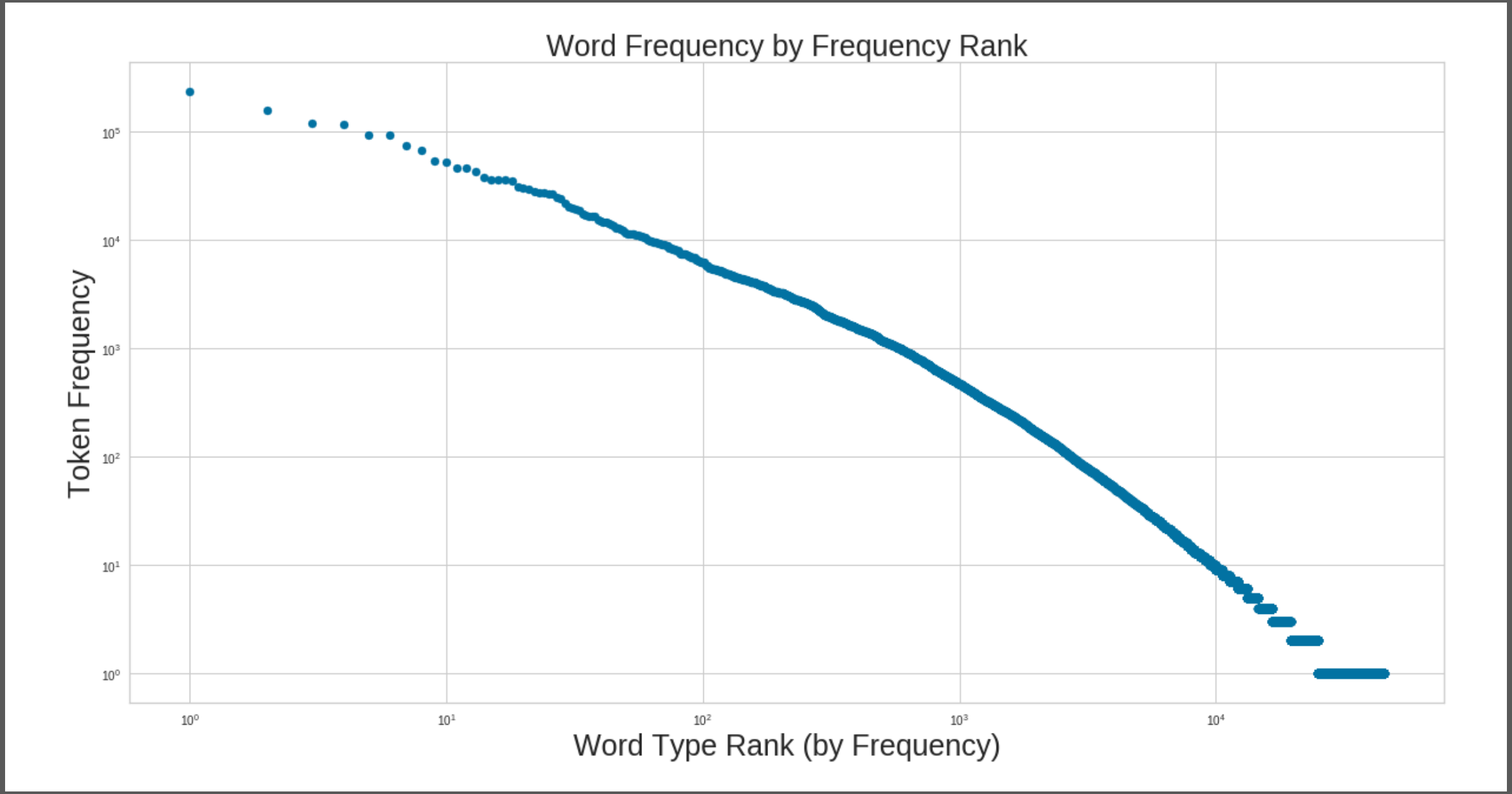
Corpus Analysis


Frequency Distribution of Top 40 tokens
the
and
to
was
it
of
for
is
we
in
with
but
that
my
you
this
on
they
were
food
had
so
not
good
have
place
at
as
are
here
our
great
be
very
there
all
if
like
service
just
0
50000
100000
150000
200000
vocab: 45,367
word tokens: 4,157,669
hapax: 20,444
corpus

Zipf's Law

- Zipf's law states natural language corpus of utterances, the frequency of any word type is inversely proportional to its rank in the frequency table.
- So frequency of the word with rank n is proportional to 1/n. In other words, the most ranked word is around twice as common as the second ranked word, and a thousand times more common than the word with rank 100,000.)

- We can check Zipf's Law for the scraped corpus of Yelp reviews by plotting the frequencies of the word types in rank order on a log-log graph.


Word Frequency by Frequency Rank
Token Frequency
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2
10^3
10^4
Word Type Rank (by Frequency)

Part of Speech Tagging

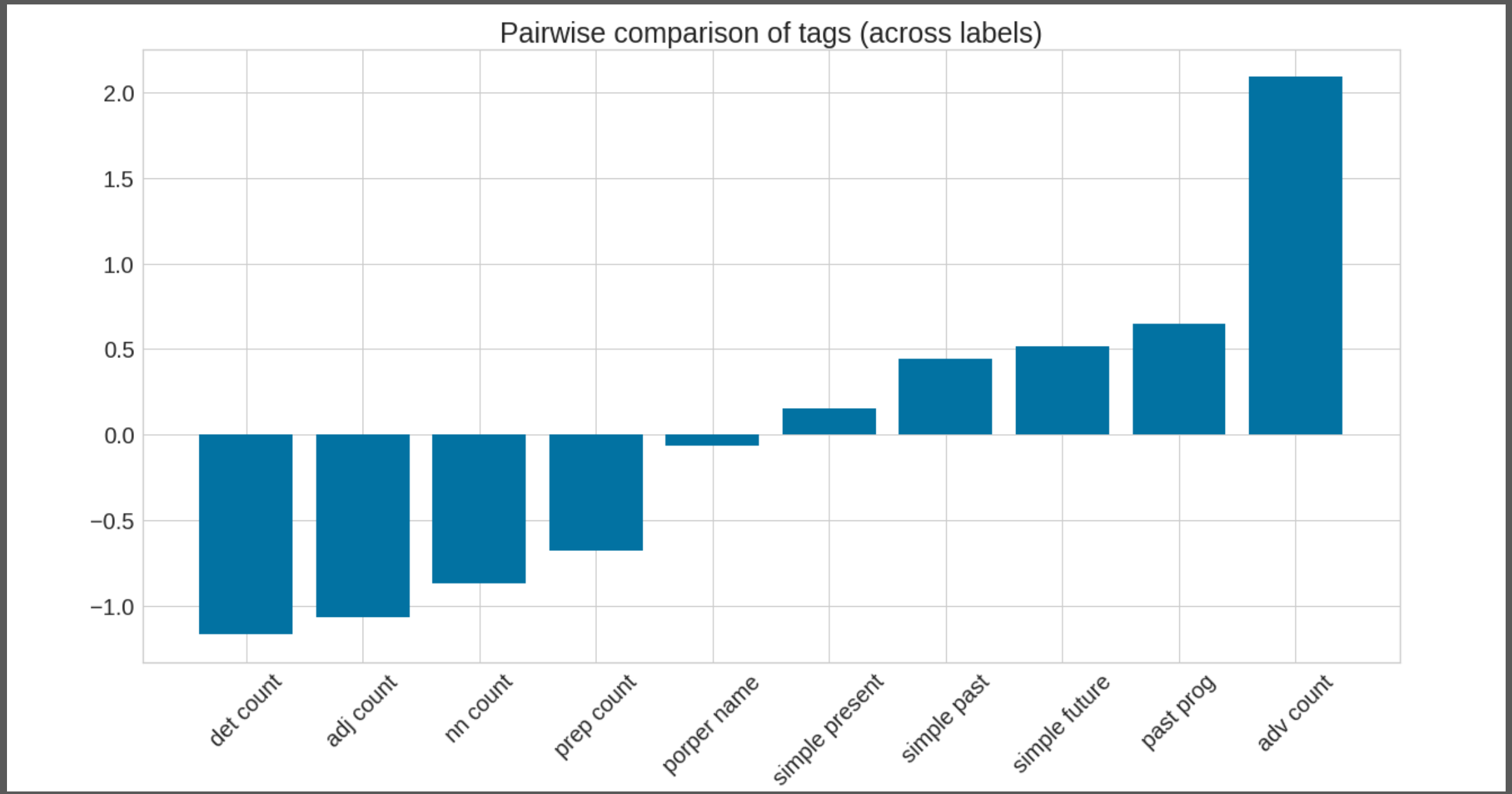
- I ran the Standford part of speech tagger on all the reviews.

- This software is a Java implementation of the log-linear maximum entropy part-of-speech taggers described in Toutanova et al (2003).

POS Pairwise comparison

- For the corpus analysis I did a pairwise comparison of the POS tags across the two labels adapted from Pak & Paroubek (2010):

$$P_{1,2}^{T} = \frac{N_1^T - N_2^T}{N_1^T + N_2^T}$$


Pairwise comparison of tags (across labels)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
det count
adj count
nn count
prep count
porper name
simple present
simple past
simple future
past prog
adv count

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Predictive Modeling: Logistical Regression

Features

- Review length (by charrachters)
- Week of the year (based on post date)
- Day of the week (based on post date)
- Is the city referenced in the post?
- POS frequency
  - Determiner,
  - Adjective
  - Natural noun
  - Preposition
  - Simple past
  - Simple Future
  - Past progressive
  - Adverb

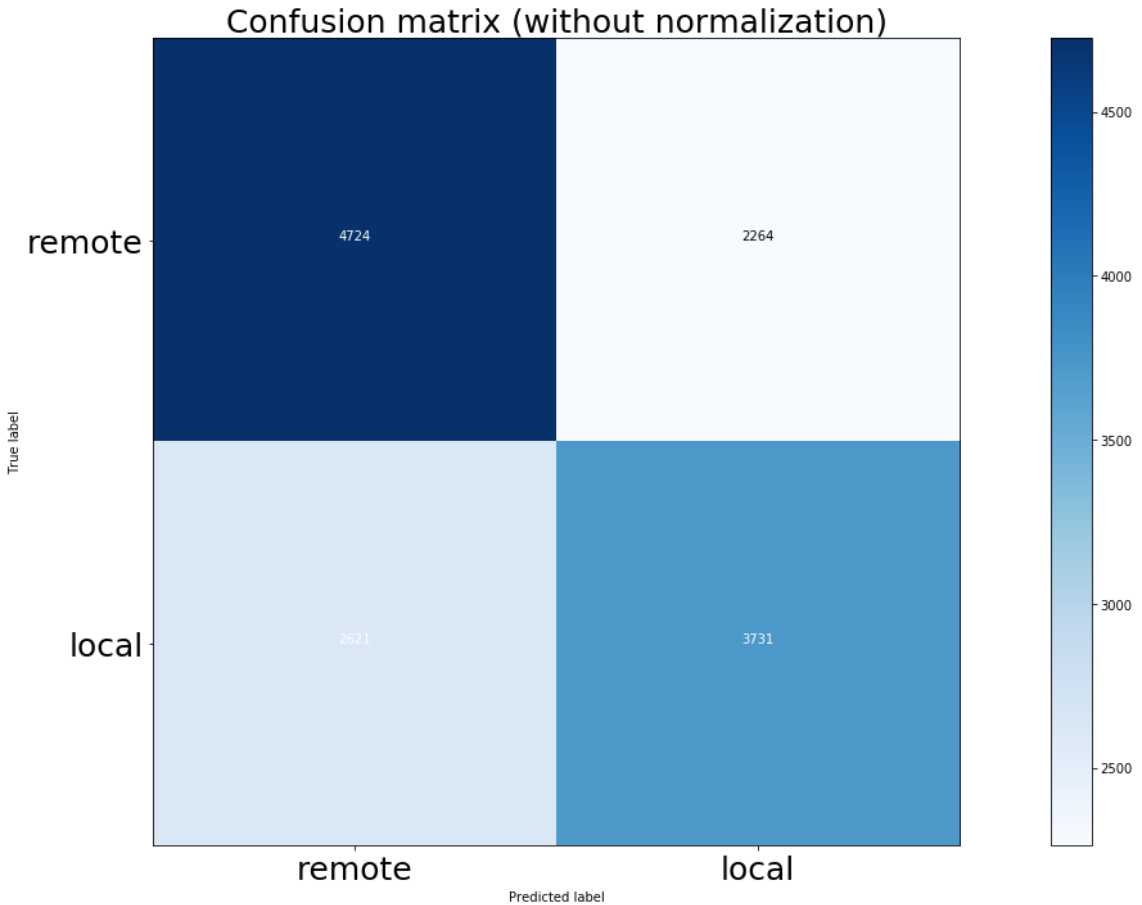
N-gram Model

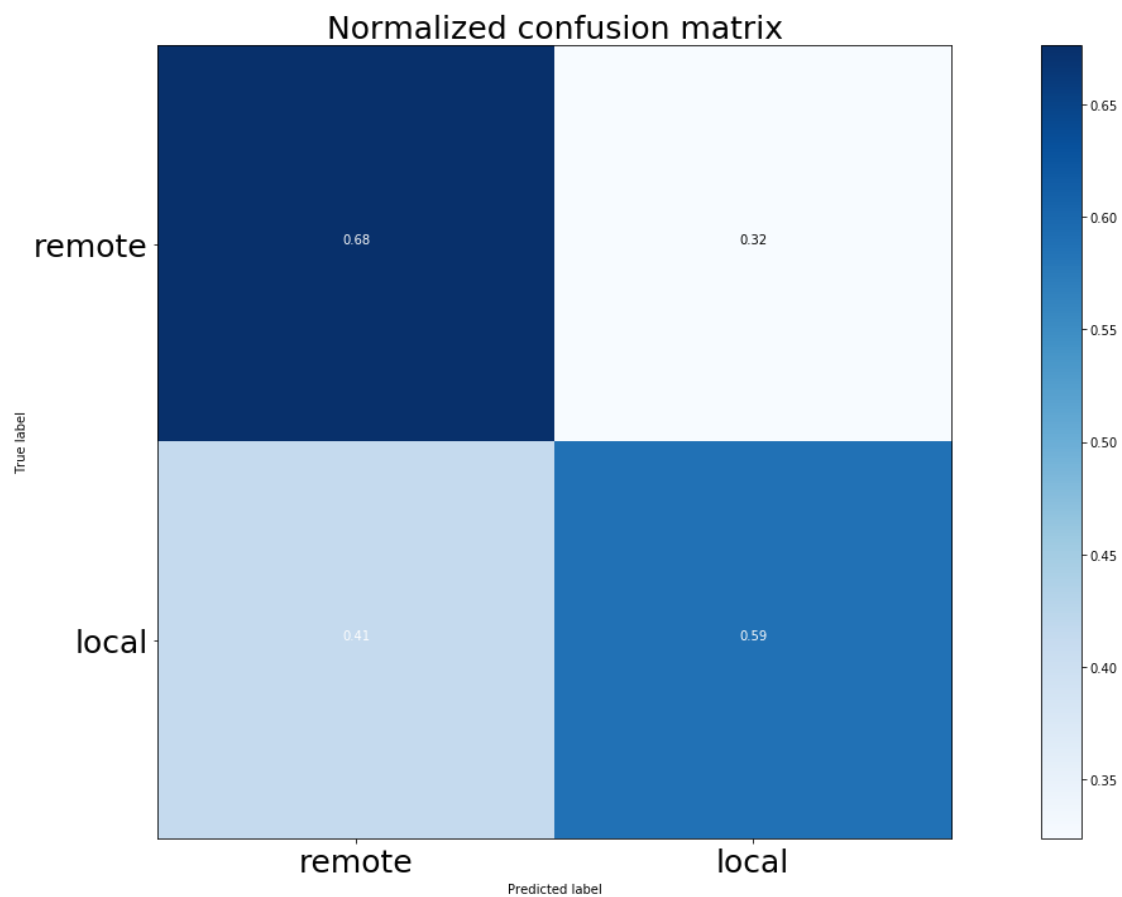
- I also created an n-gram model (both unigram and bigram) that produced the relitave frequency of a word mentioned for each word in the corpus.
- In order to increceaccuracy and reduce noise I addapted Pak & Paroubek's (2010) strategy of calculating salience.

$$salience(g) = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1 - \frac{\min(P(g|s_i), P(g|s_j))}{\max(P(g|s_i), P(g|s_j))} \tag{11}$$

UNI-GRAM	SALIENCE	LABEL
disney	0.883721	remote
hotel	0.812500	remote
nyc	0.810526	remote
often	0.694444	local
parking	0.685000	local
patio	0.653846	local
prefer	0.692308	local
spots	0.666667	local
today	0.727273	local
trip	0.916667	remote
usually	0.701389	local
vegas	0.758974	remote

BIGRAM	SALIENCE	LABEL
quick and	0.875000	remote
waiter was	0.875000	local
and even	0.875000	remote
clean and	0.857143	remote
is nice	0.857143	local
and for	0.857143	local
taste the	0.857143	local
everything is	0.857143	local
not be	0.846154	local
in vegas	0.846154	remote


Confusion matrix (without normalization)
remote
local
4724
2264
2621
3731
True label
remote
local
Predicted label
4500
4000
3500
3000
2500


Normalized confusion matrix
remote
local
0.68
0.32
0.41
0.59
True label
remote
local
Predicted label
0.65
0.60
0.55
0.50
0.45
0.40
0.35