On Google autocomplete, not all developing countries are the same

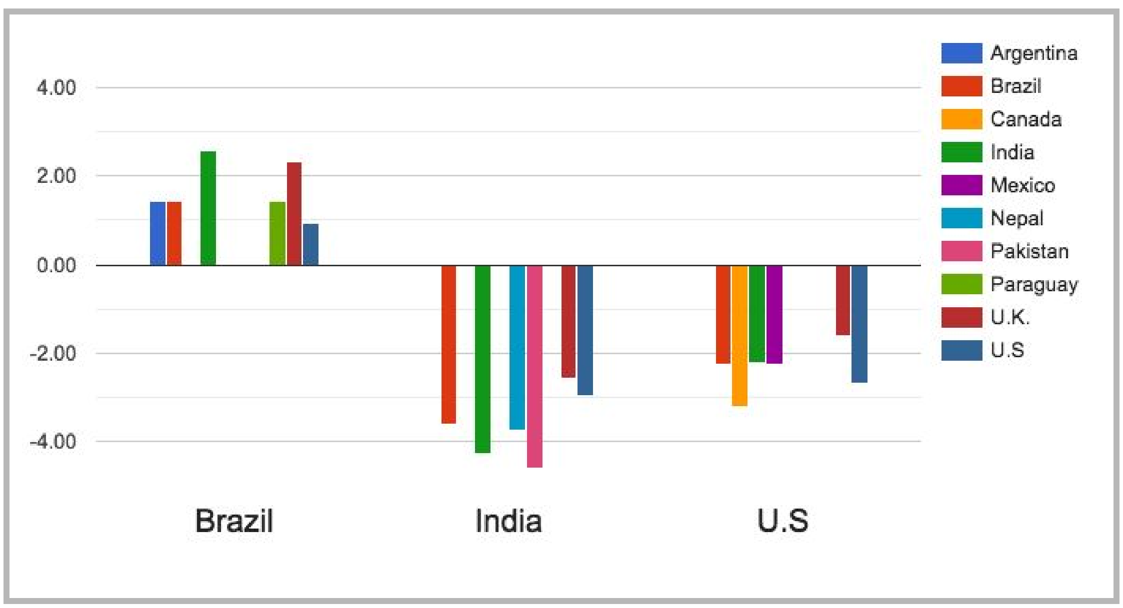
People look at Brazil more positively than India, according to the search giant's suggestions

Daniel Trielli and Karthik Abinav

Brazil and India are similar in plenty of ways. They are both very large countries, they're regional powers in their own continent, and they're at a similar phase of economic development. But when it comes to perception of international internet users, they are quite different. Data extracted from Google autocomplete suggestions shows that Brazil is held in a higher regard than India when they are being searched from other countries.

According to Google, its autocomplete suggestions – those sentences that pop up when you start writing a query – are extracted from actual trending search terms[1]. That means they can be an accurate look into the preconceived notions people have about a variety of subjects[2]. By simulating searches from specific top-level domains, we can see how those perceptions play out according to each nationality and about different countries.

Depending on who's doing the query, Brazil is "a good place to visit" or "violent"; while India is "better than China" or "dirty". The United States is "the best country in the world", but also "racist". Nepalese think Indians are "good programmers" and Americans think Brazilians "gamers are toxic", while at least six countries think Americans are "loud".



Manual analysis of Autocomplete suggestions

[1] https://support.google.com/websearch/answer/106230?hl=en
[2] http://www.tandfonline.com/doi/abs/10.1080/17405904.2012.744320

On average, however, Brazil has a more positive score than India – and the United States, for that matter. We looked at 1.384 autocomplete suggestions for searches about these three countries that would have been made to users from different places. Scoring them as positive or negative, averaging them out, and weighting these results by its place in the suggestion ranking it's possible to see how different these nationalities are seen by the world. Of the three countries, Brazil has the best image: an average score of 1.7 in a scale from -20 to 20. India, on the other hand, has a slightly negative score, of -3.6. And the U.S. is way below, with an average score of -2.4.

Power dynamics and geopolitical rivalries can affect the suggestions made by Google. Take India's relations with Pakistan, for instance. Ever since their independence in 1947, they engaged in multiple conflicts, and that hostility has affected their perception of one another. That is quite visible when we look at the suggestions on autocomplete when users in Pakistan ask Google "why India is". For these users, the top three autocomplete suggestions are "afraid of Pakistan", "against Pakistan" or "jealous of Pakistan". The fourth suggestions is also pretty straight-forward: "why India is the worst country". Remember, these suggestions are being generated by the Pakistani users themselves.

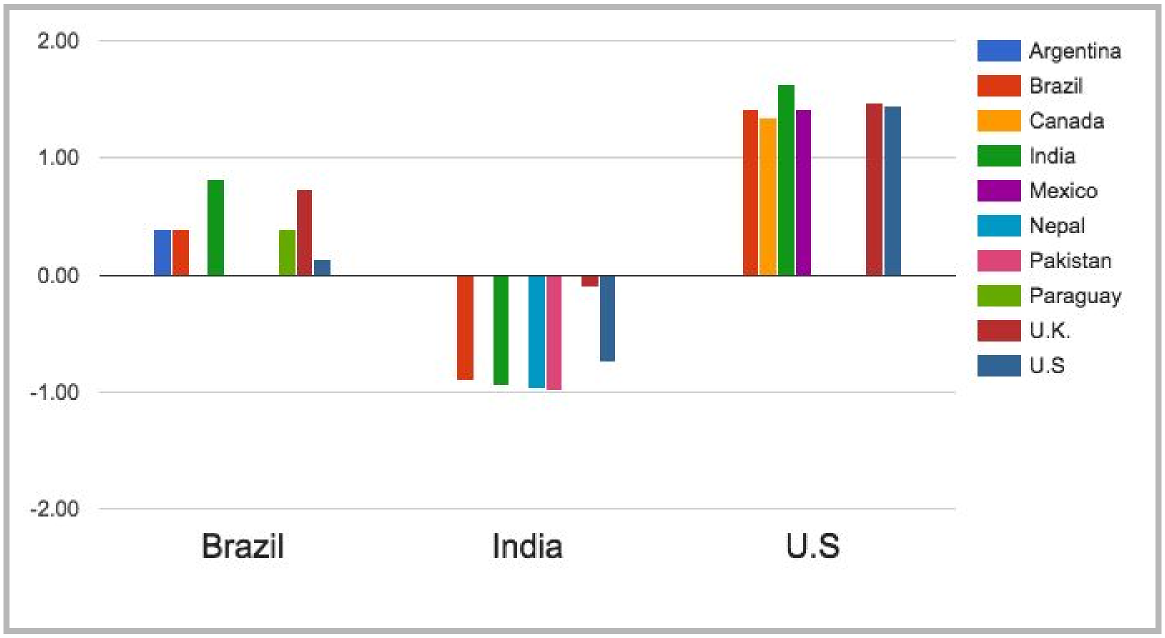
With -4.6, Pakistan had the worst average score for searches about India among the five countries analyzed (the others were Brazil, Nepal, U.K. and the U.S.). The least negative was its former imperial controller: even though British searches generate suggestions such as "Why are indian tourists so rude" or "Why is India so corrupt", they also come up with "Why is India the best" or even "Why is India better than Pakistan."

Brazil, on the other hand, has no negative scores when considering averaged searches from five countries (Argentina, India, Paraguay, United Kingdom and the U.S.). Autocomplete suggestions are more favorable in India, with an average score 2.6. For them, Brazil is "the best country in the world", "amazing" and "important", even as some of them want to find out why "inflation is so high" or the country is "so poor" or "so violent". Even its regional rival, Argentina, wonders "why Brazil is good in football" or "a good place to visit" and, on average, has a positivity score of 1.4 in regards to its historical nemesis in sports.

It's also interesting to note that the average outside perception of a country matches the image it has of itself. Brazil's autocomplete suggestions about Brazil are positive (1.4) and India's results about India are negative (-4.3).

The same can be observed when looking at suggestions about the United States. When analyzed by five other countries and itself, the U.S. has an average score of -2.4, very similar to its self-image (-2.7). The only country that has a worse picture of the U.S. among the ones we analyzed is Canada (-3.2), in part because Canadians are curious as to why are American "cops are so violent" and American "schools are so expensive."

An automated look into an extra 6.222 results yielded similar results. Using a method called sentiment analysis, we were able to give a positive or negative score for every possible permutation of the queries, including less natural searches such as "brazil is why" or "united states why is". Granted, it's not a natural way of searching, but in the variable ways people use the internet, is worth looking at. The results follows the same pattern of Brazil with a positive image and India with a bad score. But the problem with this type of automatic analysis is that it tends to give false positives to innocuous sentences, such as "why united states is called the united states" or "why is united states eeuu in spanish", and that skews the results.



Automatic analysis of Autocomplete suggestions

But no matter who's doing the analysis, whether it is an algorithm or a person, some autocomplete suggestions are unclassifiable. Queries such as "Why is Brazil freaking out", for instance, confound any cultural or geopolitical explanation.

Methodology

1) How did you collect the data you have obtained?

Our data collection process was done through a call to the *Google Autocomplete API* [1]. We set the browser to *Chrome*. The high-level Domain was varied based on the *country* and the language was set to *English*. The search was done in an anonymized fashion i.e. no particular user history was associated with it. We ran the code on two different occasions with a gap of two weeks apart and took only those suggestions that intersected the two. The motivation behind doing this was to filter out suggestions due to a spike of event activity. Having said that, it must be noted that, there were no significant differences between the two collection trials.

2) What were the various queries that was used for this article?

For this study we used the following six query template.

- Why is {country}
- Why is {nationality}
- Why are {nationality} men
- Why are {nationality} women
- Why are {nationality} tourists
- Why are {nationality} cities

Here, {country} refers to India, Brazil and United States. Similarly, {nationality} refers to Indians, Brazilians and Americans. Additionally, we used permutations of the words in each of the queries. This is because the suggestions given for the query "Why is India" is significantly different from the one given for "Why India is".

3) What were the countries chosen as evaluators?

For each of the three target countries (India, Brazil and United States) we considered the following countries as evaluators.

- **India**: India, Pakistan, Nepal, United States, United Kingdom, Brazil
- **Brazil**: Brazil, Paraguay, Argentina, United States, United Kingdom, India
- **United States**: India, Brazil, Canada, Mexico, United States, United Kingdom

For each country, we chose the two other target countries. Additionally, we chose a country for which there is some kind of socio-political tension (e.g. Argentina and Brazil, India and Pakistan). We then chose one country where the target country has an influence upon (e.g. Brazil and Paraguay, India and Nepal) and finally a neutral country, United Kingdom.

4) What are the various metrics you used to generate your score?

We used a combination of two metrics for evaluating the results. The first one was a manually generated score and the second one was a based on score generated from sentiment analysis. For the manually generated score, we rated each of the collected auto-suggestion with a score of +1 for positive, 0 for neutral and -1 for negative. We understand that the score can be highly subjective and in most "grey" area cases, the suggestion were marked as neutral. The score for sentiment analysis was based on the score generated by the NLTK toolkit [(2)]. The tool marked each of the suggestion with a real number between -1 to +1. The more closer the score was to -1, the more negative the sentiment of the statement was. For both the scores, we weighted them by where they occur in the suggestion list. Note that for any user using Google, the first one or two autosuggestion is the most critical. Hence, a suggestion coming higher in the list is more "important" for studying bias as opposed to the later ones. The first suggestion was given a weight of 20, the next was given a 19 and so on. We obtained the number 20, since the code we used returns at most 20 autocomplete suggestions.

5) What is your rationale for using two different metrics?

The sentences of autocomplete suggestions are short. Hence, the sentiment analysis doesn't perform the best. Sentiment Analysis heavily depends on context and usually works well on large documents. Hence, we wanted to corroborate the results from sentiment analysis with a manual verification. The reason we had three levels of granularity for the manual score was because of the high level of subjectivity involved with a manual score. At best, we can classify things as positive, negative and neutral. It is hard to classify a sentence as more positive and less positive without inducing a lot of subjectivity.

6) What does a "positive", "neutral" and "negative" sentiment mean?

When we classify sentences we look at how a particular sentence mean in the given context of the country. Note that the sentiment analysis given by NLTK does not account for this. For example, "Why India is afraid of Pakistan" is rated as neutral by NLTK. However, we rate this as negative. In the context of the tension between India and Pakistan, saying India is "afraid" of Pakistan is giving a negative sentiment of India as a country.

So, let us define what constitutes as "positive", "neutral" and "negative".

Positive: We say a statement is positive, if the country(ies) used in the sentence is replaced by a completely different set of country(ies) and the statement is still looked upon favorably or neutral. In other words, the sentence is inherently positive and not because of the country in subject.

Negative: Similarly, we say a suggestion is negative, if replacing the country(ies) in the statement with a different set makes the statement negative or neutral. This says that, in the context of the current country(ies), the statement is spreading some kind of stereotype.

Neutral: Any suggestion which doesn't fall into the above two categories is rated as neutral. Hence, while interpreting the results, the reader should be aware that we have been conservative in our approach to avoid being too subjective.

7) How can reproduce the results you have obtained?

Unfortunately, you will not be able to reproduce the results *exactly* as we have obtained in this study. However, you can run our code available at this link to collect the data. The link also contains the code used to obtain the score using the NLTK sentiment analysis library. To generate the manual score, the reader will be able to use the above definitions to get a close approximation to the scores we obtained. This is because, even though we tried to keep the definitions as concrete as possible, some suggestions (especially the ones that are mildly positive or negative) might end up being marked as neutral.

8) I think language is as important as the high-level domain. Why have you not considered it?

We agree with you that the language of search is equally important. However, the Google Translate API is paid service and our resources wasn't sufficient to use it. In our future post, we will definitely consider this!

9) Why did you choose Chrome as the browser? Is there a difference between what browser you use, as far as auto-suggestions is concerned?

Chrome is the most used browser worldwide(3) capturing more than half the market share. Unfortunately, Google shows different autocomplete suggestions based on the browser. Hence, we selected the most popular browser and did our analysis on that.

10) Who was in charge of what in the project?

Both Karthik Abinav and Daniel Trielli discussed the goal of the project and the methodology. In the division of labor, Karthik was in charge of building and running the code and collecting data, as well as writing up the methodology. Daniel Trielli analyzed the data, scoring the results and wrote the reports.

11) I have more questions which has not been answered here. How can I contact you?

That is awesome! Please email us and we will get back to you with answers.

Footnotes:

1) http://suggestqueries.google.com/complete/search?client=chrome&q=YOURQUERY
2) http://www.nltk.org/howto/sentiment.html
3) http://www.sitepoint.com/browser-trends-january-2016-12-month-review/