

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 classifying yelp reviews as being written by a tourist.

Fields

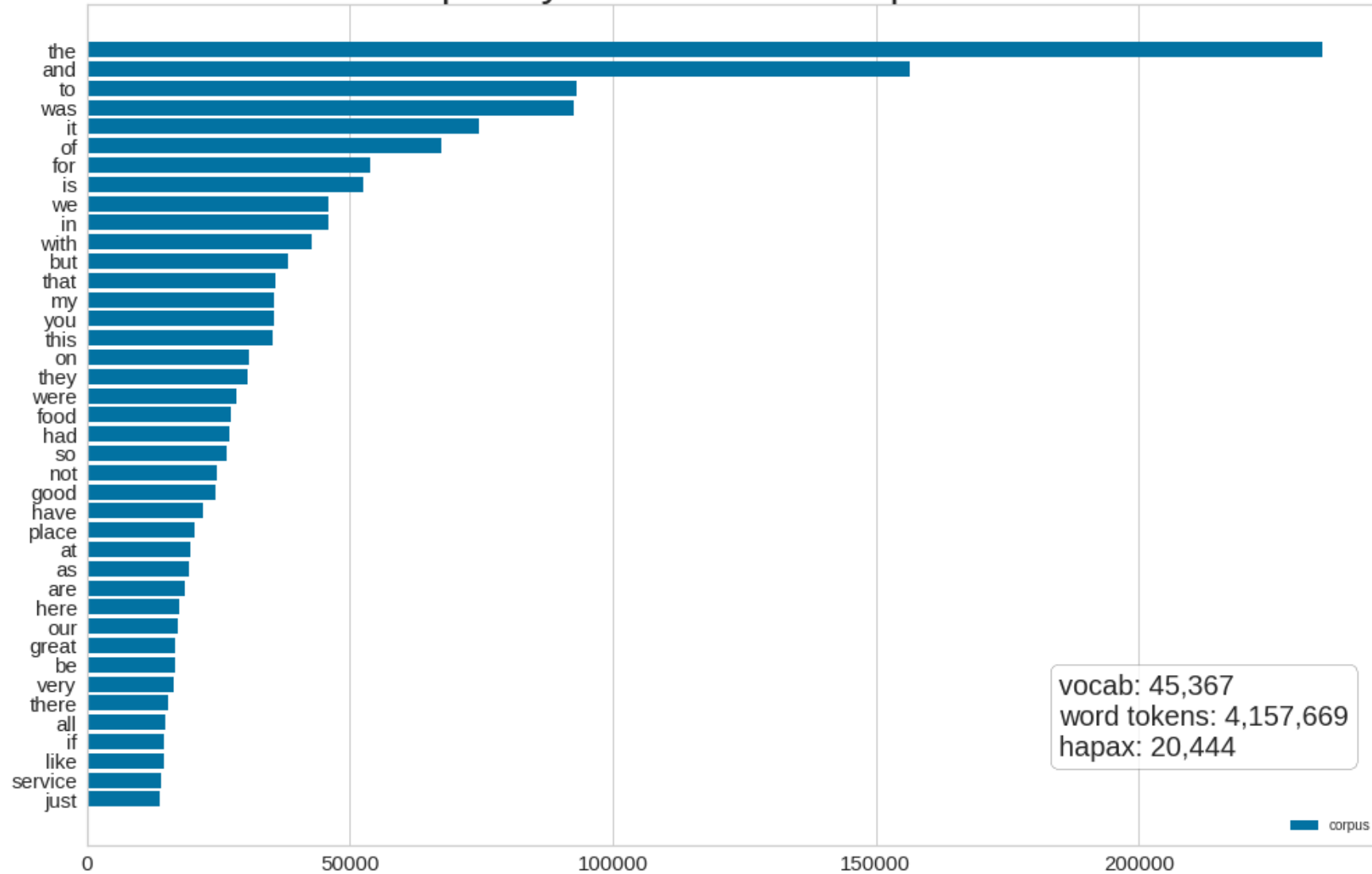
- Business:
 - City
 - Name
 - Average Star Rating
 - State
 - Zip code
- Review:
 - User Location
 - Cool
 - Funny
 - Useful
 - Date
 - Star Rating
 - Reviewer ID

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

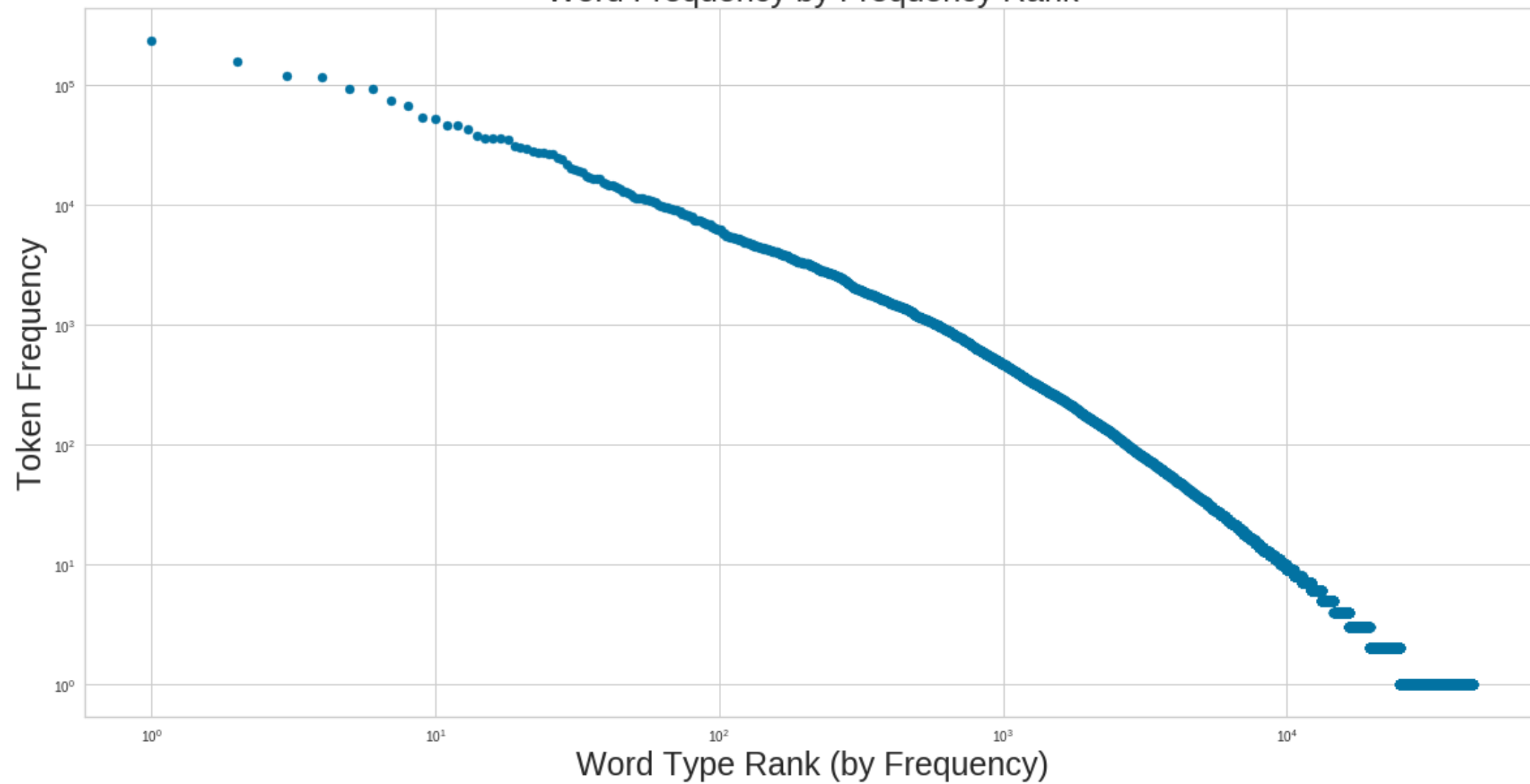
Frequency Distribution of Top 40 tokens



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



Part of Speech Tagging

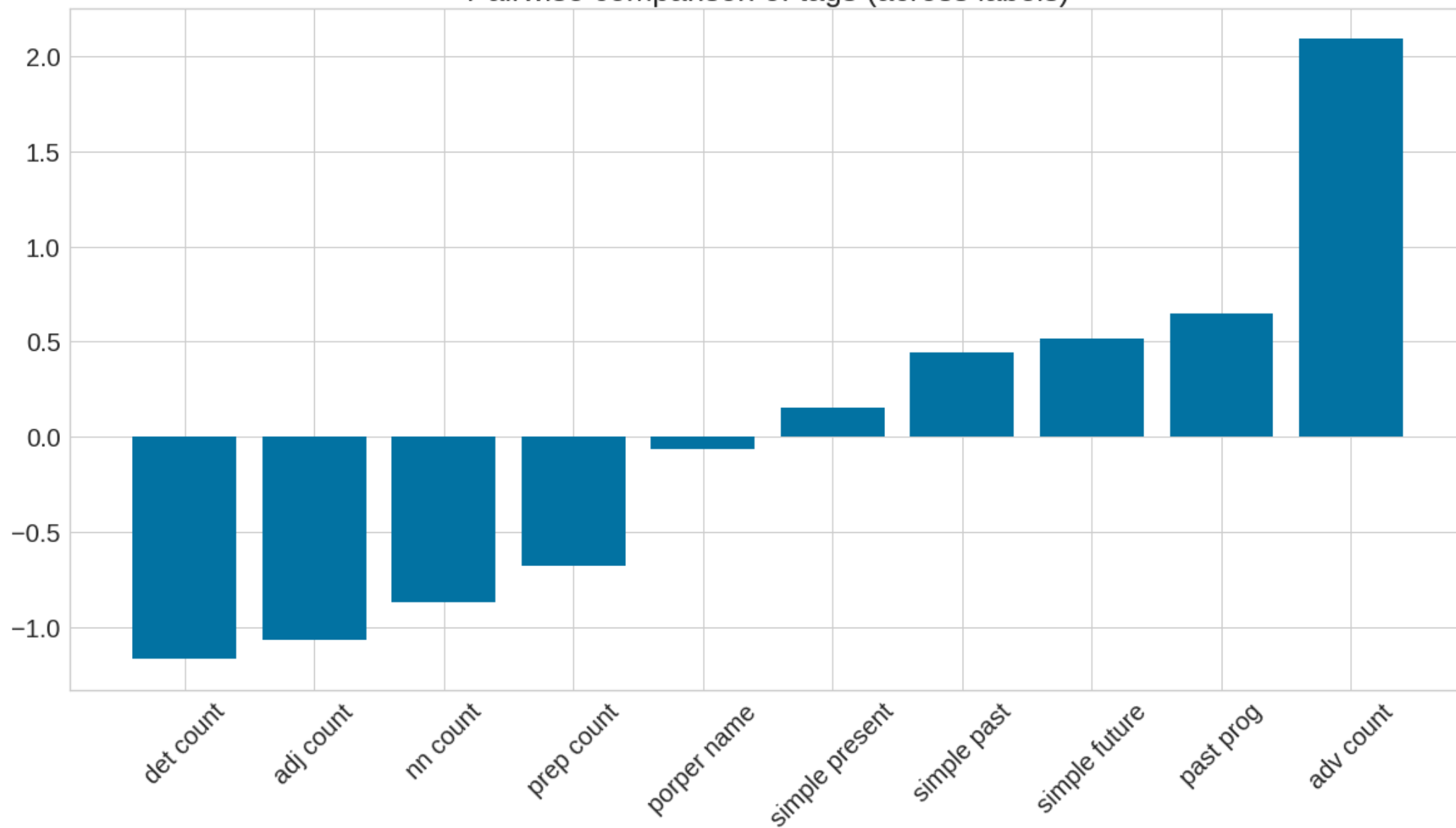
- I ran the Stanford part of speech tagger on all the reviews.
- This software is a Java implementation of the log-linear maximum entropy part-of-speech taggers described in Toutanova et al (2003).

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)



Predictive Modeling: Logistical Regression

Features

- Review length (by charracters)
- 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 relative frequency of a word mentioned for each word in the corpus.
- In order to increase accuracy and reduce noise I adapted Pak & Paroubek's (2010) strategy of calculating salience.

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

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

