Do Tourist Review Differently?
An NLP Classification Task

Tyler Wilbers

Outline

- The Goal
  - My Hypothesis
- Creating the Corpus
  - Web scraping
  - Corpus analysis
- Training a Yelp review classifier
  - Feature Extraction
  - Increasing Accuracy
- Results

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.

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.

- In what follows, I will present a corpus analysis of reviews scrapped from yelp.com in order to confirm this hypothesis.
- Show results of a predictive model that can classify whether a Yelp review written in English was written by a local to the region.

The Goal: Concept

1/14/2016

1 check-in

Although I sense danger, for there's now a bar downstairs; I'm pretty stoked about about family owned bar with good taps, Whiskey drinks, food and atmosphere--finally. We've been here for years and this is the first decent chill spot close to the Gates and Kosciusko stop. B52 bus is a block away. The twins (no you weren't that drunk) and a Chicagoan bartender were super abputwayttentive and the pretzel dude was off the chain.

Was this review …?

Useful 3 | Funny 1 | Cool

→ Local

6/10/2014

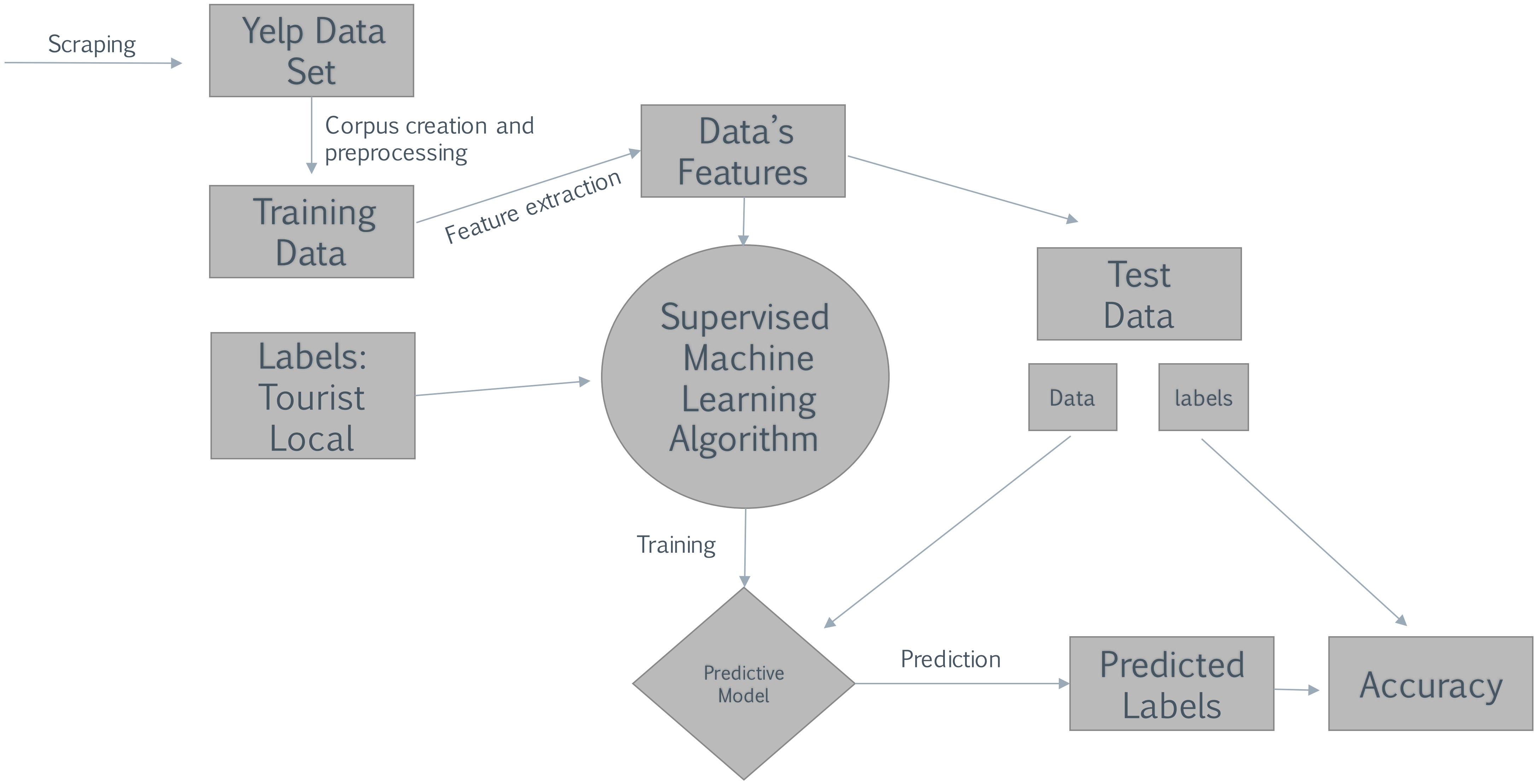
We entered the restaurant at 1:30 on a Sunday and it was pretty packed. We were still seated quickly though and service was perfect. I can't wait for my next trip to south Georgia so I can stop at Bay South for lunch.

Was this review …?

Useful 4 | Funny | Cool

→ Remote

The Goal: Workflow



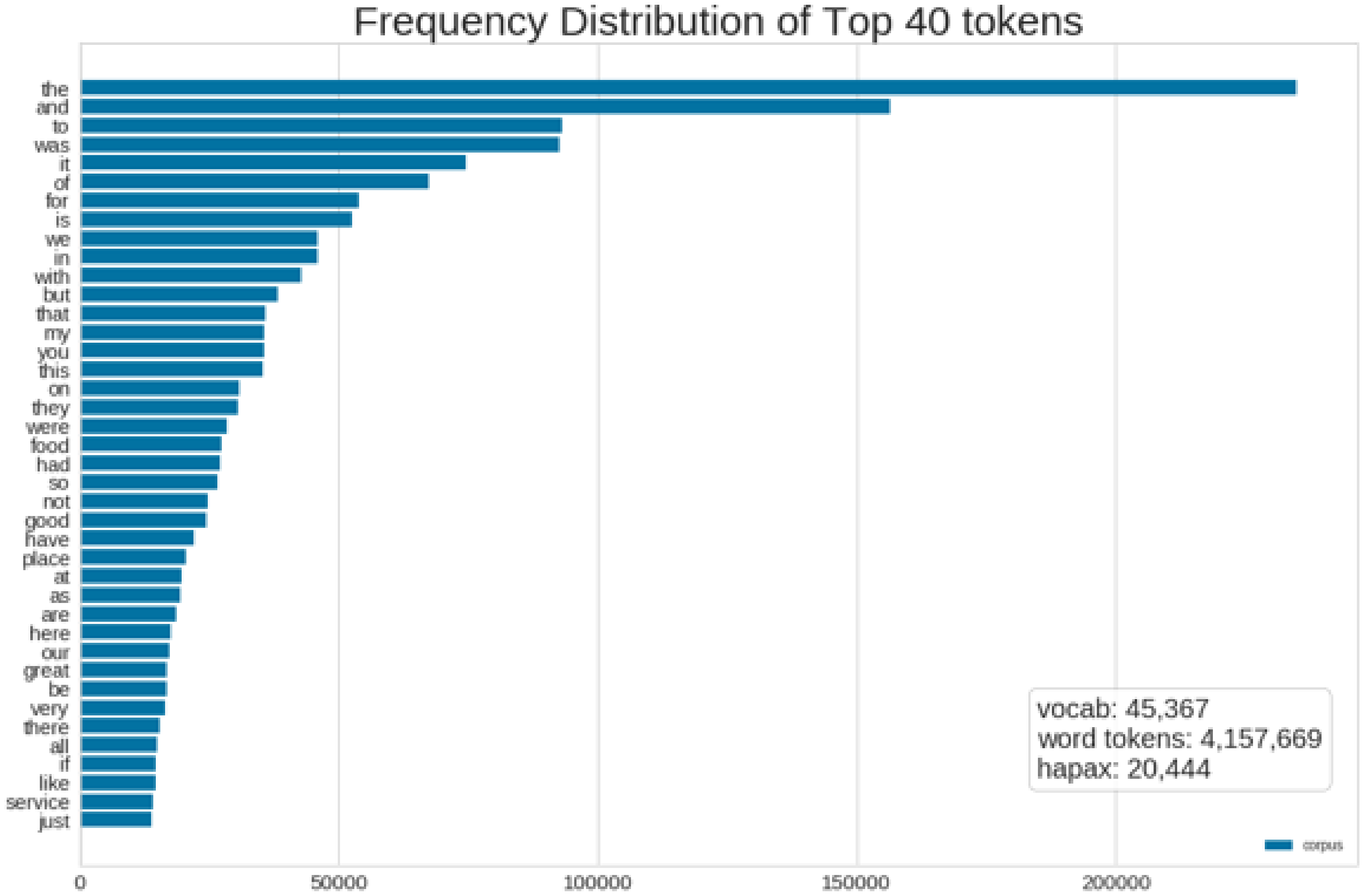
The Corpus: Web Scrapping

- I scrapped 53,000 reviews across 42,800 URLs from the 1,000 most reviewed restaurants from the top five most visited areas in the USA.
  - New York City : 8662 reviews
  - Los Angeles: 6859 reviews
  - Chicago: 15796 reviews
  - Las Vegas: 9725 reviews
  - Orlando: 12015 reviews

The Final Corpus

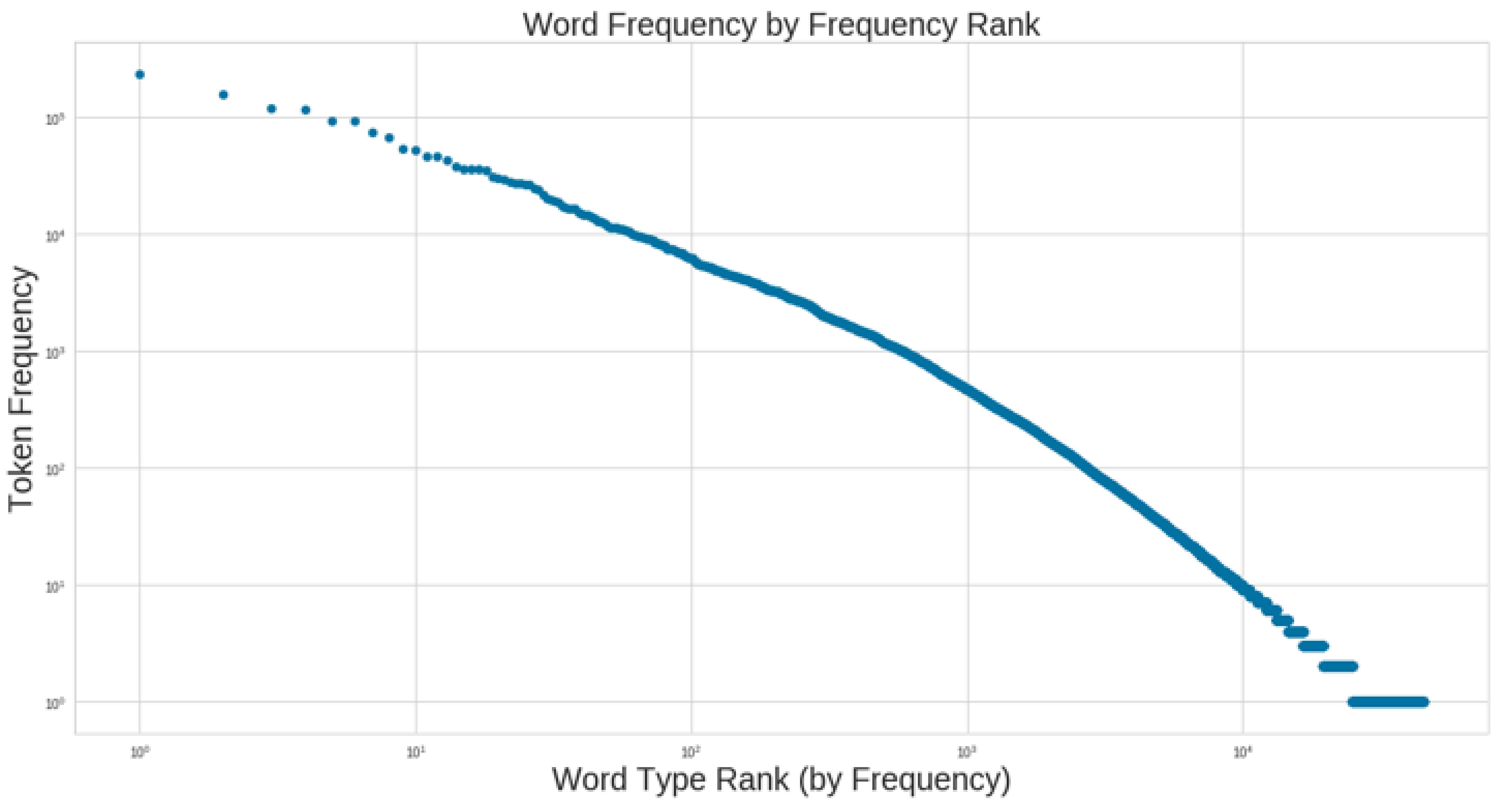
- In order not to overrepresent cities with more reviews I took a sample of the total number of scrapped reviews.
  - 6,859 reviews from each area
- Then I made sure to equally represent local and tourist reviews.
- This filtration process made the final corpus around 32,000 reviews.

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
vocab: 45,367
word tokens: 4,157,669
hapax: 20,444
corpus
0
50000
100000
150000
200000

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 1,000.
- We can check Zipf's Law on 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
Word Type Rank (by Frequency)
10^0
10^1
10^2
10^3
10^4
10^5

Part of Speech Tagging

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

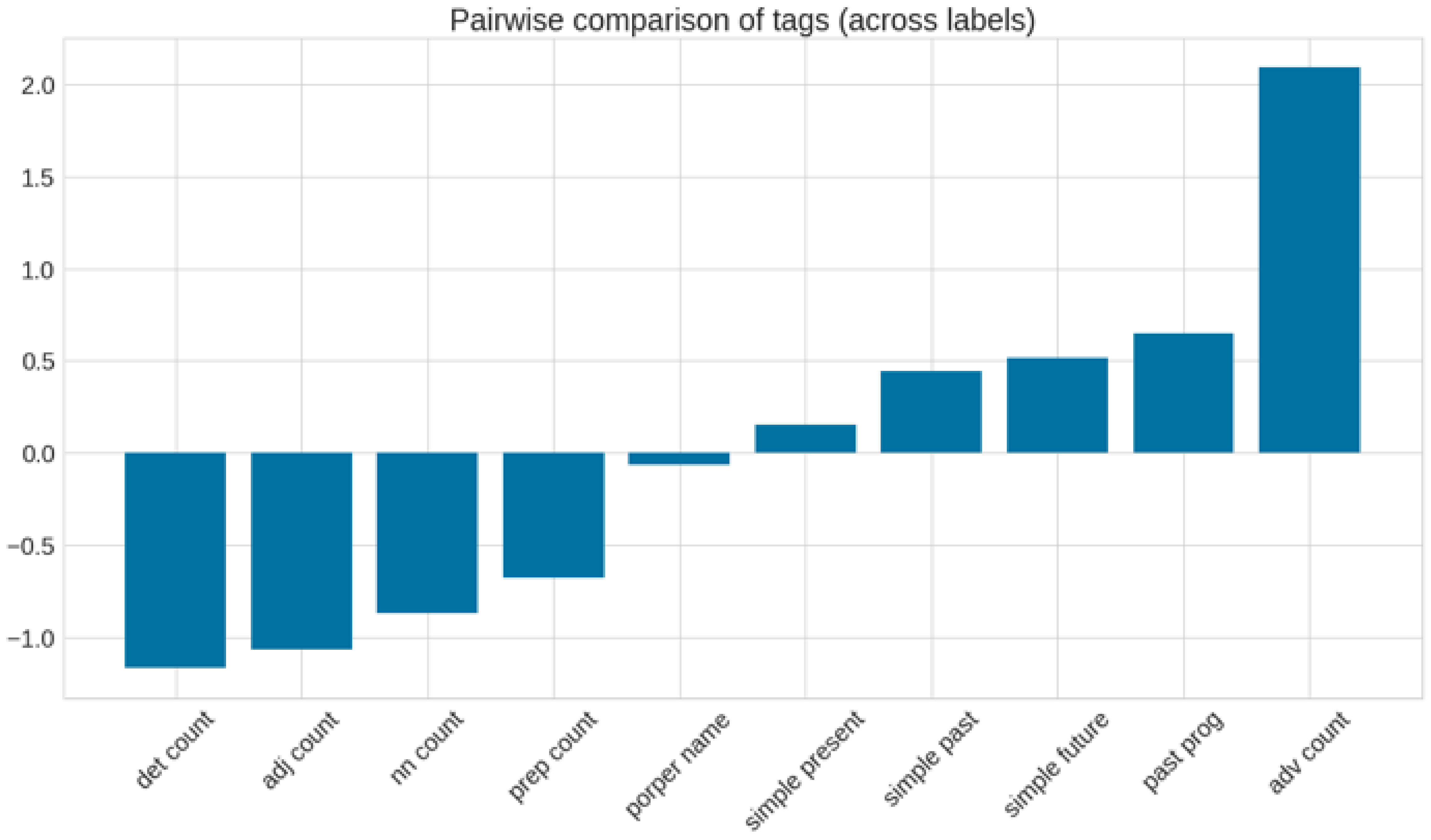
- https://nlp.stanford.edu/software/tagger.shtml

POS Pairwise comparison

- Following Pak & Paroubek (2010) I implemented a pairwise comparison of the POS tags across the two labels.
- This helped identify the part of speech tags that would be good indicators for the classifier.

$$P^T_{1,2} = \frac{N^T_1 - N^T_2}{N^T_1 + N^T_2}$$

Where NT1 and NT2 are the numbers of tag T occurrences in local and tourist reviews, respectively.


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

Predictive Model:
Logistical Regression

Yelp Review Features

- Review length (by characters)
  - On average local's write longer reviews
- Day of the week
- Week of the year:
  - Categorical variable that rages over ever week in the year (1-52).
  - Intuition: reviews written during certain times are more likely to be remote.

Linguistic Features: City Mentioned

- Location mentions:
  - Binary feature based on whether the location of the business is mentioned in the review.

  Prediction:
  - Local reviews are less likely to mention the city the review is in because they see it is more implied knowledge.

  Results:

Linguistic Features: Length

- Review length:
  - Character count of the review.

Results:

Linguistic Features: Tense, Aspect, and POS

- Proper Noun Count (NNP, NNPS)
- Noun Count (NN, NNS)
- Preposition Count (IN)
- Tensed Verbs Count
  - Count Past Participle (VBN)
  - Count Simple Past (VBD)
  - Count Simple Present (VBP, VPZ)
- Adverb count (RB, RBR, RBS):
  - The number adverb occurrences in a review.

Introduce Concepts

- Decision (Adda et al., 1998):

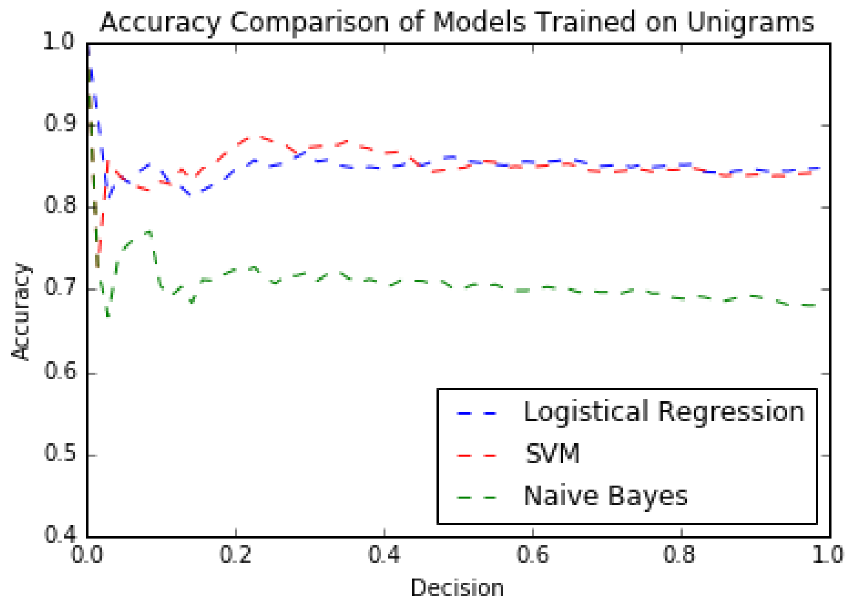
$$decision = \frac{N(\text{retrieved documents})}{N(\text{all documents})}$$

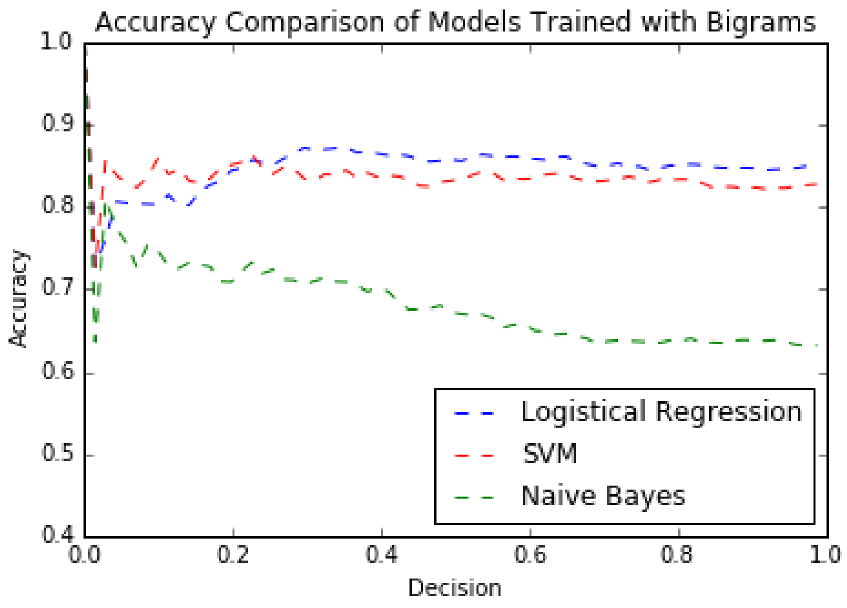
- Accuracy (Manning and Schütze, 1999):

$$accuracy = \frac{N(\text{correct classifications})}{N(\text{all classifications})}$$

Linguistic Features: N-grams

› Preliminary testing showed that bigrams and unigram models yielded similar results across multiple ML algorithms:





Increasing Accuracy

- To discriminate common n-grams I used Pak's (2010) strategy of introducing a salience threshold:

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

- Suppose that a n-gram occurs twice as often in remote reviews.
  - The salience measure for that n-gram would be the one minus the sum of probability of that gram appearing in a local review over the sum of the probability of that gram appearing in a remote reviews (ie 0.5).

Increasing Accuracy

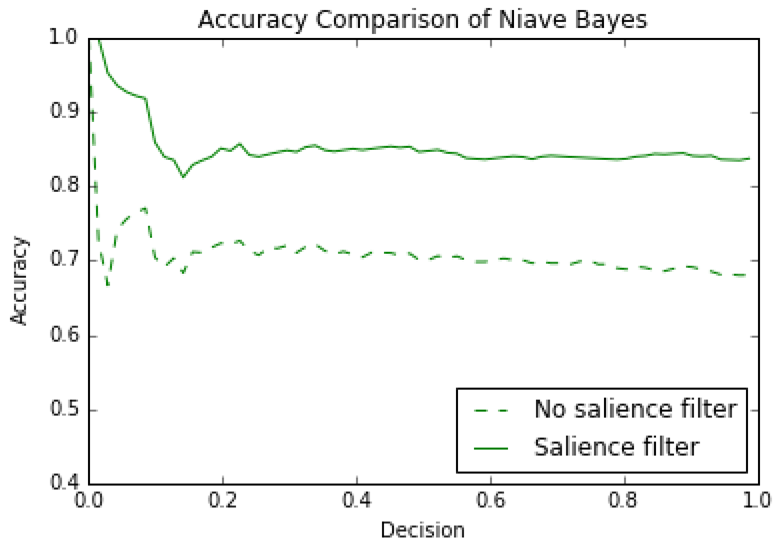
- The following are some examples of unigrams with high salience:

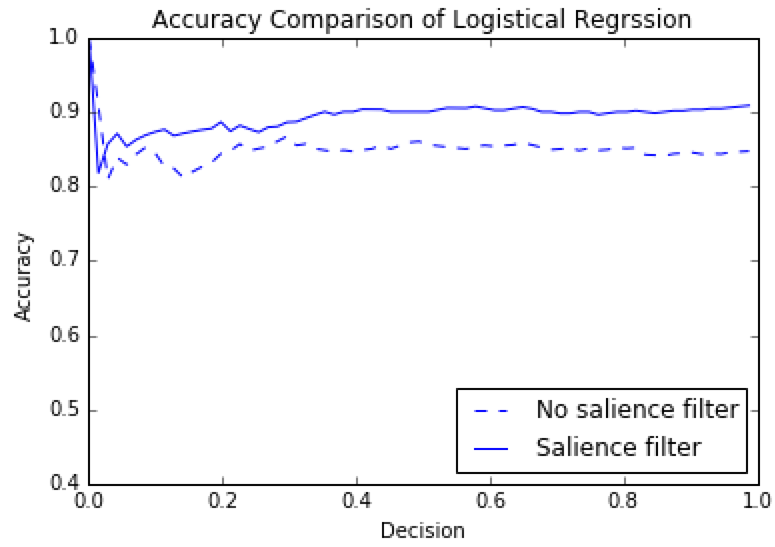
unigram	Salience
wedding	.909
hotel	.967
coupons	.909
golf	.941
tuesday	.875
staying	.939

bigram	Salience
charlotte airport	.857
sunday buffet	.857
time visit	.833
new york	.8
ranch dressing	.9
never bad	.818

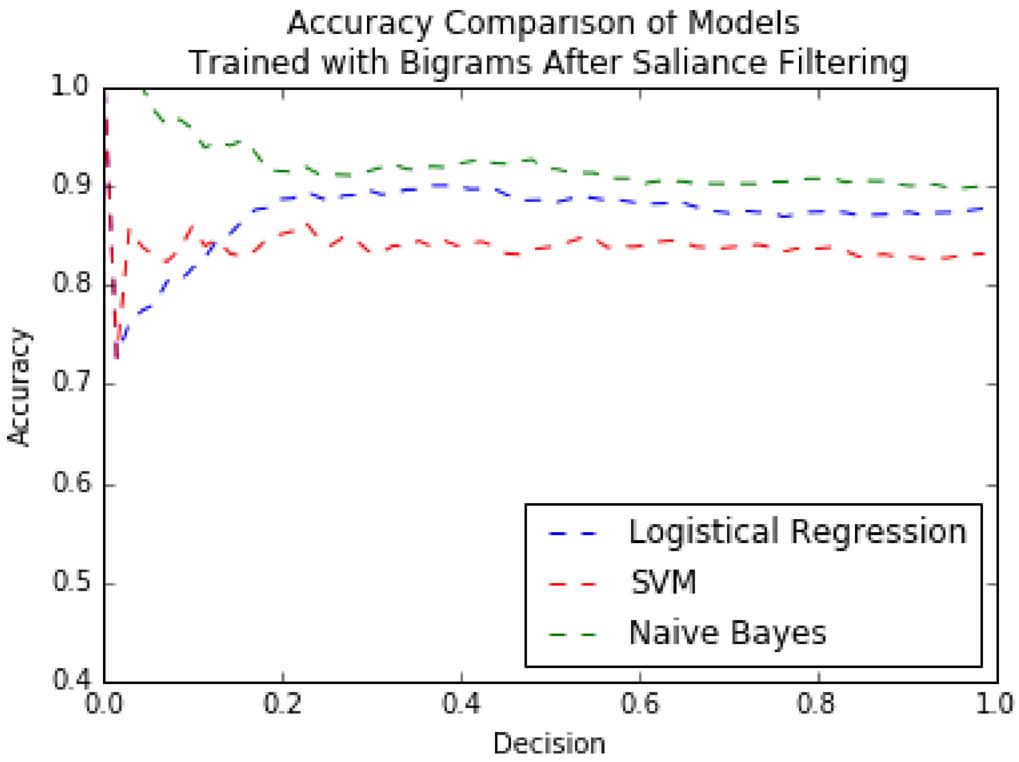
Unigram Results

- By using a salience threshold, ϑ, I was able to eliminate common n-grams.
- Before filtering the average salience was .738 with a standard deviations of .362.
- Setting the ϑ *to.* .65 helped me to significantly improve accuracy.

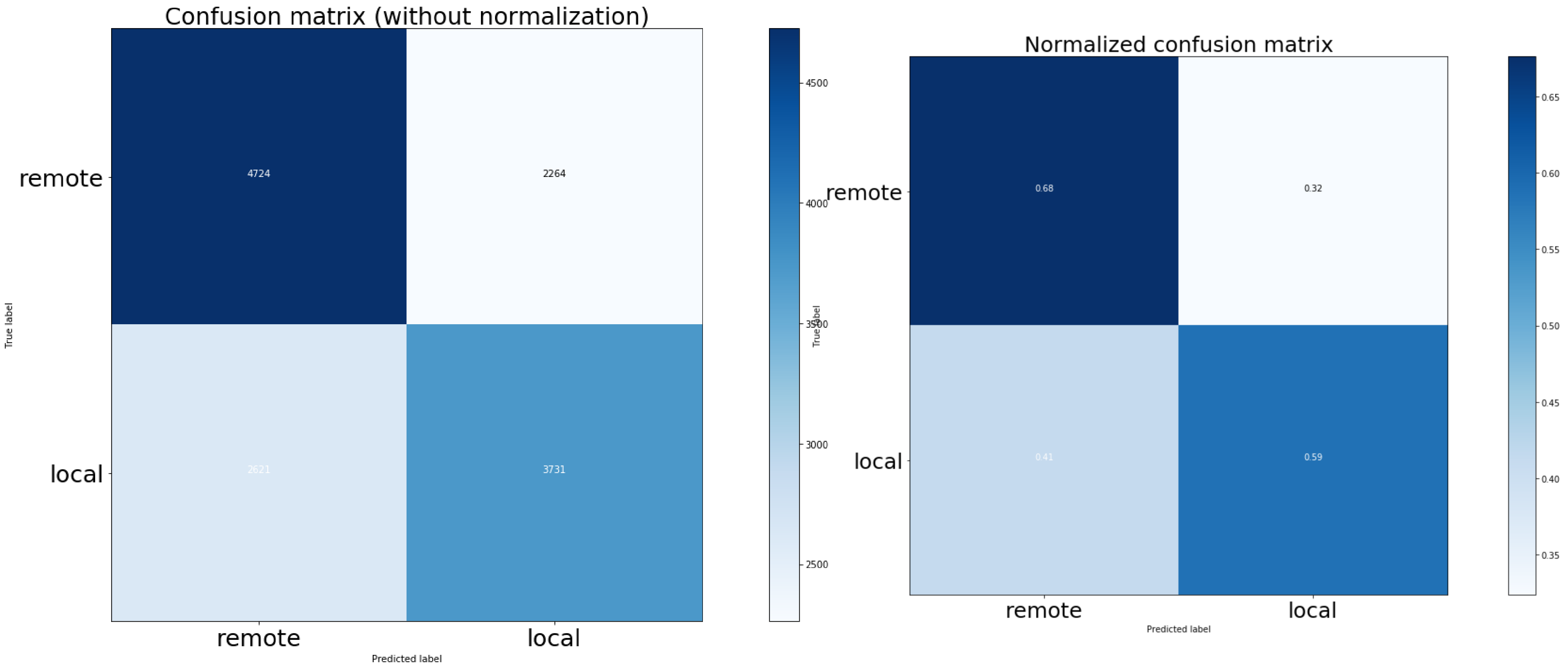




Bigram Results



Confusion matrices after predicting test set



Citations

- G. Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Rajman. 1998. The GRACE French part-of-speech tagging evaluation task. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, LREC, volume I, pages 433–441, Granada, May.
- Christopher D. Manning and Hinrich Schutze. 1999. Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA.
- Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).