Data Science for Linguistics: Usage of Positive 'Anymore'

Synthesized by **Xiaonan Hu & Liqi Zhu**

'Anymore' is a typically negative polarity item (NPI), however, speakers of some dialects use it in positive contexts to represent *nowadays* or *from now on*. This phenomenon always occurs in some varieties of North American English, and is theorized from Irish or Scots-Irish sources.

To examine the geographical feature, as well as contextual meaning, it is essential to collect vast data of sentences containing positive ‘Anymore’. And because it’s kind of dialectal usage, the datasets had better include an indicator of geographical information and be from natural speaking contexts.

Our projects primarily collected multiple datasets from different websites, including social media like Twitter, Quora and various forum websites. Then, to label the data of positive ‘Anymore’, we basically defined a list of *NPIs* and *Negative Triggers* to clean up our origin datasets. We did some analysis to some of the content involved in determining ‘positive or semi-positive anymore’. And at last, we conducted researches regarding comparisons of geographical usage frequencies and general sentiments of sentences with ‘Anymore’.

Through analyzing the data getting from *subreddits* based on geographic location, we found **Iowa** has the highest frequency of positive ‘Anymore’ among all the observed states. And **Missouri** is the second one, followed by Wisconsin, Texas, Nebraska, Pennsylvania and Arizona *(Fig. 1. Distribution Based on State)*. To some extent, the result confirms the geographic distribution given by Professor Abtahian’s lecture. And further, by conducting statistical test, we can draw the conclusion that the overall frequency within the given circle of isogloss is significantly higher than the one outside the circle *(Fig 2. Percentage of Positive Anymore Based on Inside vs. Outside of the Given Isogloss).*

We also gathered data through web scraping from combined forums that fulfill our expectations. We focused on three sites: *Straight dope message board, Body building.com, and City Data.com*. These sites have thriving communities that spend most of their time off topic, and all three have a user’s location listed next their name on every post made. After labeling all the data points, we used the k-nearest neighbor algorithm (k=25) to classify each pixel of the map to obtain a distance weighted map *(Fig. 3. Distance Weighted Map Based on Positive and Negative Anymore Usage)*. According to this map, we can see high rates of positive ‘Anymore’ usage in states center on **Iowa** and **Pennsylvania**. To some extent, it is similar to the result got from *subreddits*, further confirms given geographic distribution.

This synthesized discussion focused on the geographic distribution of positive ‘Anymore’ usage. Basically, the results are consistent with theorized and former research. However, there’s some deviation, we found usage in **Texas** from both datasets is mentioned, and the reason is unclear that need further research.

Fig. 1. Distribution Based on State

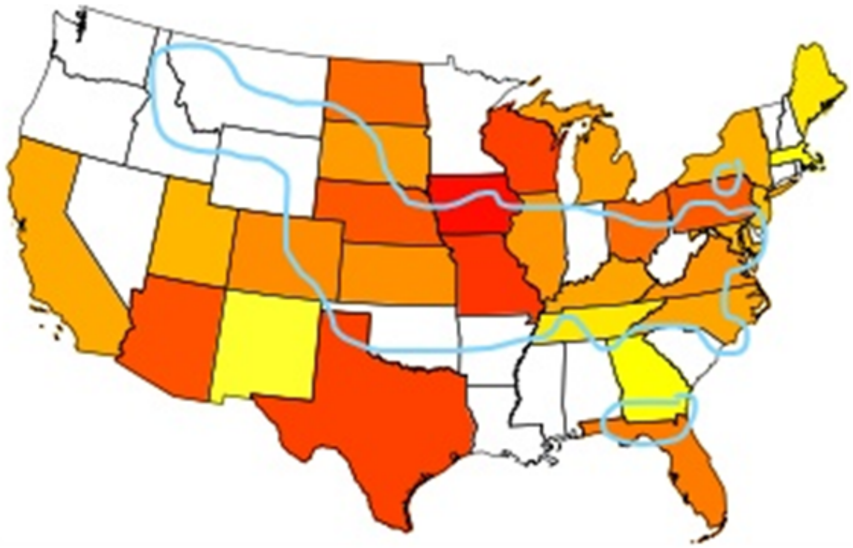


Fig 2. Percentage of Positive Anymore Based on Inside vs. Outside of the Given Isogloss

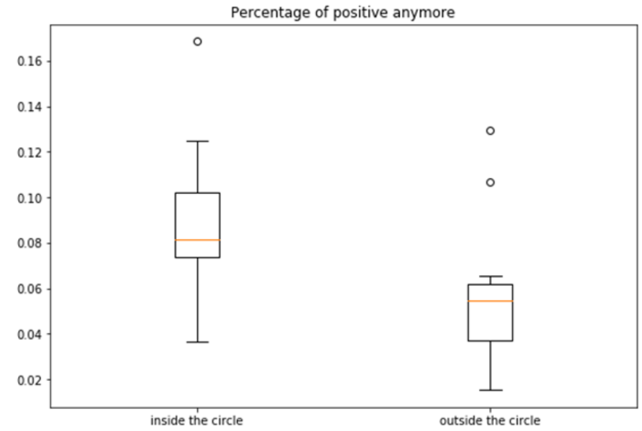
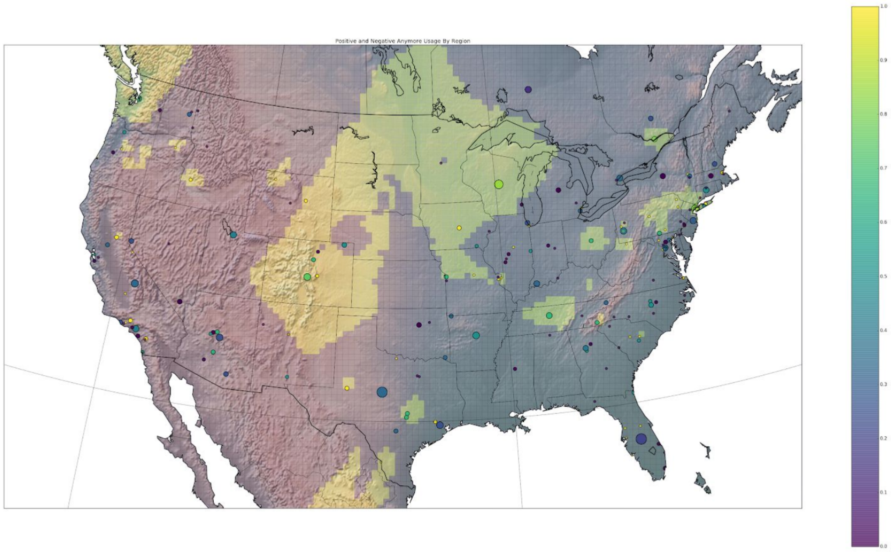


Fig. 3. Distance Weighted Map Based on Positive and Negative Anymore Usage


Positive and Negative Anymore Usage By Region
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0